Hypothesis Generation and Pursuit in Scientific Reasoning

NYRUP, RUNE

Hypothesis Generation and Pursuit in Scientific Reasoning

Rune Nyrup

Abstract:

This thesis draws a distinction between (i) reasoning about which scientific hypothesis to accept, (ii) reasoning concerned with generating new hypotheses and (iii) reasoning about which hypothesis to pursue. I argue that (ii) and (iii) should be evaluated according to the same normative standard, namely whether the hypotheses generated/selected are *pursuit worthy*. A consequentialist account of pursuit worthiness is defended, based on C. S. Peirce's notion of 'abduction' and the 'economy of research', and developed as a family of formal, decision-theoretic models.

This account is then deployed to discuss four more specific topics concerning scientific reasoning. First, I defend an account according to which explanatory reasoning (including the 'inference to the best explanation') mainly provides reasons for pursuing hypotheses, and criticise empirical arguments for the view that it also provides reasons for acceptance. Second, I discuss a number of pursuit worthiness accounts of analogical reasoning in science, arguing that, in some cases, analogies allow scientists to transfer an already well-understood modelling framework to a new domain. Third, I discuss the use of analogies within archaeological theorising, arguing that the distinction between using analogies for acceptance, generation and pursuit is implicit in methodological discussions in archaeology. A philosophical analysis of these uses is presented. Fourth, diagnostic reasoning in medicine is analysed from the perspective of Peircean abduction, where the conception of abduction as strategic reasoning is shown to be particularly important.

HYPOTHESIS GENERATION AND PURSUIT IN SCIENTIFIC REASONING

Rune Nyrup

A THESIS SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Department of Philosophy

Durham University

March 2017

Table of Contents

List of Illustrations

Declaration

I confirm that no part of the material contained in this thesis has previously been submitted for any degree in this or any other university.

Chapter 6 is based on joint work with Donald E. Stanley. The individual responsibilities for this work are as follows: Stanley developed the clinical case in Section 6.3, the medical examples throughout and the discussion of strategies for hypothesis generation in Section 6.2.2. My primary contributions are the philosophical framework in the remainder of Section 6.2, the critique of existing accounts of diagnosis in Section 6.4 and the account of diagnosis as strategic reasoning in Section 6.5.

All other material is the author’s own work, except for quotations and paraphrases which have been suitably indicated.

Statement of Copyright

Acknowledgements

To write a thesis takes, if not a whole village, at least a sizeable intellectual community and a supportive social network. Throughout my PhD, I have been lucky enough to be able to draw on the support, help and insight of many friends and colleagues both in Durham, at CHESS (the Centre for Humanities Engaging Science and Society) and the Department of Philosophy, and elsewhere.

My first thanks are to my primary supervisors, Julian Reiss and Nancy Cartwright. They have guided me through my PhD-studies and read many drafts of my work, from early drafts of papers to nearly-finished chapters. Although they only have one entry each in my bibliography, the ideas I develop in this thesis owe a huge amount to my discussions with them, and I am immensely grateful for their help and support over the past 3½ years. I would also like to thank Alison Wylie, who supervised Chapter 5 and read my early musings on analogies in archaeology, and Wendy Parker, who supervised Chapter 4 and read parts of the thesis during the final push. All of their comments and suggestions are hugely appreciated.

In the Department of Philosophy at Durham University, I would also like to thank Simon James, Ian Kidd, Andreas Pantazatos, Emily Thomas, Matthew Tugby and Peter Vickers, who have all offered invaluable advice and help in one way or another. I owe special, additional thanks to Simon, Wendy and Alison, who organised a mock-interview on short notice, and to Ian for feedback on several job applications. I would also like to thank Juha Saatsi (University of Leeds), with whom I have often been able to discuss my work and who has provided crucial career advice and support.

The philosophy postgraduates in Durham have offered countless opportunities, both academic and casual, to learn from others, discuss my ideas and relax with friends. I have had the chance to do the same with the archaeologists, more often than not at The Vic on a Friday. I reserve a special mention to the other CHESS-students whom I have worked alongside for all or most of my PhD: William Peden, Anna de Bruyckere, Sarah Wieten, Erin Nash, Chien-Yang Huang, Anthony Fernandez and Donal Khosrowi. I have learnt a lot from them and they have made my time in Durham much more enjoyable.

Donald E. Stanley showed me that Peirce’s views on abduction can be useful for understanding medical diagnosis, and it was he who proposed that we collaborate on the paper on which Chapter 6 is based. I would like to thank him for the chance to explore a different kind of scientific reasoning, and look forward to working with him in the future.

Jennifer Peacock was the first to introduce me to the fascinating world of archaeology. Our many discussions about the topic have been a tremendous help in thinking about analogies in relation to archaeological method and theory. I would also like to thank her for all of her support over the last three years.

Special thanks go to the Peacocks, for welcoming me and for many dinners and teas.

Finally, I would like to thank my parents, Inge and Kim Nyrup, without whom this thesis would not have been possible. They have always trusted and encouraged me, even when I explained that I wanted to study philosophy at University. Their love and support means the world to me.

Introduction

Understanding scientific reasoning from a normative perspective is a major aim for philosophy of science. Often, when philosophers discuss scientific reasoning, the implicit assumption is that the normative standard for evaluating such reasoning is whether it provides reasons for *accepting* a theory or hypothesis, i.e. reasons for regarding it as (in some sense) an established piece of scientific knowledge: something which can be regarded as true, accurate or in some other sense a reliable representation of the world. However, even adopting a broad understanding of ‘acceptance’, deciding which hypotheses should be accepted does not exhaust the kind of reasoning which scientists must (and do) engage in. In this thesis, I examine two further, closely related kinds of reasoning, namely: (i) reasoning concerned with the *generation* of new hypotheses and (ii) reasoning about which hypothesis should be prioritised for *pursuit*, i.e. for further testing and development.

In the 1940s and 1950s, many philosophers of science rejected the idea that anything normatively interesting can be said about these questions. How hypotheses are generated and decisions about which of them to pursue were thought to belong to what Hans Reichenbach had called “context of discovery” rather than the “context of justification”. According to Reichenbach, and many other philosophers drawing on his work, only the latter is amenable to any kind of normative, philosophical analysis. During the 1950s and 1960s, philosophers and historians of science such as Mary Hesse, Norwood Russell Hanson and Thomas Kuhn started to challenge this orthodoxy. They highlighted many aspects of scientific reasoning which, although usually deemed part of the context of discovery, seemed both interesting and possible for philosophers to analyse. However, while many came to reject—rightly, in my view—the injunction against exploring and theorising about these kinds of reasoning in philosophy, most continued to assume that reasoning concerned with generation and pursuit could still be understood in roughly the same terms as reasoning about acceptance, i.e. in terms of the likeliness or plausibility of hypotheses.

It is one of central claims of this thesis that this assumption is mistaken. Instead, I defend an account of reasoning concerned with the generation and pursuit of scientific hypotheses inspired by Charles Sanders Peirce's mature view of the form of reasoning he called *abduction*. On this view, both kinds of reasoning answer to the same normative standard, namely the 'Economy of Research'. In brief, the *pursuit worthiness* of a theory or hypothesis depends on how much we can expect to learn from pursuing it, and whether this would be a cost-effective use of the limited resources available for research. It is this normative standard which should be used to evaluate the generation, as well as the selection, of hypotheses for pursuit. While likeliness or plausibility is still a relevant consideration on this account, it is far from the only one.

The first two chapters of this thesis develop and defend this account. Chapter 1 is a historical review of how the distinction between acceptance and pursuit emerged in twentieth-century philosophy of science. In it, I argue that this distinction was implicit in the works of Karl Popper, Thomas Kuhn and Imre Lakatos. I then present my interpretation of Peirce's mature account of abduction, and discuss its reception in view of Reichenbach's distinction between the context of discovery and the context of justification. From this, I draw out some themes which will inform my discussion in the rest of the thesis.

In Chapter 2, I develop my account of pursuit worthiness from a systematic perspective. I start by addressing two prima facie objections to the project of developing a normative account of pursuit: (i) that there are no interesting normative constraints on what scientists should pursue and (ii) that the normative constraints on pursuit are of the same kind as acceptance. I argue that plausible responses to these objections can be given within a general consequentialist approach to pursuit worthiness. Based on this, I develop a family of formal decision-theoretic models, before showing how these can represent a number of factors relevant to evaluating the *epistemic* aspects of pursuit worthiness and the trade-offs which exist between them.

The remaining four chapters draw on the first two to discuss different kinds of scientific reasoning.

I start, in Chapter 3, with explanatory reasoning. Here, I develop an account according to which the explanatoriness of a hypothesis (i.e. how satisfying the explanations are which it would provide, if it were true) provides reasons for pursuing the hypothesis, rather than reasons for accepting it. This account, which I call *the Peircean view*, avoids the problems facing *explanationism*, i.e. the view that explanatory considerations can provide some reason for accepting a hypothesis, often summarised in the so-called "inference to the best explanation". Furthermore, I criticise a number of empirical arguments for explanationism.

In Chapter 4, I discuss the role of analogies in science, utilising a case study involving the development of the liquid drop model of the atomic nucleus. I argue that, at least at some stages of its development, the fact that the liquid drop model was based on analogies with macroscopic water drops provided some reason for pursuing it further. I consider a number of possible accounts of how analogies can justify pursuit, proposing that in cases like the liquid drop model, analogies allow scientists to transfer already well-understood modelling frameworks and use them to construct explanations in a new domain.

Chapter 5 analyses the longstanding methodological debate within archaeology concerning the proper us of analogies. I show that a distinction between using analogies

(i) to provide reasonns for accepting an interpretation, (ii) to generate new interpretative hypotheses and (iii) to provide reasons for pursuing these is implicit in some of the archaeological literature. However, many archaeologists still discuss analogies as if they represent a single problem. I present a philosophical analysis of these uses of analogy within archaeology, arguing that each use is associated with different adequacy criteria and potential problems. I illustrate how this framework helps clarify the methodological debate by applying it to a case study from Roman archaeology.

The final chapter, based on a paper co-authored with an experienced physician, discusses medical diagnostic reasoning. In it, we analyse a detailed clinical case study, showing how decisions about pursuit often have downstream consequences for later stages of inquiry, e.g. by producing clues for further hypothesis generation. Due to the open-endedness of the interaction between generation and pursuit, we argue that reasoning about which diagnostic hypotheses to pursue cannot be captured in the simple decision-theoretic model currently popular in the medical literature. Instead, we argue that diagnostic reasoning can better be understand in terms of strategic reasoning.

These four chapters do not amount to an overreaching argument. They are, to a large extent, self-contained studies which aim to illuminate each type of scientific reasoning in its own right. But they are not a disparate collection of unconnected essays either. The chapters are bound together by drawing on a set of common themes developed in Chapters 1 and 2, and on the general account of hypothesis generation and pursuit defended there. At the same time, I hope to have avoided the opposite vice, of simply letting each chapter be a naïve and repetitive application of my general account. While the starting point for each chapter is the distinction between acceptance, generation and pursuit, the issues raised in each context by this distinction differ. In writing these chapters, I have found that this distinction highlights and clarifies issues which have not been fully articulated in the existing literatures. At the same time, my thinking about the distinction, as articulated in the first two chapters, has been informed by issues encountered when using it within the concrete debates of the later chapters. I hope readers of this thesis will find both the general framework and the discussions in the later chapters as fruitful and illuminating as I have.

Chapter 1. The Emergence of the Acceptance/Pursuit Distinction

1.1. Introduction

The key conceptual tool in this thesis is the distinction between reasons for *accepting* hypotheses and reasons for *pursuing* hypotheses. The purpose of this chapter is to outline the historical background to the emergence of the distinction between acceptance and pursuit in twentieth-century philosophy of science, and to highlight some lessons for later chapters.

While there are earlier examples of philosophers and scientists discussing arguments for pursuing rather than accepting theories (see Achinstein 1993: 90-5), the distinction between acceptance and pursuit emerged in the recent philosophy of science literature during the 1970s. The significance of this distinction was not so much that it introduced a completely new topic into philosophy of science; rather, it clarified and made explicit issues that were already implicit in two distinct (but often intertwining) strands of post-positivist philosophy of science in the twentieth century. First, the debates over the accounts of science developed by Karl Popper, Thomas Kuhn and Imre Lakatos and, second, the reception of Charles Sanders Peirce's notion of abduction as a "logic of discovery" and the ensuing debate over the so-called "context of discovery".

In this chapter, I provide an overview of both lines of development. I start by giving a preliminary characterisation of the acceptance/pursuit distinction, before showing how this distinction can help us make sense of the Kuhn-Popper-Lakatos debates and the reception of Peircean abduction. My aim is not to give a detailed account of the evolution of the distinction, but rather to highlight some of the ways the distinction was implicit in these classical debates. This will provide historical context for a systematic discussion in

Chapter 2 of recent work on pursuit since the distinction became an explicit topic of research in the 1980s.

1.2. The Acceptance/Pursuit Distinction

For the purposes of this chapter I will adopt a broad preliminary characterisation of the distinction between acceptance and pursuit. As I will use the term, to accept a hypothesis is to regard it, to some degree, as a piece of established scientific knowledge (cf. Franklin 1993a: 253), whereas to pursue a hypothesis is to spend time and resources testing it, calibrating its empirical parameters, developing it theoretically (e.g. by resolving conceptual problems or drawing out its implications) or applying it to new domains. More succinctly, to pursue a hypothesis is to *work* on it. Thus, acceptance is a matter of having or adopting some positive epistemic attitude towards a hypothesis, whereas pursuit is a practical activity.

It is worth noticing at the outset that this characterisation glosses over a number of nuances regarding exactly what is involved in "accepting" a theory or hypothesis. First, the scientific realism debate is often characterised partly in terms of this question (e.g. van Fraassen 1980: 6-13; Godfrey-Smith 2003: 175-179). Scientific realists claim that acceptance involves regarding a hypothesis as true, or at least partially true, likely to be true or in some other sense descriptively accurate. Anti-realists instead take acceptance to involve something else, e.g., regarding the hypothesis as empirically adequate (van Fraassen *ibid.*) or treating it as if it were true because it has the highest problem-solving power (Laudan 1977: 108). Second, some draw a distinction between *accepting a hypothesis*, in the sense of taking it as a premise in theoretical or practical reasoning, and *believing a hypothesis*, in the sense of holding it to be true (e.g. Dawes 2013). Third, 'acceptance' is often discussed as if it denoted a single, unified type of attitude. But, as

Daniel McKaughan (2007) points out, scientists can and often do adopt a number of different attitudes towards different theories, with a wide scope for variation in both doxastic modality (unqualified belief, that a theory is more likely than not, that it is *prima facie* plausible, …) and semantic content (literal truth, empirical adequacy, closer to the truth than known alternatives, …). Finally, as Parker (2010) and Elliot and McKaughan (2014) argue, rather than asking whether a given model or hypothesis should be accepted *simpliciter*, it is in many contexts more relevant to ask whether one should regard the model as adequate for certain purposes. Since this attitude also involves relying on the model for certain practical purposes, it can reasonably be characterised as a form of acceptance, although it differs from the kind of acceptance usually discussed by philosophers of science.

In this chapter, I intend to be neutral with regards to all of these nuances. Generally speaking, I will take ‘acceptance’ to cover any type of positive epistemic attitude towards the theory, and ‘pursuit’ any attempt to work on the theory in order to further evaluate whether it should be accepted in this broad sense. Likewise, I will not assume any particular answer to what the appropriate units of acceptance and pursuit are in science, e.g. specific claims or hypotheses about the world, other representational entities (models, simulations), broader theoretical frameworks (e.g. classical mechanics vs. quantum mechanics) or larger sociological units (paradigms, research programmes, …).[1] What is important for my purposes is that, given any reasonable notion of acceptance, one can distinguish between (a) accepting a hypothesis and (b) trying to find out whether one should accept it, i.e. pursuing it. Correspondingly, one can distinguish between (a’) reasoning concerned with whether a hypothesis should be accepted and (b’) reasoning

[1] Later in this thesis it will be relevant to focus on more narrowly constrained notions of acceptance and pursuit. I will highlight this when relevant.

concerned with whether it is reasonable to pursue it. One may, of course, argue that the two forms of reasoning coincide or that they are in some sense of the same kind, but they are at least conceptually distinct.

1.3. Popper, Kuhn, Lakatos and Their Critics

1.3.1. Popper's Logic of Scientific Discovery

In a nutshell, the core aim of Karl Popper's 1934 book *Logik der Forschung* (translated as 1959/1992 *The Logic of Scientific Discovery*) is to formulate an account of science which explains how empirical science differs from mere speculation (1959/1992: 10-12), given that this cannot be achieved by any inductive account of science. Although, to my knowledge, he never thought of it in those terms, many aspects of Popper's methodology can be interpreted as primarily giving an account of rational pursuit rather than acceptance.[2]

The starting point for Popper is his well-known anti-inductivist stance, which is of course a thesis about acceptance. Popper rejects all 'inductivist' methodologies, i.e. accounts which take the rationality of empirical science to be based on giving reasons for the truth of scientific theories. He regards the problems facing these views as decisive (1959/1992: 6), and denies that there can be positive reasons to regard a theory as true or even probably true. According to Popper, the best we can do is to submit our theories to severe empirical tests, and reject those theories which we judge to have false empirical consequences (19-20, 65-7).[3] Thus, as regards acceptance (at least in the sense of

[2] This has also been noticed in passing by Schindler (2014: 495), who remarks that a strict Popperian could be characterised as someone who never believes in any theories but only ever pursues them.

[3] The emphasis on judgement is important to Popper, since he denies that theories can strictly speaking be falsified by comparing them to experience. The reason is that Popper thinks there can only be logical or rational relations between sentences. Thus, there can be no rational relations between theories and psychological states such as experiences. For this reason, he takes a theory to be falsified if it is inconsistent with those 'basic sentences' we regard as established empirical facts. When to accept basic sentences on

regarding theories as true or probably true), Popper's methodology is primarily negative: there are no positive reasons for accepting theories, only reasons for rejecting them. The best we can say of any theory is that it has stood up to all tests—so far.

Popper does however provide other kinds of positive advice. Since science cannot give us positive reasons for accepting theories, he instead claims that the aim of science should be to find theories which (a) have as much empirical content as possible but (b) have not actually been empirically falsified (95-7).[4] To unpack this claim, notice first that, for Popper, the empirical content of a theory consists in its *falsifiability*, i.e. of how many different possible empirical results it is inconsistent with. His motivation for this account is, that the more possible empirically distinguishable situations a theory rules out, the more precisely it circumscribes the actual empirical world. Given this account, we can restate the aim of science of science to be to formulate theories which permit those, and only those, empirical situations we actually encounter.

Popper's anti-inductivism of course prevents him from saying that we can ever have reason to think that we have *attained* this goal. Nonetheless, scientists can still adopt methodological rules which allow them to make *progress* towards this goal. In particular, he argues that when proposing new theories (or revising old ones) scientists should formulate theories which (i) are as yet unfalsified, i.e. consistent with all empirical facts known so far, but (ii) still increase the empirical content (i.e. the falsifiability) of their overall theoretical system. This is the reason, according to Popper, that scientists should and often do prefer theories which predict novel kinds of empirical phenomena (62-3),

the basis of experience and when to reject them as (say) unreliably produced or as a mere artefact is ultimately a conventional decision (21-22, 74-94). Furthermore, since it is always logically possible to resist the potential falsification of any particular statement, e.g. by suitably revising one's auxiliary assumptions, Popper insists that we must adopt conventions as part of our empirical methods that forbid those types of manoeuvres (19-20, 32-4, 61-2). (See also Lakatos 1970/1978: 20-31 on the conventionalist elements in Popper's view).

[4] See also Popper (1957/1972) for a similar account.

theories which make more precise predictions or theories which have a wider scope than previous ones (105-7). The reason is that these types of theories will tend to be more falsifiable. Of course, by 'preferring', Popper does not mean that scientists have reasons to accept the theory as true or regard it as more probable than its competitors. Instead, since the main practical implication of 'preferring' theories of this kind is that they should be prioritised for further testing and revision, it is natural to interpret Popper's recommendations as simply giving directions for pursuit, rather than any kind of acceptance.[5]

This interpretation also helps make sense of one otherwise peculiar claim of Popper's, namely that the most preferable theory is often the *less* probable one. His argument, briefly, is that the more falsifiable a theory is, the more possible ways there are for it to be wrong. For instance, since the theory that (Q) all planets move in circles entails that (S) all planets move in ellipses, Q is less probable than S.[6] But since any falsification of S will also falsify Q, but not *vice versa*, Q is more falsifiable than S (105-8). Now, if preference is here taken to involve some kind of acceptance, the claim that we should prefer the less probable theory seems baffling.[7] However, if it is instead taken to mean that we should prioritise the less likely theory for testing, the recommendation seems much more reasonable, at least given Popper's overall methodological view. If the only way empirical testing can help us make progress towards the goal of science is by refuting false theories, it makes sense that we should prioritise the more improbable (but not yet falsified) theories for testing.[8]

[5] Lakatos (1968/1978: 171-181) also argues that the main consequence of theories being "accepted" on this Popperian model is that it is accepted "for serious criticism", in other words that it has "testworthiness"; see Section 1.3.3 below.

[6] This assumes that some types of non-circle ellipses have a non-negligible probability.

[7] Popper of course denied that probability can be a measure of acceptability, so he saw this as a welcome result.

[8] I will reconstruct this argument more formally in Chapter 2, Section 2.6.

While we can thus make sense of some aspects of *The Logic of Scientific Discovery* by interpreting it in terms of pursuit, I do not claim that this interpretation covers all parts of Popper's writings. He sometimes claims, especially in his later work (esp. 1972, ch. 1),[9] that his account provides a non-sceptical solution to the problem of induction, in addition to giving a non-inductive account of the rationality of empirical science. His claim is, briefly put, that if one is forced to choose one theory for acceptance out of a range of competing options, for instance for the purposes of a practical application, one should choose the most severely tested and highly falsifiable hypotheses (what he calls 'highly corroborated' theories). Since, according to Popper, this is the most (indeed the only) rational way to empirically investigate a theory, accepting the most corroborated theory is the most rational thing one can do.

Here Popper clearly moves beyond the pursuit-interpretation I have proposed of his earlier views: choosing to rely on one theory is a form of acceptance, not pursuit. However, it is also a rather implausible claim. As many critics (e.g. Salmon 1981; Godfrey-Smith 2003: 67-70) have pointed out, the claim that it is rational to accept and use highly corroborated theories for practical applications is either trivial or unconvincing, depending on what the contrast-class is supposed to be. If Popper merely means that we should prefer highly corroborated theories to those that have been tested *and falsified*, this is of course correct. It is more rational to rely on a possibly true theory

[9] Additional evidence that Popper's view may have shifted in his later work comes from the fact that in at least two places in the 1959 English translation of *The Logic of Scientific Discovery* where Popper talks of theories being accepted, the original German does not bear out this interpretation. First, compare: "It may now be possible for us to answer the question: How and why do we accept one theory in preference to others?" (1959/1992: 91) and "Hier können wir nun auch die Frage beantworten, in welcher Weise die jeweils bevorzugte Theorie ausgezeichnet wird." [We can now also answer the question, in which way a given, preferred theory is distinguished] (1934: 64). Second: "What compels the theorist … is almost always the experimental *falsification* of a theory, so far accepted and corroborated" (1959/1992: 90) and "[…] ist fast immer die experimentelle *Falsifikation* einer als bewährt anerkannten Theorie" [… is almost always the *falsification* of a theory accepted [or: recognised] as corroborated) (1934: 63, original italics, underlining added].

than to rely on a definitely false one—but this is trivial. If, on the other hand, he also means that we should prefer highly corroborated theories to other unfalsified theories, his answer is unconvincing. Popper's official account is that this is because the most corroborated theory has stood up to the most severe standards of criticism possible (1972: 22): being highly falsifiable but resisting actual falsification. But since Popper's solution is meant to side-step the problem of induction, having withstood severe criticism cannot be taken to show that a theory more likely to be true or give us reasons to think it will be reliable for practical applications, and Popper explicitly denies this (1972: 21-3). Popper insists that there is nothing *more* rational than to prefer theories on the basis of severe criticism (1972: 27). But as Salmon (1981: 120-1) points out, the problem is that Popper has not given any account—and indeed seems to deny that there could be any account—of why this is *any more* rational than so many other possible but clearly unreasonable ways of choosing between the theories (flipping a coin, picking the theories with the fewest letters in it is name, etc.).

In my view, this problem arises because Popper conflates or confuses preferring theories for the purposes of further pursuit with preferring them for some kind of acceptance (viz. for practical purposes). The confusion arises because of his insistence that he can give a non-sceptical but anti-inductivist solution to the problem of induction. But, although Popper's bid to having solved the problem of induction thus fails, we may still be able to salvage insights from his account if we interpret it as an account of rational pursuit, rather than acceptance.

1.3.2. Kuhnian "Theory Choice"

In the closing chapters of *The Structure of Scientific Revolutions* and his well-known 1977 paper "Objectivity, Value Judgement, and Theory Choice" from *The Essential Tension*, Kuhn argues that scientific “theory choice” (or “paradigm choice” in *Structure*) is guided by both objective and subjective criteria. In particular, he denied that ‘the evidence’, i.e. the number of empirical successes and failures of the competing theories, is sufficient to determine the choice between them. In *Structure,* he even claims that scientists often embrace new paradigms “*in defiance of the evidence* provided by problem-solving”; instead, scientists adopt a new paradigm because they “*have faith* that the new paradigm will succeed with the many large problems that confront it, knowing only that the older paradigm has failed with a few” (1962/1996: 157-8, emphases added). In the 1977 paper, he tries to develop this account in more detail. Kuhn explains that scientists do rely on a number of common criteria—he mentions accuracy, consistency, scope, simplicity and fruitfulness as examples of standard criteria—in choosing between theories (1977: 321-2). However, he argues, scientists supporting different theories will disagree about how these criteria should be interpreted, and how they should be weighed against each other when they support different theories. Therefore, these criteria cannot form a shared, objective basis for theory choice (323). Rather, they function as *values* which *influence* (without *determining*) the choice of which theory to prefer (331).

As Šešelja and Straßer (2013: 11; following Hoyningen-Huene 1993: 239) highlight, part of Kuhn’s motivation stems from the kinds of competing theories or paradigms he has in mind. In the passages where he emphasises the role of “faith”, “personal and inarticulate aesthetic considerations” and “persuasion” (rather than argument) in theory choice, Kuhn is usually talking about the choice of between “an established theory and an upstart competitor” (1977: 322, cf. 331). Part of Kuhn’s argument in these passages is that it takes time and effort to develop a new paradigm to “the point where hard-headed arguments can be produced and multiplied” (1962/1996: 158) in favour of the newcomer. Thus, scientific revolutions, i.e. the replacement of an old paradigm by a new one, are only possible if scientists can be convinced to work on the new paradigm even though there are at present strong arguments in favour of the old paradigm. If scientists only pursued the theories that account best for the evidence at the time, this would stifle scientific progress.[10]

Furthermore, Kuhn develops this point to argue that it is a *good thing* that scientists disagree about how to interpret the values he claims guide theory choice. In most cases, the new competitor will not succeed and an explanation of apparent anomalies will be found within the old paradigm. If all scientists based decisions about theory choice on identical criteria, this would have one of two unattractive consequences: "With standards for acceptance set too low, [the scientists] would move from one attractive global viewpoint to another, never giving traditional theory an opportunity to supply equivalent attractions. With standards set higher, no one … would be inclined to try out the new theory… I doubt that science would survive the change" (1977: 332). Thus, according to Kuhn, it is crucial to scientific progress that theory choice is based on criteria which permit rational disagreement.

Since Kuhn here talks about "standards for acceptance", the upshot of his argument can appear baffling. For instance, the claim that a mere "faith" in the future promise of a theory or its fruitfulness can be a good reason to prefer it "in defiance of the evidence" over a well-established alternative seems clearly unreasonable if interpreted as a claim about which theory should be accepted. However, if we focus on the practical implications highlighted by Kuhn of 'choosing' between theories or paradigms, the choice in question seems to be better characterised as concerning pursuit rather than

[10] In some places in *Structure*, e.g. when comparing scientific progress to artistic and theological progress (1962/1996: ch. 13), Kuhn seems to throw doubt on whether there can be any progress which is not strictly paradigm-relative. I here follow e.g. Godfrey-Smith (2003: ch. 5) in interpreting Kuhn as regarding scientific revolutions as crucial to the ability of science to achieve progress.

acceptance.[11] For instance, in *Structure,* Kuhn notices that debates about whether to embrace a new paradigm or continue working within an old paradigm "are not really about relative problem-solving ability … Instead, the issue is which paradigm should in the future guide research on problems many of which neither competitor can yet claim to resolve completely" (1962/1996, 157-8). These choices, Kuhn claims, "must be based less on past achievement than on future promise" (*ibid.*). In other words, scientists are not choosing which paradigm should be accepted as true or even as the best problem-solver. Rather, they are choosing which paradigm should be pursued to provide potential solutions to unsolved problems.

This interpretation also helps make some of the otherwise puzzling implications of Kuhn's view seem more reasonable. First, when "theory choice" is interpreted as concerned with pursuit rather than acceptance, it seems much more plausible that we should focus on future promise rather than past achievements. After all, we may in some cases reasonably judge that, although a given theory/paradigm has been able to explain all previous anomalies, it has now exhausted its problem-solving power. Meanwhile, even if a new theory has not yet had any significant empirical successes, there can be good reason to suspect that it could solve the new problem, partly because it is still relatively undeveloped. Second, by distinguishing pursuit from acceptance, we can see that it can be reasonable to pursue a theory one does not accept. Scientists can pursue the new theory/paradigm because of its future promise and at the same time accept the established theory, or at least consider it the best available theory for practical applications, until there are reasons to think the upstart rival can in fact realise its promise. Thus, when Kuhn says that the new paradigm is embraced "in defiance of the evidence", we can interpret him to

[11] This is argued in detail by Šešelja and Straßer (2013). Sarkar (1983: 145) also makes this observation in passing.

mean that the new theory/paradigm is pursued although the evidence still favours accepting the currently dominant paradigm.

Although this reconstruction helps make sense of many of Kuhn's otherwise controversial views, it is unlikely that Kuhn himself was clear on the distinction. For instance, he suggests in the 1977 paper that each scientists should choose between theories by deciding what probability to assign to theories given the available evidence, where the evidence also includes considerations e.g. of fruitfulness and simplicity (1977: 328-9). Although this claim is made in the context of a rather ironic "hypothetical dialogue" between himself and one of his detractors, Kuhn only seems to be objecting to the idea that there is any kind of shared, objective algorithm which determines what this probability should be. He does not object to the basic idea that the theory choice, in the sense he is discussing, can be thought of in terms of deciding which is the most likely theory. This contributes to the impression that the early adoption of a theory or paradigm is always irrational: after all, how can it ever be rational to assign a higher probability to a theory based merely on an unproven "faith" in its future problem-solving power or on "personal aesthetic considerations"? The problem, in my view, is that Kuhn here conflates reasons for accepting a theory with reasons for pursuing it, thus implicitly assuming that any form of theory choice must involve regarding a theory as more likely. As I will argue later, reasons for pursuit can come apart in important ways from reasons for acceptance.

1.3.3. Lakatos and the Methodology of Scientific Research Programmes

Lakatos presents and develops his "methodology of scientific research programmes" as a modification of Popper's falsificationism able to answer objections highlighted by Kuhn and Feyerabend. I will here focus on two of Lakatos' main innovations.[12]

First, Lakatos rejects (what he calls) the 'naïve' falsificationist idea that an empirical test consists in simply comparing the results of an experiment with a given theory in order to determine whether the experiments falsifies the theory or not (1970/1978: 31). His motivation for rejecting this idea stems from the historical observations, highlighted by Kuhn, that scientists often seem to hold scientific theories which face known anomalies, either where they stick to an old theory after an apparent anomaly has been found or a new theory is adopted despite apparently conflicting with some of the available evidence. To avoid classifying such cases as irrational, Lakatos argues that theoretical appraisal is always directed towards series of theories, rather than individual theories in isolation. Whether a new theoretical development (or sometimes "problem-shifts") should be accepted as scientific or not can only be evaluated in relation to its place in a series of theories (*ibid*. 33-4). Proposed changes to an existing theory should be 'accepted' only if they are "theoretically progressive", i.e. if they make some novel predictions in addition to those made by the existing theory. Furthermore, a series of theories should be 'rejected' (that is, abandoned) only if it is superseded by an "empirically progressive" new series. A theoretical change, T_2, is empirically progressive relative to an existing theoretical series, T_1, only if (i) T_2 makes some novel predictions in addition to those made by T_1 and (ii) some of those novel predictions can be verified (35-6).

[12] Lakatos argues that some of his modifications were anticipated by Popper. However, Popper (1974: 999-1004) strongly denied that Lakatos had interpreted him correctly. I will focus on presenting Lakatos' views and ignore these exegetical disputes.

Second, Lakatos realises that the above requirements are vulnerable to a version of the tacking paradox (46): they allow for the construction of a theoretically progressive series simply by conjoining the existing theory with a "low-level hypothesis" (*ibid.*), e.g. one which merely states the existence of some new phenomenon, without any connection to the preceding theory. Lakatos wants to avoid classifying these and similar trivial modifications as progressive. Therefore, he also requires that theoretical changes should be evaluated in relation to their place in an overall *research programme*. This research programme consists, first, of a *negative heuristic*, which specifies a *hard core* of assumptions that no theoretical modifications should change. Second, research programmes have a *positive heuristic*, which suggests a range of natural or permissible modifications or auxiliary hypotheses which scientists should instead attempt to adopt in order to ward off potential refutations of the hard core. As long as a research programme is able to accommodate new anomalies by theoretical changes which are suggested by the positive heuristic, and avoids revising its hard core, a temporary lack of empirical progress (i.e. *verified* novel predictions) is acceptable. For Lakatos, as long as a research programme occasionally makes successful novel predictions, this is sufficient to turn the preceding "chain of defeats – *with hindsight* – into a resounding success story" (49).

It should be clear that Lakatos' methodology is, to a large extent, concerned with pursuit. Research programmes are defined by how they guide and restrict the direction of new research. The positive heuristic indicates in which direction new research should preferably proceed, while the negative heuristic forbids the pursuit of certain kinds of theories, namely those that would lead to revisions of the hard core of the research programme. However, on a superficial reading, especially of his classic paper "Falsification and the Methodology of Scientific Research Programmes" (Lakatos 1970/1978), it may still seem that the first aspect of Lakatos' methodology is concerned with the acceptance and rejection of the latest step in a theoretical series, since he uses these terms to formulate his account. However, it should be noticed that Lakatos is deliberately employing ‘acceptance’ and ‘rejection’ in a non-standard sense. To indicate this, he qualifies these (and many other) terms with inverted commas to indicate that he does not understand them in their usual sense. However, Lakatos is occasionally very clear e.g. that his use of terms like ‘falsification’ and ‘refutation’ should in no way be taken to imply a *disproof*, in an ordinary sense, or evidence against the truth of the latest step in a theoretical series; it only means that research programme has been superseded by a more progressive one (37, esp. note 5). The “pragmatic meaning of ‘rejection’ [of a research programme] … means *the decision to cease working on it*” (70, note 4). Likewise, while Lakatos takes successful novel predictions to be ‘verifications’ of a research programme, he adds in a footnote that “of course, a ‘verification’ does not *verify* a programme: it shows only its heuristic power” (51, note 4, original emphasis).

To understand how he uses these terms, one needs to look more closely at the practical implications of ‘accepting’ a theory in Lakatos’ sense, something which he himself does explicitly towards the end of his 1968/1978 paper “Changes in the Problem of Inductive Logic”. Here, Lakatos identifies three senses of ‘acceptability’, distinguished by subscripts. The first two, ‘acceptability$_1$’ and ‘acceptability$_2$’ correspond to what above was called theoretically and empirically progressive theoretical changes. A theory (i.e. the latest step in a theoretical series) is ‘acceptable$_1$’ if it is bold, that is, if it makes more novel predictions than the previous theory. It is ‘acceptable$_2$’ if some of these predictions are borne out. Lakatos explicitly states that the main implication of acceptance in either sense is that the research programme should be pursued further: “above all, this *acceptance$_1$* is acceptance for serious criticism, and in particular for testing: it is a certificate of testworthiness” (*ibid.*, 171, original emphasis). Similarly, “The scientist ‘*accepts$_2$*’ a theory for the *same* purposes as he *accepted$_1$* the bold theory before the tests” (175, original emphasis). The only difference is that an accepted$_2$ research programme is “regarded as a supreme challenge to the critical ingenuity of the best scientists” (*ibid.*), since it is more difficult to eliminate. The reason is that a merely bold (accepted$_1$) theoretical change can be eliminated by showing that all of its novel predictions are false, or by being superseded by an even bolder, new theory which also contains the previous theory’s predictions. If it is furthermore accepted$_2$, i.e. if any of its novel predictions are confirmed, it can only be rejected by being superseded by a new theoretical change which makes some further *successful* novel predictions beyond those of the current stage of the research programme (177).

Lakatos is clear that neither form of acceptance has anything to do with believing in the truth of a theory. For instance, one can ‘accept$_2$’ a theory (in the sense of the latest stage of a research programme) even if many of its predictions have been conclusively falsified, and it is thus known to be false (178). For similar reasons, it can be rational to simultaneously ‘accept$_2$’ mutually inconsistent theories. While these consequences may seem puzzling on the usual understanding of ‘acceptance’, they are fairly unproblematic if it understood that, for Lakatos, to accept a theory (in this sense) is merely to subject its associated research programme to further pursuit.

The two first kinds of acceptance are sharply distinguished from the third kind, ‘acceptance$_3$’. Lakatos is clear that only acceptance in this last sense has any implications for the future performance of a theory (182). He accepts that philosophy of science needs to say something about why scientific theories should be trusted in practical applications. Unlike Popper (cf. Section 1.3.1 above), he also accepts that neither his nor Popper’s account is able to address this problem if they insist on being strictly anti-inductivist (Lakatos 1974/1978: 159-167). While he thinks that ‘acceptance$_1$’ and ‘acceptance$_2$’ are rational and important forms of theoretical appraisal, he grants that they have no implications for any kind of ‘acceptance$_3$’, i.e. to accepting a theory in the usual sense, whether as true, empirically adequate or for a practical purpose. To construct a notion of ‘acceptance$_3$’ within his account (182-188), Lakatos proposes (“tentatively”) that one can adopt the additional principle—as a metaphysical conjecture—that the current state of a *long* theoretical series, i.e. a series which is the result of long chains superseding previous theories, should be regarded as closer to the truth and thus more likely to be reliable. One can then identify an ‘acceptable$_3$’ theory by taking the conjunction of all currently ‘accepted$_1$’ and ‘accepted$_2$’ theories, and weakening them (e.g. by restricting their scope) in such a way that they become consistent and no longer have any falsified predictions. He immediately points out such ‘acceptable$_3$’ theories will no longer be ‘acceptable$_1$’ or ‘acceptable$_2$’, since they make fewer predictions than their predecessors. However, he takes this to be unproblematic in this context because “here we do not aim at scientific growth but at reliability.” (183)

While there are many questions one could raise about this account of acceptance, even for a practical purpose, I shall not dwell on them here. For my purposes, I merely want to highlight that Lakatos’ distinction between ‘acceptance$_1$’/‘acceptance$_2$’ and ‘acceptance$_3$’ corresponds to the distinction between what I am calling pursuit and acceptance. Within the core parts of Lakatos’ system, research programmes are only ‘accepted’ (in the first two senses) in order to be tested and developed further, i.e. they are merely pursued. To construct any notion of acceptance (in the third and usual sense), he has to introduce some kind of independent (and, he stresses, merely conjectural) inductive principle. Furthermore, only this last kind of acceptance aims to identify theories which can be reliably applied. The rationale or guiding aim for the first two kinds of appraisal is that they are conducive to “scientific growth”, by which Lakatos means the proliferation of bold theories which can be further tested, thus (occasionally) leading to the discovery of new empirical facts.

1.3.4. The Post-Lakatosian Emergence of the Acceptance/Pursuit Distinction

So far, I have tried to show how the writings of Popper, Kuhn and Lakatos can be interpreted in light of the distinction between acceptance and pursuit. This distinction was brought out most explicitly in Lakatos' methodology, although still couched in the somewhat confusing terminology of 'acceptance' and 'rejection', and heavily peppered with inverted commas and subscripts. In subsequent commentary on these debates the distinction was formulated more clearly, and highlighted as important independently of the Popper/Kuhn/Lakatos debates. To round off this section, I will highlight two notable and influential cases.

The first to discuss the distinction (in writing) using the terms 'acceptance' and 'pursuit' was Laudan (1977: 108-114), although it seems to have been formulated independently by Laudan and Philip Quinn by at least 1973.[13] Laudan's motivation for introducing the distinction was to counter the argument (which he associated with Kuhn, Lakatos and Feyerabend) that science is irrational because scientists sometimes adopt theories which clearly have less empirical support than their competitors, or adopt multiple incompatible theories at once. As argued above, once the acceptance/pursuit distinction is made clear, there seems nothing irrational about scientists accepting the

[13] In an unpublished paper from 1973 Aldolf Grünbaum writes: "As Laurens Laudan and Philip Quinn have independently pointed out to me, we must be mindful here of the distinction between the rationality and irrationality of *belief* in a hypothesis on the one hand, and the rationality or irrationality of *pursuing* some kind of provisional research work on it, on the other" (quoted from Lakatos 1978c: 217). Laudan (1977: 234, note 38) in turn writes that "My analysis here [of the acceptance/pursuit distinction] owes much to discussions with Adolf Grünbaum". Although Quinn (1972) does not explicitly mention an acceptance/pursuit distinction, he discusses Lakatos' methodological appraisal of research programmes in terms of whether they should be "pursued".

dominant theory while pursuing incompatible newcomers. Laudan wanted to argue that there are rational standards for pursuit (he took Feyerabend and Lakatos to deny this) and that these can come apart from those governing acceptance.[14]

Around the same time, Ernan McMullin (1976: 422-5), commenting on the role of theoretical fruitfulness in Lakatos' methodology, drew a distinction between two kinds of theory appraisal: *epistemic appraisal* and *heuristic appraisal*. The first, *epistemic appraisal*, concerns whether the theory should be accepted in a realist sense, e.g. is there reason to think the theory conforms "reasonably well to the structure of the real" or is "one warranted in accepting the existence of the theoretical entities it postulates" (422). The second, *heuristic appraisal*, concerns the theory's "research-potential for the future". Questions relevant to the heuristic appraisal of a theory include: "How likely is it to give rise to interesting extensions? Does it show promise of being able to handle the outstanding problems (inconsistencies, anomalies, etc.) in the field? Is it likely to unify hitherto diverse areas or perhaps open up entirely new territory?" (423-4). McMullin points out that these two kinds of evaluation are distinct and can come apart in important ways: theories with high potential for future research may have a low epistemic status exactly because of their many as-yet-untested suggestions. Conversely, a well-established theory may hold little potential for further development because most of its promise has now been successfully borne out. He emphasises that heuristic appraisal is "of enormous importance in the planning of scientific work" (424) and criticises Lakatos and other falsificationists for having blurred the distinction between these two kinds of appraisal.

[14] Specifically, Laudan proposed that scientists should accept the theory which has the highest total problem solving power, while they should pursue the theory which has the highest current rate of new problem solutions. I discuss this account in Chapter 2, Sections 2.7.2 and 2.8.1.

1.4. Abduction and the Discovery/Justification Distinction

The debate over whether it is possible for philosophy to give any interesting normative account of the so-called “context of discovery” or whether philosophy should focus exclusively on the “context of justification” emerged during the 1950s, 1960s and 1970s in response to Hanson’s claim that Peircean abduction provides a “logic of discovery”. The discovery/justification distinction had been formulated by Reichenbach during the 1930s and by the late 1950s the distinction seems to have become a more or less self-evident principle among philosophers influenced by Reichenbach and logical empiricism more generally. So when historicist critics of the received views of logical empiricism framed their views as challenging this distinction—e.g. Hanson (1965: 60-1) and Kuhn (1962/1996: 9)—supporters of the received view reacted with a mixture of dismissal and frustration. For instance, Herbert Feigl (1970: 4) wrote: “I confess I am dismayed by the amount of—it seems almost deliberate—misunderstanding and opposition to which [the discovery/justification] distinction has been subjected in recent years”. Wesley Salmon, a student of Reichenbach, recalls that when he first read Kuhn’s *Structure* “I was so deeply shocked at his repudiation of the distinction between the context of discovery and the context of justification that I put the book down without finishing it” (1990: 325).

While both sides of this debate, in particular Hanson and Salmon, initially seemed to take the distinction for granted, it became increasingly clear towards the end of the 1970s that the discovery/justification distinction is rather ambiguous and conflates a number of distinct issues. Consequently, some philosophers proposed new distinctions which aimed to better capture the issues at stake in the debate, some of which introduced a “context of pursuit” or categories closely related to it. Despite this, many assumptions that Reichenbach had implicitly associated with the discovery/justification distinction continued to be influential.

I will start by discussing Peirce's writings on abduction and presenting what I take to be the most plausible interpretation of his mature view. I then trace the developments of Reichenbach's discovery/justification distinction. Finally, I discuss how this distinction influenced the reception of Peircean abduction, first by Hanson and Salmon in the 1960s, and then further on by other philosophers in the 1970s and 1980s.

1.4.1. Peirce on Abduction[15]

Throughout his career, C.S. Peirce argued for the existence of a third kind of inference or reasoning, in addition to deduction and induction, which he called abduction.[16] However, during his 50 years of writing about the topic, his account of what distinguishes abduction from the traditional forms of inference changed repeatedly. Commentators usually distinguish two main phases in Peirce's thinking about abduction: the first comprising his writings between 1860 and 1890, and the second, mature period emerging post-1890.[17]

In his early view, Peirce regarded abduction and induction as two separate kinds of probable inferences, differing in the kinds of premises they rely on, and the kinds of conclusions they support. Whereas induction infers general laws from observed regularities, abduction infers causes or explanatory hypotheses from their observed effects. However, the two inferences are similar in that both provide some degree of non-demonstrative, probable support for their conclusions.

[15] Parts of this section is based on the Peirce exegesis which I contributed to Stanley and Nyrup (*forthcoming*).

[16] Peirce's terminology varies. He also sometimes calls this third form of inference *retroduction*, *hypothesis*¸ *hypothetic inference*, or *presumption*. I follow commentators in simply using the term abduction (cf. Fann 1970: 5, note 19).

[17] This division was introduced by Burks (1946). Burks regards the period between 1890 and 1900 as a transitionary period between Peirce's early and mature views. See also Fann (1970), Niiniluoto (1999) and Psillos (2011a) for overviews of the development of Peirce's thought on abduction.

Post-1900, Peirce came to regard his previous discussion as badly confused. For instance, in a 1902 manuscript, he concedes that his earlier conception of abduction "had necessarily confused two kinds of reasoning", since "probability proper [i.e. empirically grounded probabilities][18] had nothing to do with the validity of Abduction, unless in a doubly indirect manner" (CP2.102).[19] Here, and in other writings around the same time, Peirce redraws the distinction between abduction and induction in terms of the role they play in scientific inquiry. He now classifies as inductive all inferences which provide empirically based probable support for a hypothesis. These include both arguments from random samples to general statistical regularities, which he sometimes calls 'quantitative induction', and inferences which support a hypothesis by confirming its observable consequences, which he calls 'qualitative induction', the latter resembling what he had earlier called abduction.

In his mature view, abduction instead becomes an inference, or line of reasoning, by which a new hypothesis is in some sense *introduced* into scientific inquiry. This does not (in itself) provide any kind of probable support for the hypothesis, except insofar as a hypothesis introduced by abduction can subsequently be supported through successful inductive testing. Thus, in a 1903 lecture series at Harvard, Peirce writes that abduction:

> is the only logical operation which introduces any new idea. … Its only justification is that from its suggestion deduction can draw a prediction which can be tested by induction, and that, if we are ever to learn anything or to understand phenomena at all, it must be by abduction that this is brought about. No reason whatsoever can be given for it, as far as I can discover; and it needs no reason, since it merely offers suggestions. (CP5.171).

[18] Peirce around this time tends to reserve the term 'probability' for empirically grounded judgements about how likely a hypothesis is to be true. He distinguishes this from 'likelihood', by which he means *a priori* judgements about likeliness. Peirce thinks the latter just expresses our prejudices and so should generally not be trusted.

[19] Following standard conventions, Peirce (1932-58) is cited in the format 'CP[volume].[paragraph]'. As these volumes are not chronologically organised, I will sometimes add the approximate year of writing to these citations.

Despite this merely suggestive role, Peirce emphasised that abduction is a form of reasoning, that it involves giving reasons and that there is a difference between good and bad abductions.

In some places, Peirce still seems to regard abduction as providing some kind of epistemic support for a theory. For instance, in the same lectures, he characterises abduction as “the operation of adopting an explanatory hypothesis” (CP5.189), arguing that it follows the following inference schema:

> The surprising fact, C¸ is observed;
> But if A were true, C would be a matter of course,
> Hence, there is reason to suspect that A is true.

Here, abduction is supposed to provide some “reason to suspect” that the explanatory hypothesis is true, suggesting that abduction provides at least a weak form of epistemic justification. In his Lowell lectures also given in 1903, Peirce similarly characterises abduction as “any mode or degree of acceptance of a proposition as a truth, because a fact or facts have been ascertained whose occurrence would necessarily or probably result in case that proposition were true” (CP5.603). Here, it seems that abduction can even provide enough justification to accept an explanatory hypothesis as true. Yet, in the preceding paragraph, Peirce insisted that “abduction commits us to nothing. It merely causes a hypothesis to be set down upon our docket of cases to be tried” (CP5.602). Here, “adopting” a hypothesis through abduction merely consists in giving it priority for testing. Furthermore, Peirce stresses in a number of places that it can be reasonable to test a hypothesis because it is easily falsifiable, rather than because it is likely to be true: “The best hypothesis, in the sense of the one most recommending itself to the inquirer, is the one which can be the most readily refuted if it is false. This far outweighs the trifling merit of being likely". (CP 1.120, c. 1896).

As McKaughan (2008) has persuasively argued, the support abduction provides for a hypothesis is first and foremost to justify giving a theory "a high place in the list of theories of those phenomena which call for further examination" (CP2.776, 1902), i.e. to provide reasons for pursuing it.[20] The normative standard guiding abduction is, according to Peirce, "economy", which in the context of research he takes to be "how, with a given expenditure of money, time, and energy, to obtain the most valuable addition to our knowledge" (CP7.140, 1879). Considerations of economy are crucial to Peirce because:

> Proposals for hypotheses inundate us in an overwhelming flood, while the process of verification to which each one must be subjected before it can count as at all an item, even of likely knowledge, is so very costly in time, energy, and money—and consequently in ideas which might have been had for that time, energy, and money, that Economy would override every other consideration even if there were any other serious considerations. In fact there are no others. (CP5.602, 1903).

Since there are only a limited amount of resources available for scientific research at any given point, scientists ought to prioritise them such that they would contribute the most to our knowledge. While considerations of how likely a hypothesis is to be true can still play some role in evaluating this, it is only in an indirect way: "the likelihood would not weigh with me directly, as such, but because it would become a factor in what really is in all cases the leading consideration in Abduction, which is the question of Economy—

[20] Thus, abduction cannot be assimilated to the modern notion of "inference to the best explanation", as pointed out by many Peirce scholars (Hintikka 1998, Minnameier 2004, Paavola 2004, Campos 2011 and Pietarien and Belucci 2014).

Economy of money, time, thought, and energy" (CP5.600). Furthermore, how considerations of likeliness affect the decision to pursue a hypothesis varies. In some cases, a hypothesis may be so unlikely that it is not worth spending energy on: "if a man came to me and pretended to be able to turn lead into gold, I should say to him, 'My dear sir, I haven't time to make gold" (*ibid.*). Peirce allows that if a hypothesis has a "marked probability of the nature of an objective fact, it may in the long run promote economy to give it an early trial" (CP6.534, 1901). However, the opposite may also be the case: if a hypothesis can be easily tested and "promises not to detain us for long, unless it be true", then "Sometimes the very fact that a hypothesis is improbable recommends it for provisional acceptance on probation" (CP6.533). Finding out which hypotheses are false can in itself be a valuable contribution, partly because it "leaves the field free" for further investigations (CP1.120-21, 1896). Furthermore, Peirce argues that the resources invested in pursuing a hypothesis will gradually yield diminishing returns of new knowledge so that, at some point, it will no longer be worthwhile to pursue it further (CP1.122). As I will argue in the next chapter, Peirce's account can plausibly be reconstructed in decision-theoretic terms.

A remaining tension in Peirce's characterisation of abduction concerns in what sense abduction is supposed to "offer suggestions" or "introduce" hypotheses into inquiry. Sometimes he characterises abduction as the process of *generating* or *formulating* a new explanatory hypothesis. In the Harvard lectures, Peirce claims that "Abduction consists in studying facts and *devising* a theory to explain them" (CP5.145, emphasis added) and that "Abduction is the process of *forming* an explanatory hypothesis." (CP5.171, emphasis added). At other times, however, he seems to characterise abduction as an inference where a hypothesis is (in some sense) *adopted* because it could potentially explain an otherwise puzzling set of phenomena. When abduction is written as an inference schema, such as the one above, the inferred hypothesis (‘A’) is explicitly mentioned in the premises, and so it seems to have been formulated *before* the abductive inference was made (Frankfurt 1958, Kapitan 1997). Furthermore, Peirce usually characterises reasoning as controlled and voluntary thinking, remarking in the 1903 Harvard lectures that “To criticize as logically sound or unsound an operation of thought that cannot be controlled is no less ridiculous than it would be to pronounce the growth of your hair to be morally good or bad” (CP5.109). But since we do not control which ideas occur to us, the question is how abduction could “suggest” ideas to us and still be considered an inference subject to any kind of normative criticism.

There are different proposals for how to reconcile the tensions between these two characterisations of abduction. McKaughan (2008) argues that the interpretation of abduction as generative reasoning should be rejected. However, some commentators (Fann 1970; Curd 1980; Psillos 2011a) have however argued that the tension can be resolved by construing abduction as a “dual process” (Psillos 2011a: 133), one which encompasses both the generation of a hypothesis and its adoption for pursuit. As Fann (1970: 42) notes, simply coming up with a new hypothesis is easy; the challenge in generating or formulating a new hypothesis is not merely to think of any hypothesis whatsoever. Peirce illustrates this point with the following example:

> Consider the multitude of theories that might have been suggested. A physicist comes across some new phenomenon in the laboratory. How does he know but the conjunctions of the planets have something to do with it or that it is not perhaps because the dowager empress of China has at that same time a year ago chanced to pronounce some word of mystical power or some invisible jinnee may be present” (CP5.172)

The problem for the physicist here is not to generate new hypotheses. Rather, since he cannot examine every conceivable hypothesis, the problem is to come up with a *good* hypothesis, one that it is worth considering further. The normative criteria for *generating* good hypotheses in this context are the same as the criteria for *adopting* a hypothesis for pursuit. Although the concrete thought processes in generating and adopting hypotheses of course differ, it makes sense to classify them together under the label 'abduction' because they aim to satisfy the same normative standards, distinguishing them from induction and deduction.

This interpretation also suggests an answer to the objection that generative reasoning is not subject to control. What we can control are decisions about *how* to attempt to formulate new ideas, and such choices can be subject to normative criticism on the basis of how effectively they lead to theories that are worth pursuing. A physicist who simply starts freely associating all sorts of possible causes of a puzzling phenomenon, in the manner parodied by Peirce above, would rightly be criticised by her colleagues for wasting their time. Even if she cannot strictly speaking control which ideas occur to her, she can still exert some control on the direction of her thoughts, choosing for instance to focus on known physical causes which it would be possible to test. In some contexts, more systematic heuristic strategies may be available or, if the desiderata on a satisfying problem solution are sufficiently constrained, it may be possible to deductively derive all hypotheses that could be of interest. The choice to adopt such methods can also be evaluated in terms of how effectively and efficiently they will generate an adequate range of pursuit worthy hypotheses.

Part of the choice here also concerns *when* or *whether* to generate new hypotheses and *when to stop*. Scientists at any given stage of inquiry will only be able to effectively consider a limited range of hypotheses and so it will often be reasonable to stop generating hypotheses once a few good candidates for pursuit have been found. Allan Franklin (1986: ch. 1) describes one such an example. In the 1950s, particle physicists were faced with a puzzling phenomenon: for certain observed decay patterns, the principle that all particles have a unique mass indicated that the decay products stemmed from the same particle, whereas the principle of parity conservation ruled this out (8-10). At a conference in 1956 where the problem was discussed, physicists proposed several possible explanations (34-5). After a range of possibilities had been discussed the chair "felt that the moment had come to close our minds" (35), i.e. to start thinking about how to test the salient proposals rather than generate new ones. The following year, experiments designed to test whether parity conservation is violated in these interactions confirmed (much to the surprise of most physicists at the time) that this was indeed the case. In this case, at least, the decision to stop generating new hypotheses seems reasonable for the same reason that it was reasonable to start pursuing the already generated hypotheses: the latter represented a more cost-effective use of their time and resources. Here, again, the normative criteria for generation overlaps with those for pursuit.

1.4.2. Reichenbach on the Context of Discovery and the Context of Justification

While Peirce's account of abduction can thus be interpreted as providing a plausible, unified framework for thinking about the generation and adoption of hypotheses for pursuit, the relevance of such an account was obscured for much of the twentieth century. Many philosophers argued that such questions concerned the "context of discovery" and was therefore not of relevance to philosophy, which they claimed is only concerned with the "context of justification". Exactly what the distinction is supposed to be was however unclear. It was used to refer to a number of distinct issues which were often conflated by both proponents and critics of philosophical accounts of "discovery". Distinctions that were subsumed under the discovery/justification distinction included:[21] (a) the process of making a discovery vs. providing reasons in favour of its truth e.g. in a scientific publication; (b) the generation, invention or formulation of a theory vs. trying to determine its truth value; (c) the actual (e.g. historical, psychological, sociological) processes of science vs. the rational reconstruction or normative analysis of those processes. Furthermore, the distinction was used to argue a number of different conclusions, including (i) that there can be no normative account of the generation of scientific theories; (ii) that normative philosophical analyses are essentially different from the empirical analyses of e.g. history or sociology of science; (iii) that only logical factors, as opposed to e.g. psychological or sociological factors, can play a role in normative analyses of science; and (iv) that the normative standards for the justification of scientific theories are independent of historical or social context.

The distinction between the context of discovery and justification was coined by Hans Reichenbach (1935b, 1938a, 1938b), and many of the conflations are apparent in his work. While variants of (some of) these distinctions can be found throughout the nineteenth century (Hoyningen-Huene 1987: 502-3), most debates from the 1960s onwards were based on Reichenbach's distinction and inherited many of its ambiguities. I will outline how the distinction occurs in Reichebach's work, before discussing how it influenced the reception of Peirce's account of abduction and the subsequent debates over the discovery/justification distinction.

The origin of Reichenbach's discovery/justification distinction can probably be located at a specific event, namely a conference attended by the Vienna circle and its associates in Prague, 31 August to 2 September 1934.[22] At this conference, Reichenbach

[21] See also the lists compiled by Nickles (1980c: 8-9) and Hoyningen-Huene (1987, 2006).

[22] The conference was a pre-meeting to the 1935 *International Congress for the Unity of Science* in Paris. A summary of the Prague conference was published in in *Erkenntnis* vol. 5 (p. 1-2), followed by papers

presented a summary of his views on probability (1935a) and gave the opening talk for a debate on induction.[23] Reichenbach, who had recently developed his solution to the problem of induction, argued that the problem of induction is philosophically significant and can be solved using probability theory. This argument was challenged by Popper, Carnap and Neurath, and Reichenbach seems to have formulated the discovery/justification distinction in the course of these debates. Two short discussion notes subsequently published in *Erkenntnis* mention the distinction, one by Popper (1935), who attributes the distinction to Reichenbach, and one by Reichenbach himself.

Reichenbach's note is entitled "On the Induction Machine".[24] It explicitly targets Neurath, who held that the problem of induction is a pseudo-problem, since there are no systematic rules for choosing between empirically equivalent theories. According to Neurath, this can only be done on the basis of conventional decisions, which have to be made on the basis of value judgements.[25] Reichenbach wanted to resist this argument. First, he argues that when scientists choose between theories, it cannot be a mere matter of convention, since these choices guide our predictions about the future.[26] Second, he claims that observational facts do point to some theories being more reliable than others and that "the procedure which science here uses [to choose between theories] is fundamentally rationalisable". In support of the latter claim, Reichenbach draws a

from the speakers at the conference and a number of discussion notes by participants.

[23] *Erkenntnis* vol. 5, p. 2.

[24] "Zur Induktionsmaschine", Reichenbach (1935b). All translations of the original German in the following are mine.

[25] Neurath (1913) is the original source of this argument. Don Howard (2006) has argued that this debate was part of a broader disagreement between the left- and right-wing of the Vienna circle concerning the role of values and politics in science and scientific philosophy. Neurath held that scientists had to rely on "auxiliary motives", i.e. political priorities, in making empirically underdetermined choices between theories, whereas Reichenbach tried to minimise the role of values in theory choice.

[26] This argument—that since we rely on induction for action we need an account of what makes some choices of theory more rational than others—is developed in more detail in Reichenbach (1938b: §38). It was also used later e.g. by Salmon (1981) to criticise Popper's methodology (see section 1.2.1. above).

distinction between two different parts of scientific work: “the process which the individual researcher uses in the discovery of new theories” and “the process in which he presents his theory publicly”. He calls the former “the process of discovery” [*Auffindungsverfahren*][27] which, he adds, “is hardly rationalisable, any more than the guessing of riddles is.” The latter, which he calls “the process of justification” [*Rechtfertigungsverfahren*] (172) or “context of justification” [*Rechtfertigungs-Zusammenhang*] (173), is on the other hand governed by rules – namely the principle of induction – which are in principle rationalisable, even though the presentations of scientists are never developed in complete rigour. He concludes by suggesting that just as it is in principle possible to build a “deduction machine”, one could equally build an “induction machine”.

Two points are worth noting. Firstly, as the distinction is drawn here it is between two different kinds of scientific *activity* rather than, for instance, between normative and merely descriptive analyses of science. The “process of justification” refers to the activity of publicly defending a theory, presumably in an academic context such as a talk or a paper, while the “process of discovery” seems to cover everything else a scientist might do to get to the point where they are able to formulate such a public defence. Second, Reichenbach’s purpose in drawing the distinction is to highlight that scientists sometimes make rule-governed decisions, based on observations, about which theories to accept. When Reichenbach compares the process of discovery to riddle guessing, the purpose mostly seems to be to grant his critics that not *all* aspects of science are rule-governed and rationalisable, in order to highlight one aspect which, according to Reichenbach, clearly is. Should it be replied that the process of discovery *is* more rational than mere

[27] *Verfahren* can alternatively be translated as “procedure” or “method”. To avoid the connotation of systematicity in these terms, I have here chosen the more neutral “process”.

riddle guessing, this would not weaken Reichenbach's main point, viz. that the process of justification is rationalisable.

Reichenbach, however, soon came to use this distinction to block other kinds of objections. This happened in a (1938a) reply to Ernest Nagel's (1936) review of Reichenbach's (1935c) *Warscheinlichkeitslehre*. In this review, Nagel questions whether Reichenbach's probabilistic account of induction adequately describes scientific reasoning. First, according to Nagel: "many physicists frankly admit that the notion of a theory being probable has no fixed, "objective", meaning for them; a careful search of scientific treatises reveals that the probability of theories is not discussed in them" (508). Second, he claims that "eminent men of science repeatedly assert that a theory is found satisfactory by them partly on esthetic grounds, partly because they know of no alternative theory, and partly because the consequences of the theory have been tested in accordance with a definite technique" (*ibid.*), citing Einstein as example (513). In effect, Nagel argues that even in the context that Reichenbach had claimed to be the rule-governed and rationalisable part of science, viz. public presentations in scientific treaties, scientists do not seem to follow the rules set out by Reichenbach.

In his reply, Reichenbach first argues that philosophers of science should not accept the authority of scientists with regards to their own epistemology: "a philosopher should carefully avoid asking a man of science why he believes in his theories. What we obtain by such an inquiry is a kind of religion for personal use, but not a philosophic argument." (Reichenbach 1938a: 34). The reason, argues Reichenbach, is that while scientists are experts in *applying* epistemological concepts, this does not necessarily enable them to analyse those concepts at a satisfactory level. While physicists may sometimes believe theories on "esthetic grounds", he claims that "scientific theories are better, mostly, than the epistemology combined with them; the esthetic taste of great physicists coincides in an astonishing way, with the postulates of the principle of induction" (35). Reichenbach does not want to dissuade scientists from relying on their aesthetic taste, if it in fact helps them identify good theories. However, he insists that insofar as these methods do work, "there *will be* better reasons" for accepting the theory, and "It is the task of the philosopher to show, by analytic methods, the inductive relations which justify a good hypothesis in respect to observed facts" (*ibid.*).

To summarise this argument, Reichenbach then introduces the discovery/justification distinction: "The *context of discovery* is to be separated from the *context of justification*; the former belongs to the *psychology of scientific discovery*, the latter alone is to be the object of the *logic of science*." (36). He cites his 1935 note, but does not provide a definition of the distinction. Instead, he illustrates it by arguing that the same distinction also applies in mathematical cases. In a geometrical problem, whether a proposed solution is correct is determined by the deductive relations defined by the problem. The context of justification analyses how and whether these deductive relations justify the solution. This should be distinguished from "The way we find the solution" which "remains to a great extent in the darkness of productive thinking, and may be influenced by esthetic considerations, or a feeling of "geometrical harmony"" (36). Reichenbach points out that "Nobody would here, in spite of this psychological fact, propound as a philosophical theory that the solution of geometrical problems is determined by esthetic points of view" (36). The distinction for inductive problems is then supposed to be analogous: while the way that scientists "find" theories may involve all sorts of psychologically important ways of thinking, this is irrelevant to the philosophical task of analysing the inductive relation between theories and observed facts.

We can see that Reichenbach here starts to conceptualise the "context of justification" in terms of a specific kind of analysis, i.e. normative analysis by the "logic of science”, rather than in terms of a specific kind of scientific activity, viz. defending a theory in public. He still understands this context as a specific subset of scientific activity—it is the “object” of the logic of science—but whether a given consideration counts as part of the context of justification now seems to be determined by whether it is amenable to normative analysis. For instance, Reichenbach remarks that of the considerations that Nagel claims scientists use to choose between theories, “I find that only the third argument presented there [the theory having been tested] belongs to this context [of justification]” (37), presumably because this is the only type of argument he regards as normatively plausible. Rather than identifying the context of justification as specific part of science and claiming that it is rule-governed, and thus amenable to philosophical analysis, Reichenbach instead seems to define the context of justification as those parts of science that are amenable to such analyses. Notice that it here becomes important for Reichenbach’s argument that the context of discovery is irrational, or at least non-rational, since this is what allows him to dismiss objections to his account of induction, e.g. “esthetic” arguments, as irrelevant to a normative, philosophical analysis.

This construal of the discovery/justification distinction was further developed in §1 of Reichenbach’s *Experience and Prediction* (1938b), which became the *locus classicus* for the distinction in later discussions. His main purpose here is to distinguish epistemology from empirical studies of science, such as psychology, sociology or history. He starts by noticing that they have the same starting point, namely the actual psychological or social processes which lead scientists to accept a theory. However, while psychology and sociology study these processes empirically, Reichenbach claims that epistemologists instead construct *rational reconstructions* (borrowing the term from Carnap 1928): they try to “construct thinking processes in a way in which they ought to occur” by constructing “justifiable sets of operations which can be intercalated between the starting-point and the issue of thought-processes" (1938b: 5). Although, as Reichenbach notices, scientific thinking rarely conforms to these rational reconstructions in practice, reconstructions show how the conclusions reached by scientists could in principle be justified logically. Since the rational reconstructions provide the normative standard, it will "never be a permissible objection to an epistemological construction that actual thinking does not conform to it" (6). On the other hand, rational reconstructions can be used to criticise actual thinking by showing that "certain chains of thought, or operations, cannot be justified", i.e. by showing that "it is not possible to intercalate a justifiable chain between the starting-point and the issue of actual thinking" (8).

To explain the notion of rational reconstructions, Reichenbach then introduces the distinction between context of justification and context of discovery, following his 1935 definition, as the difference between "the form in which thinking processes are communicated to other persons instead of the form in which they are subjectively performed" (6), which he also takes to correspond to "the well-known difference between the thinker's way of finding his theorem and his way of presenting it before a public" (*ibid*.). Although Reichenbach suggests that "epistemology is only occupied in constructing the context of justification" (7), he immediately points out that that "Even in the written form scientific expositions do not always correspond to the exigencies of logic or suppress the traces of subjective motivation from which they started" (7). Thus, even this is "only an approximation to what we mean by the context of justification" (*ibid*.).

At this stage, Reichenbach's definition of the context of justification makes it practically indistinguishable from rational reconstructions (or logic of science): strictly speaking, the *true* context of justification is *constructed* by philosophers. The part of scientific activity—public presentations— which his earlier definition of the context of justification marks out, only approximates these ideal rational (re)constructions. In proposing rational reconstructions, philosophers attempt to formulate the ideal context of justification which is only imperfectly approximated by the scientific practice of publicly presenting and defending a scientific theory.

This definition of the context of justification, however, introduces some ambiguity into what exactly the context of discovery denotes. On the one hand, it can encompass the whole contrast class to the ideal, rationally reconstructed context of justification, i.e. all parts of scientific thinking as described by empirical disciplines including the imperfect approximations to the context of justification contained in public presentations. On the other hand, the context of discovery retains the connotations of being whatever is not concerned with publicly justifying the acceptance of a theory. In particular, "discovery" is associated with the "finding" of a theory.

These ambiguities in turn meant that Reichenbach had an ambivalent attitude towards thinking which aims to introduce a theory. As Nickles (1980c: 10-15) and Curd (1980: 210-11) have pointed out, Reichenbach sometimes suggests that the generation of new theories prior to empirical testing can also be captured by his rational reconstruction, i.e. his probabilistic account of induction. In the reply to Nagel, Reichenbach argues that "even before the test [of a theory] there must be facts on which the theory is based; and there must be, also before the test, a net of inductive relations leading from the facts to the theory—else the theory could not be seriously maintained." (1938a: 37).[28] Reichenbach does not disagree with Peirce that there can be good reasons for adopting a theory for further testing. However, he maintains that, insofar as it is reasonable to propose a theory before testing, this must be because it already has some probabilistic

[28] Reichenbach is most likely alluding to Peirce's claim that there is a significant difference between abduction and induction. Nagel (1936: 508) mentions Peirce in passing. Apparently in response to this, Reichenbach remarks: "I admire Charles Peirce … but just his remarks concerning what he calls "abduction" suffer from an unfortunate obscurity which I must ascribe to his confounding the psychology of scientific discovery with the logical situation of theories in relation to observed facts" (1938a: 36).

support from our empirical background knowledge. So, unlike Peirce, Reichenbach does not think there is a principled difference between the kind of support a theory can have before and after testing. He also uses this account to avoid the conclusion that scientists, such as Einstein or Newton, were simply making guesses which were no more rational than so many other possible proposals when they developed their theories (1938b: 381-2). Reichenbach agrees that Einstein's theorising was a significant achievement and argues that this can be captured by his theory:

> Why was Einstein's theory of gravitation a great discovery, even before it was confirmed by astronomical observations? Because Einstein saw—as his predecessors had not seen—that the known facts indicate such a theory; i.e., that an inductive expansion of the known facts leads to the new theory. (382)

Reichenbach is here relying on the discovery/justification distinction understood as a distinction between empirical descriptions of actual scientific thinking and normative rational reconstructions. Whereas Einstein's actual thought processes might not follow any systematic rules, we can still judge them rational because it is possible to rationally reconstruct an argument from the known facts to the theory. Similarly, Reichenbach argues that when scientists claim to be guided by conceptions of "natural hypotheses" or the "harmony of nature", insofar as they are rational, this will be because their decision to propose those theories can be rationally reconstructed on the basis of his inductive principle (403).

At other times, however, Reichenbach dismisses reasoning involved in developing a hypothesis on the basis that it belongs to the context of discovery and is therefore philosophically irrelevant. For instance, in his *Philosophic Foundations of Quantum Mechanics,* Reichenbach at one point discusses the lines of theorising that led to the formulation of the modern quantum theory. He here maintains that since we could not know in advance how to develop quantum mechanics from classical mechanics, this "could not be found by logical reasoning" (1944: 66). Even though physicists such as de Broglie, Schrödinger and Heisenberg "felt obliged to adduce logical reasons for the establishment of their assumptions" and this "apparently logical line of thinking was an important tool in the hands of those who were confronted by the task of transforming ingenious guesses into mathematical formulae" (66), this is merely part of the context of discovery. The analogies employed by Schrödinger may show a hypothesis plausible and this can be "an excellent guide within the context of discovery" (71), but Reichenbach quickly adds that these analogies rely on assumptions specific to the one-particle case which cannot be assumed to hold generally. Therefore, he concludes that the epistemic support for quantum mechanics rests solely on its empirical success, as judged within the context of justification. Since these analogies were merely a tool for the discovery of the theory, Reichenbach concludes that they are normatively uninteresting. He remarks that discovery "runs through 'series of inferences which are deeply veiled by the darkness of instinctive guess'" (1944: 67, quoting a letter from Schrödinger) and so going into "an exact analysis of Schrödinger's ideas would lead us too far from the purpose of a merely logical analysis with which this book is concerned" (ibid.). While Reichenbach recognises that certain kinds of considerations may show a hypothesis plausible and that this can play an important role in scientific reasoning, he dismisses trying to provide any deeper philosophical analysis of these lines of reasoning.

To summarise, Reichenbach's distinction between the context of discovery and context of justification evolved between 1935, when it was a distinction between two different kinds of scientific activity, and 1938, when it became a distinction between normative and descriptive analyses of science. However, many of the earlier connotations continued to influence Reichenbach’s thinking. First, he often (though not always) assumed that reasoning concerned with generating hypotheses belongs to the context of discovery and therefore is unrationalisable. Second, he assumed that the context of justification consists of rational reconstructions in terms of the relations of deductive or inductive support. He does not seem to have considered the possibility that reasons for “seriously maintaining” a hypothesis prior to testing could consist in anything else than it being inductively supported by known empirical facts. As we shall see below, these assumptions continued to influence the post-positivist debates during the 1960s and 1970s.

1.4.3. Hanson and Salmon on the ‘Logic of Discovery’

An early and influential critic of the logical empiricist use of the discovery/justification distinction was N. R. Hanson who, in a series of papers (1958, 1960a, 1965), argued that philosophers can and should try to analyse the ‘logic of discovery’. What exactly this means was something he struggled with. In the earliest paper, Hanson sometimes seems to be interested in the *process* by which new hypotheses are formulated in scientific practice (1958: 1083). Thus, he criticises both “inductionists”, for proposing that new hypotheses are proposed through induction-by-enumeration (1958: 1080-1), and hypothetico-deductivists for suggesting that formulating a hypothesis is dependent on “intuition, hunches, and other imponderables” (1083). While he denied that it was possible to formulate “a manual to help scientists make discoveries” (1073), he proposed that Peircean retroduction—understood as reasoning from a surprising phenomenon to an explanatory hypothesis—gives an account of how scientists “catch” their hypotheses. Similarly, in a later paper (1965), he argues that, while the formal criterion for successful retroduction is the same as for successful hypothetico-deductive testing (namely to show that the hypothesis entails empirically confirmed phenomena), they differ temporally and thus in the reasoning task facing a scientist (1965: 54). In the hypothetico-deductive case, the task is to derive testable predictions from a given hypothesis, while in the retroductive case, the task is to reason backwards from a given phenomenon to a hypothesis capable of explaining it.

However, at times, Hanson also denied that he wanted to give an account of the process of formulating a new hypothesis.[29] In many places, he instead claims that his aim is to argue that there is a logical difference between "reasons for accepting an hypothesis H" and "reasons for suggesting H in the first place" (1958: 1073). Amongst the latter he includes things like analogical arguments, symmetry considerations, simplicity and "aesthetic elegance" (1958: 1078; 1965: 61). He argues that these kinds of arguments had been neglected by philosophers because they were deemed part of the context of discovery; as we saw above, Reichenbach did sometimes dismiss giving a normative account of analogical arguments for this reason. Since Hanson seems to accept that these types of arguments belong to the context of discovery, he took himself to be arguing for a logic of discovery. The difference, according to Hanson, between the two kinds of reasons is that "reasons for suggesting H in the first place" merely "make H *a plausible conjecture*" (1958: 1074). He argues that, since these kinds of reasons make it reasonable to suggest a hypothesis but "could never *by themselves* establish an H", they "must be different in type" (1079) from reasons for accepting a hypothesis.

In responding to Hanson, Salmon (1967) concedes that the "standard answer", i.e. simply ruling that there can be a logic of discovery, "is, nevertheless, a very disappointing one" (111). He agrees with Hanson that it would be unsatisfactory if "logical analysis can

[29] In a critical commentary, Schon (1959) pointed out that Hanson (1958) seemed to discuss both *reasons* for suggesting hypotheses and the *processes* by which hypotheses were formulated. Hanson (1960b) acknowledges having been unclear and states that he wishes only to give an account of the former.

be used for dissection of scientific corpses, but it cannot have a role in living, growing science" (111-12). To answer this worry, he distinguishes: "(1) thinking of the hypothesis, (2) plausibility considerations, and (3) testing and confirmation", explaining: "There is, presumably, a time between first thinking of a hypothesis and finally accepting it during which we may consider whether it is even plausible. At this stage we are trying to determine whether the hypothesis deserves to be seriously entertained and tested or whether it should be cast aside without further ceremony" (113-14). Thus, Salmon agrees with Hanson that plausibility considerations play an important role in evaluating whether theories should be further tested, but criticises him for sometimes conflating these with the process of first thinking of a hypothesis. The latter he identifies with "discovery" or "psychology of discovery" (114), indicating that the process of formulating hypotheses is not a topic for logical analysis. Salmon goes on to argue that plausibility considerations can be naturally fitted into a Bayesian account of scientific reasoning by identifying them with estimates of the prior probabilities of hypotheses (118). He also argues that the kinds of arguments Hanson highlights (analogies, symmetry considerations, etc.) can provide legitimate reasons for regarding a hypothesis as having a higher prior probability (125-129).

Both Hanson and Salmon come very close to the account Reichenbach (sometimes) gives of how new hypotheses can be normatively evaluated (i.e. rationally reconstructed). All three agree that there can be reasons for "seriously maintaining" or "suggesting" a hypothesis prior to any testing of it and that the same kind of reasons can *also* contribute as reasons for accepting the hypothesis after its testing (Hanson 1958: 1079). Finally, they agree that the reason why it is reasonable to use these types of arguments to provide pre-test support for a hypothesis is that there are empirical ("inductive") reasons for regarding them as relatively reliable (Hanson 1958: 1079; Salmon 1967: 126-28). This type of pre-test assessment essentially corresponds to what I call reasons for pursuing a hypothesis. However, their account differs from Peirce's, since Hanson and Salmon both seem to uncritically accept Reichenbach's assumption that the reasons which support the pursuit of a theory are simply weak reasons for accepting it.

1.4.4. New Distinctions: The Discovery/Justification Debate Post-1970

This account influenced many subsequent attempts to move beyond the discovery/justification distinction. We have already seen that Salmon takes one such step by limiting "the psychology of discovery" to the initial thinking of a hypothesis and including plausibility considerations explicitly in the logic of justification. Carl Kordig (1978: 114) similarly distinguishes between (1) the "initial thinking" of a hypothesis, (2) its plausibility and (3) its acceptability. Kordig also construes plausibility as relevant to whether a hypothesis should be pursued: "Hypotheses are initially plausible prior to test. They are worthy of further consideration, though not yet acceptance. Consideration of one hypothesis rather than another is often reasonable. After initially thinking of an hypothesis, and yet before its test, good reasons often support its plausibility. They support its further exploration, its being seriously entertained" (115). Furthermore, he thinks plausibility and acceptability rely on the same kinds of reasons: "Good reasons are relevant to both plausibility and acceptability. They support acceptability. Prior to experimental test, they also support plausibility" (*ibid.*).[30] Finally, like Salmon and (sometimes) Hanson, Kordig maintains that these reasons are not relevant to the initial thinking of a hypothesis: "Good reasons are not required to think … Plausibility and

[30] The main way Kordig's position differs from Salmon's is that Kordig does not identify acceptability with probability and does not identify plausibility with prior probability.

justification require reasons. Initial thought does not. Initial thought is prior to plausibility and justification." (114).

This last assumption is challenged by a further refinement introduced by Robert McLaughlin (1982), who recasts the distinction as one between (1) the invention of hypotheses, i.e. the "initial thinking" or construction of a hypothesis, and (2) their appraisal. The latter includes (2a) enhancement arguments, corresponding to Hanson and Salmon's the plausibility considerations, and (2b) the confirmation of hypotheses through testing. McLaughlin's central argument is that the invention of hypotheses often happens through what he calls an "advancement" argument (77-8). He takes these to rely on the same kinds of premises as enhancement/plausibility arguments, i.e. simplicity, analogy or symmetry considerations. In fact, the only difference is that in advancement arguments, these considerations are used to guide one's invention of a new hypothesis rather than to support an already formulated hypothesis. Furthermore, McLaughlin argues that the rationale for generating hypotheses in this way—e.g. by formulating hypotheses on analogy with known phenomena or by through assuming symmetries—is the same as for enhancement arguments: hypotheses generated in this way will already have a plausibility argument in its favour, namely the one it was generated through. There is no normatively significant distinction between advancement and enhancement arguments. Consequently, this removes "the attraction of the thesis that epistemology is properly concerned solely with the rational reconstruction of appraisal" (78). McLaughlin in effect points out that if one allows a normative account of pre-test evaluation, there is no reason why the same normative account cannot also be applied to evaluate methods, heuristics and choices concerned with the invention or generation of new hypotheses. Since the purpose of generating new hypotheses is presumably to subsequently evaluate whether the hypotheses thus generated should be accepted, the normatively optimal methods of generation are just those that generate hypotheses worth spending time trying to evaluate, e.g. through further testing. McLaughlin retains the assumption, however, that reasons for pursuit can be reduced to reasons for acceptance, i.e. to showing the hypothesis to have a high degree of prior probability: “the goal of a law discovery process, in science at any rate, is surely the discovery of hypotheses which will turn out to be highly-confirmed in the context of appraisal; that is, *plausible* hypotheses” (96).

This assumption, the last remnant of the ambiguities introduced by Reichenbach’s distinction, was challenged by Larry Laudan (1980) and Martin Curd (1980). Laudan argues that the previous debate had been muddled by conflating what Laudan calls the “context of pursuit” with the discovery of hypotheses in the narrow sense “as concerned with ‘the *eureka* moment’, i.e., the time when a new idea of conception first dawns” (174)—in my terms, with questions concerning hypothesis generation. Laudan agrees with Salmon that Hanson’s ‘logic of discovery’ leaves the generation of hypotheses unanalysed and rather concerns which hypotheses are worthy of pursuit. However, Laudan argues that Hanson’s critics had been wrong to assume that considerations regarding pursuit therefore belong to the “context of justification”, i.e. that they provide reasons for acceptance (*ibid.*). Nonetheless, Laudan still argues that methods for hypothesis generation can only be justified if they show a hypothesis more likely. Since he doubts that there are methods which will reliably generate likely hypotheses, he is sceptical towards giving normative accounts of hypothesis generation.

Like the previous authors, Curd (1980) recognises that the traditional context or logic of discovery conflates the prior assessment of theories and methods for the generation of theories (203). As Laudan did, he furthermore points out that the most relevant form of prior (i.e. pre-test) appraisal concerns whether hypotheses are worthy of pursuit, rather than their prior probability (203-4). Finally, like McLaughlin, Curd argues that the criteria for the prior appraisal of theories also provides a basis for rationally reconstructing (normatively evaluating) the "inferences scientists make in reasoning to their hypotheses" (205), i.e. for evaluating the choices and strategies for generating new hypotheses.

1.5. Conclusion: Better Distinctions

In this chapter, I have reviewed two strands of twentieth-century philosophy of science, outlining how the distinction between pursuit and acceptance emerged through these debates. To conclude, I want to highlight a number of insights which will inform my discussion in later chapters. While the discovery/justification distinction ended up conflating a number of independent issues, when disentangled it contains three orthogonal types of distinctions that are still worth preserving.

The first is the distinction between a descriptive analysis of scientific practice and a normative analysis of whether and to what extent that practice is reasonable. While certain forms of philosophical naturalism deny that the descriptive and the normative can be neatly separated, for the purposes of this thesis I will assume (though not defend in any detail) that it makes sense to normatively analyse and evaluate scientific practice.

The second distinction concerns which type of scientific activities is being analysed. In this chapter I have discussed three main types: (1) the generation of theories, hypotheses or models, (2) adopting one or more of these for further pursuit and (3) accepting them as, in some sense, correct. Notice that I avoid the terminology of 'contexts' here. In my view, this has the misleading connotation that these activities are in some sense separated, e.g. as occurring at different times or "stages" of scientific research or in different conversational contexts. But, of course, they are often intertwined and overlapping in practice. Furthermore, as McLaughlin, Curd and (sometimes)

Reichenbach recognised, there are no principled reasons against subjecting any of these types of activities to normative analyses.

In some cases, further subdivision of these categories can be useful. As already mentioned, acceptance covers a range of different attitudes one can take towards a hypothesis, including regarding it as true, partially true, empirically adequate or adequate for some practical purpose. While it will sometimes be important to distinguish these, grouping them together under the label ‘acceptance’ can be useful for the purposes of contrasting them with pursuit. What distinguishes these forms of acceptance from pursuit is a willingness to rely on the theory or model as a premise for practical or theoretical reasoning and some commitment to how well it fits the world. Pursuit does not entail such a commitment. Instead, pursuit usually aims to find out how well the item being pursued fits the world or, more generally, *whether* it is reasonable to rely on it in practical or theoretical reasoning.[31] Pursuit can cover a number of activities, including testing the hypothesis, refining it empirically (e.g. determining empirical constants through experiments)[32] developing it theoretically (e.g. by clarifying its core concepts, extending it to apply to new phenomena) and so on. Within the category of ‘generation’ I include choices concerned with generating new hypotheses, including using specific heuristics or strategies for generating hypotheses, focusing on developing specific theoretical ideas further and deciding *not* to generate further hypotheses. As the last point indicates, the generation and pursuit of hypotheses are to some extent continuous with each other.

Third, one can distinguish which types of normative criteria or standards are being used to evaluate the activity being analysed, both from the descriptive perspective, i.e. on

[31] Thus, pursuit may aim to develop theories or models which are helpful for particular practical problems or for building technologies, rather than ones which represent the world accurately in any sense. I say more about the distinction between epistemic and practical goals for pursuit in Chapter 2, Section 2.4.1.

[32] As van Fraassen (1980: 73-74) and Franklin (1986) point out, not all experiments aim to test a hypothesis; some are better characterised as developing the theory by empirical means.

the basis of which criteria do scientists themselves make decisions, and from the normative perspective, i.e. which criteria should be used to evaluate whether a given activity is reasonable or not. As argued in this chapter, there are important normative criteria to consider apart from the probability of a hypothesis or its epistemic support more generally. Following Peirce, choices about pursuit and generation of hypotheses should not be evaluated according to the same criteria as acceptance. These decisions should instead be evaluated according to the ‘economy of research’, i.e. how much epistemic output they promise to return for the limited resources available for research. I will defend and develop this account in more detail in the next chapter.

Chapter 2. Reasoning about Pursuit: A Decision-Theoretic Approach

2.1. Introduction

As we saw in the previous chapter, during the 1970s philosophers of science started emphasising that in addition to reasoning about whether a theory should be accepted or rejected, there is a distinct modality of theory appraisal: whether it should be pursued, i.e. whether one should spend time and resources testing and developing it further. As pointed out, many earlier writers had tended either to focus exclusively on acceptance or else conflated pursuit and acceptance. Yet today, 40 years later, philosophers of science have still paid relatively little attention to the normative standards governing when it is rational to pursue a theory or not (with some notable exceptions to be discussed below). By contrast, discussions of acceptance and related notions, such as confirmation and different types of inductive inferences (hypothetic-deductive, inference to the best explanation, etc.), have been lively throughout.

The purpose of this chapter is to present my positive account of reasoning about pursuit and to compare it to other accounts in the literature. It defends a broadly decision-theoretic approach to justification for pursuit, inspired by C.S. Peirce's mature account of abduction and the 'Economy of Research'. In brief, I analyse justification for pursuit as based on how to achieve the highest epistemic output from the limited resources available for scientific research. While this idea may seem obvious or trivial, I shall argue that taking it seriously highlights important lessons which have often been overlooked in previous discussions of pursuit. Firstly, it corrects some common but, on reflection, clearly mistaken assumptions about pursuit. Secondly, it highlights some shortcomings and lacunae in recent, more nuanced accounts.

This chapter is structured as follows. I start by discussing two general objections to developing a normative account of pursuit: in Section 2.2, that there are no (or only very weak) normative constraints on what it is rational to pursue and, in Section 2.3, that reasons for pursuing a theory are identical to, or at least very closely connected to, reasons for accepting it. In both cases I argue, first, that the assumptions are descriptively mistaken about scientific practice and, second, that there are relatively intuitive arguments which show them normatively mistaken as well. In Section 2.4, I argue that the replies to these objections can be captured by a consequentialist conception of pursuit worthiness. Next, I show how this can be developed into a family of relatively simple decision-theoretic models (Section 2.5), highlight some of the merits of these (2.6), and discuss some ways they can be extended and modified (2.7). Finally, in Section 2.8, I compare this approach to other extant normative accounts of pursuit, highlighting ways in which my account corrects and complements these before I conclude, in Section 2.9, by discussing some limitations of my approach and avenues for further development.

2.2. Are There Any Normative Constraints on Pursuit?

That there could be an interesting normative account of pursuit is sometimes dismissed on the grounds that there are supposedly no, or only very weak, normative constraints on what scientists should pursue. In its simplest form, this idea can be stated as the claim that scientists should be free to pursue whatever they want. In this section, I will first highlight some examples from scientific practice where scientists recognise that there are normative constraints on pursuit and give reasons for why theories should be pursued or not. This makes developing a normative account of pursuit *prima facie* reasonable. I then consider some principled arguments against there being normative constraints on pursuit which I argue are unpersuasive.

2.2.1. Examples from Scientific Practice

The notion that pursuit is normatively unconstrained is at odds with scientific practice. First, scientists do in fact sometimes defend the pursuit of theories, something which would be superfluous if there were no normative standards for pursuit. For example, Achinstein (1993) discusses Niels Bohr's early quantum theory, published in three papers from July to November 1913. In these papers, Bohr notices that Rutherford's atomic model faced certain theoretical problems. To resolve these, Bohr proposes several at the time radical proposals for revising the model. These include the assumptions that electrons can only move in certain stable orbits and that radiation is only released when an electron "jumps" between these orbits, violating classical electrodynamics. Finally, Bohr derives an expression from the revised model for the wavelength of light emitted from a hydrogen atom similar in form to the experimentally determined Balmer formula (Achinstein 1993: 95-98).

As Achinstein points out, Bohr does not claim his theory to be correct or even probable. Furthermore, in letters to a colleague, Bohr acknowledges its speculative character: "For the present I have stopped speculating on atoms. I feel it is necessary to wait for experimental results" (Bohr to Mosely, 21 November 1913, quoted from Achistein 1993: 98). While the positive results gave him "hope to obtain knowledge of the structure of the systems of electrons surrounding the nuclei in atoms and molecules", he also emphasised that he did not think this was achieved by "the result which I mean I can obtain by help of my poor means, but only of the point of view … which I have been led to by considerations such as those above" (Bohr to Hevesy, 7 February 1913, Achinstein *ibid*.). However, Bohr did regard his articles as a defence of the model: he was trying to convince his colleagues that it was worth pursuing, by testing the model's predictions experimentally and trying to further develop its “point of view”, i.e. its general idea. Since Bohr gives arguments in favour of the model, he recognises that his peers, at least implicitly, endorse some constraints on what it is worth pursuing.

Secondly, scientists in a given field only pursue, and indeed *could* only pursue, a limited number of theories. They need to prioritise their time, efforts and resources and consequently regard some proposals as clearly a waste of these. For example, McKaughan (2007: 20-24, 291) points out that the particle physicist Steven Weinberg in his *Dreams of a Final Theory* admits that the theories which he proposes and works on are “of limited validity, tentative and incomplete” (Weinberg 1992: 13). Weinberg nonetheless insists that these theories are “worth taking seriously” (Weinberg 1992: 103); despite their likely flaws, he thinks that it is reasonable for him and his colleagues to spend their time working on these theories. He also makes clear that not every tentative theory has this status:

> I receive in the mail every week about fifty preprints of articles on elementary particle physics and astrophysics, along with a few articles and letters on all sorts of would-be science. Even if I dropped everything else in my life, I could not begin to give all of these ideas a fair hearing. So what am I to do? Not only scientists but everyone else faces a similar problem. For all of us, there is simply no alternative to making judgements as well as we can that some of these ideas (perhaps most of them) are not worth pursuing. (Weinberg 1992: 50)

As Weinberg points out, this is a general problem: due to the limited time and resources, scientists are forced to make judgements about which theories should be prioritised. We saw in the previous chapter that Peirce also took abduction to be based on economic considerations. In fact, Peirce at one point gives an argument very similar to Weinberg’s. Having claimed that abduction shows that a theory should be given “a high place in the list of theories of those phenomena which call for further examination”, he pre-empts an objection:

> If this is all his conclusion amounts to, it may be asked: What need of reasoning was there? Is he not free to examine what theories he likes? The answer is that it is a question of economy. If he examines all the foolish theories he might imagine, he never will (short of a miracle) light upon the true one. (CP 2.776)

As the Bohr case illustrates, scientists do not make these judgements blindly: they consider some theories more reasonable to pursue than others and give arguments for why their preferred theories should be pursued.

The fact that scientists do in fact recognise the need to make such judgements, and present arguments for doing so, indicates that this is an aspect of scientific reasoning which is subject to normative constraints which philosophers can attempt to spell out. As with any aspect of scientific practice, it can of course turn out that no plausible normative account can be given. Weinberg’s judgements about pursuit worthiness may simply boil down to idiosyncratic preference. However, Weinberg (and Peirce) do highlight a plausible basis for making normative judgements about pursuit: there is only a limited amount of time and resources available for research, so scientists should focus on those theories that promise the most epistemic output for the time and resources invested in them.[33] I conclude that it is reasonable to try to formulate such a normative account of pursuit unless there are principled reasons why it could not succeed. I will now consider some candidate reasons for this.

[33] As I argue below (Sections 2.3.2 and 2.6), even if scientists had infinite resources, there are further normative constraints on pursuit, namely whether anything interesting could be reliably learned from pursuing the theory.

2.2.2. Principled Arguments Against Normative Constraints on Pursuit

Lakatos seems, in response to criticism from Feyerabend (1970), to have adopted a rather deflationary attitude towards the implications of his own account. Feyerabend notices that, in laying out his account Lakatos had accepted the lesson from Kuhn, viz. that a theory should not be rejected as soon as the first anomaly appears. Research programmes should be given the opportunity to overcome temporary setbacks and should not be rejected when they are merely "intermittently degenerating" (Lakatos 1970/1978: 48-49). However, according to Feyerabend, it is always possible that "what looks like a degenerating problem shift may be the beginning of a much longer period of advance" (1970: 215). Thus, he argues, there is no time at which one should take the fact that a research programme is 'degenerating' as a sufficient reason to abandon it; the terms 'degenerating' and 'progressing' should simply be seen as "verbal ornaments" without any practical force (*ibid*.). Somewhat surprisingly, given my discussion in Chapter 1 (Section 1.3.3), Lakatos at least sometimes accepted this argument. For instance, at one point he remarks that "One may rationally stick to a degenerating programme until it is overtaken by a rival *and even after*. What one must *not* do is to deny its poor public record… It is perfectly rational to play a risky game: what is irrational is to deceive oneself about the risk" (1971/1978: 117). On this version of Lakatos' account, as long as scientists are open about the fact that they are pursuing a degenerating research programme, nothing can be said for or against the rationality of pursuing it.

As Musgrave (1976: 478) argues, this argument is not very convincing: just because a research programme which currently looks unpromising *might* improve in the future, it does not follow that we cannot say *anything* about which of them it is most reasonable to pursue. He uses an analogy to illustrate the point: suppose we have to choose which road to travel around a mountain and we know that, while one of the two is usually free, the other is often blocked. The fact that either road might be blocked by some unpredictable avalanche within the next five minutes does not mean that we cannot have good reasons to choose one road over the other. Similarly, Quinn (1972: 144-5) argues that it not is *always* "perfectly rational to play a risky game" as long as one is honest about the risk. Quinn illustrates the point with the following analogy: suppose a hunter has a choice between hunting for tigers and hunting for rabbits. While a tiger skin is more valuable than a rabbit skin, the hunter knows that a merely wounded tiger is very dangerous. Since she is not a particularly good shot (she only has a 50% chance of killing the animal), it is clearly more rational to hunt for rabbits than tigers, even if she is perfectly honest about the risk she is taking.

Musgrave's and Quinn's analogies illustrate that even if the future direction of science is uncertain, this does not entail that there are no normative constraints on which theories should be pursued. However, both analogies involve some judgement of how likely the options considered are to lead to successful outcomes (the road being unblocked, killing the tiger, etc.). A committed anti-inductivist would be sceptical of such predictions, so part of Lakatos' reason for accepting Feyerabend's argument may have been his official anti-inductivist stance.[34] Musgrave (1976: 480-2) argues that anti-inductivists can still resist the argument by pointing out that a (theoretically) progressive research programme has more interesting *potential* outputs, even if we do not know how likely they are to obtain. A progressive research programme raises new and interesting problems to investigate, whereas a degenerating research programme merely promises to accommodate the anomalies generated by other research programmes. As an example, Musgrave mentions that Lavoisier's new chemistry, with its oxygen theory of acidity,

[34] Feigl (1971) argues against Lakatos that predicting which research programme will be worth pursuing in the future based on their past performance seems to rest on a form of inductivism.

predicted that oxygen can be extracted from all acids. Even if this prediction turned out false, this would in itself constitute an interesting new discovery. By contrast, Priestley's phlogiston programme at best (according to Musgrave) offered to formulate uninteresting ad hoc explanations of anomalies. Thus, Musgrave argues, even if we have no way of predicting how likely either of two research programmes are to be successful, we can still base decisions about pursuit on comparisons of the worst- and best-case consequences of pursuing a given theory or research programme.

Of course, someone who takes the development of science to be *completely* chaotic, or a radical inductive sceptic who holds that all predictions about the future are equally uncertain, could deny that there is any reason for saying that progressive research programmes are more likely than degenerating ones to lead to interesting new discoveries. If the future direction of research is radically unpredictable, there is no basis for Musgrave's claim that the potential outcomes of degenerating and progressive research programmes differ. Both would have the same range of potential consequences, namely a completely open, unpredictable one. Thus, if Feyerabend's argument is interpreted as relying on such a radically sceptical premise, the argument is valid.[35] However, there is no reason why we should accept this premise. As long as we can make some reasonable predictions about which directions of research are more likely to be successful, or has the most interesting potential outcomes, this radical interpretation of Feyerabend's argument can be resisted.

[35] It is doubtful that Feyerabend intended this interpretation. As scholars of his work have pointed out (e.g. Kidd 2015), he generally sought to challenge overly monistic and rationalistic assumptions about science rather than to advocate radical scepticism. His remark that one cannot "run a complex and often chaotic business like science by following a few simple and 'rational' rules" (1970: 215), is most plausibly taken to deny that science is guided by *a few simple* rules which are 'rational' in the sense of being independent of values and context.

A different type of argument for resisting normative accounts of pursuit focuses on the claim that *individual* scientists should be free to pursue whatever they want. Affirming this claim sometimes seems to have been Lakatos' main concern. For instance, he at one point claims that although scientists are free to pursue whatever they want, journal editors should still refuse to publish articles from degenerating research programmes and research foundations should deny them funding (1971: 174). The idea seems to be, then, that judgements about pursuit only apply to the scientific community as a whole, where these should guide decisions about editorial policies and funding, rather than to individual scientists (Musgrave 1976: 478-80). Two possible motivations for this claim suggest themselves: (i) maintaining a plurality of active research approaches or (ii) securing the autonomy of individual scientists in organising their own research.

As we saw previously (Section 1.3.2), the idea that scientific research benefits from pursuing a plurality of projects was also one of Kuhn's motivations. It can be motivated by a less radical interpretation of Feyerabend's argument: degenerating research programmes are sometimes revived, while progressive ones sometimes run into insurmountable problems. It is therefore unwise for science to put all its eggs in one basket. There should be room for some scientists to pursue less promising projects, to ensure that viable alternatives are available if the dominant approach starts degenerating, without this being labelled as irrational.[36] While this is a valid point, restricting evaluations about what should be pursued to the community level in the way Lakatos proposes is an inadequate response. For one thing, Feyerabend's argument seems also to apply at the community level: if there are good reasons for individual scientists not to abandon degenerating research programmes (since they might stage a comeback), why

[36] This point has more recently been emphasised e.g. by Chang (2012: 221) and Šešelja and Straßer (2014: 3112-13). The idea of course goes at least as far back as Mill's defence of free speech in *On Liberty* (1859/2003).

should the same reasons not apply to decisions made by the enforcers of the community level judgements, i.e. science funders and journal editors? If the goal is to ensure that a sufficiently broad range of research topics is pursued, severely restricting funding and publishing opportunities seems counter-productive. Conversely, if journal editors and funders do have good reasons to deny degenerating research programmes *any* support, why are these not also reasons for individual scientists to refrain from pursuing them?

A better solution is to distinguish community level and individual level judgements about which theories or research programmes are worth pursuing. As Šešelja and Straßer (2013: 13) argue, the community level judgement that a theory is worthy of pursuit does not exclude other theories from being worthy of pursuit as well. Community level judgements recommend that *someone* should pursue a given theory, while individual level judgements concern how specific individuals should prioritise their efforts. A group of scientists can agree that a certain range of theories are all worth pursuing without implying that each of them should pursue the full range. Typically, the efforts of individual scientists will be best spent pursuing one or only a few theories or research programmes. But from the fact that T is the best theory for a given scientist to pursue it does not follow that T is the best theory for all other scientists to pursue. This approach thus supports a reasonable pluralism with regards to what should be pursued, while at the same time allowing for there to be some normative constraints on pursuit: there is still room for some theories to be deemed not worth the efforts, either of the research community as a whole or of individual scientists.

This brings us to the second concern, that a normative account of pursuit would infringe on the autonomy of scientists to organise their own research. However, passing normative judgement on an activity does not in itself infringe on the autonomy of a person to persist in that activity. Even if we judge that a scientist is not spending their efforts optimally, it does not follow that they should be forced to change their priorities. Compare the rational evaluation of beliefs or the ethical evaluation of actions: we may judge that an individual is being irrational in his beliefs or unethical in his actions without thereby saying that he should be forced to change his beliefs or actions. The possibility of passing normative judgement is consistent with the general principle that everyone should have the autonomy to make up their own mind, and that reasonable people can disagree about such judgements (at least within certain limits). In the same way, formulating a normative account of pursuit does not in itself have implications for what individual scientists should be allowed to do. There are of course important questions about the best way to organise scientific research and research funding, e.g. whether scientific autonomy and self-organisation is beneficial, or some degree of central planning is feasible. While normative accounts of pursuit may inform these debates, giving such an account does not in itself entail any particular solution. It may still be that the best approach is to let each scientist organise their own research according to where they believe their efforts are best spent.

Suppose a scientist agrees that they would make a more significant contribution to science if they worked on a different topic, but decides to work on something they find more personally fulfilling, more consistent with their ethical commitments or simply less boring. Again, there are important questions here, partly about how the epistemic priorities of science should be weighed against the broader values of individual scientists, and partly about the extent to which individual scientists can be expected to sacrifice their own fulfilment for the common good. Both of these points raise important and interesting questions which I will, however, not have much to say about here. My focus in this chapter is to formulate an account of the normative criteria for judgements about pursuit, rather than discussing the practical implications of this account.

2.3. Are the Normative Constraints for Pursuit the Same as Those of Acceptance?

A second challenge to developing a normative account of pursuit is the objection that the normative constraints on pursuit are not interestingly different from those governing acceptance. If this were the case, there would be no need to develop a *separate* normative account of what justifies pursuing a theory, distinct from an account of what justifies accepting it. I will in this section consider several different versions of this objection.

2.3.1. Should One Pursue the Most Well-Supported Theory?

The strongest version of this objection is to claim that normative standards for pursuit coincide with those for acceptance: scientists should simply pursue those theories which they currently have the most reason to accept. Sometimes this claim seems to stem from a lack of recognition of the distinction between acceptance and pursuit, as for instance the conflation of acceptance and pursuit in Kuhn's writings on theory choice. Kuhn often seems to simply assume that if scientists work on a theory then they also accept it—even if, as he notes, this happens "in defiance of the evidence".

As Laudan (1977: 110) noticed, one can avoid the conclusion that scientists in such cases accept theories on irrational grounds by recognising that there are different normative standards for acceptance and pursuit. At least in some cases, scientists draw a similar distinction in order to defend their work on theories which are clearly not yet well-supported enough to be accepted. For instance, Laudan (1977: 112-113) and Laurie Whitt (1990) highlight that many of the nineteenth-century chemists who worked on Daltonian atomic theory were reluctant to accept it. Some explicitly denied that they accepted the theory as true and merely regarded it as a "useful ladder" (Whitt 1990: 467). Others, such as Berzelius, did extensive work on determining atomic weights but also highlighted the many theoretical and empirical problems of the theory. His stated purpose was not to argue that the theory could not overcome these problems but "to lay open all the difficulties of that hypothesis that nothing might escape our attention calculated to throw light on the subject" (Whitt 1990: 468). Berzelius did not pursue the atomic theory because he already accepted it, but in order to find out *whether* the theory should be accepted.

While scientists do not always differentiate their attitudes towards a theory this clearly, the case shows how they are at least sometimes willing to adopt distinct normative standards for acceptance and pursuit. However, some philosophers who do recognise the conceptual distinction between acceptance and pursuit still question the coherence of doing so. Kitcher (1990) gives the following argument: if we accept that it is rational for a person to accept the better-supported theory because "that person's aim is to achieve true beliefs", then "it appears that the person should also pursue the better-supported theory, since pursuing a doctrine that is likely to be false is likely to breed more falsehood" (8).[37] Instead, Kitcher argues, it is "Only if we situate the individual in a society of other epistemic agents" that it can be "rational for someone to assign herself to the working out of ideas that she (and her colleagues) view as epistemically inferior" (*ibid*.). To Kitcher, working on an epistemically inferior theory is only rational in the context of a division of cognitive labour where someone else is pursuing the more likely theory. Working on a poorly supported theory is only justified by the fact that this helps the scientific community hedge its bets. Thus, for scientists to work on an epistemically inferior theory they have to be either "epistemic altruists", who consign themselves to pursuing an inferior theory for the greater good of the community, or motivated by non-epistemic motives (e.g. the fame associated with being the first to solve given a problem).

[37] Rueger (1996: 268-9) also seems to accept this argument.

Kitcher's argument is unconvincing.[38] First, it is not generally the case that pursuing a false theory will "breed more falsehood". For example, one way to pursue a theory is to subject it to further testing. As long as one's methods of conducting and interpreting these tests are reliable, there is no reason why testing a theory which is likely false should be likely to produce false results. It may be likely that the test will show *that* the theory is false, but that result will itself be a truth. Furthermore, developing, refining or "working out [the] ideas" of a theoretical model currently thought to be false might still bring about more truth, either by allowing scientists to determine more clearly why it is false or to identify things the model gets right which are overlooked by the best-supported model.

Second, there is no reason to think that pursuing the most well-established theory is always an effective way to generate new truths. An important factor in deciding which theory to pursue concerns how much potential for further development it has. As McMullin (1976: 424) points out, the future potential of a theory need not correlate with its current degree of epistemic support. A well-established theory may have exhausted its future potential, whereas a new and untested theory may hold a lot of promise, exactly because of its untested potential. Even if we grant that working on the remaining problems within an epistemically superior theory is more likely to be successful, working on the inferior theory may still lead to more truth by virtue of having a much larger stock of potential new insights and thus more chances of discovering new truths. In addition, if Musgrave (1976) is right that for many of the problems suggested by a new theory, even

[38] The situations which Kitcher discusses later in the paper in relation to the division of cognitive labour are different from those targeted by this argument. Rather than whether to pursue a poorly supported theory, Kitcher is here concerned with cases where scientists have to choose whether to try and solve a problem *using* the resources of a poorly supported theory. In these cases, Kitcher's claim seems more plausible: employing e.g. the experimental procedures associated with a poorly supported theory may (plausibly) be more likely to lead to further falsehoods.

negative results would constitute an interesting discovery, a higher likelihood of making correct predictions may not even count in favour of the better-supported theory.

2.3.2. Are Reasons for Pursuit Weak Reasons for Acceptance?

A weaker version of the claim that the normative standards for acceptance and pursuit overlap is that reasons for pursuing a hypothesis are just weak reasons for acceptance. This idea is typically expressed through equating the pursuit worthiness of a theory with its plausibility, and then arguing that reasons for the plausibility of a theory are of the same kind as reasons for its acceptability. As described in Chapter 1 (Sections 1.4.3 and 1.4.4), this account was often proposed within the debates over Hanson's (1958) 'logic of discovery', including by Hanson himself, Salmon (1967), Kordig (1978) and McLaughlin (1982). This was generally based on a tacit assumption, rather than an explicit argument, that good reasons just are reasons for acceptance. They do not seem to have considered the possibility, pointed out by Curd (1980), that arguments for pursuing a hypothesis could be evaluated according to a distinct normative standard.

In addition, some cases of pursuit from scientific practice also throw doubt on this account. There are arguably cases where scientists pursue theories which they do not even regard as plausible. Allan Franklin (1993b) discusses the case of the so-called "fifth force" hypothesis, a theory that there might be an additional force in nature. This was supposed to be a weak force, about 1% the strength of gravity, with an effective range around the order of 100 metres (101). This theory was tested extensively during the late 1980s and was regarded as conclusively refuted by the end of 1990 (93). Nonetheless, some experimentalists continued to carry out the experiments (101-105). As Franklin explains (123), this was not because the physicists believed there was still a chance of the original fifth force hypothesis being confirmed. Rather, they regarded the experimental work as interesting in itself, partly because it allowed them to determine how strongly the evidence supported rejecting any hypothesis, such as the fifth force hypothesis, which postulated short-distance divergences from Newton's law of gravitation.

It may be argued that in this case, the physicists were not strictly speaking pursuing the fifth force hypothesis anymore, but were rather using the same experiments to pursue other possible theories postulating corrections to the force of gravity at small distances. Even so, the physicists did not seem to regard such hypotheses as plausible in any stronger sense than that they were not ruled out by the previous experimental results. Establishing this does not require an independent plausibility argument. The reasons for continuing this experimental work was rather, as Franklin (123) argues, that the experiments used to test the fifth force hypothesis also provided an effective means of testing the possible scope for other hypotheses postulating a correction to gravity. Since these experiments were already set up, it was relatively cost effective (both in terms of time and money) to continue running them. They mainly hoped to be able to refine the relevant experimental techniques, e.g. by discovering possible sources of noise or systematic error, but had they detected an inexplicable anomaly, this would of course also have been a very interesting result.

As this example illustrates, although plausibility judgements may sometimes provide reasons for pursuing a hypothesis, there are other relevant considerations. These include how easily and reliably the hypothesis can be tested, how costly it would be to pursue a given line of research and how interesting the potential results would be (Franklin 1993b: 122). As I will argue below, these considerations can be captured by a decision-theoretic framework. It might be replied that although plausibility is not the only relevant factor, some minimal degree of plausibility is at least a necessary condition on pursuit. Unless there is some chance of a theory or hypothesis being true, there seems to be little point in investigating whether it is true. This is however only correct if subject to a number of qualifications.

First, as pointed out with the fifth force hypothesis, if a hypothesis is sufficiently cost effective, the degree of plausibility needed can amount to little more than not already being conclusively falsified.

Second, one can pursue a theory which is known to be false either to develop it into a theory which could be true or to find out if it contains some partial truths not captured by other theories. Thus, plausibility is only a necessary condition on pursuit if one stipulates that it need not be the first-order hypothesis *H* itself which is plausible, but that it can also be some relevant meta-hypothesis such as "*H* contains some partial truths" or "*H* can be developed into a plausible theory".

Third, even if it is granted that a minimal degree of plausibility is necessary for a hypothesis to be worth pursuing (subject to the above qualifications), this point works both ways. If it is correct that there is little point in pursuing a hypothesis which has no chance of being true, then, equally, there is little point in pursuing a hypothesis which has no chance of being false. In both cases, the problem is presumably that nothing new can be learned about the truth of the hypothesis since this is already completely settled. In some cases, one might support pursuing a hypothesis by showing that it has *more* plausibility than previously thought. However, one can equally support a hypothesis by showing that there is more reason to *doubt* it than previously thought. Of course, the two are often connected: reasons for doubting a previously well-established theory will typically also make some competing alternatives more plausible. For instance, part of the reason physicists began pursuing the fifth force hypothesis were the results of certain experiments which seemed to indicate a divergence from Newton's law of gravitation.

This both provided some reason to doubt Newton's law and gave the fifth force hypothesis some plausibility.

The kernel of truth in the proposal that plausibility is a necessary condition for pursuing a hypothesis is that there needs to be *some* reason to think that something new can be learned from pursuing it. I now want to introduce a more systematic normative account of pursuit which captures this insight together with other observations from this and the previous section.

2.4. The Consequentialist Approach to Pursuit Worthiness

The acceptance of a theory concerns the question of what the world is like: whether the hypothesis true, or partially or approximately true, of the world (or the observable parts of the world). In contrast, decisions about pursuit concern which course of action to pursue (McKaughan 2008: 454; cf. Kapitan 1992, 1997). The normative basis for judgements about pursuit should therefore be construed in terms of practical rationality. More specifically, as Šešelja, Kosolosky and Straßer (2012: 53) point out, it is natural to construe pursuit worthiness in terms of a broadly consequentialist or goal-oriented conception of practical rationality. On this account, the pursuit worthiness of a hypothesis should be evaluated in terms of how well the expected consequences of doing so contributes to achieving the goals of research (whatever these are), compared to how one could have otherwise spent the available time and resources.

One reason for adopting this approach is that it is already implicit in most of the discussion reviewed above. It is clearly the basis for the 'Economy of Research' argument, which Weinberg and Peirce use to make that point that it is necessary to make judgements about pursuit worthiness. This is also the thinking underlying Franklin's account of the fifth force hypothesis: after the hypothesis had been refuted, it was still reasonable to carry out the experiments despite modest the expected epistemic gains, because the experiments were already set up and thus still cost effective. Finally, the arguments against there being normative standards for pursuit (distinct from those on acceptance), made by Feyeberabend/Lakatos and Kitcher, all assume that if there were normative constraints, these would be related to the expected outputs of pursuing the hypothesis. Feyerabend's argument (in the radical interpretation) implicitly relies on this assumption to conclude from the premise that we cannot predict the future development of science that there are no normative constraints on pursuit. Similarly, Kitcher's argument assumes that the pursuit worthiness of a hypothesis should be evaluated in terms of whether this is likely to result in more falsehoods or truths.

2.4.1. Drawing Distinctions

To start developing this into a more systematic account, it is useful to draw attention to some distinctions pointed out by Šešelja, Kosolosky and Straßer (2012). Saying schematically that it is rational for *Y* to pursue *X* if, and only if (or to the extent that), pursuing *X* contributes to achieving the goals *Z*,[39] they distinguish different ways to instantiate the variables *X*, *Y* and *Z*. Each of these corresponds to a different kind of pursuit worthiness judgments.

First, we can distinguish between different agents, *Y*, for whom the judgment of pursuit worthiness is made. In particular, we can distinguish between whether we are making a community level judgement regarding what should be pursued by *someone* within a scientific field or whether the judgement concerns what an individual scientist

[39] This schema is adapted, with some changes, from Šešelja, Kosolosky and Straßer (2012: 53).

should pursue.[40] As argued above (Section 2.2.2), both types of judgements are legitimate but should not be conflated.

Second, we should distinguish which type of item we take to be the object, *X*, of pursuit. In this thesis, I will mostly consider cases where *X* is some kind of theoretical representation, i.e. hypotheses, theories or models which purport to represent some target phenomenon. Here, the aim is usually to find out how well that hypothesis represents or at least predicts the relevant aspects of the target. In other cases, scientists are primarily pursuing some technological development, aiming to develop a specific kind of instrument or technique or to refine and calibrate an existing one. For example, part of the reason for continuing the gravitational experiments in the fifth force case, according to Franklin, was to improve experimental techniques.

Third, we can distinguish the type of goals, *Z*, that we are evaluating the situation in terms of. Two distinctions are relevant here. First, we should distinguish between (a) evaluating a case from the internal perspective, i.e. assuming the goals endorsed by the agent making a decision about pursuit, and (b) evaluating the case from the external perspective, i.e. using the goals we (the evaluators) think the agent *ought* to have. Second, we can distinguish between whether the goals refer narrowly to the *epistemic* or *intellectual* goals—e.g. accepting theories that are close to the truth or obtaining explanations of puzzling phenomena—or whether it includes a broader set of moral and political values and goals as well. Following Šešelja et al., I will refer to analyses taking into account only the former, narrow set of goals as concerned with *epistemic* pursuit worthiness and analyses taking into account the latter, broader set of goals as concerned

[40] One can also draw more fine-grained distinctions between different levels, e.g. science as a whole, disciplinary fields, research groups and individual scientists. The same point applies here.

with *practical* pursuit worthiness. I will mainly focus on epistemic pursuit worthiness in this thesis.[41]

2.4.2. Relevant Factors

On the consequentialist conception of pursuit worthiness, the kinds of considerations which can be used as reasons for or against a pursuit worthiness judgment can, in principle, include anything which is relevant to estimating the outcomes of pursuit. While these will presumably vary between contexts, some general suggestions can be made. As argued above, although the likeliness of a hypothesis has some relevance, it is not the only factor. Summarising Peirce's view, McKaughan (2008: 457) concludes that the most important considerations in deciding whether to prioritise a hypothesis for pursuit are "factors like our time, resources, and value of the estimated payoff in comparison to other courses of action. … If we estimate that testing the hypothesis will be *easy*, of potential *interest*, and *informative*, then we should give it a high priority". Independently, Franklin concludes from his case studies that "the decision to pursue an investigation seems to depend on a weighting of at least three factors: the interest of the hypothesis, its plausibility, and its ease of test" (1993b: 122). He also mentions factors to do with conserving resources, such as "recycling expertise" and continuity with already ongoing research programs (Franklin 1993b: 101).

In addition to these factors, Peirce emphasises that since "very rarely can we positively expect a hypothesis to prove entirely satisfactory", it is important to consider the "effects upon other projects" of pursuing the hypothesis, that is, "we must always consider what will happen when the hypothesis breaks down" (CP7.220).

[41] Exactly what the epistemic goals of science are differs between realists and anti-realists. I discuss the extent to which this makes a difference to my account in Sections 2.5.3 and 2.7.2.

What Peirce has in mind here are strategic considerations of how learning that a hypothesis is false can inform later stages of inquiry.[42] He draws an analogy with playing twenty questions, where the most strategic yes/no questions are those that narrow down the field as much as possible regardless of what the answer will be. Similarly, it can in some cases be worth testing a hypothesis simply to "clear the field". Peirce also argues that it can be worth testing a simpler hypothesis because this makes it easier to interpret how the results differ from the predictions of the hypothesis, thereby providing suggestions for how to formulate a better hypothesis. Finally, as Musgrave (1976) points out, we should also bear in mind that an experimental result which serves to falsify a hypothesis can sometimes constitute an interesting discovery in itself.

If reasoning about pursuit consists of comparing different candidates for pursuit in terms of these kinds of factors, it raises the question of how those factors should be weighed against each other. In practice this usually will be a matter of informed judgement. However, in order to clarify the underlying logic of these decisions, it can be useful to think of pursuit worthiness in terms of simplified, idealised decision-theoretic models. I will now develop such a model which captures many of the factors discussed above and which will be particularly useful for my purposes in this thesis.[43]

[42] Recent interpretations have for this reason characterised Peircean abduction as a form of "strategic reasoning" (Hintakka, 1998, Paavola 2004, Pietarinen and Belucci 2014). This type of strategic considerations will be particularly relevant for our discussion of diagnostic reasoning in Chapter 6.

[43] The models developed here draw on and extend the model presented in Nyrup (2015). Decision-theoretic models of pursuit worthiness have also previously been developed, although in a different direction to the one taken here, by Kukla (2001, 2010: ch. 1) and Harp and Khalifa (2015). Similar models have also been used to analyse decisions about whether to test uncertain diagnostic hypotheses (Pauker and Kassirer 1980); I discuss these further in Chapter 6.

2.5. Decision-Theoretic Models

2.5.1. Causal Decision Theory

The model developed here will be based on the framework of causal decision theory.[44] Suppose an agent can choose between a set of possible actions, $A = \{a_1, a_2, \ldots\}$,[45] and that we are interested in a certain range of potential consequences of those actions. To avoid double-counting, we need to partition the latter into a set, $C = \{c_1, c_2, \ldots\}$, of exhaustive and mutually exclusive sets of *total* consequences. Thus, each c includes all relevant changes brought about by a given action. Suppose, furthermore, that the probability that each action brings about a given c depends on the background state of the world. We thus distinguish a set of exhaustive and mutually exclusive possible background states of the world, $S = \{s_1, s_2, \ldots\}$, which obtain independently of the actions of the agent. For each possible state, s, there is a probability, $\Pr(s)$, of that state obtaining and for a given action, a, there is a probability, denoted $\Pr(c \mid s, a)$, that choosing a will bring about the consequence c given that the state of the world is s. The possible total states of the world that can result from the actions are called the *outcomes*, each of which is associated with a utility. In the cases I am interested in, the utility of a given consequence can depend on the uncertain background state. I will therefore represent the outcomes explicitly as a conjunction of the consequences and the background states in the utility assignments: the utility of the outcome resulting from the consequence c and background state s is denoted $\mathrm{u}(c, s)$. Given these definitions, the principle of rational action stipulated by causal decision theory can be stated as follows. The agent should

[44] Exactly how to interpret this framework is the subject of debate (e.g. Joyce 1999), related to the so-called Newcomb problem (Nozick 1969). Since I do not consider Newcomb-style situations in this thesis, I will assume that the correct solution to these problems does not have any implications for my applications of the framework here.

[45] Here, and throughout the rest of this chapter, I adopt the convention that lower-case letters denote members of the corresponding upper-case set. Thus, x denotes a member of the set X.

choose the action a which maximises the expected utility of the action, EU(a), defined as:

(1) $$\mathrm{EU}(a) = \sum_i \big(\Pr(s_i) \times \sum_j \big[\mathrm{u}\big(c_j, s_i\big) \times \Pr(c_j \mid s_i,\ a)\big]\big)$$

It will sometimes be useful instead to consider the expected *change* in utility, ΔEU(a), of an action a, i.e. how much we expect the utility to change relative to the current state of the world. If we knew the utility of the current state, u(s_0), we could simply define this as ΔEU(a) = EU(a) – u(s_0). However, since I have assumed that there can uncertainty about which state currently obtains, the utility of the current situation may be unknown. Instead, let u(s_i) be the utility of the world remaining unchanged when the background state is s_i. We can then compare the change of utility that would occur for each state, defined as Δu(c_j, s_i) = u(c_j, s_i) – u(s_i). The expected change in utility of performing an action a can then be defined as:

(2) $$\Delta\mathrm{EU}(a) = \sum_i \big(\Pr(s_i) \times \sum_j \big[\Delta\mathrm{u}\big(c_j, s_i\big) \times \Pr(c_j \mid s_i,\ a)\big]\big)$$

Under this definition, the ordinal ranking of actions according to ΔEU(a) is equivalent to EU(a). That is, ΔEU(a_1) >/= ΔEU(a_2) if and only if EU(a_1) >/= EU(a_2).[46] To see this, start by noticing that, since u(c_j, s_i) and u(s_i) can be separated into separate sums, we can rewrite (2) as:

[46] I use the abbreviation 'x >/= y' to represent that these equivalences hold both for $x > y$ and $x = y$.

$$\Delta\mathrm{EU}(a) = \sum_i \big(\Pr(s_i) \times \sum_j [\mathrm{u}(c_j, s_i) \times \Pr(c_j \mid s_i,\ a)]\big) - \sum_i \big(\Pr(s_i) \times \sum_j [\mathrm{u}(s_i) \times \Pr(c_j \mid s_i,\ a)]\big) = \mathrm{EU(a)} - \sum_i \big(\Pr(s_i) \times \sum_j [\mathrm{u}(s_i) \times \Pr(c_j \mid s_i,\ a)]\big)$$

Now, since u(s_i) is the same for all j, it is also the case that $\sum_j [\mathrm{u}(s_i) \times \Pr(c_j \mid s_i,\ a)]$ = $\mathrm{u}(s_i) \sum_j \Pr(c_j \mid s_i,\ a)$. Furthermore, because the c_j are exhaustive and mutually exclusive, $\sum_j \Pr(c_j \mid s_i,\ a)$ = 1. Putting this together, we get that:

$$\Delta\mathrm{EU}(a) = \mathrm{EU}(a) - \sum_i \big(\Pr(s_i)\, \mathrm{u}(s_i)\big).$$

Notice that this is analogous in form to ΔEU(a) = EU(a) – u(s_0) and that $\sum_i \big(\Pr(s_i)\, \mathrm{u}(s_i)\big)$ constitutes a plausible substitute for u(s_0): it is simply the average utility weighted by the uncertainty of the background states. Finally, since $\sum_i \big(\Pr(s_i)\, \mathrm{u}(s_i)\big)$ is independent of a, it is a constant for all actions in a given decision problem. Thus, the following equivalences hold: ΔEU(a_1) >/= ΔEU(a_2) if and only if EU(a_1) – $\sum_i \big(\Pr(s_i)\, \mathrm{u}(s_i)\big)$ >/= EU(a_2) – $\sum_i \big(\Pr(s_i)\, \mathrm{u}(s_i)\big)$ if and only if EU(a_1) >/= EU(a_2). This concludes the proof.

2.5.2. Pursuit Worthiness in Causal Decision Theory

We can apply this framework to the question of pursuit worthiness by taking the set of possible actions to be defined in terms of the different candidate objects of pursuit. Let p_x denote the action of pursuing x, where x can be any of the different possible objects of pursuit (hypotheses, models, theories, research programmes, technological developments, etc.) which Šešelja, Kosolosky and Straßer (2012) distinguish. I will assume that it is possible to pursue more than one item at once, so that x can be either a single item or a set of items.

This allows us to define several relevant notions of pursuit worthiness within this framework. As Šešelja, Kosolosky and Straßer point out (55-59), 'pursuit worthy' can both be used in a *comparative* sense: x_1 is more pursuit worthy than x_2; and in a *non-comparative* sense: x is pursuit worthy independently of how it compares with other candidates for pursuit.

Defining a comparative notion of pursuit worthiness in the present framework is straightforward. We can say that the item x_1 is *more pursuit worthy* than x_2 if, and only if, EU(p_{x1}) > EU(p_{x2}). Building on this, we can say that a given consideration (factor, argument, line of reasoning, …) provides *some* reason for pursuing x (in a given context) if it makes x more pursuit worthy (in that context), i.e. if it increases our estimate of EU(p_x).

To define a non-comparative notion of pursuit worthiness, a natural suggestion is that x is pursuit worthy only if it is expected to make a positive contribution towards achieving our goals. We can say that an item x is *minimally pursuit worthy* if and only if ΔEU(p_x) ≥ 0. It should be noticed, however, that just because x is pursuit *worthy* in this sense it does not follow that x should be pursued by anyone. On a consequentialist conception of pursuit worthiness, decisions about pursuit are ultimately comparative: one should pursue the line of research which has the highest expected utility. Thus, we can say that x is *absolutely pursuit worthy* if, and only if, either p_x is the action with the highest expected utility or x is part of a set of items, X, such that p_X is the action with the highest expected utility.[47] As long as there is some available action a for which ΔEU(a) ≥ 0, e.g.

[47] One could invoke other principles of rationality, apart from utility maximisation, to propose different notions of pursuit worthiness. For instance, on a precautionary principle, minimal pursuit worthiness would require that the worst potential consequence of pursuing x is not unacceptably bad. Similarly, on a maxi-

doing nothing, minimal pursuit worthiness is a necessary condition on absolute pursuit worthiness.

Notice that these definitions are flexible enough to accommodate the distinctions identified by Šešelja et al.

First, as previously noted, the potential actions can include pursuing any of the different kinds of items that can be an object of pursuit.

Second, depending on which kinds of goals we focus on (e.g. practical or epistemic) we can include a different set of consequences and interpret the utility assignments accordingly. Similarly, depending on whether we are adopting an internal or an external perspective on pursuit, we can choose to interpret the utilities as those endorsed by the agent or those endorsed by the evaluator. The same goes for the interpretation of the probabilities (a point not mentioned by Šešelja et al.): we can either interpret these as the agent's own estimates or according to the estimates of the evaluator, depending on whether we are interested in giving an internal or external evaluation of pursuit worthiness.

Third, the 'agent' in question can both refer to individual scientists and to larger research communities. Consistent with Šešelja et al.'s argument, the expected utilities of individual scientists need not correspond to those of their overall community. Furthermore, we can allow actions to include the pursuit of multiple different items at once. Thus, it is possible for a community of scientists to agree that the research community as a whole should pursue a broad range of hypotheses, while each scientist decides to pursue a single hypothesis. As mentioned in Section 2.2.2, there are

min principle, unconditional pursuit worthiness would be defined in terms of maximising the utility of the worst potential consequences. Comparative pursuit worthiness would depend on the utility of worst consequences. However, for the case of epistemic pursuit worthiness, which is my main focus in this chapter, expected utility maximisation provides a plausible reconstruction of the factors highlighted by previous commentators; see Section 2.6.

complicated issues concerning whether individual scientists should always pursue one of the hypotheses endorsed by the community. The decision-theoretic framework will be able to express this debate in terms of whether individual scientists must evaluate their decisions about pursuit in terms of the goals and consequent utility assignments they endorse for the community-level analysis. As mentioned, I will not go deeper into these issues in this thesis.

While the above shows that it is possible to formulate a notion of pursuit worthiness in decision-theoretic terms, this in itself arguably adds little to the informal consequentialist conception presented above. The foundations of causal decision theory are still subject to a number of debates, and expected utility theory is not uncontroversial as a general account of practical rationality. Furthermore, the mathematical precision of this framework can give a misleading impression of the vague and rough estimates of potential benefits or harms which are no doubt often the only available basis for deciding whether to pursue a theory. I do not want to claim that scientists always conform to, or even approximate, the strict requirement of maximising expected utility, nor that it generally would be better if they tried to do so.

Rather, I believe that the advantage of this framework is that it allows us to develop certain restricted and simplified models which provide a useful, normative perspective from which to think about pursuit worthiness. More specifically, I will develop a family of models focused on the epistemic aspects of pursuit worthiness. As I will show in Section 2.6, a relatively simple model allows us to capture and clarify the underlying logic of many of the points concerning epistemic pursuit worthiness discussed previously.

2.5.3. A Decision-Theoretic Model of Epistemic Pursuit Worthiness

To develop this model, I will start by focusing on cases where the potential objects of pursuit are a set of hypotheses, H, understood here as fairly specific claims about the world. While a similar analysis would apply to other representational items such as models or more general theoretical frameworks, focusing on hypotheses provides the simplest case. Furthermore, I will restrict my focus to choices between the pursuit of single hypotheses; I discuss how to accommodate the possibility of pursuing multiple hypotheses in Section 2.7.3.

Since the model is meant to capture epistemic pursuit worthiness, I will only focus on the potential epistemic consequences of pursuing a given h. In fact, I will restrict the relevant consequences to consist only in the set of epistemic states, $ES(h) = \{es_1(h), es_2(h), \ldots\}$ which the agent can be in concerning h. By an epistemic state concerning h, I mean the agent's total state of knowledge, broadly construed, regarding h. In the simplest case, I will restrict the model to just three potential consequences of pursuing h:

1. We get strong enough evidence in favour of h to accept it.
2. We get inconclusive evidence and so suspend judgement about h.
3. We get strong enough evidence in favour of h to reject it.

I will denote these as $acc(h)$, $sus(h)$ and $rej(h)$, respectively. More complicated models could include additional, more nuanced epistemic attitudes (strong acceptance, tentative rejection, etc.) or define them in terms of different degrees of belief. Further epistemic states could also be introduced by distinguishing how much evidence the agent has for the epistemic attitude. However, for most of my discussion the above three states will suffice.

Finally, I will assume that the relevant background states for evaluating the utility of a given epistemic state concern a set of possible truth values of h, $T(h) = \{t_1(h), t_2(h), \ldots\}$. The current model thus assumes a form of axiological scientific realism, i.e. that part of the aim of scientific inquiry is to discover the truth.[48] This is not uncontroversial; scientific anti-realism is often characterised as rejecting this premise (e.g. van Fraassen 1980: 6-13; Laudan 1984: ch. 5; Godfrey Smith 2003: 175-179). In Section 2.7.2, I will consider how the model can be restated in anti-realist terms. Here, however, I will assume realism since it provides a simple and intuitive model. The range of relevant truth values of h could, in principle, include a range of degrees to which h could be partially or approximately true. However, in the following I will assume that we only need to distinguish between whether the hypothesis is true or false, denoted h and $\neg h$.

Given these assumptions, we can write the general expected utility functions for the pursuit of a hypothesis h as follows:

$$(3) \qquad \mathrm{EU}(p_h) = \textstyle\sum_i \big(\Pr(t_i(h)) \times \sum_j \big[\mathrm{u}\big(es_j(h), t_i(h)\big) \times \Pr(es_j(h) \mid t_i(h),\ p_h)\big]\big)$$

If we restrict the possible background states to h being true and false we get:

$$\begin{aligned}(4) \qquad \mathrm{EU}(p_h) = \quad & \Pr(h) \times \textstyle\sum_i [\mathrm{u}(es_i(h), h) \ \times \Pr(es_i(h) \mid \ h, p_h)] \\ + \ & \Pr(\neg h) \times \textstyle\sum_i [\mathrm{u}(es_i(h), \neg h) \ \times \Pr(es_i(h) \mid \neg h, p_h)]\end{aligned}$$

Finally, the special case where we only consider the three attitudes mentioned above, this becomes:

[48] This is a somewhat naive formulation of axiological realism, but sufficient to indicate my intent here. See Lyons (2005) for critiques of this formulation and attempts at more sophisticated ones.

$$
\text{(5)} \qquad \mathrm{EU}(p_h) = \Pr(h) \times \begin{bmatrix} \mathrm{u}(acc(h), h) \times \Pr(acc(h)|\ h, p_h) \\ +\mathrm{u}(sus(h), h) \times \Pr(sus(h)|\ h, p_h) \\ +\mathrm{u}(rej(h), h) \times \Pr(rej(h)|\ h, p_h) \end{bmatrix}
$$

$$
+ \Pr(\neg h) \times \begin{bmatrix} \mathrm{u}(acc(h), \neg h) \times \Pr(acc(h)|\ \neg h, p_h) \\ +\mathrm{u}(sus(h), \neg h) \times \Pr(sus(h)|\ \neg h, p_h) \\ +\mathrm{u}(rej(h), \neg h) \times \Pr(rej(h)|\ \neg h, p_h) \end{bmatrix}
$$

The model described by equation (5) will form the basis for most of my discussion in this and the next section. I shall refer to it as the *Simple Model*.

In the Simple Model, we only distinguish six possible outcomes of pursuit: (i) accepting a truth, (ii) suspending judgment about a truth, (iii) rejecting a truth, (iv) accepting a falsehood, (v) suspending judgment about a falsehood and (vi) rejecting a falsehood. Due to my focus on epistemic pursuit worthiness, in interpreting the utilities I will only take into account the *epistemic* value directly associated with these outcomes. I understand this to mean something like how much we would have learned about the world in each of the six outcomes, and how intellectually valuable or important this would be. I will not assume any specific account of this kind of value. However, I will take the assumption of axiological realism to at least entail that, *ceteris paribus*, accepting a truth and rejecting a falsehood are both preferable to suspending judgement, which in turn is better than accepting a falsehood or rejecting a truth. Specifically, the following inequalities are assumed to hold: u(*acc*(*h*), *h*) > u(*sus*(*h*), *h*) > u(*rej*(*h*), *h*) and u(*acc*(*h*), ¬*h*) < u(*sus*(*h*), ¬*h*) < u(*rej*(*h*), ¬*h*).

These utility assignments ignore the moral, political and other practical consequences of adopting the different attitudes (taking these into account would be relevant in a model of practical pursuit worthiness). More critically, I also do not take into account the costs of pursuing h, such as the time, effort, money and other resources spent pursuing *h*. These are obviously a crucial aspect of Peirce and Weinberg's 'Economy of Research' argument for the importance of pursuit worthiness judgement and, more generally, for the consequentialist approach to pursuit worthiness endorsed above. I leave them out for now for two reasons. First, it is not clear whether the value of the time and resources spent are commensurable with the epistemic value of learning that a hypothesis is true. Second, many of the model's merits, which I discuss Section 2.6, do not involve the costs of pursuit. I will return to these complications in Section 2.7.3.

When interpreting this model, it is important to bear in mind that the agent will usually already be in some epistemic state regarding *h*.[49] While the agent's epistemic state is presumably known, the utility of this state depends on whether the hypothesis is, in fact, true or false. Thus, it is often more relevant to consider the potential epistemic gains of pursuing *h*, $\Delta\mathrm{EU}(p_h)$. Since the total current state is a combination of the agent's current epistemic state and the unknown background state, let us say that, if the agent is currently in epistemic state $es(h)$, then $\mathrm{u}(s_i) = \mathrm{u}(es(h), s_i)$. Notice that this entails that for all *i*, $\Delta\mathrm{u}(es(h), s_i) = 0$. In other words, those outcomes where we would not change our epistemic state, and so have learned nothing relevantly new, are given zero weight.

To illustrate the idea, suppose in the Simple Model that the agent currently suspends judgment about *h*. We thus set $\mathrm{u}(s_i) = \mathrm{u}(sus(h), s_i)$. This, together with the inequalities assumed above, entails that for the six possible outcomes, the following hold:

[49] A possible exception is when we are considering the pursuit worthiness hypotheses that have not yet been generated; as argued in Chapter 1 (Sections 1.4.1 and 1.5), pursuit worthiness is usually the normative standard for evaluating hypothesis generation. However, we can get around this by introducing an epistemic state consisting in having no attitude towards *h*.

(i) $\Delta\text{u}(acc(h), h) = \text{u}(acc(h), h) - \text{u}(sus(h), h) > 0$

(ii) $\Delta\text{u}(sus(h), h) = \text{u}(sus(h), h) - \text{u}(sus(h), h) = 0$

(iii) $\Delta\text{u}(rej(h), h) = \text{u}(rej(h), h) - \text{u}(sus(h), h) < 0$

(iv) $\Delta\text{u}(acc(h), \neg h) = \text{u}(acc(h), \neg h) - \text{u}(sus(h), \neg h) < 0$

(v) $\Delta\text{u}(sus(h), \neg h) = \text{u}(sus(h), \neg h) - \text{u}(sus(h), \neg h) = 0$

(vi) $\Delta\text{u}(rej(h), \neg h) = \text{u}(rej(h), \neg h) - \text{u}(sus(h), \neg h) > 0$

This gives us the following expression for the expected change in utility of pursuing *h*:

(6)
$$\Delta\text{EU}(p_h) = \Pr(h) \times \begin{bmatrix} \Delta\text{u}(acc(h), h) \times \Pr(acc(h)|\ h, p_h) \\ + \Delta\text{u}(rej(h), h) \times \Pr(rej(h)|\ h, p_h) \end{bmatrix}$$
$$+ \Pr(\neg h) \times \begin{bmatrix} \Delta\text{u}(acc(h), \neg h) \times \Pr(acc(h)|\ \neg h, p_h) \\ + \Delta\text{u}(rej(h), \neg h) \times \Pr(rej(h)|\ \neg h, p_h) \end{bmatrix}$$

Since, recall, that ΔEU(p_h) is equivalent to EU(p_h), to evaluate the pursuit worthiness of *h* in this case, we only need to consider outcomes where the agent obtains sufficient evidence to accept or reject *h*.

To take another case, suppose we are considering whether to pursue a hypothesis we already accept, so that u(s_i) = u(*acc*(*h*), s_i). Then, we instead get:

(7)
$$\Delta\text{EU}(p_h) = \Pr(h) \times \begin{bmatrix} \Delta\text{u}(sus(h), h) \times \Pr(sus(h)|\ h, p_h) \\ + \Delta\text{u}(rej(h), h) \times \Pr(rej(h)|\ h, p_h) \end{bmatrix}$$
$$+ \Pr(\neg h) \times \begin{bmatrix} \Delta\text{u}(sus(h), \neg h) \times \Pr(sus(h)|\ \neg h, p_h) \\ + \Delta\text{u}(rej(h), \neg h) \times \Pr(rej(h)|\ \neg h, p_h) \end{bmatrix}$$

In this case, both Δu(*sus*(*h*), *h*) and Δu(*rej*(*h*), *h*) will be negative, while Δu(*sus*(*h*), ¬*h*) and Δu(*rej*(*h*), ¬*h*) will be positive. Thus, the first term of (7) is always negative and the second term always positive. Here, decreasing Pr(h)—and, equivalently, increasing Pr(¬h)—will always increase ΔEU(p_h). In more informal terms, if the agent were to obtain some reason to suspect that h is less likely than previously thought, this will provide *more* reason for pursuing it; the evidence increases the probability that the agent is currently in the highly undesirable state of mistakenly accepting a falsehood, thus providing some reason to investigate whether this is the case.

2.6. Merits of the Simple Model

An attractive feature of the Simple Model is that it captures many of the factors relevant to evaluating the pursuit worthiness of a hypothesis, surveyed in Section 2.4.3. Furthermore, it calls attention to some plausible factors left out by previous commentators.

The unconditional probabilities Pr(h) and Pr(¬h) represent, respectively, how likely the hypothesis is to be true or false at the stage of inquiry where pursuit is being considered. They can thus either represent the initial plausibility of the theory before any testing has been done or its posterior probability in light of previous testing where we are considering whether to pursue h further. It is this factor which e.g. Salmon takes arguments for pursuing a hypothesis to manipulate. In the Simple Model, then, the plausibility or probability of a hypothesis does play *some* role (to be explored in more detail) in determining the pursuit worthiness of a hypothesis.

The conditional probabilities represent how likely we are, given that the hypothesis is true (or false, respectively), to obtain sufficient evidence to accept, reject or suspend judgement about h. Thus, Pr($acc(h)$ | h, p_h) can for example be taken to represent how likely we are to get *reliable* evidence in favour of h, while Pr($acc(h)$ | ¬h, p_h) represents how likely we are to get *misleading* evidence in favour of h and similarly, *mutatis* *mutandis,* for other permutations of epistemic attitudes and truth values. This corresponds to the testability of a hypothesis which seems to be at least part of what Peirce, McKaughan and Franklin have in mind when they talk about how ‘easy’ it is to test a hypothesis. It also highlights, plausibly, that ‘testability’ covers several distinct considerations: the testability of a hypothesis includes both how likely we are to get reliable evidence, for or against *h,* as well as how likely we are to get misleading evidence. This seems right: if, for example, the available experimental procedures have some significant chance of producing misleading results, this is surely relevant to deciding whether to pursue a hypothesis.

The utilities can be taken to represent the sense in which pursuing a hypothesis can be “informative”, “interesting” or “of potential interest”. The Simple Model represents this as consisting of two components.

The first is the intrinsic epistemic value of being in a given epistemic state, which depends (in line with the assumption of axiological realism) on the truth value of the hypothesis. I have also assumed that rejecting a falsehood is more valuable than suspending judgment. This corresponds to Musgrave and Peirce’s point that knowing that *h* is false can have some epistemic value, even if we would perhaps prefer to learn that it is true (the model is neutral on which of these is more valuable, if any).

The second component concerns how informative pursuing the hypothesis will be. As explained above, one should take into account the agent’s current epistemic state and consider whether pursuing the hypothesis is likely to change the epistemic state in a favourable direction. In the Simple Model, this point is, for instance, illustrated by the case, described by equation (7), where the hypothesis is already accepted (similar remarks apply to the case where the hypothesis is rejected). Here, if the hypothesis is in fact true, there is no scope in this model to learn more by pursuing the hypothesis further. If the hypothesis is false, on the other hand, pursuing it would have a large scope for being informative: both suspending judgment and rejecting the hypothesis would represent an improvement in the agent's epistemic situation. McKaughan arguably alludes to this factor when mentioning that pursuit should be informative. The Simple Model has the merit of making this point explicit and representing it directly in relation to the other factors.

In addition to representing the different factors relevant to pursuit, the model also allows us to reconstruct some of the arguments regarding pursuit presented above. The simplest is the observation that, contrary to what Hanson and Salmon assumed, not all reasons for pursuit are reasons for acceptance. The probability of the hypothesis plays some role, but it not the only relevant factor. For instance, in the Simple Model one can, all things being equal,[50] increase EU(p_h) by increasing the estimated value of learning whether h is true, i.e. by increasing u($acc(h)$, h) or u($rej(h)$, $\neg h$), or by showing that pursuing it is more likely to generate reliable evidence, i.e. by changing the estimates of the relevant conditional probabilities.

Next, the Simple Model also illustrates why, in some cases, showing that a hypothesis is *less* likely can be a reason to pursue it. As we saw in Chapter 1, Peirce argued that this is sometimes the case (Section 1.4.1). Furthermore, Popper can be interpreted as claiming that the more pursuit worthy hypothesis is *always* less probable (cf. Section 1.3.1). The Simple Model allows us to identify several circumstances under which this holds. One such case was identified at the end of Section 2.5.3 and mentioned earlier in this section: namely, that if a hypothesis is already accepted, then reducing its probability will, all thing beings equal, increase ΔEU(p_h). Here, showing the hypothesis

[50] Here, and in the rest of this chapter, it is necessary to add the "all things being equal"-clause to rule out cases where more than one quantity is modified at the same time.

less probable makes it more likely that pursuing h will be informative, namely by revealing that the agent had mistakenly accepted h.

Peirce and Popper are, however, unlikely to have had this case in mind. Both usually suppose that before pursuing a hypothesis, we neither accept nor reject it. Let me start by reconstructing Popper's reasoning. Suppose that the agent currently suspends judgment about h. We should then focus on equation (6). Due to his anti-inductivism, Popper can be construed as denying that we could ever obtain sufficient evidence to rationally accept h. Thus, we set $\Pr(acc(h) \mid h, p_h) = \Pr(acc(h) \mid \neg h, p_h) = 0$, reducing (6) to:

$$\begin{aligned}\Delta\mathrm{EU}(p_h) &= \Pr(h) \times \Delta\mathrm{u}(rej(h), h) \times \Pr(rej(h) \mid h, p_h) \\ &\quad + \Pr(\neg h) \times \Delta\mathrm{u}(rej(h), \neg h) \times \Pr(rej(h) \mid \neg h, p_h)\end{aligned}$$

Since we assume that $\Delta\mathrm{u}(rej(h), \neg h) > 0 > \Delta\mathrm{u}(rej(h), h)$, decreasing the probability of h will here, all things being equal, increase $\Delta\mathrm{EU}(p_h)$. Conversely, increasing the probability of h will all things being equal decrease $\Delta\mathrm{EU}(p_h)$. Thus, the Simple Model shows why, given Popper's other commitments, he was right to claim that more improbable hypotheses also tend to be more pursuit worthy.[51]

Peirce, however, did believe that we can have reasons for accepting a theory (at least in the long run: he also stressed that we often have to reject many false hypotheses before reaching the true one). However, the more general point still holds that whether increasing $\Pr(h)$ make h more or less pursuit worthy depends on how the utilities and conditional probabilities in (6) balance out against each other. More precisely, decreasing $\Pr(h)$ will increase $\Delta\mathrm{EU}(p_h)$, all things being equal, if and only if:

[51] I here ignore Popper's view that the probability of any general theory is zero. If the hypotheses compared all have the same probability, viz. zero, this would make irrelevant the claim that less probable hypotheses are more pursuit worthy.

$$(8)\qquad \begin{gathered} \Delta\mathrm{u}(acc(h), h) \times \Pr(acc(h)|\ h, p_h) \\ + \Delta\mathrm{u}(rej(h), h) \times \Pr(rej(h)|\ h, p_h) \\ < \\ \Delta\mathrm{u}(acc(h), \neg h) \times \Pr(acc(h)|\ \neg h, p_h) \\ + \Delta\mathrm{u}(rej(h), \neg h) \times \Pr(rej(h)|\ \neg h, p_h) \end{gathered}$$

Consider now one of Peirce's arguments that the improbability of a hypothesis can be a reason for pursuing it:

> if there be any hypothesis which we happen to be well provided with means for testing, or which, for any reason, promises not to detain us long, unless it be true, that hypothesis ought to be taken up early for examination. Sometimes the very fact that a hypothesis is improbable recommends it for provisional acceptance on probation. (CP6.533)

Part of Peirce's reasoning here is, of course, related to the costs, specifically in terms of time. But it also relies on the testability of the hypothesis: the reason why the hypothesis would not detain us long, unless true, is presumably that if it were false, this would be easy to show. Suppose, then, that the probability of getting reliable evidence is much higher if the hypothesis is false than if it is true, i.e. that Pr(rej(h) | ¬h, p_h) >> Pr(acc(h) | h, p_h), and that the remaining quantities roughly balance each other out. In this case, (8) would be satisfied. Thus, the more improbable h, the more reason to pursue it.

Finally, the Simple Model also provides a reconstruction of my argument, made in Section 2.3.2, regarding the claim that plausibility is a necessary condition on pursuit worthiness. As argued, the kernel of truth in this claim is that a hypothesis is only (epistemically) pursuit worthy, if there is some reason to think that something new can be learned from pursuing it. Notice that if, in equation (7), h is certain to be true, i.e. if Pr(h) = 1 and Pr(¬h) = 0, only the first, negative term is given any weight and thus ΔEU(p_h) is guaranteed to be negative. For parallel reasons, ΔEU(p_h) is also guaranteed to be negative if we instead take h to be certainly false and have already rejected it. In other words, in the Simple Model a hypothesis is only minimally pursuit worthy if 1 > Pr(h) > 0. While simplified, this provides an illustration of the point that even if some minimal degree of plausibility is a necessary condition on pursuit, it also the case that a pursuit worthy hypothesis should have some chance of being false. I say simplified, because I also argued that it can be worth pursuing a hypothesis known to be false in order to determine if it contains some truth or to develop it into a potentially true theory; these possible outcomes are not represented in the Simple Model and are thus not taken into account in the above inequality.

2.7. Extensions and Modifications of the Simple Model

2.7.1. More Epistemic States and Truth Values

The Simple Model can be extended and modified in a number of ways. The most obvious has already been indicated in Section 2.5.3, namely to include more epistemic states among the potential consequences and more possible truth values of the hypothesis in the possible background states. These modifications will allow the model to represent a more nuanced range of considerations. It is worth noticing, however, that the logic of the arguments reconstructed in Section 6 is preserved under many of these modifications.

Consider first Peirce's argument that improbability can be a reason for pursuit. Suppose we add more epistemic states and so use equation (4). It is still the case that increasing Pr(h) increases EU(p_h), all things beings equal, if and only if the following holds: $\sum_i[\mathrm{u}(es_i(h), h) \times \Pr(es_i(h) | h, p_h)] > \sum_i[\mathrm{u}(es_i(h), \neg h) \times \Pr(es_i(h) | \neg h, p_h)]$. As long as there are some positive utilities in the right-hand-side of this inequality, there will still be some distribution of conditional probabilities for which decreasing Pr(h)

increases EU(p_h), all things being equal. Popper's argument is of course just a special case of this point. Similarly, if we add further truth values, and so use (3), we can still compare the $\sum_j[\mathrm{u}(es_j(h), t_i(h)) \times \Pr(es_j(h) \mid t_i(h),\ p_h)]$. For any truth value $t_i(h)$, as long as at least one $\mathrm{u}(es_j(h), t_i(h)) > 0$, there will be ways to assign the conditional probabilities such that increasing Pr($t_j(h)$) increases EU(p_h), all things being equal.

Next, consider the arguments which concerning how informative pursuing h would be. These depend on there being, for each truth value $t_i(h)$, some epistemic state $es_j(h)$ which would be the most valuable if $t_i(h)$ is the case. Suppose we start in a state $es_x(h)$ which is optimal for $t_y(h)$. Then, $\Delta\mathrm{u}(es_x(h), t_y(h)) = 0$ and for all $i \neq x$, $\Delta\mathrm{u}(es_i(h), t_y(h)) < 0$ and thus $\sum_j[\Delta\mathrm{u}(es_j(h), t_y(h)) \times \Pr(es_j(h) \mid t_y(h),\ p_h)]$ is guaranteed to be negative. In this case, then, reducing Pr($t_y(h)$) will increase ΔEU(p_h), all things being equal. If we furthermore assume that Pr($t_y(h)$) = 1, then ΔEU(p_h) = $\sum_j[\Delta\mathrm{u}(es_j, t_y(h)) \times \Pr(es_j \mid t_y(h),\ p_h)]$ which, again, is guaranteed to be negative. We here recover the argument that there must be some minimal chance that pursuing h will be informative in order for h to be minimally pursuit worthy.

Under some circumstances, we may still want to say that we could always learn *something* more by pursuing h. If nothing else, one might argue, it is always of *some* value to get a bit more evidence in favour of h. The model can be adapted to reflect this: we can distinguish epistemic states in terms of the total amount of evidence relevant to h which the agent has, in a given situation, represented as an unbounded continuum of epistemic states where the optimal state can only be approached but never reached. Under this assumption, the condition that the agent is already in the optimal epistemic state is never satisfied and the results of the preceding paragraph do not hold. So, if we accept this assumption, even a completely certain hypothesis may still be minimally pursuit worthy. Notice that the parallelism between h being completely certain and completely implausible still holds here: if it is always valuable to get a bit more evidence in favour of h¸ equally, it would always valuable to get a bit more evidence against h.

2.7.2. Anti-Realist Axiologies

The Simple Model relies on the assumption of a realist axiology. The two main anti-realist alternatives are empiricist and problem-solving axiologies.

According to the empiricist axiology, defended primarily by van Fraassen (1980), science aims to discover theories which are empirically adequate, i.e. which are true for all of their observable parts. Accepting a theory here consists in regarding it as empirically adequate. Adopting this axiology does not change the structure of the above models. One can adapt the Simple Model to this axiology by simply changing the background states from h being true or false, to it being empirically adequate or inadequate. The assumptions of how to rank the utilities remain unchanged: accepting an empirically adequate theory is better than suspending judgment about it, which in turn is better than rejecting it, etc.

The problem-solving axiology has been defended by Laudan (1977, 1984) and Nickles (1981) and is arguably implicit in the work of Kuhn, Lakatos and some of Popper's later writings. The general characteristic of this view is that science aims to achieve as much problem-solving power as possible. For a number of reasons, the models developed here are less suited to this axiology.

First, since a single hypothesis considered in isolation cannot be said to solve a problem, the problem-solving approach often only applies to larger units of science, such as paradigms (Kuhn), research programmes (Lakatos) or research traditions (Laudan). My models, by contrast, focus on individual hypotheses. One can of course switch hypotheses for, say, paradigms as the object of pursuit. However, it is less natural to ascribe semantic values such as truth or empirical adequacy to these units, especially if we take them to be partly constituted by systems of practice or know-how. We may, however, be able to identify some constituent parts of a paradigm—e.g. the core theories accepted at a given time—which could be assigned semantic values.

Second, even if we identify core theories which can be ascribed semantic values, many proponents of the problem-solving axiology would still deny that we can assign probabilities to these theories being true. For them, an important motivation for the problem-solving axiology is that they do not think we will ever reach a point from which we can tell whether our paradigms (research traditions/programmes, etc.) are true or likely to be true. Laudan (1984: 51-3), for instance, criticises the realist axiology on the grounds that it proposes a "utopian goal" which we can never tell if we are making progress towards, viz. learning the truth.

Third, related to the preceding point, the value of obtaining a certain level of problem-solving power is often construed as independent of whether our current theories are true or at all close to the truth. Laudan (1977: 109) explicitly argues that this is a virtue of the problem-solving account, since this allows us to determine whether science is making progress.

Finally, it is not clear whether there is anything in the problem-solving axiology which corresponds to the theoretical acceptance of a theory, distinct from (a) choosing to pursue the theory and (b) accepting it for the purposes of some practical application. For instance, as discussed in Section 1.3.3, the only kind of 'acceptance' on Lakatos' view which is distinct from pursuit is what he calls 'acceptance$_3$', consists in applying a theory to a practical problem. There is no natural way to represent this view within my models, as they aim to capture epistemic pursuit worthiness rather than practical pursuit worthiness. Laudan does distinguish theoretical acceptance from pursuit. According to him we should accept the research traditions which have the highest current problem-solving power within a given domain (1977: 109). But even for Laudan, the main implication of accepting a research tradition is to apply it for practical purposes, in particular to use it for developing experiments.

For these reasons, the best way to construct a decision-theoretic model based on the problem-solving axiological framework would be as follows. The objects of pursuit would be research programmes and the outcomes consist in different levels of problem-solving power a given research programme could achieve. The expected utility of pursuing a research programme would be measured in terms of the expected problem-solving power which could be achieved through pursuing that research programme.[52]

The structure of such a model would differ significantly from the ones developed in Section 2.5.3. First, the utility of achieving a certain level of problem-solving power would not depend on the background states. Second, the relevant background states would not be different possible truth values, but whatever factors which might be relevant to estimating how much problem-solving power could be achieved through pursuing a given research tradition. These factors would presumably include a wide variety of factors not intrinsically linked to the truth or empirically adequacy of core theories of the research tradition.

While the problem-solving axiology is still consistent with the overall consequentialist approach to pursuit worthiness advocated here, it is not captured by the decision-theoretic models developed above. The realist axiology will usually provide the most natural framework for discussing the problems I am interested in. Therefore, I will not develop decision-theoretic models based on the problem-solving approach in any more detail in this thesis.

[52] Laudan does not adopt this proposal; instead, he proposed that one should pursue the research tradition with the highest current rate of added problem-solving power. I discuss this in Section 2.8.1.

2.7.3. Costs and Multiple Hypotheses

The utilities in my model of epistemic pursuit worthiness do not take into account the costs of pursuing the hypothesis. As mentioned in Section 2.5.3, one reason for this is that it is not clear whether the costs, in terms of the resources and time available to do scientific research, are commensurable with the epistemic value of learning more about the world. One attractive feature of the Simple Model (and the extensions discussed in Section 2.7.1) is that changes in utility—e.g. the added utility of accepting a true theory we had previously suspended judgment about—can be interpreted as representing how much we would have learned about the world. This is feature that allows it to represent the factor McKaughan (2008) calls informativeness, which is intuitively relevant to decisions about pursuit. This factor would be obscured if u($acc(h)$, h) instead represented some combination of the value of accepting the truth *and* the costs of finding this out.

One way to incorporate costs while avoiding this problem would be to assume that the total utility of a given outcome could be written as a linear combination of the epistemic value of the situation and its costs. If we let c(p_h) be the costs of pursuing *h* we could, for example, replace each utility in equation (3) by one of the form: u[$ej_j(h)$, $t_i(h)$, p_h] = u($ej_j(h)$, $t_j(h)$) + c(p_h). If we assume that the costs of pursuit are independent of the epistemic state reached, we can rewrite (3) as:

(9)
$$\mathrm{EU}(p_h) = \sum\nolimits_i \left(\Pr\big(t_i(h)\big) \times \sum\nolimits_j \left[\mathrm{u}\left(es_j, t_i(h)\right) \times \Pr(es_j \mid t_i(h),\ p_h)\right]\right)$$
$$+ \sum\nolimits_i \Pr\big(t_i(h)\big) \times \mathrm{c}(p_h)$$

If we furthermore assume that the costs are independent of the truth value of *h*, this reduces to:

$$(10) \qquad \mathrm{EU}(p_h) = \sum_i \left(\Pr\big(t_i(h)\big) \times \sum_j \left[\mathrm{u}\left(es_j, t_i(h)\right) \times \Pr(es_j \mid t_i(h),\ p_h)\right]\right) + c(p_h)$$

Here, c(p_h) will typically be a negative number although we could, in principle, imagine cases where a research project is expected to be so profitable that we want to say that pursuing *h* makes a net contribution to the available resources.

A different way to incorporate costs into a comparative notion of epistemic pursuit worthiness, which does not assume commensurability of epistemic value and costs, is to define the pursuit worthiness of a hypothesis in terms of the expected utility gain per unit of resources it would cost to pursue it. If c(p_h) represents the total costs of pursuing *h*, we could say that h_1 is more pursuit worthy than h_2 if, and only if, EU(p_{h1})/c(p_{h1}) > EU(p_{h2})/c(p_{h2}). However, this has the undesirable consequence that we can no longer translate freely between expected utility, expected change in utility and pursuit worthiness. Since c(p_h) is not the same for all hypotheses, it is not the case that EU(p_{h1})/c(p_{h1}) > EU(p_{h2})/c(p_{h2}) if and only if ΔEU(p_{h1})/c(p_{h1}) > ΔEU(p_{h2})/c(p_{h2}). For example, for hypotheses where ΔEU(p_h) is negative, higher costs increase ΔEU(p_h)/c(p_h) and would thus *increase* pursuit worthiness if we define it in terms of this quantity. By contrast, higher costs always reduce EU(p_h)/c(p_h).[53]

A better approach may be to bring in costs as an external constraint on the decision problem. Here, the decision problem facing the agent is how to best spend a given amount of resources, including time and manpower. The agent then needs to consider the different

[53] I here assume that both the costs and the utilities are expressed as positive quantities. For this reason, this quantity can also not be applied in any plausible way to cases where pursuing *h* would make a net contribution of resources.

sets of hypotheses that could be pursued for that amount of resources and choose the set which has the highest expected (epistemic) utility.

This raises the question of how to represent the pursuit of multiple hypotheses in the framework. The natural way to do this is to include all possible combinations of truth-values of the relevant hypotheses in the background states, and to let the consequences include each possible assignment of epistemic states to the hypotheses. To illustrate this in terms of the Simple Model, this would give us the following expression for the expected utility of pursuing the set of two hypotheses $H = \{h_1, h_2\}$:

$$\begin{aligned}(11)\qquad \mathrm{EU}(p_H) = {} & \Pr(h_1\&h_2) \times \textstyle\sum_i[\mathrm{u}(es_i(H), h_1\&h_2) \times \Pr(es_i(H) \mid h_1\&h_2, p_H)] \\ & + \Pr(h_1\&\neg h_2) \times \textstyle\sum_i[\mathrm{u}(es_i(H), h_1\&\neg h_2) \times \Pr(es_i(H) \mid h_1\&\neg h_2, p_H)] \\ & + \Pr(\neg h_1\&h_2) \times \textstyle\sum_i[\mathrm{u}(es_i(H), \neg h_1\&h_2) \times \Pr(es_i(H) \mid \neg h_1\&h_2, p_H)] \\ & + \Pr(\neg h_1\&\neg h_2) \times \textstyle\sum_i[\mathrm{u}(es_i(H), \neg h_1\&\neg h_2) \times \Pr(es_i(H) \mid \neg h_1\&\neg h_2, p_H)]\end{aligned}$$

Here, the $es_i(H)$ cover the six different ways to assign the three epistemic attitudes to the two hypotheses (accepting both, accepting h_1 and rejecting h_2, etc.). Thus, we take into account twenty-four different outcomes. Notice that the utilities here need not be equal to the sum of their constituents. For instance, we do not assume that u($acc(h_1)\&acc(h_2)$, $h_1\&h_2$) = u($acc(h_1)$, h_1) + u($acc(h_2)$, h_2). There may be synergy effects between the two hypotheses, such that learning either would not be particularly interesting but where it would be very interesting if we knew that both were the case. The analogous point holds for the conditional probabilities: we allow for the possibility that pursuing the two hypotheses together may either increase or decrease the probability of obtaining (say) reliable evidence for either. For instance, we do not assume that Pr($acc(h_1)\&acc(h_2) \mid h_1\&h_2, p_H$) = Pr($acc(h_1) \mid h_1\&h_2, p_{h1}$) × Pr($acc(h_2) \mid h_1\&h_2, p_{h2}$).

This way of incorporating costs into the model allows for the costs to be relevant to whether a hypothesis is pursuit worthy in the absolute sense. A hypothesis is absolutely pursuit worthy for a given agent if, and only if, it is part of the set of hypotheses *H* which maximises EU(p_H), while staying within the constraints imposed by the resources available to the agent. It does not, however, provide a way to make costs of pursuit relevant to comparative pursuit worthiness of hypotheses.

How to best incorporate costs into a comparative definition of pursuit worthiness within this model will not be crucial in this thesis. For comparative pursuit worthiness, all I will require for my purposes is that hypotheses which are costlier to pursue are less pursuit worthy, all things being equal.

2.8. Comparison with Other Accounts of Pursuit Worthiness

Having presented my general consequentialist approach to pursuit worthiness, and shown how this can be modelled in decision-theoretic terms, I now want to remark on how this compares with other, recent accounts of pursuit worthiness.

2.8.1. Lakatos' and Laudan's Backwards-Looking Accounts

As discussed in Section 2.7.2, because Lakatos and Laudan assume a problem-solving axiology, their accounts are not naturally captured by the decision-theoretic models developed in this chapter. Here I want to highlight a further difference, namely that both of their accounts of pursuit worthiness are backwards-looking, in the sense that they evaluate pursuit worthiness in terms of the past performance of research programmes/traditions.

As we saw in the previous chapter, Lakatos' methodology of scientific research programmes is most plausibly interpreted as an account of pursuit. On this account, one should pursue those research programmes which, until now, have been empirically progressive (at least intermittently) and have developed in accordance with their positive and negative heuristic. It is thus the past development of the research programme which determines if it should be pursued further. On Laudan's account, one should pursue the research tradition which is currently accumulating problem-solving power at the highest *rate*.

As McMullin (1976) pointed out with regards to Lakatos and Whitt (1990, 1992) argues in more detail, basing pursuit purely on past performance seems misconceived. Assuming that we aim to achieve as much problem-solving power as possible, what is relevant is whether pursuing a research programme/tradition will generate more problem-solving power *in the future*. Laudan seems to be aware of this issue when he claims that his account is "making explicit what has been implicitly described as "promise" or "fecundicity"" (1977: 112). However, strictly speaking, his account only looks at whether the theory has "*recently shown* itself to be capable of generating new solutions to problems at an impressive rate" (111, emphasis added). As Whitt argues (1990: 478-9), we sometimes have good reasons to think that a research programme which has made little progress so far has potential for significant future growth, say, because it is newly formulated. This is also the intuition Lakatos (1970/1978: 70-1) expresses when he insists that "budding" research programmes should be "sheltered for a while from a powerful established rival". Similarly, Laudan remarks that "It is common knowledge that most new research traditions bring new analytic and conceptual techniques to bear on the solution of problems … which, particularly over the short run, are likely to pay problem-solving dividends" (1977: 111). As Whitt (1992) argues, this is not just the case for *new* research programmes. There are several features of research programmes, such as having unused conceptual resources, which we can use to evaluate the future performance of a research programme at any given stage of development.

The problem raised here of course draws on the consequentialist approach to pursuit worthiness, on which I have based my account in this chapter. Laudan and Lakatos might resist adopting this approach, since they are sceptical of our ability to predict the future performance of research programmes. However, as we saw in Section 2.2.2, unless we implicitly assume that the past performance of the research programme (or some other factor) gives us reason to predict something about its future performance, the radical interpretation of Feyerabend's argument remains sound. If we cannot predict *anything* about the future performance of different research programmes, it is difficult to see why we should rationally prefer to pursue some of these rather than others.

2.8.2. Achinstein's Contextual Schema

Based on his case study of Bohr's 1913 atomic model, Achinstein (1993) criticises what he takes to be Peirce's abductive schema for pursuit worthiness and proposes an alternative, more contextually sensitive schema. The schema criticised by Achinstein is the following:

> "*T* is reasonable to pursue if
>
> (a) There is some set of observable phenomena that *T* if true would correctly explain
>
> (b) *T* satisfies some general methodological criteria (e.g. simplicity, consilience)
>
> (c) The practical costs of pursuing *T* are reasonable" (1993: 108).

Achinstein raises a number of problems for this schema. First, he argues that the requirement that there is *some* set of phenomena which *T* can potentially explain is too permissive. Even if we suspect those phenomena to have some explanation, if there is no reason to think *T* is the correct explanation, then "the theory may be of insufficient interest to pursue, especially if there are other more promising theories" (109). Second, he argues that this schema does not take into account that scientists are often interested in answering specific questions, rather in simply obtaining some explanation or other. Bohr did not just set out to construct an arbitrary explanation of the spectral lines in atomic radiation. Rather, Achinstein argues, Bohr was interested in explaining them in terms of a theory of the structure of atoms.

Achinstein concludes that this schema lacks "mention of arguments for individual assumptions", as well as "information about questions to be raised, instructions to be satisfied, and what justification there is for doing so." (110). To address this, he proposes the following, more context-sensitive schema:

> "Theory *T* is reasonable for scientist *S* to pursue if:
>
> (a) There is a set of questions that *S* seeks to answer and a set of instructions that *S* seeks to impose with respect to these answers.
>
> (b) *S* is justified in raising these questions and imposing these instructions, and in believing that *T* provides answers to these questions in ways which satisfies these instructions." (111)

By "instructions", Achinstein means the different constraints which guide what counts as satisfying answers. These include "general methodological criteria", the empirical condition that an acceptable theory needs to satisfy (e.g. be consistent with certain phenomena) and practical constraints, such as that “testing is economical given the resources available” (*ibid.*). His schema is, he notes, highly context-dependent: exactly what counts as an interesting question or an adequate answer will depend on the specific situation. He seems to regard this as a virtue of his revised schema.

Many of the virtues of Achinstein’s schema are also captured by my decision-theoretic account. First, the spirit of his schema is in line with the consequentialist approach taken here: there is a certain goal which the scientists wish to reach—obtaining adequate answers (as specified by the “instructions”) to the questions they are interested in—and they are justified in believing that *T* can satisfy this goal. In terms of the Simple Model, his schema focuses on cases where pursuit is motivated by the possibility of learning that an interesting theory is true—i.e. on cases where u($acc(h)$, h) has a high epistemic value and where Pr($acc(h)$ | h, p_h) and Pr(h) are high enough to give this significant weight. Second, the decision-theoretic approach is also able to accommodate the context-sensitive nature of judgements about pursuit. Agents in different contexts may ascribe different epistemic value to learning whether a certain theory is true, for instance because they interested in different kinds of questions. Similarly, estimates of the probabilities in the model may differ due to difference in contextually available information.

One advantage of decision-theoretic models with structures similar to the Simple Model is that they can cover a wider range of arguments for pursuit. In particular, they are able to represent cases where pursuit is motivated by the desire to show that a theory is false. In this respect, Achinstein’s schema is more restricted; it gives a more specific account of the epistemic value of theories, namely their ability to answer interesting questions. My decision-theoretic approach is neutral on this, but it is consistent with this account. Finally, Achinstein claims that “A balancing of factors is required” (111), i.e. of the factors mentioned in his schema. He takes this to be something that varies contextually as well. For instance, "If the scientist regards it as very important to answer certain questions so that certain empirical conditions are satisfied, and if *T* is the only theory known to do so, it may be reasonable for that scientist to pursue *T* despite the fact that it does not yet have independent warrant or a known means of testing" (*ibid.*). However, his schema gives no account of what the structure of these trade-offs look like. It is only in virtue of the vagueness of his schema that Achinstein can take the existence of these trade-offs into account. In contrast, my decision-theoretic models represent the structure of at least some of these trade-offs explicitly.

2.8.3. Šešelja and Straßer's Potential Justification Account

According to a recent proposal by Šešelja and Straßer, "a theory is epistemically worthy of further pursuit to the extent it is potentially epistemically justified" (2014: 3113) i.e. "to the extent that it can be shown to have a promising potential for contributing to those epistemic goals that determine theory acceptance" (3115).

The motivation for this approach comes from a focus on the large-scale development of science. They argue that the epistemic goals of science are "(a) to gain adequate and accurate knowledge about the world and (b) scientific knowledge should be robust" (3135). By robust, they mean that the body of scientific knowledge is able to resist "perturbations" (3112), i.e. situations where our currently accepted theories run into an intractable anomaly or other crises, some of which may be severe enough to warrant rejecting the currently dominant theory. "These crises", they argue, "we do not want to face empty-handed" (*ibid.*). The solution, they propose, is to pursue a number of potentially epistemically justified theories alongside the dominant one.

In specific terms, Šešelja and Straßer flesh out the notion of potential justification within a coherentist framework, adapting BonJour's (1985) criteria for the coherence of a theory to formulate a sense in which a theory can be *potentially* coherent. The details of Šešelja and Straßer's account are not crucial here. The most significant deficit, from my perspective, is that their focus is on the potential for developing a theory into a satisfying candidate for acceptance. Thus, like Achinstein, their account neglects cases where pursuit is motivated by the aim of learning that a theory is false. It is a merit of my decision-theoretic analysis that it is able to account for these types of cases.

It may be possible for Šešelja and Straßer to extend their approach to account for these cases. They might allow that a theory can be worthy of pursuit in virtue of its potential *incoherence* (or more generally, if it has a significant potential negative justification). This seems especially relevant if it is the currently accepted theory which shows a high degree of potential incoherence: this corresponds, in the Simple Model, to a case where h is accepted but where $\Pr(\neg h)$ is high enough to make it worth finding out whether it is mistakenly accepted. This could also be seen to contribute to the robustness of scientific knowledge: if the currently accepted theory shows signs of running into insurmountable problems, it may be worth exposing these sooner rather than later so that the defective theory can be jettisoned and more promising, alternate candidates developed.

This is not to suggest that Šešelja and Straßer's account can simply be subsumed under my decision-theoretic models. For one thing, while they stipulate the overall robustness of scientific knowledge as an epistemic goal in addition to obtaining accurate or correct theories, I have interpreted the epistemic utilities as only concerned with the latter. On their account, a theory can be pursued without it being currently interesting to know whether it is true or false; rather, it can be valuable in virtue of the robustness it would provide at potential later stages of inquiry.

2.8.4. Esperable Uberty, Fertility, Future Promise and Strategic Reasoning

A number of commentators have pointed out that part of what motivates the pursuit of a theory is often a belief that it will, in a certain sense, be fruitful or fertile of future developments. McMullin calls this 'potential fertility' of a theory (as contrasted with its proven, past fertility), it is what Whitt's (1992) 'indices of theory promise' are supposed to track, while French (1995) refers to it using the Peircean term 'esperable uberty'.[54] The idea is often taken to build on Lakatos' idea that the development of research programmes is guided by a 'positive heuristic'. However, already in 1953, Hesse argued that a physical theory "carries with it suggestions for its own extension and generalisation …, some of which will be misleading and some of which will be useful for further progress" (1953: 212-3). Later, in *Models and Analogies in Science*, she furthermore argued that these "suggestions" are often contained in the guiding analogy of theories and this is the reason why it is reasonable for scientists to pursue theories based on analogies (Hesse 1966: ch. 1).

The idea behind these concepts is that, often, an important reason to further develop a hypothesis, theory or model is that this may lead to the formulation of new models or hypotheses which might be true, rather than that it is likely to be true in its current formulation. For example, we saw in Section 2.2.1 that Bohr did not regard his original atomic model as very likely to be true. Yet, as he wrote to Hevesy, he nonetheless hoped

[54] Peirce (e.g. CP 8.833-8) sometimes tried to cash out the difference between deduction, induction and abduction in terms of a distinction betweeb the 'uberty' and the 'security' (i.e. reliability) of the inferences. While deduction is completely secure, it has little uberty since it merely draws out the implications of existing ideas. Abduction, by contrast, has very little security but this is made up for by its ability to lead us to new ideas, i.e. its high degree of uberty. (Induction is supposed to be somewhere in between the two).

that the “point of view” presented by his model would enable physicists “to obtain knowledge of the structure of the systems of electrons surrounding the nuclei in atoms and molecules” (Achinstein 1993: 98). After Bohr presented the original model it went through a succession of revisions, prompted by various anomalies that were discovered. Strikingly, he and many other physicists regarded it as a major success for the model that it could be *revised* to account for these anomalies.[55]

That this kind of fertility can be an important factor in deciding whether to pursue a theory, though not the only one, is plausible. As Peirce argues (cf. Section 2.4.3 above), when thinking about pursuit in consequentialist terms, or as he says, in terms of ‘Economy’, it is very important to “consider what will happen when the hypothesis breaks down” (CP7.220). These considerations are not represented in my models of pursuit worthiness, since the models only consider the epistemic utility of accepting or rejecting the hypothesis itself. One can of course interpret the utility of correctly rejecting a hypothesis, u(*rej*(*h*), ¬*h*), to include the value of potential future developments. But this would merely conflate a number of distinct considerations into this quantity without adding any clarity to their structure. To represent the structure of these considerations, one would in principle have to include consequences regarding the acceptance or rejection of every future potential development of the theory. But, since a given theory often has an open-ended scope for further development, it would be somewhat arbitrary what to include among these. Furthermore, these considerations are iterative: new developments of a theory may modify its potential for further developments, and may in turn be revised at even later stages of inquiry. It is not clear that trying to include all of these factors in a

[55] This case was discussed by both Lakatos (1970/1978) and McMullin (1968, 1985); see Schindler (2017) for a recent philosophical analysis of the developments of Bohr’s model and the issues raised by Lakatos and McMullin.

model of the type developed here would bring much clarity and I will, at any rate, not try to do so.

In some cases, this problem may be alleviated by a judicious formulation of the hypothesis *h* which is the object of pursuit. For instance, in the Bohr case, rather than taking *h* to be the exact model Bohr presented in his 1913 papers, we take it to refer to the "point of view" which Bohr took the model to represent—e.g. we can take *h* to be a hypothesis of the form "the electrons in atoms are structured roughly along the lines represented by *this* model". This hypothesis is still capable of being rejected, as it indeed was after Schrödinger, Heisenberg and others developed modern quantum mechanics. However, I do not claim that this solution can always be applied or that it allows us to capture all relevant considerations even in cases similar to Bohr's model.

This points to a more general limitation of the decision-theoretic analyses, namely that they are ill equipped to take into account the way that choices about pursuit can influence later stages of inquiry.[56] For instance, the range of hypotheses or theories that are well-formulated enough to be considered serious contenders at a given time, either for pursuit or acceptance, is influenced by the lines of research which have been previously pursued. Similarly, decisions about which hypotheses to pursue now will affect the range of evidence available at later stages as well as which experimental techniques have been developed and calibrated enough to be considered reliable. I do not take these suggestions to be exhaustive. While these factors can often be relevant to decisions about pursuit, they are often too indeterminate to be interestingly captured by a decision-theoretic analysis.

In this thesis, I will refer to considerations about how pursuit will affect later stages of inquiry as *strategic reasoning*. These will play an important role in later chapters,

[56] Some of the ways mentioned here that choices about pursuit affects later stages of inquiry has also been highlighted by Elliot and McKaughan (2009) in the context of debates about the role of non-epistemic values in science.

especially the discussion of medical diagnosis in Chapter 6. For now, I merely highlight them as a limitation of the decision-theoretic framework.

2.9. Conclusion: Advantages and Limitations of the Decision-Theoretic Approach

Let me conclude this chapter by taking stock of my argument. I started by arguing for a general consequentialist approach to pursuit worthiness on several grounds. First, it provides a plausible way to answer a number of common objections to the project of spelling out a distinct, normative account of pursuit. Second, at least for some cases, this approach provides a natural way to interpret what motivates pursuit in scientific practice. Next, I showed how this general approach can be spelled out in terms of decision-theoretic models which allow us to represent several factors independently highlighted as important to pursuit, as well as to explicate a number of plausible arguments regarding how these trade off against each other in determining the (epistemic) pursuit worthiness of a hypothesis. Specifically, I have argued the Simple Model (i) captures the argument that higher probability can sometimes *lower* the pursuit worthiness of a hypothesis and (ii) highlights the importance of considering cases where pursuit is motivated by the aim of *refuting* the hypothesis.

In Sections 2.6 and 2.7 I highlighted a number of limitations of the decision-theoretic models. I regard the following as particularly important. First, the models only aim to capture what I have called epistemic pursuit worthiness, i.e. the extent to which pursuit is motivated by a desire to learn more about the world (in terms of either a realist or an empiricist axiology). They does not take into account pursuit aiming to solve practical problems or, more generally, reasons for pursuit stemming from political or ethical values. Second, even within the scope of broadly epistemic goals, the models are only meant to capture goals having directly to do with the epistemic state the agent has concerning a hypothesis. They do not include, for instance, Šešelja and Straßer's goal of robustness. Third, the models do not include all relevant consequences of pursuit, such as the downstream effects on later stages of inquiry that I noticed above.

Given these limitations, what is the value of these models? In my view, they provide a useful normative perspective from which to think about pursuit. In calling this a normative perspective, I do not mean to imply that the models are prescriptive of scientific practice, nor do I think they describe a rational ideal for reasoning about pursuit. Rather, I consider them idealised models of some factors relevant to pursuit worthiness and the structure of trade-offs between these. As such they are one tool, among others, for theorising about pursuit from a normative perspective. As with any tool, the proof of its merit will be whether it is useful in application. In the rest of this thesis, I will show how these models can be applied, together with other ideas about pursuit that I have developed in this and the preceding chapter, to illuminate certain problems in general philosophy of science, and in specific methodological debates within archaeology and medicine.

Chapter 3. A Peircean View of Explanatory Reasoning

3.1. Introduction

The form of reasoning called *inference to the best explanation* (IBE) has attracted much attention in philosophy of science, and beyond.[57] Briefly, IBE is an inference where the fact that a hypothesis is (in some sense) the best available explanation of one or more empirical phenomena justifies accepting the hypothesis as true (or at least more likely or closer to the truth than its competitors; I discuss these qualifications in Section 3.2 below). Many proponents of IBE also endorse the more general view, which I shall here refer to as *explanationism*, that being a good (potential) explanation generally provides some reasons for the truth of a hypothesis.[58] Explanationism usually involves both a descriptive and a normative claim. On the descriptive side, explanationists argue that *explanatory reasoning*, i.e. reasoning about what would constitute a good explanation of a given range of phenomena, plays an important role in scientific practice. This supposed ubiquity of explanatory reasoning in scientific practice is in turn taken to motivate the normative claim that explanatory reasoning provides a rational or reliable guide to the truth.

The idea that explanation plays an important role in scientific reasoning predates its current popularity by at least a hundred years, namely in Peirce's writings on abduction from the 1860s onwards.[59] One of the features which Peirce thought distinguished abduction from his other two types of inferences, induction and deduction, is that abduction is guided by the aim of identifying hypotheses which can explain some

[57] Outside of philosophy of science, IBE has for instance been used to spell out the position of coherentism (e.g. BonJour 1985, Lycan 2012) in general epistemology and is often invoked in metaphysics in order to defend the methodological soundness of metaphysical theorising (e.g. Lewis 1986). See Day & Kincaid (1994), Minnameier (2004) and Saatsi (2017) for further examples and discussion. I here focus on IBE and explanationism as accounts of scientific reasoning.

[58] Explanationism is sometimes also referred to as explanatory coherentism (e.g. Lycan 2012).

[59] Thagard (1978: 77) also ascribes the idea to David Hartley, Whewell, Leibniz and Descartes.

otherwise puzzling phenomenon. However, as we saw in Chapter 1 (Section 1.4.1), Peirce's mature view of abduction differs significantly from the contemporary notion of IBE. Explanationists take IBE, and explanatory reasoning more broadly, to provide reasons for accepting hypotheses and thus regard it as a species of inductive or ampliative reasoning. By contrast, Peirce held that only empirical investigations can justify accepting a hypothesis. Abduction gives us no reason to regard a hypothesis as true, except insofar as it leads to subsequent empirical testing. Rather, while Peirce agreed that abductions should guide the choice (and generation) of hypotheses, he only understood this in the sense of choosing which hypothesis to investigate further, i.e. which hypothesis to pursue.

I present two arguments in this chapter. First, inspired by Peirce's account of abduction, I want to defend a pursuit worthiness account of the justificatory role of explanatory reasoning in science, and IBE in particular, which I shall call *the Peircean view*.[60] Drawing on the normative account of pursuit worthiness from Chapter 2, I argue that explanatory reasoning and IBEs can plausibly be taken to provide reasons for pursuing a hypothesis, rather than reasons for accepting it. Second, I present a negative argument, challenging the empirical motivation for explanationism.

In support of the Peircean view, I argue that it avoids two well-known problems for explanationism. First, there does not seem to be any connection, at least *prima facie*, between a hypothesis being a good potential explanation and its truth. So why should it be any more likely to be true simply because it would be a good explanation if it *were* true? Call this the *truth-connection problem*. Second, judgments about what counts as a satisfying explanation often seem to rely on subjective or aesthetic criteria, such as elegance or harmony. Furthermore, what counts as a good explanation not only varies

[60] Though this name signals the Peircean inspiration of my view, my aims in this chapter are not exegetical. I do not, for instance, claim that the view defended here is identical to Peirce's considered views on abduction, much less that it captures everything he ever wrote about it.

between different scientific fields; it also (as Kuhn pointed out) changes through time within the same field. But it is unclear how these kinds of subjective and variable judgements could serve as the basis for a rational or reliable form of inference. I will call this the *subjectivity problem*.[61] Explanationists usually respond by trying to deny the premise of these problems, e.g. arguing that the criteria of explanatory goodness are not completely subjective and that there *are* good reasons to regard them as a guide to the truth. By contrast, the Peircean view side-steps these problems altogether. Even if being a good explanation relies on subjective criteria which have no connection to the truth, there are still good reasons to pursue the best explanation. Next, in Section 3.2 I will present explanationism, before introducing the two problems in Section 3.3 and explaining how they challenge explanationism. In Section 3.4, I will show how the Peircean view avoids them.

My negative argument challenges empirical arguments for explanationism. In the second half of the chapter, Sections 3.5 and 3.6, I argue that many of the examples of explanatory reasoning in scientific practice cited by explanationists do not support the empirical premises which are required to support their view. To the extent that explanatory considerations played a role in these cases, it is either (i) doubtful that explanatoriness was used as a guide to truth, rather than as a guide to pursuit worthy theories, or (ii) if it was used as a reason for acceptance, it often guided scientists away from the truth.

Whereas most discussions of empirical arguments for explanationism are framed within the scientific realism debate, focusing on whether they are dialectically effective against antirealists, my criticism in this chapter is independent of the realism debate: it

[61] Lipton (2004: 142-3) refers to these two problems as “Voltaire’s objection” and “Hungerford’s Objection”, respectively. I prefer the labels given here as they wear the heart of the objection on their sleeves.

remains a problem for explanationism even if realism is assumed true. It does, however, have some implications for the realism debate as it reveals a tension between explanationist defences of realism, also known as the "no-miracles argument" (NMA), and the usual realist responses to the so-called "pessimistic meta-induction" (PMI) against realism. The most promising realist responses to PMI—focusing on novel predictive success and shifting to a form of selective realism—only exacerbate the problems I identify for explanationism, and thus throw doubt on the inference form (IBE) that is supposed to power NMA. I consider the implications of my argument for the realism debate in Section 3.7.

3.2. Explanationism, Its Motivation and Refinements

The first to introduce the term 'inference to the best explanation' (in the modern debate at least) was Harman (1965), in an article arguing that IBE provides a more fundamental form of non-demonstrative inference than enumerative induction. He claims that the inference he calls IBE "correspond[s] approximately to what others have called "abduction," "the method of hypothesis," "hypothetic inference," "the method of elimination," "eliminative inference," and "theoretical inference"" (88-9). However, since he takes these labels to have "misleading suggestions" (89), he prefers the IBE-terminology.

Specifically, Harman characterises IBE as an inference where one infers the truth of a hypothesis *H* from the fact that *H* would explain the available evidence. He then notices that since multiple hypotheses can often explain the same evidence, one needs to somehow judge which is the better explanation. Such judgments he takes to be based "on considerations such as which hypothesis is simpler, more plausible, which explains more, which is less *ad hoc*, and so forth" (89). Harman does not specify further what he means by a ‘better explanation’, but he takes it to depend, at least in part, on non-empirical criteria, such as simplicity, scope of explained phenomena and non-ad hocness.

The idea that IBE provides a legitimate form of inference, distinct from enumerative induction, which relies on non-empirical criteria quickly became popular within philosophy. This was motivated by several factors. First, it was supported by Quine’s argument (also used by Harman) that theories are underdetermined by the empirical evidence and that non-empirical reasons therefore play a crucial role in choosing which theories to accept (Quine 1953b, c; cf. Lycan 1985: 159, note 2). Second, IBE also seemed to give a plausible rational reconstruction of the observation that non-empirical criteria play an important role in ‘theory choice’, as highlighted by Kuhn (1977) and others (e.g. Buchdahl 1970). Third, the supposed ubiquity of IBE within science suggested a respectable strategy for naturalistically inclined philosophers to argue for metaphysical conclusions: if scientists use IBE to support theories that go beyond the empirical evidence, and philosophy is in some sense continuous with science, then supporting metaphysical theories through the same form of argument appears more respectable.[62] In philosophy of science, in particular, this was seen to provide a strong argument for scientific realism, the so-called ‘Ultimate Argument’ or ‘No-Miracles Argument’ associated e.g. with Putnam (1975) and Boyd (1983) (cf. van Frassen 1980: 34-50; Psillos 1999: ch. 3).

I will return to the question of how to understand the notion of ‘the best explanation’ in the next section. First, I want to highlight some refinements to Harman’s original formulation that have been proposed in the recent literature.

[62] E.g. Armstrong (1983: ch. 5) argues for the existence of laws of nature on the grounds that they provide the best explanation for the existence of regularities (cf. van Frassen 1989: 138-142). Saatsi (2017) provides many further examples from metaphysics, philosophy of mathematics and meta-ethics.

In its simplest form, an IBE is supposed to proceed as follows: a scientist is faced with range of competing hypotheses, each of which could potentially explain some set of empirical phenomena. Comparing the quality of the explanation offered by these hypotheses, she then infers the best explanation along the lines of the following schema (e.g. Lycan 1985: 138; Psillos 2002: 614):

> (IBE1) *D* is a set of empirical phenomena (data, facts, observations, etc.).
>
> (IBE2) The hypothesis *H* explains *D*.
>
> (IBE3) No other available competing hypothesis explains *D* as well as *H*.
>
> (IBE-C) Therefore, *H* is (probably) true.

An influential objection to this simple schema is van Fraassen's (1989: 142-3) "best of a bad lot" problem. The scientist is only comparing the *available* competing hypotheses, the hypotheses that have actually been formulated so far. Since the range of possible explanations of *D* is vast, most competitors to *H* have not yet been formulated and perhaps never will be (146). We know that most possible explanations of *D* will be false and there does not seem to be any reason to think that the true hypothesis will be among the hypotheses considered so far. Thus, even if being the best explanation is indicative of the truth, we cannot conclude that being the best *available* explanation is.

Explanationists have usually replied to this objection by restating the conclusion of an IBE as a comparative claim about the available hypotheses. There are a number of possible ways to do this. One is to say that being a good explanation only makes a hypothesis *more likely* to be true (van Fraassen 1989: 145-6 mentions this possibility), so that the conclusion of an IBE is rather:

> (IBE-C’) *H* is more likely to be true than the available competing hypotheses.

This does not imply that *H* is more likely to be true than not. A different strategy is instead to claim that being the best available explanation is an indicator of *closeness to the truth* (Douven 2011: §2), such that the conclusion of an IBE becomes:

> (IBE-C’’) *H* is closer to the truth than the available competing hypotheses.

Again, *H* can be *closer* to the truth without being close to it in absolute terms. The two strategies can, furthermore, be combined to yield conclusions such as:

> (IBE-C’’’) *H* is more likely than the competing hypotheses to be the closest to the truth out of the available hypotheses.

Each of these construals of IBE avoids the charge that we are assuming an implausible ability to ensure that the true explanation will be among the potential explanations we can consider at a given time. These responses still allow explanationists to endorse the simple version of IBE in the special cases where we do have good reason to think that the correct hypothesis is among those generated.

A further refinement, proposed by Lipton (2004: 148-63), attempts to give an argument for why we should expect the correct hypothesis to be generated (at least in the long run). It rests on the observation that explanatory reasoning does not just guide the *selection* of hypotheses but also informs the *generation* of available hypotheses. This seems plausible. As Hanson and Peirce highlighted, scientists often generate new hypotheses by considering what could possibly explain an otherwise puzzling phenomenon. Of course, scientists do not try to generate every conceivable hypothesis. Rather, they rely on their background knowledge of what tend to be regarded as good explanations within their field to formulate a limited range of potential explanations which, in light of this background knowledge, seem promising. Building on this observation, Lipton argues that if explanatory considerations can be used to reliably rank the available hypotheses in terms of their likeliness (as van Fraassen grants for the purposes of this objection), we also have a reason to think that they are somewhat reliable in generating new hypotheses (2004: 151ff). After all, Lipton reasons, the hypotheses which make up the background knowledge guiding the generation of new hypotheses were ranked highly by previous applications of IBE. In other words, Lipton's argument is that repeated applications of a reliable selection criterion, along with subsequent hypothesis-generation informed by the results of this selection, will tend to increase the reliability of the generation step as well. Thus, in the long run at least, we have good reason to assume that IBE will guide us towards true hypotheses.

My purpose here is not to endorse Lipton's argument. For one thing, one may question the premise that generating hypotheses based on background theories which are merely likelier or closer to the truth than potentially very poor competitors, should increase the reliability of the generative process.[63] Rather, I want to highlight a few aspects of these refinements to the simple version of IBE which will be relevant to my further discussion.

First, the responses to the "best of a bad lot" objection differ in terms of whether they take being the best explanation as a reason to regard the hypothesis as the likeliest,

[63] Stanford (2006) argues, based on case studies, that we have empirical reasons to think that scientists often fail to generate the correct explanations.

or the otherwise closest to the truth (compared to the available competitors). What they have in common is that the fact that *H* is the best available explanation does not always provide reason for regarding *H* as true, but rather for regarding it as having some truth-related property such as being more likely or closer to the truth. Since the objections I will discuss below are unaffected by these distinctions I will, for ease of exposition, often not distinguish them carefully in the following. When I talk of explanatory reasoning providing *reasons for the truth* of a hypothesis or being a *reliable guide to the truth*, this should be read as also covering these other truth-related properties.

Second, the above refinements naturally lead explanationists to the claim that being a better explanation can serve as a guide to evaluate any given hypothesis in terms of its likeliness or truth-closeness, rather than merely the best explanation. For one thing, the comparative conclusions (IBE-C')-(IBE-C''') are supposed to be justified independently of what the set of competing hypotheses are. Suppose that out of a given set we remove the best explanation. Then, the second-best explanation will be the best explanation in this new, smaller set and thus the most likely/closest to the truth in the new set (though of course no more likely/close to the truth than the explanation which was removed from the original set). This can be repeated until we have ranked all hypotheses from the original set. Besides, it seems a natural extension of IBE to say that, just as the best explanation is the likeliest (or closest to the truth), the second-best explanation is the second-likeliest, and so on. For these reasons, I shall interpret explanationism not just as committed to (some form) of IBE being correct, but to the more general view that the 'quality' of an explanation provides some reasons for its truth.

Finally, as Lipton's proposal highlights, explanatory reasoning can play a role in both the selection and generation of hypotheses. While most discussions of IBE focus on the former, similar questions arise about the reliability of generating hypotheses through explanatory reasoning. Although I shall focus on hypothesis selection in this chapter, the Peircean view can equally be applied to account for generative uses of explanatory reasoning.[64]

3.3. Two Problems for Explanationism[65]

In the preceding section, I took for granted the idea that potential explanations can be 'better' than each other. However, it is important to recognise that the slogan "infer the best explanation" conceals an important distinction between two ways an explanatory hypothesis can be better than its competitors (Lipton 2004: 59-65). In one sense, a hypothesis can be better simply because we think it is more likely or closer to the truth than its competitors. For instance, we may be able to rule out, or show highly improbable, all plausible alternative explanations in light of our available evidence and accepted background theories. Here, the remaining hypothesis would be the likeliest available explanation, and in this sense the best. However, as defenders of explanationism such as Lipton (2004: 60-62) and Psillos (2002: 617) point out, if explanationism merely recommends inferring the hypothesis which we already think is most likely or closest to the truth, it would be a fairly uncontroversial but also rather uninteresting position.[66] What motivated the interest in IBE is that the *explanatory* quality of the competing hypotheses is supposed to give us an independent, non-empirical criterion for choosing between hypotheses. It is this feature which (i) allows IBE to provide a solution to Quine and Harman's worries about the empirical underdetermination of theories, (ii) makes it an

[64] See also my discussion of the possibility of normative accounts of generative reasoning in Chapter 1, Sections 1.4 and 1.5.

[65] This section and the next draw on and expand the argument presented in Nyrup (2015).

[66] As the "best of a bad lot" objection shows, even this inference is not completely uncontroversial, depending on what is concluded on the basis of an IBE. See also Achinstein's (1990) criticism of 'the only game in town' inferences.

account of the use of non-empirical criteria of scientific theory choice and (iii) would make IBE a means for supporting metaphysical theories. For the purposes of my discussion in the rest of this chapter, I will ignore the trivial interpretation of IBE.

The more interesting version of explanationism I examine here, then, brackets what we believe about the truth of the hypotheses and instead focuses on their explanatory qualities. Let us say that the *explanatoriness* of a hypothesis *H* consists in how satisfying *H* would be *qua* explanation if it were true.[67] This will generally depend on the number and quality of explanations that *H* would provide if it were true.[68] Since the quality of an explanation is often taken to consist in how much understanding it provides, we might also say that the explanatoriness of *H* consists in the amount of additional understanding *H* could potentially afford us. There are of course different accounts of explanation (causal, unification, etc.), and these emphasise different theoretical virtues (simplicity, unification, coherence, elegance, quantitative precision, specifying a mechanism, etc.) as being characteristic of good explanations (Thagard 1978; Lycan 2002: 414-16; Lipton 2004: 122). Often, these criteria overlap with the theoretical virtues that Kuhn (1977) highlighted as important to theory choice. However, since the arguments of this chapter will not depend on any particular view of explanation or understanding, I will bracket these details and simply assume that it makes sense to distinguish between more and less explanatory hypotheses. Whichever accounts of these matters suffice for explanationism will be equally good for my argument.

[67] 'Explanatoriness' here corresponds to what Lipton (2004: 59) calls 'loveliness'.

[68] This subjunctive formulation is also motivated by Lipton's view (2004: 57-8) that only true hypotheses can be genuine explanations. Some philosophers deny that explanation requires truth (e.g. van Fraassen 1980) or even hold that achieving understanding sometimes requires *sacrificing* truth (Cartwright 1983: ch. 2). For the purposes of this chapter I will follow most explanationists in assuming that successful explanation requires truth. Notice that in my decision-theoretic argument for why explanatoriness justifies pursuit (Section 3.4 below), if an explanatory hypothesis can be valuable even if it is false, this only strengthens the argument.

Given this focus, the core claim of explanationism is that having a high degree of explanatoriness can give us some additional reason for the truth of a hypothesis. Of course, explanationists do not claim that explanatoriness should trump all other considerations; there may be other independent empirical or non-explanatory theoretical reasons which tell against the truth of the most explanatory hypothesis (Lipton 2004: 61). Thus, they still hold that which hypothesis counts as "the best" explanation in the IBE inference schema is determined by the likeliness (or truth-closeness) of the hypotheses. What attenuates the charge of triviality here is the claim that explanatoriness can serve as a *guide* to the truth of a hypothesis, in addition to any other non-explanatory considerations. I will call this the *truth-guidance claim*.

The truth-guidance claim is what makes explanationism interesting but, when interpreted as a normative claim, it is also the source of some of the most pressing problems for explanationism. These were voiced already by Reichenbach (1938), in response to Nagel's (1936: 508) observation that scientists often accept hypotheses on partly the basis of non-empirical "esthetic grounds":[69]

> It may be true that a physicist believes in his theory because he thinks it to satisfy esthetic standards; but I do not see any reason why *we* should believe in predictions which are based on esthetic arguments; or why a *technician* should do so. I do not see any relation between esthetic qualities and predictional qualities—and the latter are what a good theory must have. The beauty and harmony of a theory is a matter of taste; it should be easy to construct theories of an extreme beauty which are obviously false. … I cannot accept the esthetic argument as anything connected with the validity of scientific theories in an objective sense; i.e. as an argument which makes the acceptance of a theory justifiable (Reichenbach 1938a: 34-5, original emphasis).

[69] Nagel raised this objection in a review of Reichenbach (1935c). See my discussion in Chapter 1, Section 1.4.2.

Reichenbach here raises two distinct objections, both of which have been echoed by later critics of explanationism. First, explanatoriness (like Reichenbach's "esthetic qualities") seems too subjective to provide a plausible, objective guide to the truth—they are a "matter of taste" and therefore not objective enough to make a theory justifiable. This corresponds to what I, in Section 3.1, called the subjectivity problem. This is also the objection alluded to by Hacking's (1984: 167) quip that IBEs are just inferences "from what makes our minds feel good". Lipton (2004) similarly notices that if explanatoriness (like beauty) is merely "in the eye of the beholder" (143),[70] then it becomes unclear how it could provide a reliable guide to truths about the world.

Second, Reichenbach points out that there is no obvious logical or conceptual connection between the explanatoriness of a hypothesis and its truth. This is what I call the truth-connection problem. Why should the fact that a hypothesis would be a good explanation if it *were* true have any implications for whether it *is* in fact true? Just like there are many beautiful and harmonious theories which are clearly false, there are many false theories which would provide very good explanations if they were true. As Lipton notes, with a nod to Voltaire, "Why should we live in the loveliest of all possible worlds?" (2004: 144); to assume so seems worryingly close to a form of wishful thinking. Furthermore, this is not just an abstract logical possibility. As Duhem (1954) and later Laudan (1981) highlighted, the history of science contains many theories which, in their respective times, were regarded as excellent explanations of the same phenomenon. Since these explanations are mutually incompatible, most of them must be false (Cartwright 1983: 89-91).

[70] Since the Irish writer Margaret Hungerford is thought to be the first to use the phrase "beauty is in the eye of the beholder", Lipton calls this problem 'Hungerford's Objection'.

To be clear, I do not regard these objections as knock-down arguments against explanationism. What they highlight is that the truth-guidance claim cannot simply be assumed. Explanationists need to give some positive argument for it. Of course, explanationists have proposed a number of solutions to these problems. In effect, these attempt to deny the premise of the problems by arguing that the criteria for explanatoriness are not completely arbitrary, and that explanatoriness *can* be a reliable or rational guide to the truth.

I will consider several arguments that explanationists have proposed for the latter claim in Sections 3.5 and 3.6. As I argue there, these face serious problems. First, I want to argue that for the Peircean view, these problems do not arise at all: even if explanatoriness is completely subjective and unconnected to the truth, it can still provide reasons for the pursuit of a hypotheses. Thus, the Peircean view side-steps the problems altogether.

3.4. How Explanatory Reasoning Justifies Pursuit: The Peircean View

According to the Peircean view, IBE is primarily an argument for pursuing a hypothesis rather than an argument for accepting it. More generally, having a high degree of explanatoriness provides some reason for pursuing a hypothesis. In this section, I will first present an account of how explanatory reasoning justifies pursuit before showing how this allows the Peircean view to avoid the problems outlined in the preceding section.

In a nutshell, my account of why explanatoriness provides reasons for pursuing a hypothesis *H* is that this makes it epistemically valuable to learn that *H* is true. First, consider how this account works in an IBE. We start from the premise that *H* would provide the most satisfying explanations (or would provide the most understanding) out of a set of rival explanations, if it were true. Thus, if we were to learn that *H* is in fact true, this would be an epistemically valuable outcome, and indeed the optimal epistemic outcome as far as explanation is concerned. Suppose, then, that everything else is held equal between a set of rival hypotheses: the costs of pursuing them are the same, we regard it as equally likely that pursuing them would give us reliable evidence for or against them, all other expected epistemic outcomes of pursuing them are equal, and so on. In this case, assuming the decision-theoretic approach to pursuit worthiness defended in Chapter 2, scientists would be justified in pursuing the most explanatory hypothesis.

To illustrate this account, consider the following analogy. Suppose a team of treasure hunters know of a large treasure which could be buried on one of two islands, I_1 and I_2. As far as they know the treasure is equally likely to be on either island, but they only have the resources to send an expedition to explore one of them. However, they do know that due to the acidity of the soil on I_2 the treasure is likely to be significantly damaged if buried there. They estimate that if the treasure is instead buried on I_1, it could be worth up to twice as much as if it were buried on I_2. Assume this does not give them any further information about where the treasure is, or how difficult or expensive it would be to recover. In this situation, it would be more rational, for obvious decision-theoretic reasons, to send the expedition to explore I_1 rather than I_2.

To spell out my argument in more detail, notice first that the epistemic goals of science include more than simply knowing as many truths as possible. As Kitcher puts the point:

> Tacking truths together is something any hack can do. … The trouble is that most of the truths that can be acquired in these ways are boring. Nobody is interested in the minutiae of the shapes and colors of the objects in your vicinity, the temperature fluctuations in your microenvironment, the infinite number of disjunctions you can generate with your favorite true statement as one disjunct, or the probabilities of the events in the many chance setups

> you can contrive with objects in your vicinity. What we want is *significant* truth (1993: 94).

There are plenty of trivial truths out there that could be discovered and at much lower cost than the hypotheses actually pursued by scientists. The value of scientific knowledge depends on other factors beyond the amount of truths known, no matter how certain these are.

Now, what other epistemic goals are important in science is not something I need a general account of here. I only need to make two assumptions: first, that it makes sense to distinguish between hypotheses in terms of their explanatoriness—which explanationists also assume—and, second, that having better explanations is in fact more epistemically valuable, all else being equal. In other words, I assume that having good explanations or achieving understanding of the world are among the goals of inquiry, an assumption which is shared by most philosophers of science and explanationists in particular.[71] *One* way a hypothesis can be more epistemically valuable than merely being true is by being a good explanation or by increasing our understanding of one or more phenomena.

Given these assumptions, consider the situation in terms of the Simple Model of epistemic pursuit worthiness developed in Chapter 2 (Section 2.5.3). We can express the assumption that explanatoriness is *one* important epistemic goal as the claim that if h_1 is more explanatory than h_2, then $\mathrm{u}(acc(h_1), h_1) > \mathrm{u}(acc(h_2), h_2)$, all else being equal.[72] Notice now from equation (5) that $\mathrm{u}(acc(h), h)$ only occurs in one place, namely in the sum weighed by $\Pr(h)$. Since probabilities are always positive, it follows that if $\mathrm{u}(acc(h_1),$

[71] E.g. Kitcher (1993: 105ff) highlights “Explanatory Progress” as a goal pursued by science beyond mere truth.

[72] This is “all else being equal” since h_2 might be more valuable in terms of other epistemic goals besides explanatoriness.

h_1) > u($acc(h_2)$, h_2) then EU(p_{h1}) > EU(p_{h2}), all else being equal. Thus, if h_1 and h_2 only differ in terms of their explanatoriness, this gives us a reason to prioritise the pursuit of the more explanatory hypothesis.

So far, this argument shows that IBE can justify pursuit if all else is equal. In other words, explanatoriness can act as a tie-breaker when deciding which hypothesis to pursue. But, more generally, it is also clear that having a high degree of explanatoriness adds to the expected epistemic value of pursuing a hypothesis and thus provides *some* additional reason to pursue it, although not always a *decisive* reason. In deciding which hypothesis to pursue, all things considered, one should weigh the explanatoriness of a hypothesis against the relevant factors along the lines discussed in Chapter 2.[73] This is of course as it should be. It is analogous to the observation, made above in Section 3.3, that explanationists allow for explanatoriness to be outweighed by other reasons for the truth of a hypothesis.

Given this account of how explanatoriness justifies pursuit, the argument for why neither of the objections to explanationism pose a problem for the Peircean view is straightforward. Start with the truth-connection problem: since nothing in my account requires a connection between the explanatoriness of a hypothesis and its truth, the truth-connection problem does not arise. As argued in Chapter 2, although the likeliness or plausibility of a hypothesis is one important factor in deciding whether to pursue it, it is neither the only one nor always a positive reason for pursuit. By contrast, raising the utility of achieving a given epistemic state, whether in the Simple Model or any of the

[73] Notice that which hypothesis to pursue is decided after fixing our estimates of all relevant factors. If we *discover* that a hypothesis is more explanatory than we previously thought, or *change* it to become more explanatory, this can influence our estimates of other factors. So changes, say, to the plausibility of the hypothesis may outweigh any gains in explanatoriness. Analogously, for the treasure hunters, if knowing the acidity of the soil for some reason provides additional clues about whether the treasure is likely to have been buried on the island, this needs to be taken into account.

more sophisticated models, never lowers, and in most cases raises, the pursuit worthiness of a hypothesis, all things being equal.[74]

Turning to the subjectivity problem, my account is compatible with a wide range of views of what makes explanatoriness valuable, including radically contextualist or subjectivist ones. First, as mentioned, my account does not to take a stand on which criteria (unification, mechanism, parsimony, etc.) characterise good explanations or on how they should be weighed against each other. Furthermore, it is possible for these criteria to vary between different times, contexts or paradigms, as Kuhn (1962/1996; 1977) argues, without there being an objective, independent fact of the matter as to which type of explanations is best. As long as the agent evaluating pursuit can distinguish hypotheses in terms of whether they (if true) would constitute better and worse explanations, this allows them to use explanatoriness as a reason for pursuit. Two agents may disagree on which criteria of explanatoriness to use, say, because they belong to competing paradigms. If we accept, with Kuhn, that neither set of criteria is objectively more correct, the Peircean view would simply say that this is a case where reasonable people can disagree about which theory is most pursuit worthy. On the other hand, if we could formulate an objectively correct standard for explanatoriness, we could use this standard to say that while both agents are acting reasonably, from their own perspective, only one of the theories is truly more pursuit worthy. Either way, the Peircean view is consistent with both contextualist and objectivist accounts of explanatoriness.

Similarly, the Peircean view is consistent with a range of different accounts of why having good explanations or understanding is valuable. One could insist that understanding is somehow objectively or intrinsically valuable. But, equally, one could

[74] The only cases where increasing the utility of an epistemic state does not raise pursuit worthiness are (i) where the probability of achieving that outcome is nil and (ii) where the agent is already in that epistemic state.

hold that highly explanatory theories are valuable because they allow us to achieve other goals. These could be epistemic goals: for instance, Douglas (2009: ch. 5) argues that theories with 'cognitive' virtues (e.g. simplicity) are valuable because they are easier to make good predictions. Woody (2004, 2015) argues that the value of explanations consists in shaping and communicating the epistemic priorities within a given field of research. Alternatively, the value of explanatoriness could be grounded in our practical, non-epistemic goals. Kitcher (2001a: ch. 6), for example, argues that whether something counts as a 'significant' question depends on its relation to other significant theoretical problems. Ultimately, these significance networks bottom out in practically (politically, economically, ethically, …) significant problems. Finally, my argument is consistent with the view that the epistemic value of having good explanations is purely subjective, that it is simply a matter of "making our minds feel good". (I do not, however, find this a particularly plausible account of the epistemic value of explanatoriness.)

I do not intend to argue for any specific account of the value of explanatoriness. Rather, this brief survey will illustrate the range of available options, all of which are consistent with the Peircean view. In each case, as long as having better explanations is in fact valued within the perspective from which pursuit is being evaluated (either an external, objective perspective or the internal perspective of the agent), explanatoriness provides a reason for pursuit. Thus, the subjectivity problem does not arise.

3.5. Indirect Arguments for Explanationism

Although explanatory reasoning can be used to justify pursuit, as argued above, this does not mean that it cannot *also* be used as a guide to the truth. Even though the Peircean view avoids the truth-connection problem and the subjectivity problem, it does not follow that explanationists could not answer them. The Peircean view does not rule out explanationism being true as well. In this section and the next, I will criticise some of the arguments explanationists have offered for their view.

There are different possible, non-empirical strategies for defending explanationism. I will not review all of them here.[75] Rather, I will focus on arguments for the truth-guidance claim which are empirical, in the sense that they rely on descriptive claims about the use of explanatory reasoning in scientific practice. I distinguish between two kinds of arguments (Thagard 1988: 139-44; Douven 2011: §3.2): indirect and direct.

3.5.1. Spelling Out the Empirical Premise of Indirect Arguments

By indirect arguments, I mean arguments which rely on general claims about the role of explanatory reasoning in scientific practice. These, together with other general assumptions about science (usually some form of scientific realism), are taken to provide support for the truth-guidance claim.

Thagard (1988:144) states a fairly straightforward form of the indirect argument thus: "If we can show that scientific inquiry in general leads to truth, and that inference to the best explanation is a central part of that inquiry, then we can conclude that inference to the best explanation leads to truth." Similarly, Lipton (2004: 148) observes, that since "we believe that our inductive methods are pretty reliable", doubting the reliability of IBE "would thus undermine confidence in its descriptive accuracy as well". But, he notes, if this is right the argument can also be run in reverse: "My hope is rather that by this stage you are convinced of the descriptive merits of explanationism, so insofar as you believe that our actual practices are reliable, you will tend to discount [this] objection" (*ibid*.). A natural reconstruction of these arguments would be:

[75] Examples of non-empirical arguments for explanationism include (i) that it provides a rationale for independently plausible principles of inductive inference (White 2005), and (ii) that it is vindicated by a form of objective Bayesianism (Henderson 2014). I will not discuss these arguments further in this chapter.

(P1) IBE is a central part of scientific inquiry.

(P2) Scientific inquiry generally leads to approximately true hypotheses.

(C) IBE generally leads to approximately true hypotheses.

The second premise is a statement of scientific realism and, as mentioned in the introduction, most discussions of empirical arguments for explanationism are framed within that debate. Since the most common argument for scientific realism, NMA, is usually construed as an IBE (Psillos 1999: 78ff), anti-realists have complained that this argument is problematically circular—e.g., Laudan (1981: 45) complains that the NMA is the “The Realists’ Ultimate ‘Petitio Principii’”. Accordingly, most discussions of indirect arguments have focused on whether they can be defended in a non-circular way (Thagard 1988: 149-50; Psillos 1999: 81ff). For my purposes, however, I am happy to grant realism and focus solely on whether the arguments support explanationism.[76]

The real problem with the above formulation of the indirect argument is that it is invalid: it commits the fallacy of division. From the premise that scientific inquiry *as a whole* is reliable, it does not follow that any particular part of scientific inquiry is also reliable *on its own*. For one thing, it is possible that scientific inquiry, as it currently exists, contains components which *subtract* from its overall reliability. Leaving this worry aside, the premises of this argument also fail to exclude the possibility that explanatoriness plays a role *different* from being a guide to truth. The Peircean view here provides one salient alternative, namely that explanatoriness is a guide to constructing and choosing hypotheses that are worth pursuing further. As argued above, choosing which hypotheses

[76] Interestingly, Psillos (2011b: 33-4) has recently suggested that the main point of the NMA is to justify IBE to those who already accept the realist framework.

to pursue is important for increasing the epistemic output of science and explanatoriness provides one important factor in making these choices. Thus, explanatory reasoning can play a crucial role in science even if it does not contribute directly to the reliability of scientific inquiry.

To formulate a more plausible version of the indirect argument, let us try to strengthen the premises. Staying as close as possible to Thagard's formulation, I propose:[77]

> (P1*) IBE is a central part of scientific inquiry *for selecting which hypotheses to accept as (approximately) true*.
>
> (P2) Scientific inquiry generally leads to approximately true hypotheses.
>
> (P3) Scientific inquiry would only lead to approximately true hypotheses if most of its central methods for selecting which hypotheses to accept as (approximately) true are reliable.
>
> (C) IBE is (probably) a reliable method for selecting which hypotheses to accept as (approximately) true.

Or, if we want to stay closer to Lipton's formulation:

[77] Notice, the following reconstructions are not deductive arguments. Rather, they are instances of direct statistical inference, i.e. they infer the probability of a property (viz. reliability) from its statistical prevalence within the reference population (viz. scientific methods for accepting theories). I take this to be a very plausible form of inductive argument.

> (P1*) IBE is a central part of scientific inquiry *for selecting which hypotheses to accept as (approximately) true*.
>
> (P2*) Most methods used in science for selecting which hypotheses to accept as (approximately) true are reliable.
>
> (C) IBE is (probably) a reliable method for selecting which hypotheses to accept as (approximately) true.

Questions might be raised about premises (P3) and (P2*). They are certainly more controversial than a mere allegiance to scientific realism. The point I want to make, however, is that even if these premises are granted, the first premise now makes a much stronger empirical claim. Most explanationists may be happy to accept something like (P1*). However, as I argue below, it is far from clear that the case studies typically cited in favour of explanationism can actually support this premise.

Explanationists may object that my reconstruction of the indirect argument is uncharitable. However, it would be up to explanationists to propose a better formulation. I suspect that any plausible reconstruction of the indirect argument will rely on a similarly strong empirical premise. To illustrate, consider a more complicated example, namely Psillos' version of the NMA. This argument can be represented as a two-step argument (following Psillos 2011b: 23-4 and Iranzo 2008: 116, slightly restated for conciseness):

(A)

> (A1) Scientific methodology is theory-laden.
>
> (A2) These theory-laden methods lead to correct predictions and experimental successes (instrumental reliability).
>
> (A3) The instrumental reliability of scientific methodology is best explained by the background theories being approximately true (in relevant respects).
>
> (C1) Therefore, by IBE, the background theories are approximately true.

(B)

> (B1/C1) The background theories are approximately true.
>
> (B2) These theories have themselves been typically arrived at by IBE.
>
> (C2) Therefore, IBE is reliable: it tends to generate approximately true theories.[78]

Part (A) of the argument supports, via IBE, a version of scientific realism, while part (B) is supposed to show IBE reliable. Again, much discussion of this argument has focused on whether the fact that part (A) relies on IBE makes it viciously circular (e.g. Busch 2008, Iranzo 2008, Psillos 2011b). However, since I am happy to grant scientific realism, I will focus on part (B) and simply take (B1/C1) for granted.

To evaluate this part of the argument, then, we need to consider two connected questions. First, what does it mean that IBE "tends to generate approximately true theories", i.e. what does its reliability amount to? Second, in what sense are the background theories supposed to have been "arrived at" by IBE?

[78] Iranzo (2008: 116) takes this to be a deductive inference. But since the conclusion clearly intends to establish the reliability of IBE for future applications as well, it is rather an inductive projection of a statistical pattern.

According to Psillos, a reliable inference is “truth-conducive” if it “tends to generate true conclusions when fed true premises” (2011b: 24). Iranzo (2008: 116-7) further specifies that inferences are reliable if, and only if, they yield a “high rate” of approximately true conclusions (given true premises), a definition which Psillos (2011b: 30) does not challenge. But then the argument as stated is invalid: from the fact that IBE has generated *some* true theories—i.e. the background theories currently used—it does not follow that it has generated a *high rate* of true theories. To avoid this problem, we need to add a premise along the lines of:

> (B3) Most (or many) other theories arrived at through IBEs are true (except when they are fed “false premises”, e.g. misleading data or mistaken estimates of explanatoriness).

Exactly what proportion of these other theories have to be (approximately) true in order for it to qualify as a “high rate” is not crucial to my argument here.

To answer the second question, consider now the two interpretations of (B2), corresponding to explanationism and the Peircean view:

> (B2*) These theories were typically *accepted* by an IBE.
>
> (B2**) These theories were typically *selected as most pursuit worthy* by an IBE.

One might think that the attitude of the scientists is irrelevant, as long as the theories indicated as the best explanations have tended to be true. After all, whether the theories were accepted or merely pursued, if they tended to be (approximately) true, then IBE tends to lead to true theories. However, which of the two interpretations we adopt is crucial when considering which theories to include when evaluating (B3); this will determine whether we should include only those theories which scientists have *accepted* or all the theories which they have *pursued*. The plausibility of (B3) is strongly dependent on this interpretation: whether or not most of the theories scientists have rationally accepted can be construed as partially true (see my discussion of PMI below), no one would claim this of most theories scientists have pursued. On the contrary, scientists often pursue and rule-out (e.g. through testing or theoretical arguments) many false theories before they strike on the one they end up accepting. Thus, the most plausible interpretation of Psillos' argument also relies on a strong descriptive claim, (B2*), regarding the use of IBE in scientific practice, analogous to (P1*) above.

3.5.2. Is IBE Widely Used to Accept Theories in Scientific Practice?

To start examining the empirical premises required for these arguments, we need to consider which cases to use. An extensive survey is obviously beyond the scope of this chapter, but to focus on a few cases might risk the charge of cherry-picking. My strategy for avoiding this charge is to focus on a few cases which are cited by explanationists in favour of their view. By showing why I think these cases are less favourable to the explanationists than is usually supposed, I hope to indicate the kinds of problems a serious empirical case for explanationism would have to overcome.

Explanationists commonly make strong claims about the ubiquity of IBE in scientific practice, citing a number of supposed instances. A few examples include:

> Uses of the inference to the best explanation are manifold. ... When a scientist infers the existence of atoms and subatomic particles, he is inferring the truth of an explanation for various data which he wishes to account for (Harman 1965: 89).

> The goodness of explanations is a ubiquitous criterion; in every scientific subject it forms one of the principal standards by which we decide what to believe. … [For example:] The Copernican explanation of the regularities of the superior planets … The Daltonian explanation of the law of definite proportions … The general relativistic explanation of the anomalous motion of Mercury's perihelion … Spearman's explanation of the correlations among intelligence tests (Glymour 1984: 173-6).

> These sorts of explanatory inferences are extremely common. … The astronomer infers the existence and motion of Neptune, since that is the best explanation of the observed perturbations of Uranus. Chomsky infers that our language faculty has a particular structure because this provides the best explanation of the way we learn to speak (Lipton 2004: 56).

Despite these strong claims, little evidence is given that explanatory considerations actually played an important role in the scientists' acceptance of these hypotheses. All of these cases of course involve scientists trying to explain some range of otherwise puzzling phenomena, and in doing so they would consider or formulate a number of potential explanatory hypotheses. But this is consistent with the Peircean view, namely that the high explanatoriness of these hypotheses merely provided a good reason for pursuing them first, before trying out other, less satisfying potential explanations.

In general, for cases of this type to support explanationism, one has to show at least three further things. First, that *explanatoriness*, i.e. considerations regarding the comparative quality of the competing explanations, actually played a role in the reasoning of the scientists in the case. Second, that these considerations played a central role in determining which hypothesis the scientists *accepted*. Third, that the acceptance of the hypothesis on these explanatory grounds was regarded as *justified* by the peers of the scientists, i.e. that the case does not focus on an unrepresentative outlier of the general scientific judgment at the time.

To illustrate these points, let us look at the Neptune case mentioned by Lipton (also cited as an instance of IBE by Douven 2011: §1.2). During 1845-6, the French astronomer Urbain Jean Joseph Le Verrier developed a theoretical model of the orbit of Uranus which led directly to the discovery of Neptune.[79] Since Uranus' discovery in 1781, astronomers had struggled to construct a Newtonian theory of the planet which was capable of predicting its future movements. For example, in the foreword to his 1821 tables for Uranus, Alexis Bouvard reported that he had been unable to construct a Newtonian theory based on the known bodies of the Solar System which satisfied both the "ancient" observations (i.e. earlier observational records which had misidentified the planet as a fixed star) and the modern (post-discovery) observations. He decided to only use the modern observations, but remarked that "I leave to time to take care of revealing if the difficulty … really stems from the inaccuracy of the ancient observations, or if it depends on some strange and unperceived action, which could have acted on the planet." (quoted from Le Verrier 1846a: 908). Unfortunately, within a decade Uranus was once again diverging from its predicted path.

A number of explanations for this anomaly were entertained, including (i) that Uranus was being perturbed by an unknown body, such as a moon or one or more unknown planets, (ii) that Newton's law of gravitation might fail to strictly hold at large distances from the Sun and (iii) that there may be some kind of retarding medium in the outer parts of the solar system. While the recent discovery of Uranus, as well as several

[79] Le Verrier published three papers (1845, 1846a, 1846b) prior to the discovery of Neptune and his full calculations shortly after (1846c). Translations of the French in the following are mine. For the broader context and narrative, I rely on the historical accounts of Grosser (1962), Smith (1989) and Baum and Sheehan (1997/2013) and the contemporary accounts of Airy (1846) and Gould (1850).

minor planets between Mars and Jupiter, made many astronomers suspect that the perturbations stemmed from a planet beyond Uranus, most regarded the calculations necessary to predict the orbit of a planet as too laborious and uncertain to be worth the effort (Grosser 1962: 48-50; cf. Gould 1850: 9-16). However, Le Verrier set out to accomplish just this in his three papers.

His first paper (1845) reanalysed the observational data and recalculated the known perturbations on Uranus, in order to rule out that the anomaly stemmed from errors in earlier calculations. Next, in in June 1846, Le Verrier (1846a) used his new calculations to argue against the alternatives to the hypothesis of a trans-Uranian planet. For instance, he argued that the anomaly could not stem from an unknown moon orbiting Uranus since (a) this would produce anomalies of shorter period than was observed and (b) a moon large enough to produce anomalies of the observed magnitude would most likely have been discovered already. Similarly, he argued that the disturbing body could not be a planet situated between Saturn and Jupiter, since a planet large enough to influence Uranus at this location would also produce anomalies in the movement of Saturn, of which none had been observed. Most other competing hypotheses were similarly ruled out either because of their inability to explain the observed anomalies in Uranus' orbit or because they would imply other effects that could not be observed. The most significant exception (for present purposes)[80] is that Le Verrier only discusses mono-causal explanations: he does not even raise the possibility that multiple unknown bodies (say, two planets) could produce the anomaly while cancelling out their other effects. I will return to this point below.

[80] The other exception is the proposal to revise Newton's Law. Le Verrier rejects this option on the grounds that, in the past, it had always been possible to overcome apparent anomalies without modifying Newton's Law.

Finally, in September the same year, he derived a prediction of the position of this planet in the night sky (1846b). After lobbying different astronomers to test this prediction, Le Verrier finally managed to convince Johann Galle at the Berlin Observatory to look for the hypothetical planet. On 24 September, the first night Galle and his assistant Heinrich d'Arrest examined the designated portion of the sky, they spotted a light, too large to be a star and not recorded on their star map, within 1° of Le Verrier's prediction. Subsequent observations the next day showed that the light was moving. On 25 September, Galle wrote a letter to Le Verrier exclaiming "The planet, the position of which you indicated, really exists" (Galle 1846).

At what point is the explanatory inference supposed to have taken place? One candidate is in June, when Le Verrier argued for the existence of a trans-Uranian planet and first attempted to determine its orbit. Notice, however, that Le Verrier's official argument is purely eliminative: he rules out competing hypotheses on the grounds that, if they were true, they would either fail to explain the anomaly or entail further predictions which he took to be empirically false. While Le Verrier is clearly trying to identify hypotheses which (a) could potentially explain the anomaly and (b) are empirically plausible, he does not use the quality of the proposed explanations as a reason for or against their plausibility. But the latter is what explanationists need to support the truth-guidance claim.

One may plausibly argue that Le Verrier implicitly relied on something like the simplicity, parsimony or elegance of the mono-causal explanations when he ignored other alternatives (Jansson and Tallant *forthcoming*: 6). Since Le Verrier declares that his argument puts "the existence of a still unknown planet … out of doubt" (1846a: 918) and emphasises his "conviction that the theory which I have just presented is an expression of the truth" (1846b: 438), we may take him to rely on an implicit IBE. However, the fact that Le Verrier nowhere mentions the possibility of multi-causal hypotheses, and never cites the simplicity or elegance of the single-planet hypothesis as a reason in its favour, suggests that he did not expect these considerations to be particularly convincing to his audience. Furthermore, if Le Verrier can be interpreted to offer an IBE, this argument evidently failed to convince his peers; as Grosser points out, French astronomers considered Le Verrier's results an "analytical "triumph", but no French observer made the slightest move to look for the hypothetical planet" (1962: 102). It was only in September 1846, once Le Verrier had produced a precise, testable prediction that he even managed to convince Galle to *search* for the planet.[81] As his subsequent letter shows, Galle did not regard it as a foregone conclusion that the planet *really* existed. If explanatory considerations played any role before Galle's observation of the planet, it was to indicate that the hypothesis was worth investigating.[82]

Explanationists may, instead, be tempted to argue that the IBE took place *after* Galle had observed the moving light, when astronomers concluded this to be a planet. Surely there are many other possible explanations of the observed light consistent with this evidence, and IBE gives an account of how the scientists ruled these out. But this reply risks mirroring the "politician's syllogism": something must be done; this is something; therefore, it must be done.[83] Similarly: scientists must be using something to rule out alternative explanations; IBE is something; therefore, scientists must be using IBE to rule out alternative explanations. One cannot simply assume that scientists used IBE to rule out other possible explanations. For this case to provide support for empirical

[81] Le Verrier first tried, unsuccessfully, to convince George Airy, the Astronomer Royal in Greenwich, to attempt search for the planet (Grosser 1962: 102-3). The reason why Le Verrier needed to work hard to convince anyone to search for the planet is that most observatories of the time were busy producing observations for almanacs used in naval navigation. Thus, scarcity of free time disinclined most astronomers from testing what they saw as an uncertain theoretical prediction.

[82] Salmon (2001: 86) also points this out.

[83] This argument stems from an episode of the British satirical political sitcom *Yes Minister*.

premises such as (P1*), explanationists need to provide evidence that explanatoriness actually played an important role in this case.[84] In fact, it is at least as plausible that the astronomers simply relied on their background knowledge about the solar system to conclude that the light was a planet. Once Galle had pinpointed the planet, the simplicity of the single-planet hypothesis became irrelevant. To everyone at the time, given their knowledge and assumptions about the solar system, the movement and size of the light were decisive reasons to regard it as a planet. The relevant factors in the acceptance of the Neptune hypothesis were the observational evidence and the background assumptions shared by astronomers at the time, not the explanatoriness of the hypothesis.

Similar points have also been highlighted in recent discussions of Lipton's (2004) discussion of Semmelweis' investigations of childbed fever.

First, Paavola (2006) points out that the main role of explanatory reasoning in Semmelweis' initial investigations was to *generate* possible explanations which were subsequently ruled out, either by empirical reasons or theoretical arguments. Here, Semmelweis' use of explanatory reasoning supports the Peircean view, rather than explanationism.

Second, after Semmelweis succeeded in reducing the mortality rate by requiring doctors to wash their hands in chlorinated lime, the inference that childbed fever can be caused by cadaveric matter on the hands of medical students who had performed autopsies before delivering babies can be accounted for purely in terms of Mill's Methods (Scholl 2015).

[84] Some explanationists argue on non-empirical grounds that all ampliative inferences are explanatory (e.g. White 2005). If this is right, explanatoriness must of course have played some role in the inference to the existence of Neptune. But this line of argument would still not show that the Le Verrier case provides any *empirical* support for explanationism.

Third, at the one point where Semmelweis in fact employs something like an IBE, it was rejected by his contemporaries and, as Tulodziecki (2013) argues, for good reason. While many of Semmelweis' contemporaries were quite willing to accept that decomposing matter could be *a* cause of the disease, Semmelweis also argued for a stronger claim, namely that decomposing matter is the *only* cause of childbed fever. Here, one might argue that the unification and simplicity offered by this theory was the main reason for Semmelweis to accept this mono-causal theory. However, it was this claim that his contemporaries rejected. Tulodziecki concludes:

> Semmelweis simply did not provide any convincing reason to subscribe to the monocausality thesis. When Semmelweis was, reasonably, asked to perform certain experiments that could have supported his thesis, he declined, and, in addition, it was pointed out that the monocausality thesis failed to explain several salient phenomena associated with childbed fever that could be explained on a multicausal view (2013: 1074-5).

Although Semmelweis' monocausal view might be considered more simple and elegant, it was quite reasonable for his critics to reject it. If Tulodziecki's account is right, explanationists would not be wise to cite this aspect of Semmelweis' reasoning in support of their view.

Even when scientists do highlight the explanatory virtues of their preferred theory, explanationists cannot assume without further ado that they were therefore relying on IBE. For instance, Thagard (1988: 77) quotes Fresnel praising the number of phenomena explained by the wave theory of light in a letter to Arago, remarking:

> all these phenomena, which require so many particular hypotheses in Newton's system, are reunited and explained by the theory of vibrations and influences of rays on each other. (Fresnel 1866, vol. 1: 36, Thagard's translation).

This might sound as if Fresnel is appealing to the unification and explanatory power of the wave theory over the particle theory. However, Achinstein (1992: 359ff) argues that, rather than being an instance of IBE, the argument for the wave theory should be reconstructed as a (probabilistic) disjunctive syllogism along the following lines. At the time, the only two known ways of communicating finite motion (as was observed in light) were through the motion of a body or through wave disturbances in a medium. However, in order to produce the correct predictions of the phenomena cited by Fresnel, the particle theory had to rely on assumptions which were highly improbable given the available evidence. For instance, to explain diffraction, particle theorists postulated the existence of repulsive and attractive forces acting at a distance on the particles. Since all such known forces would depend on the size of the refractor, diffraction patterns would be expected to vary with the size of the refractor. But this effect could not be observed. So, since the only plausible competing hypothesis was highly unlikely given the available evidence and background knowledge, nineteenth century scientists concluded that the wave theory was most likely true. The fact that defenders of the wave theory could make similar arguments for a whole range of phenomena only strengthened their case. But this does not necessarily mean that they regarded the unificatory power of the wave theory as a reason in its favour *in addition* to these arguments.

3.6. Direct Arguments

The basic idea of direct arguments is that successful past applications of IBE provide direct evidence for the hypothesis that explanatoriness is a reliable guide to truth. A general objection to this strategy is that it risks circularity, since our judgements about which hypotheses are true, on the explanationist's own view, depend on explanatory considerations (Thagard 1988: 139-141; Lycan 2002: 421). However, it may be possible to meet this objection through an argument analogous to Kitcher's (2001b) "Galilean Strategy". Kitcher reconstructs Galileo's argument for the reliability of the telescope as follows: since we can independently check the reliability of the telescope for far-away things on Earth (e.g. by moving up close), and we have no good reason to think that pointing a telescope towards celestial bodies makes a difference to its reliability, we are justified in regarding it as reliable for the celestial domain as well. So, analogously, if there are at least *some* cases where we can verify the success of explanatory inferences by independent means, explanationists might construct an analogous argument that IBE is also reliable in cases we cannot check independently.[85] Alternatively, Douven (2002, 2005) has proposed that direct arguments can be reconstructed as a confirmation-theoretic "bootstrap" argument. He develops an account of how two successful, but co-dependent inferences—e.g. that the results of microscopy show that IBE is reliable, and that an IBE shows that microscopy is reliable—can provide unconditional confirmation of the reliability of both methods.[86] So, even if our methods for testing IBE partly depend on IBE, this need not pose an insurmountable obstacle to direct arguments.

Exactly how to formulate direct arguments is, however, not essential here. Rather, the points I want to highlight concerns whether the empirical evidence actually favours explanationism. If successful applications of IBE provide evidence in favour of the truth-

[85] Douven (2011: §3.2) cites Kitcher (2001b) as containing suggestions along the lines of a direct argument (although, as Magnus (2003: 472) points out, Kitcher does not himself apply the Galilean strategy to IBE).

[86] Douven does not endorse the direct argument but merely aims to show that it would not beg the question against scientific antirealists. He emphasises that supporting explanationism requires building an empirical case that IBE is better at producing correct predictions than mere guesswork (2005: 264) but remains neutral on whether this is the case.

guidance claim, then unsuccessful cases equally provide evidence against the truth-guidance claim. So, in order to ascertain whether direct arguments favour explanationism, we have to look at both how often explanatory considerations lead to hypotheses that are closer to the truth and how often they lead us away from the truth. Now, we cannot simply say that IBE is reliable if it leads to the truth more often than not. What proportion of positive and negative cases would support or undermine explanationism depends on what we take the underlying base-rates to be. For instance, if in a series of decisions we expect 95% of all theories to be false which are consistent with the evidence and which scientists have considered, then we might still regard IBE as truth-conducive if it allows scientists to choose a correct theory in 20% of the cases, since this still beats the 5% chance of choosing a true theory by mere guessing. However, it is unclear whether we can give any meaningful assessment of the base-rates of false theories in the set of theories scientists have considered in a given case (e.g. Magnus and Callender 2004).

To avoid this problem, I will not try to evaluate whether the evidence on balance favours explanationism or not. Instead, I will merely attempt to show that the empirical evidence underwriting direct arguments is less favourable than explanationists tend to assume. Specifically, I want to raise two problems for direct arguments.

First, direct arguments face the same problem as the indirect ones, namely that in many of the cases cited as an instance of IBE, it is unclear whether explanatoriness was used to support the acceptance of a theory. If this is the case, the evidence-base for direct arguments is slimmer and thus less conclusive than usually supposed. Proponents of the direct argument might argue that even if scientists did not actually rely on IBE, we can still consider whether doing so *would* have guided them towards the truth. However, as with Psillos' argument, we would then need to consider all cases where scientists *could* have applied an IBE, including those where they merely choose to pursue the best explanation. Including these cases is however likely to include many more cases where the most explanatory hypothesis turned out to be false.

Second, as I will now argue, even if we look at the cases where scientists accepted explanations and which realists have argued can reasonably be construed as partially true, the parts of the theories which underwrite their explanatory power are often not amongst those parts which are plausible candidates for the truth. Thus, even though scientific realists may be able to avoid the PMI, these responses do not provide evidence for the reliability of explanatory reasoning.

3.6.1. How Often Does Explanatoriness Lead to the Truth?

In many of the examples cited by explanationists, it is not clear that explanatory considerations guided scientists closer to the truth. For instance, the central explanatory posit of the wave theory (cited, as we saw, by Thagard as an example of IBE) was the ether. Before that, Newton's corpuscle theory was widely accepted as the best explanation of light. But according to our current quantum-mechanical understanding, both of these explanations are fundamentally mistaken. Similar points also apply to the Semmelweis case. As Scholl (2015: 101-2) points out, although the cadaveric matter did play *some* causal role, Semmelweis' full explanatory theory—that morbid matter is *the* cause of childbed fever—got many things wrong. For one thing, the disease is caused by bacteria rather than the morbid matter itself.

Another interesting case is Kepler's Copernican explanation of the regularities in the planets movements (Glymour 1984: 74 mentions Kepler among "realists" who accepted IBE). In a detailed case study, Lyons (2006: 545) shows that Kepler relied on a number of posits in constructing his theory which, according to our current physics, are completely wrong. These included the postulate that planets only move when forced to move, and that the sun emits rays (called the *anima motrix*) which *push* the planets *around* in their orbits (rather than *pulling* them towards the sun). Since he thought this pushing force was stronger nearer to the sun, the posits gave Kepler a neat, unified explanation for why the planets move faster at their perihelion and slower at their aphelion.

More generally, most currently accepted theories were preceded by a number of incompatible theories which relied on very different explanatory posits.[87] If explanationists are right that IBE played (or could have played) a central role in establishing the theories we currently regard as true, the direct argument would also need to take into account these past, apparently less successful applications of IBE. Even on the realist assumption that the currently accepted theories are essentially correct, the unsuccessful (supposed) applications of IBE most likely outnumber the successful ones. If we include cases where the explanatoriness of a theory motivated scientists to pursue it without accepting it, this would only make the problem more acute.

Explanationsts might argue that, even if explanatory considerations often lead us astray when it comes to general explanatory frameworks, they can still guide us towards essentially correct hypotheses which are retained even after the overall framework is rejected. Take the prediction and discovery of Neptune. Even though the Newtonian theory of gravitation has been rejected in favour of the general theory of relativity, the existence of a planet beyond the orbit of Uranus remains undisputed. But it is not clear that explanatory reasoning fares better with regards to this kind of predictions. For one thing, as Lyons (2006: 551-3) points out, although Le Verrier predicted Neptune’s existence, the same calculations also made many incorrect predictions. For instance, he

[87] E.g. the Aristotelian, atomistic, Cartesian and Newtonian theories of magnetism (Duhem 1954: 10ff), or the caloric and vibratory theories of heat (Laudan 1981: 26f, 33).

overestimated the eccentricity of Neptune's orbit by a factor of 12.5 and its mass to be more than double the currently accepted value.

Even more significantly, there is a cautionary tale for explanationists in the vicinity of the case of Neptune's discovery. After his success with Neptune, Le Verrier turned his attention to anomalies in the orbit of Mercury. By 1859 he had reached the conclusion that these could, in a similar manner to the anomalies in Uranus' orbit, be explained by the existence of one or more masses between Mercury and the sun. After interviewing the amateur astronomer Edmond Modeste Lescarbault, who claimed to have observed such a planet, Le Verrier became convinced of its existence and named it Vulcan. However, attempts by Le Verrier and others during the following decades to predict its orbit consistently failed, and supposed observations of Vulcan were highly contentious. Other explanations of the anomaly were proposed by astronomers, including the existence of a ring of smaller planetoids, a body of diffused matter around the sun or changes to Newton's law of gravitation. The Canadian-American astronomer Simon Newcomb analysed the problem in 1895, favouring the hypothesis that Newton's law is not a strict r^2 law, but should have a small constant added to the exponent (Fontenrose 1973:154-5). Fourteen years later, the American William Campbell proclaimed "The Closing of a Famous Astronomical Problem" (1909), arguing instead in favour of the diffused matter hypothesis on the grounds that it explained all of the anomalies found by Newcomb in the orbits of the inner planets. This hypothesis was, in turn, eventually rejected in favour of the explanation offered by the general theory of relativity (though some astronomers continued to defend the diffused matter hypothesis against Einstein's explanation for a number of years). Again, even if the acceptance of general relativity was a successful application of IBE, one would also have to count the many previous theories as failed explanatory inferences.[88]

Since this problem is similar to the one PMI poses for scientific realism, it might be thought that responses similar to those usually made in defence of realism can also rescue explanationism. The most promising realist responses to the challenge from the history of science are: (i) narrowing down the kinds of success that warrant realist commitments, usually stressing the ability to make successful novel predictions, and (ii) restricting the realist commitments to specific *parts* of successful theories, usually its "working posits", i.e. to those assumptions that play a substantive role in producing the empirical successes of the theory (Psillos 1999).

Regardless of whether these manoeuvres are sufficient to defend scientific realism from PMI (for discussion see Lyons 2006; Vickers 2013), they cause serious problems for direct arguments for explanationism. First, restricting our attention to cases where theories have made successful novel predictions effectively introduces a confounding variable into the direct argument, by raising the possibility that the explanatory qualities of the hypotheses are epistemically irrelevant. If we can only be confident that IBE works when the inferred theory leads to novel predictions, why think that explanatoriness is doing any of the epistemic work? Explanationists might claim that explanatoriness affords increased reliability *in addition* to that provided by the ability to produce novel predictions. However, to my knowledge, no work has been done to show this.

Second, restricting our realist commitments to the working posits of theories exacerbates these problems. For it is usually the posits which contribute to the

[88] Salmon (2001: 86) also mentions the failure of the Vulcan hypothesis as an example of a failed IBE. See Fontenrose (1973), Roseveare (1982) and Baum & Sheehan (1997/2013) for historical accounts of the fraught search for Vulcan and alternative explanations of Mercury's orbit during the late 19th century.

explanatoriness of a theory—by increasing its overall simplicity, unification or intelligibility—which are deemed "idle wheels". For instance, Psillos (1999:115-130) argues that the existence of caloric as a material substance was an idle posit for the predictive successes of caloric theories. But, as Chang (2003) points out, one of the most striking successes of caloric theory was that it could compellingly explain a wide range of thermodynamic phenomena. These explanations relied crucially on the hypothesis that caloric exists as a real material substance. Similarly, the explanatory power of wave theories of light or electromagnetism relies on the actual existence of the ether and, as Saatsi (2012) argues, this explanatory power is lost in most selective realist accounts of what wave theories got right. Again, the best realist response to these cases is probably to focus on posits necessary for making novel predictions, rather than those which are crucial to their explanatoriness. But in doing so the realist would move away from the explanationist claim that *explanatory* qualities are a reliable guide to the truth of hypotheses.

3.7. Implications for the Realism Debate

As emphasised throughout the two preceding sections, the problems I have raised for explanationism are independent of the realism debate. The most pressing problem, in my view, is not whether empirical arguments for explanationism beg the question against antirealists (interesting as that question might be), but whether an empirical case for explanationism can get off the ground at all.

That said, one of the most popular arguments for scientific realism, NMA, is usually construed as an IBE. As highlighted in Section 3.6.1, this argument stands in tension with the usual realist responses to PMI. In my view, unless some other argument for explanationism can be defended, realists should respond to this tension by giving up their reliance on IBE. The key premise of the no-miracles argument is that there are certain kinds of empirical achievements—e.g. the ability to make novel successful explanations or to sustain robust practical applications—which would be very unlikely unless some form of realism were not true. Realists usually try to shore up this premise by arguing that realism provides a better *explanation* of this success than anti-realism, while anti-realists have proposed competing explanations. It is this supplementary argument which would have to be relinquished. If we grant that realism provides a more satisfying explanation of the success of science, on the Peircean view this would merely be a reason for *pursuing* the hypothesis of realism—i.e. to continue the philosophical debates concerning realism. However, I do not claim to have presented an argument *against* realism. It may well be that the premise that success would be unlikely unless realism is true can be defended on non-explanatory grounds.

3.8. Conclusion

In the first part of this chapter, in particular Section 3.4, I have argued in favour of the Peircean view on two grounds. First, on the decision-theoretic account of pursuit worthiness defended in Chapter 2, the explanatoriness of a hypothesis generally provides reasons for pursuing it. Second, this account avoids the truth-connection problem and the subjectivity problem. As argued in Section 3.3, these are *prima facie* problems facing explanationism. Defending explanationism as a normatively adequate account of explanatory reasoning requires some argument in support of the truth-guidance claim. In the second part of the chapter, I raised several problems for the empirical arguments often adduced in favour of explanationism. If my criticism of these arguments is correct (and unless some other argument can be given), the Peircean view provides the most plausible, normative account of the role of explanatory reasoning in science.

Chapter 4. Pursuit Worthiness Accounts of Analogies in Science

4.1. Introduction

For much of the twentieth century the main focus in philosophical debates about analogies in science was whether these play any normatively interesting role in scientific reasoning. In defending the relevance of analogies, Norman Campbell (1920: ch. 6) and Mary Hesse (1953, 1966) were responding to Pierre Duhem (1954: ch. 4) and his intellectual heirs among the logical empiricists, in particular Hans Reichenbach. Although the latter critics of analogy usually admitted (grudgingly) that analogies sometimes guide the development of scientific theories, they regarded this as a mere psychological curiosity, not something that plays any interesting, normative role in scientific reasoning (e.g. Reichenbach 1944: 66-72). Arguing that analogies can serve important purposes that philosophers of science ought to account for, Campbell and Hesse (and to a lesser extent N. R. Hanson) opposed these at-the-time widely accepted views.

Today, the centre of gravity in the debate has shifted. Most philosophers interested in the issue now agree with Campbell and Hesse that analogies play an important, and philosophically interesting, role in science. Several different roles played by analogies have been highlighted (Bartha 2013: §1) and, correspondingly, have led to the development of a number of different kinds of philosophical accounts of analogies. Proponents of *justificatory accounts* take analogies to provide some degree of epistemic support, i.e. some reason to accept hypotheses, and try to explain how and when analogical arguments can provide this kind of support. Others challenge Reichenbach’s claim that generative reasoning is beyond the scope of normative theorising. For instance, Nancy Nersessian (1988; 2008: ch. 5), drawing on cognitive psychology and computational modelling, has argued that analogies can function as heuristics for developing or articulating scientific theories in ways that are both “systematic and subject to evaluation” (1988: 42). Call these *generative accounts* of analogical reasoning.

My focus in this chapter will be on a third type of account, which can be called *pursuit worthiness accounts*.[89] This chapter has two main aims. First, while analogies can also serve justificatory and generative purposes, I argue that pursuit worthiness accounts are necessary for explaining some uses of analogies in scientific reasoning, which are not captured by purely justificatory or generative accounts. Second, I want to investigate different accounts of how analogies can justify pursuit.

I start, in Section 4.2, by outlining a case study involving the early development of the liquid drop model of the atomic nucleus. In this case, I argue, physicists chose to pursue the liquid drop model despite it initially facing empirical and theoretical problems. In the remainder of the chapter, I then evaluate a number of different pursuit worthiness accounts in terms of how well they account for this case. In Section 4.3, I criticise accounts defended by Wesley Salmon (1967) and Paul Bartha (2010), according to which analogies provide reasons for pursuing a hypothesis in virtue of increasing their plausibility (understood as a weak form of epistemic support), thus subsuming pursuit worthiness uses of analogies within a justificatory account.

Instead, I propose that analogies are better seen as justifying pursuit in virtue of increasing the expected epistemic utility of pursuing a hypothesis. In Section 4.4, I consider an account proposed by Campbell where hypotheses based on analogies have a high potential for unification and are, thereby, more epistemically interesting. While this account is plausible for some cases, I argue that it does not fit the liquid drop model case. Finally, in Section 4.5, I propose an alternative account of this case according to which

[89] I borrow the terminology of ‘generative’, ‘justificatory’ and ‘pursuit worthiness’ accounts from McKaughan (2008).

analogies facilitate the transfer of an already well-understood modelling framework to a new domain of phenomena.

4.2. Case Study: The Development of the Liquid Drop Model

The liquid drop model of the atomic nucleus was developed from the late 1920s onwards, during a time when physicists were trying to extend their understanding of atoms to the structure of the atomic nucleus itself.[90] The model was first proposed in 1928-29 by George Gamow, then a Russian doctoral student visiting Western Europe, who suggested that the nucleus "may be treated somewhat as a small drop of water in which the particles are held together by surface tension" (Gamow, in Rutherford *et al* 1929: 386). In line with common assumptions at the time, he modelled the nucleus as consisting of a collection of α-particles and assumed that the nucleus is in equilibrium between the kinetic energy of the particles and the surface tension. On this basis, Gamow then tried to derive an expression for the mass defects (i.e. the nuclear binding energy) of the different nuclei.

Niels Bohr and Ernest Rutherford were enthusiastic about the model and worked to secure additional support for Gamow to continue working in Western Europe between 1929 and 1931. However, while Gamow made some progress with the model, he quickly ran into problems. Although his theoretically predicted mass defects traced a curve of the same general shape as the experimentally determined ones, it gave reasonably accurate quantitative predictions only for the lighter elements. He suspected this could be remedied by instead assuming that the nucleus also contains free electrons in addition to α-particles, as some physicists at the time suspected. However, when he tried to incorporate these into his model he ran into a major theoretical problem (the so-called Klein paradox) that he

[90] This section is primarily based on Stuewer's (1994) historical account of the development of the liquid drop model.

was unable to overcome. Consequently, by the summer of 1930, Gamow began to turn his attention elsewhere (Stuewer 1994: 78-85).

Despite these problems, the model quickly became popular among physicists. This was not because they were confident it accurately represented the nucleus. Rather, they saw it as a speculative but nonetheless promising approach which might help them answer some of the questions about the atomic nucleus which physicists were grappling with. For instance, in 1930 Rutherford wrote that Gamow's model "while admittedly imperfect and speculative in character is of much interest as the first attempt to give an interpretation of the mass-defect curve of the elements" (Rutherford, Chadwick and Ellis 1930: 534; quoted from Steuwer 1994: 86-7). The model was further developed during the 1930s, along two broad trajectories. First, following the discovery of neutrons in 1932, Werner Heisenberg, and subsequently Carl von Weizsäcker, tried to revise the model on the assumption that the nucleus contains a combination of protons and neutrons. Their aim was essentially the same as Gamow, namely to derive an empirically more accurate mass defect curve. Their efforts resulted around 1935-6 in what is today known as the Semi-Empirical Mass Formula (Stuewer 1994: 87-97).[91] Second, from circa 1936 onwards, Bohr and several of his collaborators attempted to adapt the model in order to account for artificially induced radioactivity, i.e. radioactive elements produced by bombarding stable elements with neutrons. Their explanation of this phenomenon was that the impinging neutrons resulted in an excitation of the nucleus and that the resulting vibrations caused the 'evaporation' of particles from the drop of 'nuclear fluid' (97-107).[92] Finally, in 1938-39, Lise Meitner and Otto Frisch realised that, by combining elements of both research

[91] The Semi-Empirical Mass Formula is so called because it is not derived from purely theoretical principles. Rather, it was constructed by calibrating certain empirical parameters in the revised liquid drop model to best fit the empirically determined mass-defect curve.

[92] A number of alternative (but related) analogies also influenced this line of physical theorising about atomic nuclei (Stuewer, *ibid.*).

programmes, the liquid drop model could be adapted to explain nuclear fission, a newly discovered and, at the time, highly puzzling phenomenon (107-116).[93]

From the latter part of this story, it is clear that the drop analogy not only inspired Gamow's original model, but played an important role in guiding the revisions and extensions of this model in subsequent work. There are two questions we might ask about this. The first concerns why the drop analogy suggested some revisions rather than others, that is, why these revisions seemed more natural to those who chose to work with the liquid drop model. This use of analogy is what generative accounts aim to analyse. I will say more about this in Section 4.5.

The second, which I will focus on for now, is why physicists chose to pursue the model in the first place, before there was any particular reason to think it even approximately true. We can distinguish a number of such decisions. First, after Gamow had the original idea, he chose to spend some of his limited time in Western Europe developing it into a formal model. Second, after he had presented the initial model in 1929, Bohr and Rutherford were sufficiently impressed to secure financial support to allow Gamow to continue working on the model. Third, after 1930, Rutherford continued to praise the model despite the empirical and theoretical problems it faced. Finally, during the 1930s, the model was pursued within several different research projects.

A more fine-grained analysis would be necessary to account for all the factors involved in the decisions to pursue the model in each of these cases. In this chapter, I will focus on just one question; namely, whether the drop analogy could have played any role in motivating the pursuit of the model. More specifically, I will discuss two extant pursuit worthiness accounts of analogy, arguing that these fail to plausibly account for the liquid

[93] See also Andersen (1997) on the experimental and theoretical developments which lead to the discovery of fission.

drop case, before proposing an alternative, more satisfactory account of how analogies can justify pursuit in cases like the liquid drop model in Section 4.5.

4.3. Did the Analogy Make the Model Plausible?

It might be thought that there is a straightforward answer to how the liquid drop analogy justified the pursuit of Gamow's model: although there might initially have been no grounds for *accepting* the model in 1930, the analogy helped to show that it was *plausible*. The plausibility of this analogy made it reasonable to pursue the model, or at least contributed to its pursuit worthiness.

As discussed in Chapters 1 and 2, many philosophers have assumed that providing reasons for pursuit simply amounts to showing it plausible (Hanson 1958, Salmon 1967, Kordig 1978, McLaughlin 1982). Here, reasons for regarding a model or hypothesis as plausible are seen as simply weak forms of epistemic support, not fundamentally different from reasons for its truth. Particularly influential is Salmon's (1967: 113-18) proposal that, within a Bayesian account of scientific reasoning, plausibility judgements can be understood as estimates of the prior probability of a hypothesis. Since it is necessary to make some judgement of prior probabilities to evaluate the posterior probability of a hypothesis, the Bayesian framework already requires scientists to make this type of judgement. According to Salmon, it is plausibility judgments in this sense which scientists rely on to decide "whether the hypothesis deserves to be seriously entertained and tested [i.e. pursued] or whether it should be cast aside without further ceremony" (113). One source of such plausibility judgments, according to Salmon, are analogies (127).

Whereas Salmon thus equates reasons for pursuit with estimates of prior probability, Paul Bartha's (2010) work on analogical reasoning gives a more nuanced account of their relation. Since it will be relevant to my later discussion, I will here outline some details of Bartha’s account. Following Hesse (1966: 59), Bartha endorses a *two-dimensional analysis* of analogical arguments. While many accounts only focus on *horizontal relations*, i.e. the similarities and differences between the source and target system, two-dimensional accounts also emphasise the *vertical relations*, consisting of dependency relations (e.g. causal, modal or explanatory relations) within the two domains. Building on this idea, Bartha (2010: ch. 4) defends an inference schema that may be summarised as follows:

> (BAR1) There is some structure of dependency relations *R*(a, b, c, …) between features a, b, c, … of the source system, S1. [*Prior association*].
>
> (BAR2) The target system, S2, has at least one of the features a’, b’, c’, … analogous to a, b, c, … [*Potential for generalisation*].
>
> (BAR3) S2 does not have any features which would preclude *R*’ (analogous to R) from obtaining. [*No critical difference*].
>
> *Therefore:*
>
> (BAR4) It is prima facie plausible that *R*’(a’, b’, c’, …) obtains for S2 and, *a fortiori*, that S2 has all of a’, b’, c’, ….

The first premise states that there is a “prior association” in S1, in the form of some structure of dependency relations between its features. Which kinds of dependency relations to look for varies between contexts, but a good example is how the parts of a mechanism interact and constrain each other to produce certain effects. Second, we look at whether there is a “potential for generalisation”, meaning that the target system has some features analogous to those involved in the prior association in S1. Finally, we consider whether there are any “critical differences” between the two systems, i.e.

whether S2 has any features precluding a relation analogous to the prior association from obtaining. Given these premises, according to Bartha, it is prima facie plausible to “transfer” the prior association to the target system, and thus infer that the relevant further features involved in the prior association obtain in S2 as well.

Bartha thinks that arguments of this type are often used to support hypotheses before they have been tested (2010: 6), and that they provide reasons for investigating hypotheses further (16). Like Salmon, he thinks this is because analogies support plausibility judgements. However, Bartha differs by not equating plausibility judgements with estimates of prior probability. That a hypothesis *p* is ‘prima facie plausible’, he instead takes to mean “roughly speaking, … There are sufficient grounds for taking *p* seriously” (16). This is partly an epistemic notion. A plausible hypothesis, according to Bartha, “has epistemic support: we have some reason to believe it, even prior to testing” (15), that is, it has “an appreciable likelihood of being true” (18). But he also takes plausibility judgements to have pragmatic connotations: “To say that a hypothesis is plausible typically implies that we have good reason to investigate it (subject to the feasibility and value of investigation)” (15). So, although epistemic support is important to what Bartha means by plausibility, considerations about ‘feasibility’ and ‘value’ are relevant as well. In a suggestive footnote (p. 18, note 19) Bartha also mentions that reasons for pursuit depend on epistemic support “in a decision-theoretic sense” given “contextual information about costs and benefits.” However, he adds that, absent such contextual information, “the two points are at least partially independent” (*ibid.*).

Given this elucidation of what he means by ‘prima facie plausibility’, it is consistent with Bartha’s account that analogical inferences can provide reasons for investigating a hypothesis without necessarily providing reasons for its truth. Similar to my argument regarding pursuit worthiness in Chapter 2, if feasibility and value are both relevant to the plausibility of a hypothesis (in Bartha's sense), one should be able to show a hypothesis plausible by arguing that it is more feasible or valuable to investigate it than previously thought. However, in practice Bartha tends to focus on epistemic support. For instance, he claims, "Any argument that a hypothesis is prima facie plausible … should provide reasons to think the hypothesis might be true" (18) and he follows Salmon in identifying a hypothesis's *degree* of plausibility with its prior probability (e.g. pp. 15-6, 291-302). As I read Bartha, analogies primarily provide reasons for pursuing hypotheses by providing them with additional epistemic support. Once this is established, whether we are then justified in pursuing a hypothesis all things considered depends on further 'contextual information', i.e. information in addition to that provided by the analogy, such as about the costs and benefits of pursuing it.

Focusing on the epistemic dimensions of plausibility, Bartha distinguishes two senses in which analogies can be used to show a hypothesis plausible, suggesting two different interpretations of how the analogy could have provided a reason for pursuing the liquid drop model. First, on the *modal* interpretation, that a hypothesis is *prima facie* plausible means that it has some minimal chance of being true, i.e. that it is regarded as a serious possibility which cannot be rejected out of hand. Since it is usually only worth pursuing theories that have a serious chance of being correct, or at least are broadly speaking on the right track, if an analogy can be used to show a hypothesis or model plausible which had previously been dismissed, this would contribute to its the pursuit worthiness.[94] Second, on the *comparative* interpretation, plausibility come in degrees. As noted, Bartha identifies these degrees with the prior probability of the hypothesis. Here, the idea is that the analogy *adds* to the plausibility of the model, beyond the minimal

[94] More accurately, as argued in Chapter 2 (Section 2.3.2), minimal pursuit worthiness requires that something new could potentially be learned which it would be worth learning given the costs of pursuit.

sense discussed above. It is because of this added plausibility that the theory or model stands out as a particularly promising in contrast to others which are not based on analogies. While I do not argue that analogies cannot support the pursuit worthiness of a hypothesis or model by showing it (epistemically) plausible in either of these two senses, I have some reservations regarding either interpretation as an account of the liquid drop model case.

Looking first at the modal interpretation, it is clear that Rutherford and others thought there was some chance that Gamow's model, or some suitably modified version of it, could provide a correct explanation of the phenomena they were interested in. Furthermore, the case does fit Bartha's schema. First, atomic nuclei and water drops share some features, e.g. they are both relatively stable collections of interacting constituent smaller particles (potential for generalisation). Second, in water drops this stability is due to the equilibrium between the surface tension, which results from the mutual attraction of its constituent particles, and the kinetic energy of the particles (prior association). Since there is no known reason why this account could not apply to the atomic nucleus (no critical difference), it is *prima facie* plausible that a surface tension can also be defined for atomic nuclei. However, this can at best account for Gamow's initial decision to start developing the model. After Gamow had developed a model in which it was possible to define a surface tension for the nucleus which was consistent with quantum mechanics and commonly accepted assumptions about the nucleus, the analogy no longer seems relevant. To the extent that Gamow's model at this stage was plausible, in the modal sense, it was because it seemed to be consistent with existing knowledge and perhaps because of its modest empirical success. Once this is taken into account, the fact that the central assumption of the model—that one can think of the nucleus as having a surface tension resulting from attractions between the constituent particles of the nucleus—was based on an analogy with macroscopic water drops does not seem to add anything important to whether it should be regarded as a serious possibility.

One might take the comparative interpretation to be more relevant here: perhaps the analogy still *added* to the plausibility of Gamow's model, even if it was no longer necessary to establish it as minimally plausible. However, it is less clear that the physicists in 1930 regarded the model as significantly *more* probable than so many other possible models, especially given the empirical and theoretical problems it faced at the time. Perhaps by 1938-9, when Meitner and Frisch were considering how to explain nuclear fission, the successful developments of the model by Heisenberg, von Weizsäcker, Bohr and others gave them good reason to regard the revised liquid drop model as a comparatively plausible representation of the nucleus. But here it is the empirical and theoretical successes of this research, rather than the initial water drop analogy, which showed the model more plausible. However, *before* the model had made these achievements, at the time when Rutherford praised Gamow's initial model and Heisenberg and Bohr subsequently decided to continue working on it, it is not clear that the model could claim a comparative advantage in terms of its plausibility.

Bartha might argue that, insofar as the water drop analogy played a role in these decisions, it must have been because it increased the plausibility of the model. However, this reply would not take the implications of Bartha's own conception of the relation between epistemic and practical aspects of plausibility judgements fully into account. Since being justified in pursuing a hypothesis or model depends on a number of factors apart from its epistemic support—e.g. the feasibility and value that Bartha mentions—why assume that the analogy increased the probability of the model, rather than one of these other factors? One cannot simply assume that when analogies motivate pursuing a hypothesis, the analogy must, therefore, have provided reasons for its truth.

Furthermore, as I have argued in Chapter 2, on a decision-theoretic construal of pursuit worthiness (of the kind that Bartha seems to endorse), it is not always the case that increasing the probability of a hypothesis is a reason in *favour* of pursuing it, let alone a sufficient reason. In the Simple Model of pursuit worthiness, whether increasing the probability of hypothesis being true also increases the expected epistemic utility of pursuing it depends on how the utilities and conditional probabilities are balanced. In the case of the liquid drop model, learning that the model is true would presumably be more interesting than learning that it is false—in terms of the Simple Model, that $\mathrm{u}(acc(h), h) > \mathrm{u}(rej(h), \neg h)$. After all, ruling out that this particular model is false would not be particularly interesting, while showing that it provides the correct explanation (say) of the mass-defect curve would achieve a major goal of research at the time. On the other hand, we might also suspect that it would be easier to discover a decisive flaw in the model than to conclusively show it correct, and so assume that $\mathrm{Pr}(acc(h) \mid h, p_h) > \mathrm{Pr}(rej(h) \mid \neg h, p_h)$. In this case (and ignoring the possibility of getting misleading evidence), increasing $\mathrm{Pr}(h)$ would make h more pursuit worthy if and only if $\mathrm{u}(acc(h), h) \times \mathrm{Pr}(acc(h) \mid h, p_h) > \mathrm{u}(rej(h), \neg h) \times \mathrm{Pr}(rej(h) \mid \neg h, p_h)$. The latter inequality is not automatically satisfied. Thus, even if Bartha is right that the analogy increased the probability of the liquid drop model, this does not automatically mean that the analogy increased the pursuit worthiness of the model.

To be clear, I do not intend to argue that analogies cannot sometimes be used, along the lines of Bartha's account, to support the plausibility of a hypothesis.[95] Similarly, the above considerations do not rule out that the water drop analogy could have made some contribution to the plausibility of the liquid drop model. However, in the remainder of

[95] For example, as I will argue in Chapter 5, Bartha's account is useful for understanding some of the ways analogies are used and debated in archaeology.

this chapter I will consider an alternative interpretation of the role of the analogy in motivating the pursuit of the model.

4.4. Analogies as a Guide to Unification

One proposal for an alternative pursuit worthiness account of analogies, similar to the account of explanatory reasoning defended in Chapter 3, is that analogies are a guide to hypotheses or models for which it would be epistemically valuable to know whether they are correct. In this section, I will examine a specific version of this idea, viz. that analogies indicate hypotheses that would provide increased theoretical unification, if shown true.

Campbell's defence of analogies in physics was partly based on this unificationist idea. Bartha sometimes suggests that Campbell can be interpreted as a forerunner of the plausibility account, citing e.g. Campbell's remarks that "in order that a theory may be valuable it must … display an analogy" (1920: 129) or that "*Some* analogy is essential to it [Fourier's theory of heat conduction]; for it is only this analogy which distinguishes the theory from the multitude of others… which might also be proposed to explain the same laws" (142). Bartha (2013: §2.3) equates a theory being "valuable" here with there being "grounds for taking the theory seriously" and thus he claims that "Campbell … thinks that analogies can establish this sort of *prima facie* plausibility" (*ibid*.) which Bartha is also interested in. As a historical point, this interpretation of Campbell is almost certainly false. In fact, Campbell (1920: 152) argued that theories based on mechanical analogies are more likely to be *false* than ones which merely posit generalised laws extrapolated from observed regularities. Instead, he took analogy-based theories to be valuable "simply because the ideas which they bring to mind are intrinsically valuable" (132). Part of the reason for this is that theories based on mechanical analogies offer the chance to discover laws capable of unifying quantities from previously distinct domains, e.g., heat and momentum in the case of the billiard ball model of gases. Campbell regarded it as intrinsically valuable to achieve this kind of unification, and argued that we "must balance that value against the chance of error" (152). While Campbell does not elaborate much further on these remarks, it is clear that when he calls theories based on analogies "valuable", it is not because he thinks they are more likely to be true.

The idea that the value of obtaining unifying theories has to be balanced against the risk of error fits my decision-theoretic approach to pursuit worthiness. If we agree with Campbell that it is intrinsically valuable to discover that a unifying theory is true, so that this increases our estimate of u(*acc*(*h*), *h*), this will, all things being equal, increase the expected utility of pursuing the theory. If this value is sufficiently high, it could outweigh a decreased prior probability, which would otherwise shift the weight towards the factors weighed by Pr(¬*h*) in the Simple Model. (But notice, again, that reducing Pr(*h*) does not necessarily decrease overall the expected utility of pursuit). More generally, Campbell's claim that analogy-based theories are pursuit worthy because they are more intrinsically valuable is similar to the subjectivist interpretation of the Peircean view of explanatory reasoning that I discussed in Chapter 3 (Section 3.4).

Campbell's unificationist account also fits one line of justification Bartha (2010: ch. 7) offers for his account, that it tends to promote the traditional theoretical virtues, in particular unification.[96] If we construe unification as the ability to explain a wide range of phenomena using the same basic explanatory pattern (Kitcher 1989), we can see how this fits Bartha's inference schema. Premise (BAR1) identifies the existence of the

[96] Bartha argues that analogies are also conducive to other theoretical virtues, including coherence, simplicity and fruitfulness, but he regards unification as the most central. Here, Bartha (2010: 256) recognises that, as long as we consider it valuable to achieve these virtues, this is sufficient to show a hypothesis 'plausible' in his sense of 'worthy of investigation'. However, he also suggests that his argument can be combined with the argument that the theoretical virtues are "indicators of empirical adequacy (or truth)" (*ibid*.).

explanatory pattern *R* (the prior association) in S1, while (BAR2) points out that there are a number of features in S2 that could potentially be explained by the same pattern. Since, (BAR3), there is no known reason to rule out this possibility, there is a potential for unifying the relevant features of S1 and S2 in single explanatory schema. So, if we were to discover that *R* holds for S2, we would have increased the unification of our knowledge of the world.

In my view, this unificationist idea provides a plausible account of how analogical reasoning justifies pursuit in some cases, but not in all. In cases such as the billiard ball analogy for gases or the ‘waves in a mechanical medium’ analogy for light (discussed e.g. by Hesse 1966; Nersessian 1988), they do promise to unify thermodynamical and optical phenomena (respectively), with the theoretical framework of classical mechanics. From the perspective of nineteenth-century physicists, these analogies pointed to potential increases in theoretical unification. However, this story does not work for other cases, like the liquid drop model. Although, in this instance, Bohr, Rutherford and other physicists took Gamow’s analogy to suggest a very promising line of research, this could not be because it promised to unify the physics of water drops and atomic nuclei. The liquid drop model employs modelling techniques analogous to those applied to water drops, but it was clear that the details of explanations within these two domains would be very different. Even if one might hope that an increased understanding of the atomic nucleus could eventually lead to a unified account of the two types of systems, the liquid drop model does not in itself promise to achieve this kind of unification in the same ways as the billiard ball model and mechanical ether models.

4.5. Transferring Modelling Frameworks Through Analogies

In order to account for how analogies justify pursuit in cases like the liquid drop model, I propose to look more carefully at the relationship between analogies and scientific models. So far, I have been talking as if a model of the kind Gamow developed is more or less equivalent to a hypothesis, as if the pertinent question is whether the analogy shows the model plausible or whether it would be valuable to learn that the model is true. However, in cases like the development of the liquid drop model, this way of thinking is somewhat misleading. For one thing, Gamow *knew*, or at least had good reasons to suspect, that his original model was incorrect: as mentioned, he had not included free nuclear electrons in the model even though he clearly suspected these would make a difference to the result. His strategy was of course to see if he could obtain some kind of promising results from the simpler model before attempting to develop the more complicated one. As Parker points out (2009), the fact that scientific models are often constructed using deliberate idealisations and simplifying assumptions makes it problematic to write as if it is the *model* which is tested. Rather, what scientists are interested in is typically some *hypothesis* about the fit between the model and the world (cf. Giere 2004). Similarly, when applying the decision-theoretic models of pursuit developed in Chapter 2, we need to be careful in specifying which hypothesis is pursued.

In the case of to the liquid drop model, we cannot say that Gamow and those who subsequently worked on the model pursued any specific hypothesis about the structure of the atomic nucleus. Rather, they tried to model the atomic nucleus as if it were a water drop in order to construct a potential explanation of some otherwise puzzling phenomenon—i.e. the mass defect curve for Gamow, Heisenberg and von Weizsäcker, artificial radioactivity for Bohr and his colleagues, and nuclear fission for Meitner and Frisch. They were of course still, ultimately, hoping to construct a model which provides an accurate (or at least empirically accurate) description of the nucleus. But their immediate priority was to formulate a potential explanation of the target phenomenon. Thus, what the physicists pursued in this case was the research project of *adapting a modelling framework* to the atomic nucleus for certain explanatory purposes. If we want to say that they pursued a hypothesis, it was not one of the form "the atomic nucleus has features a, b, c, … analogous to a water drop" but rather something like "modelling the atomic nucleus analogously to a water drop can lead us to formulate a (correct) explanation of phenomena x, y, z, …."

This point highlights the overlap between pursuit worthiness accounts and generative accounts, mentioned in Chapter 1. That analogies guide the gradual *development* of theories or hypotheses, rather than simply supporting a specific hypothesis, was also something which Campbell and in particular Hesse (1966: 4-5) highlighted as important to understanding the use of analogies. However, we need to separate two different questions here. On the one hand, many generative accounts focus on spelling out how a given analogy inspired or guided the development of new scientific concepts.[97] Here, the focus is on how the analogy helped scientists to formulate genuinely novel concepts which go beyond the conceptual resources of existing theoretical framework. But noticing that an analogy can be helpful for formulating new concepts does not in itself answer the question of why it was reasonable to pursue an analogy-based modelling framework in the first place.

To see how these two can come apart, consider the fact that Campbell explicitly denies that analogies are a help to develop theories:

[97] Examples include Nersessian (2002) on Maxwell's development of the concept of the electro-magnetic field, and Morgan (1997, 1999) on Irving Fisher's use a mechanical balance analogy to clarify and reinterpret the quantity theory of money.

> Analogy, so far from being a help to the establishment of theories, is the greatest hindrance. It is never difficult to find a theory which will explain the laws logically; what is difficult is to find one which will explain them logically and at the same time display the requisite analogy. … To regard analogy as an aid to the invention of theories is as absurd as to regard melody as an aid to the composition of sonatas (Campbell 1920: 130).

Now, *pace* Campbell, it might be that imposing constraints actually makes it easier to come up with genuinely novel ideas. However, the core point is that the relevant question is not how to most effectively come up with *novel* ideas, but rather how to come up with *ideas that are worth pursuing*. Sometimes, e.g. if we lack any possible explanations, coming up with genuinely novel ideas might be intrinsically desirable. But in other cases, e.g. if we are overwhelmed by too many hypotheses, we may instead prefer to *restrict* ourselves to generating hypotheses of high quality.

So why are modelling frameworks based on analogies more pursuit worthy in cases like the liquid drop model, than trying to develop potential explanations without relying on analogies? I want to propose that these frameworks are more pursuit worthy because they facilitate the transfer of a modelling framework in order to construct explanations in a new domain.[98] Now, my point here is not simply that the models constructed through this approach could potentially explain some of the phenomena scientists are interested in. This would not set analogy-based modelling frameworks apart from explanations constructed by other means. Furthermore, I do not here want to argue that explanations based on analogies are somehow more intrinsically interesting, as Campbell suggests,[99]

[98] This account is inspired by Hesse's and Bartha's idea that analogical inferences "transfer" explanations from one domain to another. Morgan (1999: 386-7) has also discussed when it is possible to "transfer" lessons learned within an analogical model to a real-world target system. By contrast, my account here focuses on transferring and adapting modelling *frameworks* to a new target system. In this respect, it is closer to Hesse's (1966: 157-177) suggestion that analogies are used in explanation to "metaphorically redescribe" the target domain in terms of the source analogy.

[99] I will take up this idea in Chapter 5 in the context of analogy-based interpretations in archaeology.

nor do I want to argue that analogy based frameworks are somehow more likely to produce *correct* explanations, since that would simply take us back to the idea that analogies show the hypothesis more probable.

A better reason is that transferring a modelling framework by analogy can often reduce the costs of pursuit, since trying to adapt an already existing modelling framework to a new domain is typically easier, and less time consuming, than developing a new one from scratch. Thus, in the case-studies analysed in terms of generative accounts, it is not so much the *novelty* of the explanations generated through analogies which made it reasonable to pursue this particular strategy, but the fact that they provided a *cost-effective* means of generating new potential explanations.

In my view, this simple cost-effectiveness account does go some way towards explaining why there are often good reasons to pursue analogy-based modelling frameworks. But I think we can say something more directly connected to the epistemic value of analogy-based explanations as well, namely that the benefit of transferring modelling frameworks through analogies can be that such modelling frameworks are themselves already well-understood.

To flesh out this idea, I will employ a distinction between *understanding-why* and *understanding-with*, drawn by Michael Strevens (2013: 513) based on recent discussions of scientific understanding. Understanding-why is the understanding of phenomenon or state of affairs in the world, e.g., the sense in which we can say whether someone understands combustion, heat conductivity or nuclear fission. It is typically achieved by grasping an explanation using some theory or model which represents the phenomenon of interest with sufficient accuracy. Understanding-with, by contrast, is the kind of understanding one can have of a theory, model or theoretical framework; the sense of ‘understanding’ employed when we say, e.g., that a historian of science understands the caloric theory of heat. Specifically, one has understanding-with to the extent that one is able to grasp and construct potential explanations based on the theory or model in question. To grasp an explanation here means to understand how the explanation works and why it would explain a given phenomenon if the theory or model accurately represented that phenomenon. As Strevens (*ibid.*) and others argue, understanding-with is a precondition for understanding-why, at least of the more interesting kind. For a scientist to understand a phenomenon through some explanation, it is not enough that the model or theory used provides an explananation of the phenomenon and that this explanation is factually correct. The scientist must also grasp how the explanation works in order to 'cash in' the potential understanding afforded by the model or theory.

This allows us to say more about why transferring a modelling framework to a new domain is a cost-effective way of constructing new explanations. It is not just that it will be quicker or easier to construct these new explanations (though that matters too) but, furthermore, that *if* this framework can be adapted to the new domain without too much modification, one will already have a large degree of understanding-with of this framework.[100] Thus, insofar as the scientists succeed in constructing new potential explanations of the phenomena of interest using this framework, little extra work is required to realise this explanatory potential. One might eventually achieve a similar understanding-with of a new, purpose-built modelling framework, but it would typically require extra effort to achieve the same levels of understanding-with.

Applying this account to the liquid drop model, it can, first, provide a rationale for why Gamow initially chose to pursue a modelling strategy based on the water drop analogy: if this modelling strategy were to succeed, it would provide a readily

[100] Plausibly, we may also say that *to the extent* the framework can be transferred without modification, this will allow scientists to preserve their understanding-with of the framework. However, this stronger claim is not strictly necessary for my argument here.

understandable model able to support easily graspable explanations of the nuclear phenomena he was interested in, in the first instance the mass defects. Second, when Gamow's initial work showed that this strategy was indeed feasible, though not initially particularly successful, this confirmed that the strategy was compatible with the theoretical framework of quantum mechanics. Thus, while it did not yet make it especially likely that the model was a *correct* or *accurate* representation of the nucleus, Gamow's model had nevertheless shown that the understanding-with provided by this modelling strategy did not require physicists to sacrifice any of their existing understanding. This further strengthened the pursuit worthiness of the model and provided a rationale for other physicists to pursue the model further during the 1930s.

Finally, this gambit (i.e. to use the liquid drop analogy as a cost-effective means to develop models with a high degree of understanding-with) was spectacularly vindicated in Frisch and Meitner's explanation of fission. As Stuewer (1994: 112-16) argues, it was because Meitner had worked in the Heisenberg/von Weiszäcker tradition in Berlin, and Frisch with Bohr in Copenhagen, that they were able to combine elements of both traditions to construct their explanation. In my terms, we can say that Frisch and Meitner were able to "pool" the understanding-with, developed separately in Berlin and Copenhagen, in order to develop the model.[101] Although Gamow, Heisenberg or Bohr could not have predicted this particular success, pursuing the modelling framework of the liquid drop model proved to be an effective means of generating a high level of understanding-with, thus enabling physicists such as Frisch and Meitner to quickly formulate potential explanations in response to surprising empirical discoveries.

[101] It is unclear whether Frisch and Meitner deliberately *chose* to pursue the liquid drop model in order to produce their explanation. In Frisch's recollection (Stuewer 1994: 114-15), it seems rather that Frisch and Meitner simply had the requisite ideas ready to mind and spontaneously brought them to bear in response to each other's suggestions.

4.6. Conclusion

In this chapter, I have considered several accounts of how analogies can provide reasons for pursuing a model (or modelling strategy) in cases like the liquid drop model. To be clear, I do not claim that the account of analogies developed here—i.e. that they provide a cost-effective means of transferring modelling frameworks with a high degree of understanding-with—is exhaustive of the use of analogies in science. First, pursuit worthiness accounts are compatible and to some extent complimentary with justificatory and generative accounts. The latter two types of accounts still capture interesting uses of analogy. However, I have argued that an adequate pursuit worthiness accounts of the liquid drop model case study cannot simply be subsumed within either of the other two. Second, as indicated, I do not regard my account as the only possible pursuit worthiness account of analogies. Nonetheless, I hope to have provided a plausible account of one way that analogies can be used in science.

Chapter 5: Three Uses of Analogy in Archaeological Theorising

5.1. Introduction

A central challenge in the epistemology of archaeology is to achieve a satisfactory resolution of what Alison Wylie (2002: 117) calls the “interpretive dilemma”. On the one hand, if archaeology is to produce interesting knowledge about the past, it needs to be able to draw conclusions that go beyond a mere “artefact physics” which only records and classifies the remains of the past (DeBoer and Lathrap 1979: 103). The intellectual value of archaeology rests, to a large degree, on its ability to draw substantial conclusions about life in past societies. On the other hand, archaeological theories should not be mere speculation without empirical grounding. But given the uncertainty of archaeological evidence, the fact that archaeological evidence is almost always partial, indeterminate and theory-laden in nature, can any interesting conclusions about the past pretend to be more than such speculations? Considered as a yes-no question, the answer to this is surely yes: as philosophers of science have argued, these sources of uncertainty are ubiquitous to all sciences, yet do not entail complete scepticism. The *challenge*, however, is to formulate a sophisticated assessment of archaeological theorising which respects the fact that archaeological evidence is in many ways insecure, and thus does not pretend to false certainty (Gero 2007), without retreating into the wholesale epistemological pessimism which leaves only a stark choice between ultra-conservative artefact physics and unrestricted speculation (Wylie 2002: 144).

One facet of this challenge concerns the use of analogies in archaeological reasoning. Whether analogies have any legitimate use in archaeology was the topic of a recurring debate during the twentieth century. On the one hand, sceptics point out that analogical inferences are notoriously uncertain (Smith 1955) and seem to carry with them unfounded, and often misleading, assumptions about the uniformity of human behaviour and culture across time and space (e.g. Freeman 1968, Gould 1980). On the other hand, the use of analogies seems practically unavoidable in archaeological interpretation. The material remains of the past do not speak for themselves, and since there are no well-established, general theories of human culture strong enough to ground direct inferences, most archaeological interpretations draw (whether explicitly acknowledged or not) on parallels and comparisons with other known cultures and societies. As Wylie (1982, 1988, 2002: ch. 9) has shown, attempts to formulate archaeological methodologies which eschew analogies simply end up re-introducing them by another name. While it is easy to find examples of mistaken and misleading analogies in the history of archaeology, Wylie argues, this only shows that it is the *uncritical* use of analogies which is problematic (*ibid.*). Thus the appropriate reaction, rather than banning analogies altogether, is to develop an improved methodological awareness of how they can be legitimately used, and how analogy-based interpretations can be criticised or strengthened—a conclusion now accepted by many archaeologists (e.g. Hodder 1982: ch. 1, Stahl 1993; Lightfoot 1995; Ravn 2011).

This chapter aims to contribute to the project of formulating a better methodological understanding of how analogies can be legitimately used in archaeology. Specifically, I want to argue that analogy can play several different roles in archaeological theorising, each of which forms a legitimate and, to some extent, necessary part of archaeology. What is important to recognise, however, is that these roles differ with regard to the criteria and the potential challenges for employing analogy adequately. Thus, I argue that in methodological discussions about the use of analogies, archaeologists should make clear *how* analogies are being used and take care not to conflate the adequacy criteria for different uses.

Specifically, I distinguish three ways analogies can be (and are) used in archaeological interpretation. First, they can be inferences which provide reasons for *accepting* an interpretation as likely to be accurate or correct. Second, seeking out analogies can be a method for *generating* new possible interpretations of archaeological evidence which can then be pursued. In pursuing an analogy-based interpretation, archaeologists investigate whether, or to what extent, there are substantive similarities between the practices of a past society and those of better known, apparently analogous societies. Third, analogies can provide *reasons for pursuing* specific interpretations. While all three uses of analogy play important roles in archaeological practice, and similar distinctions are sometimes drawn in methodological discussions, there has been a tendency to discuss the use of analogies as if it presents a single problem. As I shall argue, using analogies to generate interpretations and to motivate their pursuit each differ, in terms of their different adequacy criteria and the methodological challenges they face, from providing reasons for acceptance. However, most systematic analyses have focused on how to make analogical inferences more reliable. By distinguishing clearly between the different roles of analogies in archaeology and analysing them separately, I aim to add further nuance to the methodological debate over analogies in archaeology and clarify the epistemic status of analogy-based interpretations.

I start my discussion, in Section 5.2, by reviewing some of the key methodological debates concerned with the use of ethnographic analogies, arguing that the distinctions between the three uses of analogy are implicitly present in this literature. Next, in Section 5.3, I present a philosophical analysis of the three different roles for analogy in archaeology, and explain how the adequacy criteria and potential challenges for each of these uses differ. In Section 5.4, I then illustrate how this framework can illuminate the use of analogy in practice by analysing the use of analogies in the field of Roman archaeology, focusing on Penelope Allison's work on Pompeian household items (1999, 2001, 2009). I conclude, in Section 5.5, with some reflections on the implications of my account for the interpretative dilemma.

5.2. Reactions and Counter-Reactions to Analogy in Archaeology

In this section I will review some of the key methodological discussions of analogy in archaeology. Most of these have taken place in the context of debates about the relation between prehistoric archaeology and anthropology. For this reason, the debate has mainly focused on comparisons between prehistoric societies for which no textual evidence exists and contemporary or near-contemporary "primitive" societies known through ethnographic or anthropological studies. As will be illustrated in Section 5.4, the issues raised in these debates are also relevant for other branches of archaeology. For now, however, I shall follow the literature by focusing on ethnographic analogies in studies concerned with for prehistoric societies.

5.2.1. Background: Early Uses of Ethnographic Analogy

Ethnographic analogies have a long history in archaeology and examples of them being used, both in clearly successful and clearly problematic ways, abound.[102]

A classic, early example of analogies being used productively concerns the recognition, during the late sixteenth century, that Europe might have once been inhabited by "primitive" people, similar to the societies which Europeans were increasingly coming into contact with in the Americas and elsewhere (Orme 1981: 2-13). In particular, comparisons between the stone tools of North Americans and stone materials which had

[102] The following draws on previous comprehensive historical surveys (Orme 1974, 1981; Stiles 1977; Stahl 1993; Wylie 1985, 2002).

previously been thought to be of natural or supernatural origin (e.g. ‘thunderstones’ or ‘elf-shot’), led naturalists to claim the latter were in fact artefacts made by prehistoric people (Stiles 1977: 88; Trigger 1989: 47, 52-55). Gradually, archaeologists started to exploit these ethnographic analogies to draw more substantive conclusions about the archaeological material. For instance, during the 1830s and 1840s, the Swedish archaeologist Sven Nilsson carried out systematic comparisons of wear patterns on contemporary and prehistoric stone and bone tools in order to determine how the latter might have been mounted on now-perished wooden shafts. Many of these reconstructions were subsequently validated by the recovery of preserved tools from a waterlogged settlement in Switzerland in 1853 (Trigger 1989: 80-86).

Not all uses of ethnographic were this successful, however. Particularly infamous are interpretations based on assumptions of uni-linear cultural evolution, popular in the late nineteenth and early twentieth century. These accounts were based on the theory that cultural evolution proceeds through a definite sequence of evolutionary stages, and thus regarded contemporary hunter-gatherer societies as the literal descendants of prehistoric groups still arrested at a particular evolutionary step. These interpretations were widely criticised by later archaeologists (Ascher 1961; Wylie 2002). Not only did they face anomalies which they could only account for using implausible ad hoc explanations.[103] More generally, later archaeologists pointed out that these interpretations risked simply “assuming what one is trying to discover” (Clark 1951: 52), i.e. that prehistoric societies are similar to apparently analogous contemporary ones, rather than trying to investigate whether this is the case.

[103] Ascher (1961: 318) highlights an example from Solas (1911), namely that Australian aboriginals, who were supposed to be on the evolutionary stage characteristic of the Palaeolithic, use polished stone tools which according to Solas’ theory should only occur at the Neolithic stage. As an explanation for this, Solas ended up suggesting that they might have learned to use these tools via an earlier trade network which stretched from Europe to Australia.

5.2.2. The Analogy Debates in Twentieth-Century Archaeology

During the twentieth century, archaeologists conducted a series of debates about whether analogies can play any methodologically sound role in archaeology.[104] This was partly in reaction to the uncritical uses of analogy associated with earlier evolutionary theories (Wylie 2002: 138-41). More generally, it was motivated by an increased awareness of the diversity of human cultures, throwing doubt on assumptions of similarity between prehistoric and ethnographically known societies. From the 1950s onwards, a number of sceptics argued that, due to this diversity, ethnographic analogies could not be trusted to contribute much to archaeological interpretation. In fact, several distinct worries were raised, addressing all of the three uses of analogies mentioned in the introduction.

With regard to analogies used as inferences to accept an archaeological interpretation, the worry was that the diversity of human culture shows that analogical inferences are too unreliable to support any rational inferences about the past. This point was articulated in an oft-cited paper by the British archaeologist M.A. Smith (1955). She argued that due to the fact that "an incredible variety of codes of behaviour … actuate human conduct" (5), we cannot establish any necessary links between "the human activities we should like to know about", i.e. what human culture was like in past societies, "and the visible results that survive from them" (6). This lack of reliable linking principles, according to Smith, makes "it is a hopeless task to try to get from what remains to the activities by argument" (*ibid.*). In other words, her worry concerns whether inferences based on ethnographic analogies can provide a rational basis for drawing conclusions about past societies.[105]

[104] As Wylie points out, these debates seemed to go through 20-year cycles where general sceptical worries would replace cautiously optimistic attempts at determining the valid use of analogies, and vice versa.

[105] Smith thinks archaeology should instead focus on the claims that can be more directly inferred from material remains, such as whether bronze tools were available at a given time period. Her scepticism thus specifically concerns claims about cultural or social activities of past societies.

While Smith articulated genuine challenges to inferences based on ethnographic analogies, and while most archaeologists by this time had, rightly, become sceptical of uncritical analogical inferences, many resisted the strong conclusion that all analogical inferences are inherently unreliable and that archaeologists should therefore severely restrict their ambitions. A more optimistic line of research, summarised by Robert Ascher (1961), tried to formulate criteria for identifying more trustworthy uses of analogical inferences. Some criteria focused on what kind of analogies are most likely to resemble the target society. For example, it was argued that analogical inferences are more likely to be reliable if the societies being compared are close to each other in time and space, especially if there was historical continuity between them, or if the ethnographic source of the analogy lived under ecological circumstances similar to those of the prehistoric society investigated (Ascher 1961; Clark 1951, 1953; Childe 1956). Other archaeologists differentiated between the types of conclusions that could be drawn reliably from analogical inferences. For instance, it was argued that inferences about technically or physically constrained aspects of human culture, such as methods of production, are more trustworthy than those concerning more symbolic aspects, such as the religious meaning of an artefact (Hawkes1954).

These arguments made the reasonable point that not all analogical inferences are equally problematic and, thus, that undifferentiated scepticism about analogical inferences is not justified. However, it does not thereby show that analogical inferences which satisfy these criteria are particularly trustworthy: being better than completely unreliable does not necessarily make for very high reliability in absolute terms.[106] Consequently, proponents of the cautious use of analogies often also stressed that

[106] As we shall see in Section 5.4, Roman archaeologists criticise interpretations based on analogies within the Roman world, where the criteria of spatiotemporal proximity and continuity are satisfied to a much larger degree than in many ethnographic analogies.

archaeologists should not simply accept the conclusions of analogical inferences, even of the more plausible kind. Rather, analogies provide “an alluring inference” (Childe 1956: 56), something which should “spur the prehistorian to further effort and provide him clues for purposive archaeological research” (Clark 1953: 355). In other words, analogies do not provide sufficient grounds for accepting an interpretation, but should generally only be used to *generate* hypotheses to be further *pursued* (Orme 1974: 201).

The idea that archaeologists should treat interpretations as hypotheses to be tested through further archaeological work, e.g. field work designed to test specific hypotheses, was also an influential idea among New Archaeologists around the same time.[107] For example, Lewis Binford (1967, 1972) argued that Ascher’s (1961) criteria are not strong enough for analogy to supply unproblematic interpretations of archaeological data and that, instead, analogies should be used as “a means for provoking new types of investigation” (Binford 1967: 1).[108] According to Binford, “Analogy serves to provoke certain types of questions which can, on investigation, lead to the recognition of more comprehensive ranges of order in the archaeological data” (10). Similarly, Patty Jo Watson (1979) recognises that ethnographic analogy is a “wonderful means of generating … hypotheses” (286) which, however, “must be tested in other ethnoarchaeological situations and against the archaeological record itself” (*ibid.*).

This approach to analogies has been further developed by Ann Stahl (1993), drawing on similar ideas from Richard Gould (1978) and Wylie (1985). Stahl distinguishes between: (i) *illustrative* uses of analogy, where an ethnographic analogy is used to draw conclusions about a past society that are not evident from the archaeological

[107] “New Archaeology” was a movement, especially influential in North America, which sought to transform archaeology into a scientific discipline following Hempel’s (1966) account of natural science. See Bell (1994), Wylie (2002) and Krieger (2006) for historical accounts of this movement.

[108] As Smith (1977) argues, Binford still seems to use something like Ascher’s criteria to identify interpretations which are *more probable,* though not probable enough to accept outright.

evidence; and (ii) *comparative* uses, where an ethnographic analogy is only used as a starting point from which archaeologists should then consider to what *extent* the apparent analogue is similar to or differs from the prehistoric society in question. This corresponds to the distinction between analogies to provide reasons for accepting an interpretation, and merely using them to generate hypotheses to be further pursued. However, it goes beyond previous suggestions, since the pursuit of analogy-based interpretations, on Stahl's proposal, does not merely aim to confirm or disconfirm the interpretative hypothesis, but more generally to uncover similarities and differences between the archaeological subject and different potential analogues.

While the emphasis on testing the hypotheses suggested by analogies was to some extent motivated by a recognition of the limited reliability of even the best analogical inferences, it need not involve a rejection of the first line of argument, that we can distinguish between more and less reliable analogical inferences. Rather, it can be seen as supplementary, highlighting a second role for analogies in archaeology, namely to generate possible interpretations which can then be tested or otherwise pursued to investigate to what extent they correctly represent the past society.

However, this role for analogies has been met by a second challenge, voiced for instance by L.G. Freeman (1968) and later Gould (1980). According to this challenge, the wide range of variations that can be observed between contemporary cultures, and the fact that cultures evidently change through time, suggests that past societies might not be similar to *any* extant societies. But if this is the case, relying on ethnographic analogies to generate possible interpretations would effectively blind archaeologists to the possibility that archaeological remains could have been used in ways that are not exemplified by any known society. As Gould puts it: "Even the strongest analogies based on well controlled continuous or discontinuous models cannot inform us adequately about prehistoric adaptions that have no modern counterpart" (1980: 36).

This challenge has been met by two, mutually supporting replies. First, Peter Ucko and Andrée Rosenfeld (1967) pointed out that if the problem is that archaeologists might not think of the right hypotheses, limiting the resources for generating possible interpretations cannot be the solution. Just as archaeologists should not assume that past societies are similar to presently existing ones, the opposite assumption, i.e. that past societies are in no way similar to present ones, is equally unwarranted. In order to avoid overlooking possible forms of cultural expressions, Ucko and Rosenfeld instead recommend that archaeologists seek out as wide a range of ethnographic parallels as possible. The purpose, as Ucko later puts it, is "to widen the horizons of the interpreter" (1969: 262). They criticise alternative methods for generating hypotheses, e.g. based on what is "evident" from the archaeological data, on the grounds that these risk simply being expressions of the interpreters limited preconceptions: "To reject the variety of human experience in different conditions, past and present, for the assumption that at any given time in history one can (in whatever age one lives) be sure about the intentions of other peoples' activities is to go from the frying pan straight into the fire!" (Ucko and Rosenfeld 1967: 153).[109] Similarly, Wylie (1982; 1988) points out that there is no reason to think that Gould's alternative methodology for generating interpretations—to assume in the first place that societies will exploit their ecological context in the most efficient way and only add cultural explanations to the extent that these are necessary to explain anomalies

[109] Ucko and Rosenfeld also allude to "a postulated metaphysical system of male and female symbolism, both rooted in a certain school of analytical thought" (1967: 150), that they take to underlie some claims about what is evident from the material itself. In a nutshell, their worry is that the kind of general interpretative frameworks they allude to are simply veiled expression of contemporary prejudices about prehistoric cultures, rather than revealing anything which is genuinely evident in the archaeological material (151).

(Gould 1980: ch. 2 and 6)—is any more likely to generate the right interpretations. Again, even if analogies are not guaranteed to solve the problem of how to generate interpretations, it supplies one useful tool for overcoming problems of limited imagination.

Second, Wylie (1988: 146-147) argues that even when we have reason to believe that no currently known analogue exists for a given past society, analogies can still be useful for formulating hypotheses about *how* it differs from contemporary societies. Multiple analogies can be adapted and combined by considering how the past society being investigated could partly resemble different analogues. As she points out, Gould (1980: 30-31) himself provides an example of this when he proposes that the hunting practices of humans prior to the adoption of fire might resemble that of non-human predators. Even if this hypothetical society does not resemble any particular contemporary human or animal group, partial analogies with both of these provide a productive starting point that allows archaeologists to think beyond the already known.

The fact that analogies are an efficient means to generating potential interpretations, however, gives rise to a worry that pulls in the opposite direction: that this generates *too many* possible interpretations for archaeologists to consider. This worry is for instance raised by Orme (1974), in discussing the views of French archaeologists Annette Laming and André Leroi-Gourhan.[110] While they thought that ethnographic analogies could establish certain, vague generalisations—e.g. "primitive people are preoccupied with the sacred" (Orme 1974: 203)—they argued against their use for any other purposes. As Orme states their view, the diversity of possible ethnographic parallels means that "Any

[110] Laming and Leroi-Gourhan criticise the reliance on analogies in general, without distinguishing the different possible roles they can play as I have done here. For my purposes, what is important is that Orme here formulates an important challenge to the use of analogy, not whether Laming and Leroi-Gourhan specifically meant to make this point.

further use of analogy leads only to wild speculation" (204). Thus, one of the problems with using analogies is that there is "no means of selecting probable analogies from the great diversity of possible ones" (205). The issue raised here, then, is that the diversity of the ethnographic record provides too much leeway in the generation of hypotheses, allowing archaeologists to be distracted by too many irrelevant possibilities.

It might be thought that the problem of distinguishing probable analogies from merely possible ones can be solved by submitting them to testing. The problem, however, is that if there are too many possible alternatives, testing all of them becomes unfeasible. As Merrilee Salmon (1976) notices, for any given piece of archaeological evidence: "With sufficient ingenuity one can construct a great number of different hypotheses to fit the data" (379). Thus, there is a question of how archaeologists choose which hypotheses to focus on: "Why weren't any other hypotheses with the same implications considered here?" (379). Building on this point, Bruce Smith (1977) similarly notices that: "It is … clear, however, that scientists, including archaeologists, do not consider all logically possible hypotheses, but initially distinguish between those that are reasonable and those that are not." (604). Thus, between the generation of a possible interpretation and its testing, there has to be a stage where scientists or archaeologists try to "determine whether the hypothesis deserves to be seriously entertained and tested" (*ibid*, quoting W. Salmon 1967: 113).

For both Salmon and Smith, the reason why certain interpretations are not considered is that they are not plausible enough to merit consideration: "The alternative hypotheses which could account for the observed phenomena were so initially implausible that they were not even mentioned." (M. Salmon 1976: 379).[111] Since both

[111] Smith (1977: 604) points out that in the specific case analysed by Salmon, i.e. Longacre (1970), some alternatives were in fact considered. However, the general point still holds that Longacre did not consider, and could not have considered, every possible alternative explanation.

draw on Wesley Salmon’s reconstruction of plausibility arguments within the Bayesian framework (cf. Chapters 1, 2 and 4 above), M. Salmon and Smith identify the plausibility of a hypothesis with its prior probability. They regard ethnographic analogies as one important way to assess the plausibility of a hypothesis. On this account, then, while analogical inferences may not on their own be strong enough to warrant *accepting* a given interpretation, they may still show that it is *plausible*. This, they argue, solves an important problem in archaeological methodology, namely how to determine which hypothesis to pursue, which they also take to be the way e.g. Binford (1967, 1972) in fact uses analogies (Smith 1977: 605; Salmon 1982: 45).

For my purposes, what is important about Salmon’s and Smith’s discussions is that they highlight a third role for analogies in archaeology beyond (i) supporting inferences in favour of particular interpretations and (ii) generating possible interpretations that can subsequently be pursued, namely (iii) to provide reasons for prioritising the pursuit of certain possible interpretations. As I have argued in previous chapters (e.g. Sections 2.3, 2.6 and 4.3-4.5), however, the analysis Salmon and Smith provide of *why* analogies can play this role, viz. by increasing the plausibility of the hypothesis, is not satisfactory as a general account of pursuit worthiness. While plausibility judgements play some role, it is not the only relevant factor, and analogies can provide reasons for pursuing a hypothesis even if they do not make it any more plausible. In Section 5.3.3, I will discuss in more detail how analogies can support pursuit worthiness in archaeology.

5.3. Philosophical Analysis

The three uses of analogies in archaeological theorising identified above are not mutually exclusive—each can play important and often complementary roles in archaeological research—yet it is important not to conflate them. Because they serve different purposes, the different uses of analogy have different adequacy criteria. In the following, I analyse each of these with the aim of elucidating the purposes and adequacy criteria for each use of analogy within archaeological theorising.[112]

5.3.1. Analogies and Reasons for Acceptance

Perhaps the most straightforward use of analogies is to provide reasons for the truth of a given interpretative hypothesis or, more realistically, to help identify which of the competing hypotheses is most likely to be roughly correct.[113] The adequacy criteria for this use of analogies are correspondingly simple: they come down to whether the proposed inferences provide good reasons for the interpretation being true or correct. It is, however, also the most controversial use.

Philosophers have proposed different analyses of analogical arguments. The most simplistic understanding of analogical inferences is as a direct induction of the form:

> (DA1) An archaeological subject A and an independently known source B are similar with regards to a range of features f_1, f_2, f_3, ...
>
> (C) Thus, A and B are also likely to be similar with regards to some further feature f_n.

As Wylie (2002: ch. 9) has persuasively argued, much criticism of analogy in archaeology is motivated by the recognition that this simplistic inference schema is not, in general, valid or reliable. This is especially the case in archaeology, where the known variability

[112] Some of my discussion in the following repeats points developed in previous chapters. My focus here is on giving an intuitive account adapted to archaeology and based on archaeological examples.

[113] Notice that a hypothesis can be “the most likely”, without being particularly likely in absolute terms. Similarly, a hypothesis can be “roughly correct” while being wrong about many details.

of human culture provides positive reasons to regard the uncritical application of this inference schema as unreliable. Instead, Wylie has recommended two ways for archaeologists to move beyond these unsatisfying types of analogical inferences.

First, she has urged archaeologists to adopt a more nuanced understanding of analogical inference which focuses not simply on similarities, but on *relevant* similarities *and differences*.[114] This approach was pioneered by Mary Hesse (1966) and has more recently been developed in further detail by Paul Bartha (2010) (see the inference schema in Chapter 4, Section 4.3). On this account, an analogical inference is not a simple inductive inference to further similarities. Rather, it first identifies a "prior association", that is, some kind of causal or functional relation between a range of features in the source domain; e.g., the relations between a certain type of pottery, the cooking practices it is used for in the source domain and the patterns of wear and residue this tends to produce on the pottery could constitute a prior association. The prior association determines what count as relevant similarities and differences between the source and subject domain. Specifically, an analogical inference is only supported if (i) the similarities between the source and the subject concern the features that are involved in the prior association and (ii) they lack any *critical* differences, i.e. features that prevent the prior association from holding in the subject domain. If the source and subject have some of these similarities, and there are no critical differences, this makes it plausible to "transfer" the prior association to the target domain, that is, to infer that the archaeological subject also has the remaining features involved in the prior association—but not, notice, any features of the source domain not involved in the prior association. To continue the example, if sufficiently similar patterns of wear and residues can be found on the pottery in the

[114] One motivation to adopt this analysis of analogical inferences is to avoid the charge that analogies prevent archaeologists from taking differences into account (Wylie 1985, 2002).

archaeological subject, this makes it plausible that it was used for similar cooking practices. However, one cannot use this argument to infer, say, anything about the value of this type of pottery as a status symbol unless directly connected to the production of the relevant patterns of wear and residues.[115]

While presenting a more nuanced approach to analogical inference, this type of inference still only provides a limited degree of support. As Bartha argues, this type of inference can in itself only establish a hypothesis as *prima facie* plausible. This is a quite weak commitment, amounting to no more than that the conclusion *could* be the case, and that it is not an unreasonable proposal. A hypothesis can be plausible without being particularly likely, and several incompatible hypotheses can be regarded as plausible at the same time. The inference pattern outlined above certainly allows for this: it is possible for several interpretative hypotheses to satisfy the criterion of there being some relevant similarities, and no known critical differences, between the archaeological subject and *some* supposed analogy or other. If some hypothesis exhibits a larger number of relevant similarities, we might reasonably regard it as *more* plausible than the competitors. Even so, nothing in Hesse and Bartha's account guarantees that this adds up to more than a rather weak degree of support. For this reason, Bartha emphasises that the main practical implication of a hypothesis being deemed plausible on the basis of an analogical argument is that it should be investigated further—i.e. that it should be *pursued*. I discuss how analogies can justify pursuit in Section 5.3.3.

We need to consider, then, how to distinguish stronger forms of analogical inferences. This is Wylie's second recommendation. In general, to conclude anything

[115] Wylie (2002: 149-150) highlights another example of this type of inference, namely Curren's (1977) argument that a certain type of rib was used in pottery production (also discussed by Salmon 1982: 60-63). In this case, the inference was undermined by the existence of a critical difference, namely, as pointed out by Starna (1979), that some of the societies from which these ribs stemmed show no evidence of pottery production.

stronger than that a hypothesis is minimally plausible requires some additional argument, supported by relevant background knowledge.[116] I distinguish three inference patterns archaeologists can use and discuss to what extent analogies figure into these.

The simplest approach is to shore up the (DA1)-(C) schema above. Rather than accepting all instances of this schema, it should be restricted to those cases where our background knowledge licenses assuming a relevant supporting premise, for instance of the form:

(DA2) A and B are likely to be similar with regards to *f*-type features.

For example, we might accept inferences about the function of certain types of pottery in a society A based on an analogy with an ethnographic source B if there are independent reasons to think that A and B are likely to be similar in terms of the relevant uses of pottery. This strategy for strengthening analogical inferences was effectively what many of the cautiously optimistic defenders of analogy in the 1950s and 1960s tried to do by identifying circumstances—e.g. historical continuity, temporal and spatial proximity, ecological similarity, technologically constrained features—where we have better reason to assume this type of supporting premise (cf. Wylie 1988).

While this suffices to dispel undifferentiated scepticism about analogies, one should be careful not to overestimate the extent to which it legitimises analogical inferences. The problem is that archaeologists rarely possess enough relevant background knowledge to support anything but a very cautious inference of this form. As we shall see in Section 5.4, even when archaeologists do have a lot of background knowledge and the source and

[116] That analogical inferences are warranted by relevant background knowledge has been defended by (Weitzenfeld 1984) and Norton (*ms*). Wylie (1988) argues this is how archaeologists have in fact attempted to strengthen analogical inferences.

subject are historically and spatially close (e.g. within the same region of the Roman Empire), there are often still good reasons only to put limited trust in such inferences. This does not mean that they are never reasonable. As Salmon (1982: 58, also 78-9) points out, any functional ascription based on form can be thought of as an inference from similar to similar. For instance, the judgment that a group of very similar artefacts, found on the same site and dated to the same time-period, were used for similar purposes is essentially an analogical inference of this kind. At least in many such cases, this would be a quite reasonable inference. Analogical inferences supported by background knowledge should be evaluated on a case-by-case basis, not rejected or accepted across the board. Nonetheless, many (perhaps most) applications of direct analogical inference schema will only provide limited support for its conclusion.

A different type of inference is where background knowledge about the specific nature of the "prior association", i.e. the causal and functional relations between features in the source domain, makes it reasonable that *those* particular features are likely to co-occur. Here, the inference has the form:

> (IA1) The archaeological subject A has features f_1, f_2, f_3, …
>
> (IA2) In general, features like f_1, f_2, f_3, … are only likely to be produced as a result of f_n
>
> (C) Thus, A is likely also to have feature f_n.

An example of this type of inference is employed by experimental archaeologists, for instance in studies of cut marks on animal bones (e.g. Seetah 2008). Here, it may be found that certain patterns of cut marks ($f_1, f_2, f_3,$ ...) are unlikely to occur except if a specific technique is used for carving up the animal (f_n). Thus, if we find animal bones with these patterns, we may infer that the same butchering techniques was likely used by the ancient society which consumed the animal in question.

Although archaeologists sometimes describe this as a form of analogical reasoning, it differs from a direct analogical inference from B to A in that it does not rely on background assumptions about similarities between A and B specifically. Rather, studies of a contemporary context B, e.g. through anthropological studies or experimental archaeology, here provide support for a general account of *f*-type features, (IA2), which in turn licenses the inferences from $f_1, f_2, f_3, \ldots$ to f_n.[117] Of course, obtaining knowledge of the kind expressed by (IA2) is still difficult. This inference pattern however has the advantage over direct analogical inferences that there are feasible strategies for testing and strengthening the support for premises of the form (IA2), e.g. through additional anthropological field work or experiments on contemporary sources. By contrast, it is less clear how archaeologists should (in general) go about testing premises of the form (DA2).

A final way background knowledge may be used to support an interpretation is by ruling out other competing interpretations, thereby supporting an eliminative inference. Briefly put, eliminative reasoning supports a hypothesis *H* to the extent that (a) *H* can account for evidence *E* and (b) we can rule out, or at least showing highly unlikely, all plausible alternative accounts of *E*.[118] In archaeology, an eliminative inference can be represented through the following schema:

[117] One might say that (IA2) assumes that all contexts, including A and B, are similar with regards to the connection between the *f*-type features. The point is that this premise is licensed by the general theory of how *f*-type features can be produced, rather than knowledge about contexts A and B specifically.

[118] For a recent discussion of eliminativist reasoning, see Reiss (2015).

(EI1) The archaeological subject A has the features f_1, f_2, f_3, …

(EI2) The hypothesis that A has feature f_n could account for it having f_1, f_2, f_3, …

(EI3) All (or most/many[119]) alternative accounts to the hypothesis that A has f_n can be rejected.

(C) A is likely to have feature f_n.

Here, the crucial premise is (EI3). Whether this can be assumed will partly depend on whether one's evidence and background knowledge suffices to rule out the competing hypotheses one has considered. However, it also requires that there are good reasons to think that one has considered all serious competitors, and not simply overlooked or failed to consider some plausible alternatives.[120] This often difficult to guarantee, since it would ideally require evaluating the entire range of all possible hypotheses, including those one has not thought of yet.[121] This is where analogies can become relevant. Although it is thus difficult to evaluate the absolute strength of an eliminative inference, one can still *strengthen* it, in comparative terms, by making a serious effort to think of as many plausible competitors as possible. This is the way Ucko and Rosenfeld (1967) argued analogies should be used. I will now discuss this generative use of analogies.

[119] As Reiss (2015: 357-8) argues, the strongest eliminative arguments rule out all relevant alternative accounts, but weaker degrees of warrant can be obtained by ruling out at least some, many or most alternative accounts.

[120] It is not necessary to consider all the completely implausible alternatives, since their implausibility would provide immediate grounds for rejecting them anyway.

[121] This what Stanford (2006) calls the "problem of unconceived alternatives". It is a more general version of the worry raised by Freeman (1968) and Gould (1980), i.e. that archaeologists will often fail to generate the right hypotheses.

5.3.2. Generating Interpretations

There are at least two different reasons for using analogies to generate possible interpretations. First, they can suggest new interpretations in a context where archaeologists are looking for a working hypothesis to guide further investigations. This can happen, for instance, when a new body of archaeological evidence is discovered, such as a site or a previously unknown type of artefact. Here, archaeologists will naturally seek to formulate potential interpretative hypotheses about the new evidence. Another example could be when such new piece, or new methods of analysis applied to existing evidence, overturns existing interpretations and thus forces archaeologists to look around for other possible interpretations. In both cases, the purpose of generating interpretations is primarily to suggest candidates for further development and investigation. Thus, the adequacy criteria for using analogies to generate new candidate interpretations in this context coincide with those for using analogies to select a hypothesis for pursuit.

The second reason for using analogies to generate new interpretations is to strengthen eliminative inferences, in the way outlined at the end of the previous section. In the context of archaeology, as we have seen, the variability of human culture and the lack of strong background knowledge makes the problem of unconceived alternative interpretations particularly pressing. Often, it is reasonable for archaeologists to suspect that there are many plausible interpretations which could be applied to the same evidence. A lack of *prima facie* plausible contenders might therefore indicate a lack of imagination on the part of the archaeologist rather than a strong argument for the received interpretation. To build an eliminativist case for accepting a given interpretation, an archaeologist should make a serious effort to articulate as many *prima facie* plausible interpretations of the same evidence as practically possible.

It is this use of analogies which Ucko and Rosenfeld (1967) advocated. An example is Ucko's (1969) review of the many different kinds of funerary practices that are known to humans, and their relation to religious ideas. Ucko's paper primarily served a critical purpose, namely to highlight that many common interpretations of what funerary practices signify seem to reflect modern prejudices. For instance, Ucko (1969: 265) cites a number of counter-examples to the idea that elaborate funeral rituals indicate that the culture in question believed in an afterlife. In this way, using analogies to generate plausible alternative interpretations can serve to re-evaluate the strength of previous interpretations.

But generating alternative interpretations can also be seen to serve a more constructive role, since it highlights the kinds of alternative interpretations archaeologists would have to argue against in order to strengthen the positive case for a given interpretation. While sceptics such as Freeman (1968) and Gould (1980) correctly point out that it is difficult to positively assert that all plausible alternatives have been ruled out, one can still strengthen eliminative arguments by generating and arguing against a wider range of alternatives. Analogies provide one useful tool for formulating as many plausible alternatives as possible.

Whether analogies are used to generate alternative interpretations for critical or constructive purposes, it is important to notice that the relevant adequacy criteria differ from those relating to analogical inferences. While analogical inferences are stronger when we have reasons to expect that the source and subject are likely to be similar, the generative use of analogies discussed here aims to consider as a wide a range of alternatives as possible. As Ucko and Rosenfeld notice: "The more *varied* and the more *numerous* the analogies that can be adduced, the more likely one is to find a convincing interpretation for an archaeological fact" (1967: 157, emphasis added). Thus, one should focus on analogies that are likely to suggest interpretations which are *different* from those already considered, rather than situations that seem most likely to be similar to the subject of interpretation. Focusing on those cases that seem most likely to be similar risks being counter-productive, since it will tend to restrict attention to a narrower range of possibilities. While the alternatives should have some minimal plausibility to be worth ruling out at all, this can be achieved by focusing on analogies which conform to Bartha's schema. Beyond this minimal requirement, however, using analogies to generate alternative interpretative hypotheses for the purpose of strengthening an eliminative inference should focus on diversity and extensiveness rather than likeliness.

5.3.3. Reasons for Pursuit

To pursue an interpretative hypothesis is to investigate whether, or to what extent, it is correct or not.[122] This can be done, for instance, by testing it directly (when possible), by trying to establish the supporting premise of an analogical inference, e.g. of the form (IA2), by trying to formulate and test other competing interpretations (thus strengthening the eliminative case for it), or by developing or spelling out the interpretation in more detail. To have *reasons* for pursuing an interpretation is to have reasons to prioritise this kind of effort. Reasons for pursuit are both relevant to determining how to generate promising candidate hypotheses, as well how to prioritise already generated hypotheses.

As Merrilee Salmon (1976) and Bruce Smith (1977) point out, analogies are often employed in archaeology at this stage of inquiry, when researchers are trying to determine which interpretations deserve further attention. They take analogies to provide reasons for pursuit by showing the interpretation more likely or plausible than its competitors.

[122] I here focus on pursuit for epistemic purposes, i.e. pursuit aiming to learn more about the world rather than pursuit motivated by practical purposes.

While analogical inferences are rarely strong enough to justify accepting an interpretation outright, on Salmon and Smith's account they provide some initial degree of support which then forms the basis for choosing which interpretation to pursue. They thus take analogies to provide justification for pursuing an interpretation by virtue of raising its plausibility.[123]

As argued in Chapter 2, however, plausibility is not the only relevant consideration for pursuit worthiness. Of course, for epistemic purposes, it would be a waste of time to work on interpretations we already know are completely implausible. If someone were to propose without further evidence that Iron Age Britons had extensive, regular trade with people (say) in South America, few archaeologists would be willing to invest resources, time or effort into investigating it. Having a minimal degree of plausibility is usually necessary for a hypothesis to be pursuit worthy. However, being more plausible does not always make an interpretation more pursuit worthy. For instance, given the available textual sources and archaeological evidence, it is extremely plausible that Roman-style forts in Britain were occupied by soldiers from the Roman army around the first century A.D. But exactly because this is so plausible, we are unlikely to learn more about this particular question by pursuing it further. Instead, the interesting questions are those that remain more uncertain, such as which provinces of the Roman empire the soldiers came from and how they interacted with the local population. On the other hand, if new evidence came to light which suggested that some forts might not have been occupied by Roman soldiers after all—for instance, if a new written source claimed that certain forts were constructed and occupied by locals during this time—this would make it more

[123] M. Salmon (1982: 78-9) also mentions that some hypotheses are so likely that "further testing would be otiose" (78). She thus recognises that raising prior probability does always increase pursuit worthiness. However, unlike me, she seems to assume that *when* analogies justify pursuit, it is because they show that the hypothesis is likely enough to be worth pursuing.

interesting to investigate whether this could be confirmed. In this case, the new evidence makes the previous interpretation *less* plausible but *more* pursuit worthy.

From these considerations, we can see why we need to go beyond M. Salmon and B. Smith's account. As Bartha argues, an analogy can show that an interpretation is minimally plausible by demonstrating that it is at least possible for humans to manifest a certain type of behaviour in association with the observed archaeology. But once this minimal degree of plausibility is established, raising the probability of an interpretation does not necessarily make it more pursuit worthy. Nonetheless, I want to argue that there are other reasons why it is often reasonable for archaeologists to pursue analogy-based interpretations.

As argued in Chapter 2, the leading considerations in the pursuit worthiness of a hypothesis are (a) what could we potentially learn from pursuing it, including how likely are we to get reliable and strong enough evidence to learn these things, and (b) how interesting or valuable it would be to learn them. By 'valuable' here I mean the intellectual value of learning more about the things that interest us.

First, one should consider what kinds of questions we can expect to be able to answer through further pursuit. For example, one should consider what kind of hypotheses can be most effectively and reliably tested given the available evidence. If a site yields a large quantity of well-preserved pottery sherds, this can be a good reason to focus time and efforts on pursuing specific questions (e.g. what kind of pottery the inhabitants of the site preferred? Was it was imported or locally produced? Does it show signs of having been used in cooking?), even if many of the interpretations tested are not initially particularly plausible. In fact, having such extensive data can be a reason for pursuing hypotheses already suspected to be false, because such data may enable the researchers to rule them out conclusively.

Second, one should also consider the potential value of the interpretation, i.e. how interesting it would be to find out whether it is correct. Although questions about the kinds of material that were available at a particular time and place are often easier to reliably investigate, much of the interest in archaeology comes from deeper questions about culture and social structures in the past. For example, what kind of trading or other interactions took place between Roman soldiers and local populations in the provinces? Was society in the Iron Age dominated by political elites? How widespread was a belief in an afterlife? Even if these kinds of questions are more difficult to answer, they are sufficiently interesting and significant that archaeologists will often spend considerable efforts trying to answer them, as far as this is possible, or at least to clarify the extent of our ignorance about them. Importantly, in these cases, the lines of inquiry that archaeologists may want to pursue are both ones which could confirm a hypothesis *and* disconfirm it.[124] For instance, when excavating a farmstead in Roman Britain, it would be interesting to find evidence of trade with a local Roman fort, but demonstrating a complete *lack* of the kinds of evidence one would expect to find if such trade had taken place would also be very interesting.[125] Both outcomes would tell us something important about the interactions between Roman soldiers and the local population.

I want to suggest that it is often because they raise these kinds of intrinsically interesting questions that it is reasonable for archaeologists to investigate analogy-based interpretations. Many significant questions about human culture are comparative, that is, they concern to what extent human cultural expression is similar or different across time and space. An analogy-based interpretation will concern exactly this type of question. For

[124] More generally, as I emphasise below, the interesting question is often not *whether* a given hypothesis is correct but the *extent to which* is accurately describes aspects of the archaeological subject.

[125] How to interpret a negative result of course depends on whether we would expect this kind of trade to leave traces in the archaeological record. I assume for the purposes of this example that there are good reasons to think that some evidence of Roman trade would have been preserved on the site in question.

instance, ethnographic analogies raise the question of whether—or more generally, to what extent—prehistoric hunter gatherer societies resemble contemporary or near-contemporary groups known from anthropological or ethnoarchaeological studies. For questions of this kind, we are not only interested in evidence for or against specific interpretative hypotheses, but also evidence more generally of both similarities and differences. For example, Stahl (1993) argues that most ethnographically studied hunter-gatherer groups have been fundamentally shaped by colonialization and, therefore, many aspects of their culture are likely to be unique to the modern era. On the one hand, this throws doubt on naïve analogical inferences which simply project the culture of these groups wholesale into the past. But investigating *which* aspects of modern hunter-gatherer societies have been shaped by colonization also tells us something interesting about how past and contemporary societies are likely to differ. While archaeologists would of course also like to be able to give a positive account of what life in prehistoric societies was like, knowing how it is likely to *differ* from contemporary groups does provide interesting insights about the past.

The above observations should not be taken to indicate that analogy-based interpretations are always the most pursuit worthy hypotheses in archaeology. Rather, I have outlined some salient criteria for deciding which hypotheses should be prioritised for pursuit and explained why there is often *some* reason—though not necessarily *sufficient* reasons, all things considered—to pursue analogy-based interpretations.

5.3.4. Summary: Different Uses of Analogy and Their Adequacy Criteria

The key points of the preceding analysis, to be further illustrated in the next section, are the following.

First, there are at least three distinct uses of analogy in archaeology: (i) providing reasons for accepting a hypothesis, (ii) generating hypotheses and (iii) providing reasons for pursuit.

Second, each of these have their legitimate uses. As has been argued by defenders of analogy, wholesale scepticism about analogies in archaeology is unjustified. The key challenge is to formulate adequacy criteria for each use of analogy, i.e. criteria for when an analogy can be legitimately used in that way.

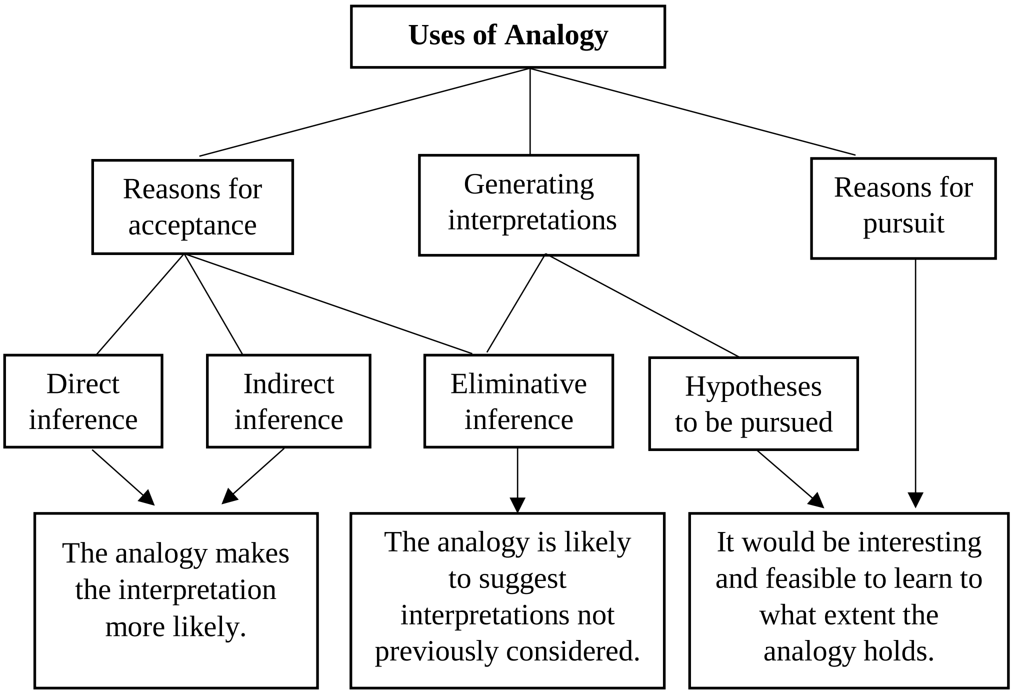
Uses of Analogy
Reasons for acceptance
Generating interpretations
Reasons for pursuit
Direct inference
Indirect inference
Eliminative inference
Hypotheses to be pursued
The analogy makes the interpretation more likely.
The analogy is likely to suggest interpretations not previously considered.
It would be interesting and feasible to learn to what extent the analogy holds.

Figure 5.1: Typology of uses of analogy. First row: three uses of analogy. Second row: sub-types of these uses. Bottom row: adequacy criteria for each use of analogies, as indicated by the arrows.

Third, different uses of analogies have different adequacy criteria. Analogical inferences for accepting an interpretation are adequate to the extent that they make the interpretation more likely. Generative uses of analogies, when used in the service of eliminative reasoning, are adequate to the extent that they suggest interpretations not previously considered. Analogies used to provide reasons for pursuit are adequate to the extent that it would be interesting and feasible to learn the extent to which the two societies are similar. I have summarised these points in the typology of uses of analogy below (Fig. 5.1). In summary, the upshot of my analysis is that archaeologists should be clear how they are using analogies and evaluate the analogies accordingly.

5.4. Case Study: Pompeian Household Artefacts

The debate on analogy, as reviewed in Section 5.2, has mostly focused on cases where analogies with ethnographically known societies are used to interpret prehistoric societies. However, the issues surrounding analogies are not specific to prehistoric archaeology: analogies are also widely used in other branches of archaeology, and although there are some differences worth noting, the same points of principle apply in these fields.

In this section, I will apply the philosophical framework developed in Section 5.3 to analyse uses of analogy in Roman archaeology. This will, first, show how the problems concerning analogies apply outside prehistoric archaeology and, second, illustrate how my framework applies to a concrete case. I start by discussing how analogies in Roman archaeology have been criticised along similar lines to ethnographic analogies. As in the prehistoric case, critics of these analogies still recognise that analogy can, and does, play a legitimate role. In proposing alternative interpretations, analogies are still used. In Sections 5.4.2 and 5.4.3, I explain how the uses of analogies by one of these critics can be seen as reasonable according to the criteria developed in Section 5.3.

5.4.1. The Reaction Against Analogy in Roman Archaeology

There are several salient differences between Roman and prehistoric archaeology. For one thing, Roman societies are not "prehistoric", since we also have textual sources from them. Furthermore, Roman architecture and artefacts are often comparatively well-preserved, whether due to the materials they are made of (stone, pottery, metal, …), the quantities in which they were produced or events such as the eruption of Mount Vesuvius which buried Pompeii. While this does mean that Roman archaeology possesses a wealth of information compared to many other branches of archaeology, this does not, as might be supposed, eliminate the need for analogies or make the need to pay critical attention to them less acute. While the analogies used in Roman archaeology differ from those discussed in prehistoric archaeology, there is no distinction of principle between the two. Interpreting the function of well-preserved artefacts recovered, say, from Pompeii still requires inferences and these will typically be based, whether explicitly or implicitly, on parallels with an apparently analogous item in a better-known context.

As sources for such analogical inferences, Roman archaeologists have not usually turned to ethnographic studies. Instead, they have tended to use two kinds of sources: either analogies with modern 'civilised' societies or analogies with other parts of the Roman world, known either through textual sources or previous archaeological interpretations.

As an example of the first type of analogy, many traditional accounts of Roman life were influenced by analogies with the upper class in Victorian Britain (Hingley 2000) or other European elites. This was often due to the fact that these interpretations were developed by members of the same elites who tended to conceive, say, of the British Empire as a modern-day equivalent of the Roman Empire and therefore tended to interpret the life of Roman elites in ways which mirrored their own, whether intentionally or not.

Another example is interpretations of household artefacts shaped by apparent analogues in modern households, often reflected in the labels given to Roman artefacts. For example, Allison (1999) criticises the use of the Italian term '*forma di pasticceria*' (pastry or confectionery mould) as a label for certain small bronze vessels. She notices that this label "suggests analogies" with moulds used in European pastry-making or possibly with moulds used by Victorians to mould delicacies such as jelly, an interpretation which "serves to link Pompeian eating habits with those of the modern European world" (p. 66). However, Allison argues, there is little evidence for this interpretation. Instead, she suggests that that they might have been used for ablutions, more specifically for pouring water over oneself, "in a manner not dissimilar to that of bathing women in the wall-painting in the bath complex" (*ibid.*) of a Pompeian house. In favour of this interpretation, she mentions that some of these "pastry moulds" are found in the vicinity of large basins independently believed to be used for ablutions. Furthermore, some of the bronze vessels are shaped as sea-shells, with a scoop-like form "suitable for pouring water over oneself" (*ibid.*).

Interpretations based on modern analogies have thus been criticised in ways similar to the criticism of ethnographic analogies in prehistoric archaeology. Because of the vast temporal distance between the source and subject, and because of the many well-known cultural changes that have occurred through the last two millennia in Europe, contemporary Roman archaeologists are rightly sceptical of inferences based on modern analogies.

However, Roman archaeologists have also criticised interpretations based on the second type of analogy, i.e. comparisons with other parts of the Roman world known either from textual sources or other archaeological investigations. While these may appear less problematic than analogies with contemporary society, Roman archaeologists have highlighted how such analogies can be potentially misleading. The problem is that they tend to rely on unfounded assumptions of similarity within the Roman world. As Allison points out, “the term ‘Roman culture’ must surely stand for what was a very multicultural society spanning many continents and centuries” (1999: 57). Thus, one should be careful about assuming, e.g., that the economic system in the North West of Roman Britain was similar to that of the Roman Empire as a whole, or even the South of Roman Britain (Peacock 2016).

Even within the same region, uncritically assuming similarity between sites can be problematic. For instance, Boozer (2015) criticises the tendency to use the well-preserved settlement of Karanis in Roman Egypt “as a “filler” when desirable archaeological evidence is lacking” (99). As an example of this, Boozer notices that in one description of the site Hermopolis Magna, where all Roman houses have been destroyed by previous digging, it is simply assumed that “We must imagine them [the missing houses] to be like those excavated at Karanis” (Boozer 2015: 99, quoting Bailey 2012). First, Boozer highlights flaws and limitations in the original studies at Karanis suggesting that the typology of Romano-Egyptian houses derived from these studies likely overlooked variability within Karanis. Second, even if it were accurate for Karanis itself, Boozer argues that the dominance of this typology in Romano-Egyptian archaeology is problematic. The default assumption tends to be that other sites will resemble Karanis, leading archaeologists to be unduly dismissive of evidence that a building on a given site diverges from the Karanis-typology.

Along similar lines, Allison (1999, 2001) points out that using textual sources from Roman authors to infer e.g. the function of an artefact will, in most cases, involve some kind of analogical inference. A written source can give us insight into how a given artefact was used or perceived in a specific time and place (usually Rome), but one cannot simply assume that these accounts are also valid for other parts of the Roman world which are temporally and spatially removed from the author.

For instance, the type of pottery known as a *mortarium*—a robust, rimmed bowl with a spout and trituration grits—became relatively common on rural sites in Britain after the Roman conquest (Cramp et al. 2011: 1339-40). *Mortaria* are described in Roman recipes as used for processing or mixing ingredients, and have sometimes been interpreted as evidence that the local population increasingly adopted characteristic Roman cooking and eating styles. However, this interpretation assumes that their function in Roman Britain was similar to that described in the written sources, an assumption which has been challenged. An alternative interpretation is that the *mortarium* was adopted by the local population because it was useful for cooking purposes that already existed in Britain prior to the Roman conquest (Cool 2004). Its increased prevalence, according to this interpretation, is simply due to it being more readily available after the conquest. Evidence supporting this interpretation includes the presence of sooting or burning on some *mortarium* sherds, suggesting that meals could have been cooked directly in *mortaria*, rather than them being used for mixing or processing ingredients (Peacock 2016; Cramp et al. 2011: 1340).

To summarise, analogies in Roman archaeology are subject to the same basic worry as ethnographic analogies in prehistoric archaeology. Even within a more narrowly defined context, such as “the Roman world”, human cultural expressions are varied and changeable. Discussions of analogies in Roman archaeology have mostly focused on uncritical uses of analogical inferences, where the archaeological subject of interpretation is simply assumed to be similar to an explicit or implicit analogue. However, as I shall show below, analogies are still being used by these critics, even if they are not always labelled as such.

5.4.2. Allison's Criticism and Use of Analogy

While many Roman archaeologists have criticised earlier interpretations for being based on uncritical analogical inferences, the interpretations which are proposed in their stead often seem to be equally based on analogies. This is not to say that these authors are being inconsistent in their attitude towards analogies, since none of them advocate a wholesale abandonment of analogies in archaeology. Rather, I want to use the framework developed in Section 5.3 to analyse how analogies are being criticised and used in the field. Specifically, I want to use Penelope Allison's (1999, 2001, 2009) discussion of Pompeian household artefacts as a case study.

That some of the interpretations Allison proposes are based on analogy is clear. To guide my discussion, I will highlight three examples. First, in the case of the bronze vessels labelled as *forma di pasticceria* (discussed above), the alternative interpretation Allison proposes refers explicitly to a scene depicted on a wall-painting in Pompeii. In this case, one might take the fact that the source and subject are so spatially and temporally close to be sufficient grounds to assume them similar, i.e. to adopt a premise of type (DA2) from Section 5.3.1. However, the analogy between the two is not perfect: as she notices in a footnote, in the painting "water is being poured from a jug by an assistant or companion" (p. 74, note 7), rather than using a shell-shaped scoop. The proposal that their "scoop-like" form makes this vessel suitable for ablutions seems instead to rely on other examples of scoops being used in that way. Allison's interpretation thus seems to combine the analogy drawn from the wall-painting with analogies to other, familiar bathing practices either from contemporary or historically known contexts.[126] This is thus an instance of the method, described by Wylie (1988), of combining multiple partial analogies to generate new interpretations.

A second example involves a type of table, mentioned by the Roman writer Varro, called a *cartibulum*. Varro mentions that when he was a boy it used to stand in the forecourts of houses with bronze vessels on and around it (Allison 1999: 61). On the basis of this description, Daremburg & Saglio (1881) used tables found in the forecourts of Pompeian houses as illustrations of *cartibula*, even though these tables often have two or three feet, and are circular rather than oblong. Allison points out that tables which fit Varro's description better are more often found in the gardens of houses in Pompeii. However, excavators have sometimes relied on Varro's descriptions to reconstruct and *relocate* these to a neighbouring 'seemingly grander' house (Allison 1999: 62). Apart from this arguably problematic practice of reconstructing the evidence to fit interpretations,[127] Allison points out that Varro was a child in the late republican period more than a century before the eruption of Vesuvius. One cannot assume that Varro's childhood is representative of all of the Roman world, or even of all of Roman Italy, across a century. Instead of uncritically assuming (or worse, constructing) a concordance between textual sources and the Pompeian objects, Allison argues that archaeologists should concentrate on assessing the relationships between the two. For instance, she wonders whether the tables found in the Pompeian forecourts could "conceivably indicate a Pompeian élite who were preserving, or mimicking, behaviours of the Roman élite from

[126] Interestingly, the scallop shell has a tradition as a Christian symbol. In late the Middle Ages, John the Baptist is sometimes depicted pouring water onto Christ's head using a small dish or seashell (this coincides with a shift away from baptism by immersion; cf. Denny 2013: 105). Whether this has influenced Allison's interpretation of the Pompeian bronze vessels, I cannot say.

[127] Due to problems of underdetermination and theory-ladeness, exactly how problematic it is to reconstruct evidence to fit an otherwise plausible interpretation is a vexed issue. In this specific case, the problem which Allison highlights is that this reconstruction was carried out uncritically, i.e. that it simply assumed, without further argument, that Varro's descriptions provide a reliably account of life in Pompeii.

a bygone republican era to establish their credentials as Roman élites?", although she worries that such an interpretation may be largely based on analogies with British colonial behaviour, rather than something which can be "validated through critical appraisal of textual information" (*ibid.*).

A final example concerns the type of pottery called *terra sigillata* or Samian ware. This is a distinctive type of red, glossy pottery found across the Roman world, often assumed to have been used as tableware. Several of the Samian ware bowls recovered from Pompeii contained food remains, probably left behind when the residents fled the eruption. Interestingly, each bowl only contained a single type of food (e.g. a whole bowl of plums). As Allison argues, this tells against the interpretation that each diner was served their own bowl of food. She suggests that this latter interpretation is based on analogies: "Assumptions that Romans ate at the table with individualised utensils that were used as sets may be based rather on funerary practices or on modern analogy than on contextual evidence." (2009: 24). Since the bowls are small enough to hold by hand, and since Pompeian dining rooms did not have space for large enough tables to facilitate buffet-style eating, Allison suggests that "This might imply communal eating habits, where the bowl is passed amongst the diners" (1999: 69). She points out that this style of eating was "common in much of Europe, and also in the United States, until at least the mid-18th century" (2009: 24).

In all three examples, Allison relies on analogies to propose alternative interpretations. However, she tends to reserve the term 'analogy' for the more problematic uses she is criticising. Now, Allison is careful to stress that she does not intend to argue that all analogy-based interpretations are wrong (1999: 72) and, as her interpretative practices demonstrate, she clearly regards some uses of analogy as legitimate. While she does not provide a systematic account of what distinguishes legitimate uses of analogy from more problematic ones, she does make some helpful methodological remarks. In her 2009 paper, Allison offers the following further characterisation of her approach: “Interrogation of the material evidence requires critical readings, and re-readings, of related textual evidence and cross-cultural ethnographic comparisons, not to directly interpret household practices but to expose the biases in our interpretations” (2009: 28). Along similar lines, she argues in an earlier paper that modern analogies “can at best be used to explore relationships between modern and ancient behaviours rather than to explain them” (2001: 194). She, reasonably, warns against using analogies in ways which smuggle the presupposition of a positive analogy into the primary data, such as giving items labels which imply specific functions. Rather, it is only once analyses of the material culture “have been rigorously carried out” that “their relationships with analogical material, textual or cross-cultural, can be explored” (2001: 201-2).

The methodology suggested here, then, is that analogies should be used to ‘interrogate’ the evidence in order to expose biases. Furthermore, archaeologists should ‘explore’ the relationship between the archaeological evidence and analogies from textual or modern sources. But analogies should not be used to ‘directly interpret’ or ‘explain’ the evidence. As noted, Allison does seem to rely on analogies when proposing alternative interpretations. This might be taken to show that her interpretative practice stands in tension with her explicitly stated methodological stance. However, in the following section, I want to use to distinguish different uses of analogy in Allison’s work, along the lines proposed in Section 5.3 and show how this can clarify her methodological recommendations.

5.4.3. Generating and Pursuing Interpretations of Pompeian Household Artefacts

It is clear that what Allison primarily objects to are uncritical analogical inferences where analogies are taken to provide sufficient reasons for accepting an interpretation. In the cases discussed, the problem is that there is not adequate background knowledge to support the premises necessary to ground an analogical inference for accepting the interpretations. First, it cannot support a premise of the form (DA2), required by a direct inference, since the archaeological subject and the analogical source cannot be assumed to be substantially similar. Second, the functions of household artefacts are not constrained enough to license assumptions of the form (IA2), required by an indirect inference. To repeat, this is not to say that these inferences are never reasonable in Roman archaeology, or indeed elsewhere. However, it illustrates that establishing background knowledge sufficient to license these inferences is difficult, especially for the more substantial questions of interest to archaeologists.

Allison's remark that analogies can be used to expose biases in the evidence (2009: 28) points to a generative use of analogies in the service of eliminative reasoning, i.e. the use of analogies recommended by Ucko and Rosenfeld (1967). This does seem to capture some aspects of Allison's use of analogies. By citing the fact that eating from communal bowls was common through much of history, she reminds us that there are other, serious alternatives to the assumption that sets of Samian ware were used for individualised dining sets. Furthermore, mentioning buffet-style dining only to quickly reject it can be seen as a step towards an eliminative argument for the 'communal bowls' interpretation. Recall also that, for this use of analogies, the point is not whether the analogies themselves are likely to be correct. There is little reason to think that Pompeian dining habits are more likely to resemble eighteenth-century European dining practices than twentieth-century ones (and Allison does not suggest so). The purpose of this generative use of analogies is rather (a) to generate alternatives to criticise existing interpretations and (b) strengthen eliminative arguments by ensuring that a wider range of plausible alternatives has been considered.

However, Allison does not attempt the kind of broad-ranging generation of alternatives which Ucko (1969) recommends. Thus, her generative use of analogies should primarily be seen as critical rather than as supporting a positive eliminative argument in favour of the interpretations she proposes. Conforming to this normative verdict, Allison primarily argues that the material and textual evidence should be critically re-examined in order to investigate whether the kinds of interpretations she suggests can be supported. In my terminology, her primary positive achievement is to provide reasons for pursuing these new interpretations, rather than to accept them.

What kind of reasons support pursuit in these cases? One factor is simply that Allison manages to throw doubt on traditional interpretations, either by citing evidence which tells against them, by highlighting that the analogical inferences behind the traditional interpretations rest on unjustified assumptions (e.g. similarity between late republican Rome and early imperial Pompeii), or by generating plausible alternatives that cannot be ruled out on the available evidence. This increased uncertainty suggests that more can be learned by re-examining the evidence, thus giving reasons to pursue this line of investigation.

However, Allison's appraisal also illustrates the point that not all proposed alternative interpretations are equally pursuit worthy. In particular, she worries that it may not be possible to validate the interpretation that Pompeian elites were mimicking earlier republican practices by considering analogies with British colonial elites. Since this kind of interpretation may simply not be something we are able to find additional evidence for or against in the existing textual or material record, it is unreasonable to pursue it. In a similar vein, Allison's 2001 paper argues more generally that archaeologists and historians should be careful to ask the right questions of their evidence, i.e. to only pursue those questions the evidence is likely to be able to answer.[128] As I have argued, an important factor in the pursuit worthiness of a hypothesis is how feasible it is to learn more about it, given the nature of the available evidence.

Finally, consider Allison's recommendation that archaeologists should 'explore the relationship' between the archaeological material and the suggested analogies. This recommendation illustrates the point that analogy-based hypotheses are often pursuit worthy because they raise interesting research questions about the similarities and differences between cultures, across both time and space. For example, the analogy Allison suggests between Pompeian dining habits and the communal dining habits of eighteenth-century Europeans not only provides a possible interpretation of Pompeian Samian ware. It also raises deeper questions about how similar or different Roman culture was to more recent periods, and ultimately to our own culture. Even if all of these analogy-based interpretations ultimately prove unsuccessful, learning that they *fail* to capture Pompeian dining habits still provides interesting insights into these deeper questions. Similar points apply to analogies within the Roman world, where interpretations are based on textual sources. Examining to the extent to which domestic life in Pompeii around the time of Vesuvius' eruption was similar (or different) to that described by earlier Roman authors can reveal interesting facts about the extent of cultural uniformity within the Roman world.

Recognising and distinguishing the three uses of analogy can help us make sense of the more nuanced uses of analogies evident in archaeological practice, such as the

[128] She also speculates that previous interpreters have taken recourse to unjustified analogical inferences exactly because the available evidence was not able to ask the kinds of questions they asked of it.

present case study. In particular, we can see that Allison’s use of analogies is not wholly negative. Although she does not claim that any of the alternative interpretations she proposes should be accepted as correct, they can be seen as highly pursuit worthy, partly *in virtue* of being based on analogies.

5.5. Conclusion: Pursuit Worthiness and the Interpretative Dilemma

I started this chapter by introducing the interpretative dilemma, of which the problem of analogy in archaeological theorising is one instance. I want to conclude by proposing that my account of analogies also suggests a more general perspective on the interpretative dilemma.

As argued in Section 5.3.1, following earlier commentators, blanket scepticism about analogies is untenable. Given appropriate background knowledge, analogies can support reasonable inference about the past and there are criteria which can guide the evaluation and strengthening of analogical inferences. However, in many cases, even comparatively strong analogical inferences will only provide moderate support for a given interpretation. In archaeology, it will often be difficult to accumulate enough secure and relevant background knowledge to provide the necessary scaffolding for strong analogical inferences, especially when it comes to broader questions regarding society and culture.

This corresponds to the more general point that archaeological theorising involves more ambiguity and uncertainty than is often expected of successful natural sciences (Gero 2007). Striving for more certainty than can be expected in archaeology invites either undue conservatism with regards to the kind of topics that get explored or exaggerated confidence in the univocality with which interpretations can be asserted. This is the threat posed by the interpretative dilemma.

As I have argued for in the case of analogies, however, allowing that archaeological interpretations are uncertain, and that there are often several reasonable interpretations of the same evidence, should not be taken to imply unrestricted speculation. Analogies often provide reasons for pursuing interpretations, rather than reasons for the truth or likeliness of the interpretation. They can do so because they raise deeper questions about similarities and differences between different social contexts. By recognising that pursuit worthiness is often a relevant dimension of evaluation for interpretations, in addition to their likeliness, we avoid the charge of unrestricted speculation. Even when archaeologists face several possible but uncertain interpretative hypotheses, they can still make reasonable decisions about which of them are most pursuit worthy, i.e. which of them holds the greatest promise for learning more through exploring the questions that interest us.

The pursuit worthiness of an interpretation is not always, or even primarily based on its likeliness. In many cases, what is learnt from pursing a hypothesis will be to discover evidence against it or even just that the evidence is more ambiguous than previously thought. This may still seem like a rather pessimistic response to the interpretative dilemma. However, for many of the deeper questions which make archaeology an intellectually interesting and valuable topic, i.e. questions about the nature of human culture and society across different time and places, even this type of negative insight has an intrinsic value. Although progress on many of these questions will consist in uncovering mistakes in previously accepted interpretations and in deepening our understanding of the kind of uncertainty and ambiguity we face in learning about our past, I want to suggest that this type of progress has a genuine value, worth pursuing for its own sake. As Joan Gero (2007) argues, archaeologists should strive to “honour ambiguity” in their interpretations, rather than paper it over. Just like learning that a given analogy-based interpretation fails is interesting because it tells us about the *relationship* between present and past cultures, learning about the limits to our understanding tells us something valuable about *our* relationship to the past.

This does not mean that archaeologists should relinquish the goal of learning as much as possible about the past, or that they should actively try to introduce unnecessary ambiguity or deliberately exaggerate the uncertainty of their interpretations. On the contrary, the only way to learn about the limits of our knowledge is to seriously try to learn as much as we can. I do not claim that archaeologists never learn more about the past, only that they should not overestimate the certainty and finality of their conclusions about it. But equally, they should be not underestimated either: honouring the ambiguity and uncertainty of archaeological interpretations only has value to the extent that it *correctly* reveals to us the extent and limits of our knowledge about the past. Thus, on the vision of archaeology proposed here, its constitutive aim is still to discover what life and culture was like in the past and what we can reasonably know about it.

Chapter 6. Generating and Pursuing Diagnostic Hypotheses

With Donald E. Stanley[129]

6.1. Introduction

How should we conceive of and evaluate the kinds of reasoning that occur in the process of medical diagnosis? Saying that it is a matter of inferring the correct (or at least the likeliest) diagnosis from the evidence available to the physician is too sparse. Diagnosis is a dynamic process involving observation, diagnostic conjectures and testing, possibly leading to new or revised conjectures. Consider, for example, the following scenario. A 54-year-old man with no previous history of chronic disease suffers sudden substernal chest pain and is rushed to an emergency room. His symptoms also include tachycardia (abnormally rapid heart rate), shortness of breath and sweating. The challenge a clinician faces in cases like this is not just to evaluate the likeliness of different possible causes of these symptoms; she also has to select which hypotheses to consider actively in the first place, which to prioritise for further testing, which can be put aside for the time being and when to initiate treatment on the basis of a given hypothesis. Additionally, all of these decisions presuppose that the relevant hypotheses have been generated and introduced into the diagnostic inquiry. The clinician does not start out considering every possible cause of chest pain known to medicine; rather, she needs to decide when and how to generate new diagnostic hypotheses, as well as when to stop.

The aim of this chapter is to present a framework for analysing the kinds of reasoning which underly medical diagnosis as it occurs in a concrete, clinical situation. Specifically, our starting point is the observation that, in addition to reasoning about the

[129] This chapter is based on a paper co-authored with an experienced pathologist, Donald E. Stanley (Stanley and Nyrup, *forthcoming*).

likeliness of candidate diagnostic hypotheses, a clinician faces two distinct types of reasoning tasks: (i) deciding how, when and whether to generate new candidate hypotheses and (ii) deciding which of these should be prioritised for pursuit, i.e. for further consideration and testing. Building on the earlier discussion of Peircean abduction and pursuit worthiness, Chapters 1 and 2, we will argue that recent Peirce scholarship which construes abduction in terms of *strategic reasoning* provides a promising framework for analysing medical diagnosis.[130]

The scope of our framework is primarily normative: we want to explicate the *reasons* which underlie diagnostic reasoning in realistic clinical situations, rather than necessarily describing the psychological processes clinicians go through. The best psychological description may often be that the clinician makes a quick, intuitive judgment, perhaps based on some unconscious heuristic. By contrast, the framework presented here aims to identify the factors which make such judgements reasonable. However, it should be noticed that although our framework is in this sense normative, we do not here aim to offer any recommendations as to whether existing practices could or should be improved. While we will offer a framework for discussing such questions, we focus in this chapter on showing how it enables us to explicate diagnostic reasoning as it occurs in current practice, rather than comparing possible changes to that practice.

A unified, normative framework for understanding clinical reasoning is currently lacking in the methodological literature. On the one hand, when hypothesis generation is addressed (e.g. Kassirer, Wong and Kopelman 2010: ch. 13) it is mainly discussed from the perspective of cognitive psychology without an underlying normative framework. On the other hand, while the so-called *threshold approach* to clinical decision-making—

[130] Previous commentators (Upshur 1997; Stanley and Campos 2013, 2016; Chiffi and Zanotti 2015) have also argued for the relevance of Peircean abduction to medical diagnosis.

currently popular e.g. in the Evidence-Based Medicine literature—is normative, it does not address the question of hypothesis generation. Furthermore, as we shall argue, because of the way hypothesis generation and reasoning about pursuit are intertwined, this neglect means that threshold models (in their current form) fail to capture all considerations relevant to whether a diagnostic hypothesis should be pursued.

Our discussion proceeds as follows. We start by outlining how lessons from the discussion of Peircean abduction earlier in this thesis apply to medical diagnosis. In Section 6.3 we use this framework to analyse a concrete diagnostic scenario before, in Section 6.4, highlighting some of the limitations of the threshold approach. Finally, in Section 6.5, we will show how the strategic reasoning interpretation of abduction can help make better sense of diagnostic reasoning.

6.2. Peircean Lessons for Medical Diagnosis

The main lessons for medical diagnosis that we want to draw from previous chapters are as follows. First, that we should distinguish (1) reasoning concerned with accepting or rejecting diagnostic hypotheses—or, more broadly, how likely different diagnoses are in light of the available evidence—from (2) reasoning concerned with generating new candidate diagnoses, and (3) selecting between and prioritising the available hypotheses for further pursuit. Second, while there are important differences between (2) and (3), they are united in virtue of answering to the same normative standard, viz. pursuit worthiness. This distinguishes them from (1), which aims to identify the likeliest diagnosis. Third, even though physicians cannot consciously control all aspects of hypothesis generation, they can still evaluate choices about *when*, *whether* and *how* to generate new ideas in terms of how conducive these are to obtaining an adequate range of pursuit worthy candidate hypotheses. Fourth, as argued in Chapter 2, some aspects of pursuit worthiness can be captured in formal decision-theoretic models. However, these are not useful for capturing what can be called ‘strategic reasoning’, i.e. considerations about what the downstream effects of pursuit are for later stages of inquiry. We will now explain how these points apply to medical diagnosis, before illustrating them in a more detailed case study in Section 6.3.

6.2.1. Selecting Differential Diagnosis

A typical diagnostic process begins when a patient arrives at a hospital or clinic and reports certain symptoms or ailments. Insofar as the situation allows it, the physician will start by interviewing the patient and performing a physical examination to gather information about the patient’s state, how long they have experienced the symptoms and their broader medical history. Based on these, the physician tries to generate one or more possible explanations for the salient aspects of the case. For example, if a patient has uncontrollable hypertension (high blood pressure) the physician may conjecture that the patient has renal artery stenosis (narrowing of kidney arteries), as this would explain that symptom.

The term ‘generation’ here should be understood in a broad sense. In most cases, medical diagnosis does not involve formulating completely novel hypotheses. Rather, it will primarily be a case of recalling already known conditions and realising that they could potentially account for the salient signs and symptoms.[131] However, this is not a sharp distinction. When facing atypical or complex cases, physicians may have to combine their knowledge of possible diseases in novel ways to explain the condition of that specific patient.

[131] Stanley and Campos (2013: 306) call the former of these “creative abduction” and the latter “habitual abduction”.

While physicians will often be able to think of a large number of theoretically possible diagnoses, it is neither practically possible nor advisable to consider every single one. Physicians need to pick out a limited number of hypotheses to focus on. The set of diagnostic hypotheses actively considered at a given time is usually called the *differential diagnosis*.[132] There are good reasons why physicians need to limit themselves to a relatively narrow differential diagnosis. First, limitations of working memory preclude working on too many hypotheses at once (Sox, Higgins and Owens 2013: 9). Second, actively pursuing too many hypotheses can lead to potentially harmful over-testing (Richardson et al 1999: 1214-15). Third, there is often only time to test a limited number of hypotheses. One can never, of course, test every conceivable hypothesis, and in an emergency situation, the number of tests that can be actively pursued is even more restricted. With a patient's health or life on the line, the physician needs to be able to effectively, rapidly and efficiently determine the likeliest cause of their ailments. This requires wisely selecting a limited range of hypotheses on which to focus.

6.2.2. Generating Candidate Diagnoses

The above arguments are often applied to the *choice* of a differential diagnosis, but similar points already apply at the *generative* stage. Just as it is inadvisable to select too broad a differential diagnosis, physicians cannot—and should not—try to generate a list of every single possible explanation before selecting a differential diagnosis. Just as physicians need to make good choices about which hypotheses to include in their differential diagnoses and which of these to prioritise for testing, they must choose how to generate

[132] Sometimes 'differential diagnosis' is instead used to refer to the process or method of considering and distinguishing different diagnostic hypotheses (e.g. Sox, Higgins and Owens 2013: ch. 2). We will only use the term in this chapter to refer to a set of competing hypotheses, rather than the process of generating or selecting between these.

possible diagnoses, as well as when to *stop*. The choice of whether to generate new diagnoses and, if so, using which strategy can be evaluated on the same grounds as choices about which hypotheses to pursue.

For example, a rather ineffective strategy for generating new diagnostic hypotheses would be to flip through a medical lexicon, hoping to chance upon diseases with symptoms similar to the ones observed in the patient. Experienced physicians will (hopefully) be able to deploy better strategies for generating hypotheses. Sometimes hypothesis generation happens almost automatically: the clinician recognises a known pattern and immediately recalls the most common, and important, diagnoses. In more atypical cases, it can be necessary to employ a more structured or directed form of thinking. Other possible strategies include (i) using one of the existing artificial intelligence programs designed to assist medical diagnosis (e.g. Isabel, DXplain) and (ii) requesting that a colleague reviews the data and offers a second opinion. We will discuss these two strategies in turn.

Current so-called ‘differential diagnosis generator’ computer programmes are based on prevalence data, weighed in terms of the signs and symptoms entered into the programme by the user together with details about the age, sex and geographic region of the patient. They indicate which conditions are most the common causes of the symptoms entered into the program and red-flag potentially life-threatening diagnoses. The ranking of the diagnoses is based on the experience of the writers of the programme and the epidemiology of diseases commonly encountered in the indicated age group, sex and geographical region. Using a computer program can be helpful for reminding a clinician of rare but dangerous conditions. However, many experienced physicians consider them of limited usefulness.[133]

First, they tend to generate a fairly long list, which is not particularly helpful in an emergency situation. Trying to work through an extensive list of possibilities is not a feasible strategy, especially when the patient is unstable, and doing so may subject the patient to unnecessary and potentially harmful over-testing. Second, physicians do not know how the programme assigns weights to each of the symptoms and the prevalence of the disease. The computer programme is based on geographically common epidemiological findings in specific diseases and populations, but this population level information cannot be translated directly to the individual case. Experienced physicians will be attuned to the concrete clinical setting (how stable is the patient, what are the urgent problems), details of the case (e.g. medical history, country of origin, foreign travel, use of drugs, smoking) and the patient's response to therapy (e.g. pain relief and normalisation of heart rate, breathing). For instance, if a patient has recently travelled in sub-Saharan Africa, this should make the physician consider anaemia (potentially caused by malaria); a sedentary life style should call attention to symptoms of coronary ischemic heart disease; a family history of diabetes in mother and grandmother would make it important to consider weight gain, hypertension, and high cholesterol levels. These facts have to be interpreted, and the physician has to judge whether or not the findings are properly perceived and integrated into the diagnostic picture. So physicians will, in any case, need to draw on their experience and insight to interpret the results generated by the programme.

[133] For a recent survey of currently available programs, see Bond et al (2012). Philosophers in the 1980s debated whether computer programs could, in principle, replace all aspects of diagnostic reasoning (e.g. Schaffner 1985, Wartofsky 1986). Here we focus on how useful currently existing programs are for the task of generating hypotheses.

Perhaps currently existing programmes could be refined to allow physicians to enter these additional pieces of information; however, this would still rely on the ability of the physician to recognise *that* a certain fact about the case is an important piece of information that needs to be taken into account. Of course, neither a computer nor a human reasoner can take a given piece of evidence into account before it is recognised *as* evidence. Perspicacious observation is here an intricate part of the reasoning process itself. There are no obvious, general constraints on the kinds of information that could be potentially relevant and, as with hypothesis generation, trying to take into account every single piece of information is not feasible, nor necessarily very efficient. While we do not want to speculate on whether any future artificial intelligence systems will be able to emulate these abilities, in their current form, differential diagnosis generator computer programmes are at best an aid to (rather than a replacement for) clinical judgement and experience in hypothesis generation.

Having a colleague review the situation can also be a way to ensure that important diagnoses are not overlooked. A difference in experience, training and background knowledge may allow the colleague to think of other possibilities. This strategy may be employed by physicians who are confused by the clinical picture, or dissatisfied with their own thought process. Although this strategy for generating hypotheses cannot guarantee to be as exhaustive as a computer program, drawing on the judgement of a colleague has the advantage of being better attuned to the concrete clinical situation, and so is more likely to generate suggestions that are reasonable and useful in context.

The experience and training of a clinician, or her colleagues, play a crucial role in several respects, both in the generation of hypotheses and the selection of a differential diagnosis. First, the clinician has to make a wise choice of which strategy for generating new suggestions strikes the right balance between expediency, exhaustiveness and quality in the concrete situation. Although one can highlight general considerations of advantages and disadvantages of different strategies, as above, the choice ultimately has to rely on the judgement of the physician. Second, it is the training, experience and background knowledge of the clinician that allows her to recognise patterns and recall possible diagnoses. Finally, in choosing which types of hypotheses to consider, the clinician needs to judge which diagnoses are most likely in the concrete case and then decide how to weigh this, e.g., against the seriousness and urgency of the disease, its testability and its treatability.

6.2.3. The Threshold Approach and Strategic Reasoning

On what grounds, then, should decisions about generation and pursuit be made? The most popular theoretical approach to the problem of choosing whether to test a given hypothesis in the medical literature is the so-called *threshold approach* (Pauker and Kassirer 1980; Djulbegovic et al 2015). This approach is based on decision-theoretic models, similar in structure to those developed in Chapter 2. In the threshold approach, the models compare a number of possible actions a physician may take vis-à-vis a given diagnostic hypothesis *H*, typically: (i) applying treatment on the assumption that the hypothesis *H* is true; (ii) applying a test for *H*, and then administering the treatment if and only if the test is positive; (iii) stop working on *H*, i.e. neither test nor treat. This model can be represented by the decision tree in Fig. 6.1. Given quantitative estimates of (a) the reliability of the test, i.e. true positive and true negative rates, (b) the likelihood of the salient consequences of treating and testing and (c) the utility of these consequences, one can derive thresholds for how probable the hypothesis needs to be in order for it to be most rational to test, treat or abandon the hypothesis.

Threshold models highlight a number of factors that should be weighed against each other in clinical decision-making, including: How reliable are the available tests? How safe/harmful are the tests? How dangerous would the disease be, if missed? How effective is the available treatment? How safe/harmful is the treatment in itself? Briefly put, on this approach, physicians have to consider whether their confidence in *H* is high enough for the potential benefits of treating the disease (if *H* is true) to outweigh the potential harms of treating or testing unnecessarily (if *H* is false).

While these factors are indeed important, they leave out what we will call *strategic considerations*, i.e. considerations concerned with possible consequences of testing a hypothesis which go beyond the direct consequences for the health of the patient. As explained in Chapter 2 (Sections 2.4.2 and 2.8.4), these are related to Peirce's observation that we also need to take into account what we might learn from pursuing the hypothesis even if it turns out to be false (CP7.220). Testing a hypothesis can have important downstream effects for later stages of inquiry, in addition to merely confirming or disconfirming the tested hypothesis.[134] For instance, an imaging study which fails to detect renal artery stenosis may also show that the adjacent adrenal gland is enlarged, a finding which would instead suggest pheochromocytoma (a tumour of the adrenal gland) as a possible cause of hypertension. At other times, it can be worth trying to rule out a potential diagnosis simply to make the diagnostic space more manageable, i.e. to pre-emptively prune off possibilities that might otherwise become relevant later on. If testing can be done reliably and without risk of harm, it can be worth trying to rule out even fairly unlikely hypotheses early on.

[134] A similar point also applies to treatment: even if a given treatment fails to alleviate a patient's symptoms, it may still provide valuable clues for further investigations. The line between treatment and testing is not always a sharp one.

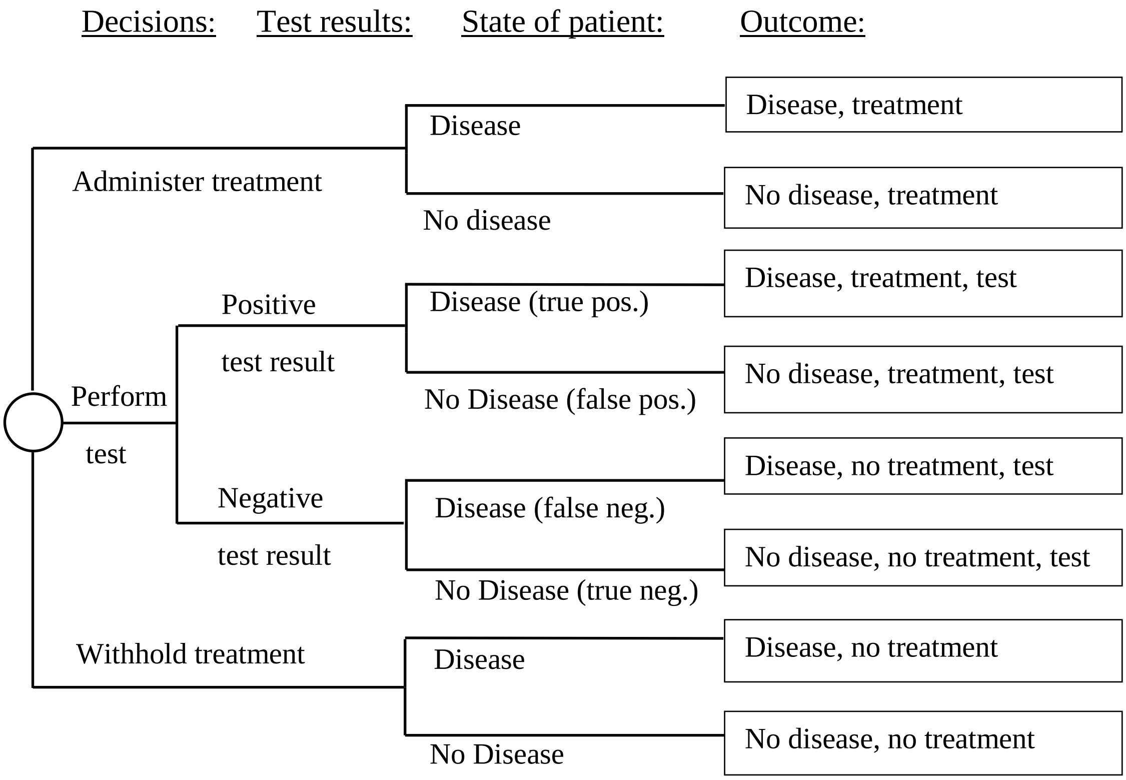

Figure 6.1: Decision tree for a threshold model. After Pauker and Kassirer (1980: 1111, fig. 1).

Strategic considerations involve reasoning about how pursuing a specific hypothesis can influence later stages of inquiry, including future generation of hypotheses. It is this dynamic and intertwining relationship between hypothesis generation and selection for pursuit which threshold models, in their current form, fail to capture. We will develop this argument further in Section 6.4. First, we will illustrate the points made in this section through a detailed clinical case study.

6.3. Clinical Case: Chest pain

The following case study has been developed on the basis of the clinical experience of one of the authors (Stanley). While we do not make any claims as to how statistically representative this scenario is, we regard it as sufficiently typical to illustrate the framework developed here. In the following, the description of clinical details of the case, given in *italics*, is distinguished from our commentary.

Scene: *At home in the Northeastern U.S.A., at 08:00, a 54-year-old man is walking down the stairs to breakfast. He suffers sudden substernal chest pain radiating to his left shoulder and back. No previous history of chest pain. No previous chronic disease. His spouse immediately calls the local emergency number.*

Emergency medical technologists (EMT) arrive in a quarter of an hour. Based on hospital protocol, an intravenous line is inserted in patient's right arm; he is administered morphine sulphate 10 mg, aspirin, beta blocker, supplemental oxygen by mask; an electrocardiogram (ECG) tracing is radioed to the local emergency room while he is in the ambulance. Sublingual nitroglycerin is given with minimal relief of pain. He receives nasal 100% oxygen but is not intubated. His respirations are more than 30 per minute and shallow. His skin is cool and clammy. Blood pressure: 110/78. The substernal pain is slightly relieved with medications and rest. EMT calls emergency triage nurse at nearest community hospital regarding middle-aged white male complaining of severe chest pain. He is breathing rapidly and perspiring.

The patient arrives in the emergency room and is seen immediately by a triage nurse. He complains of severe chest pain when descending the stairs to the kitchen; the pain persists. Nurse inquires if he has had a previous a history of chest pain. "No," the patient answers. She assesses his vital signs: heart tracing on electrocardiogram, respiratory rate, and temperature. She searches for any previous medical record in the computerised system to share with a physician. She is following protocol for chest pain and patient estimates pain at 7/10. He receives an additional 10 mg morphine sulphate *that eases his pain. On auscultation with stethoscope, a very soft holosystolic murmur*[135] *is heard over the precordium. Respirations are laboured. Tachycardia (abnormally rapid heart rate) is evident, 110 bpm.*

Initial diagnosis: *Physician arrives in the acute side of the emergency room. She decides that the likeliest diagnosis is acute coronary syndrome (ACS), i.e. a sudden restriction of blood flow from the coronary arteries into the heart, leading to cardiac ischemia (oxygen deprivation to the heart) and subsequent myocardial injury (death of heart cells). Based on the history and physical examination she orders two laboratory tests: ECG and serum cardiac enzymes.*

Commentary: How does the clinician reach her initial diagnosis? The first step is to generate one or more diagnoses capable of explaining the most salient signs and symptoms. She knows, and immediately recalls, that chest pain, shortness of breath and sweating are common symptoms of ACS, so she concludes that there is reason to suspect this diagnosis. She also knows (from prevalence studies and clinical experience) that ACS is the most common cause of chest pain in men in their fifties in this part of the country.

At this stage, rather than systematically generating a wider list of potential diagnoses, she immediately orders two tests. Her reasons can be reconstructed as follows: (i) ACS is the most common cause of the chief symptom (chest pain); (ii) ACS can cause severe damage and is life-threatening if left untreated; (iii) the ordered tests are a rapid and effective way of confirming the hypothesis: if the ECG shows patterns characteristic of myocardial damage and the blood test shows elevated levels of the enzymes an

[135] 'Holosystolic' means that murmuring sound extends over the entire contraction, ejection, and relaxation portion of the heart cycle.

ischemic heart muscle would release, this would be very strong evidence in favour of the hypothesis. The decision *not to generate* further hypotheses before taking action is based on the same kind of reasons as would justify *selecting already generated* hypotheses for further consideration.

<u>Negative results:</u> *The laboratory and ECG results are negative: the cardiac enzymes test did not show elevated levels of the relevant enzymes (c-troponin or mb-creatine kinase). The electrocardiogram shows a rapid heart rate (120 bpm) but none of the characteristics of heart disease (no elevation in the S-T segment, neither T wave inversion nor new Q wave occurrence). Both results tell against cardiac ischemia and thus against the diagnosis of ACS.*

Despite this lack of evidence, the clinician does not dismiss her initial diagnosis. Although she faces conflicting evidence, she maintains her clinical suspicion of myocardial injury caused by acute coronary syndrome. She orders the tests repeated in two hours and has the patient monitored closely for any signs of worsening condition. Because of persistent pain he receives additional morphine. Meanwhile the clinician considers alternative diagnoses which could mimic the symptoms of ACS.

<u>Commentary:</u> Why does the clinician maintain her initial suspicion? The case does not fit the textbook picture of cardiac injury caused by ischemia. However, she knows both from her own experience and epidemiological studies that atypical disease presentations are not uncommon: the pain may be referred to the jaw instead of chest, the T-waves in ECG may not develop early, etc. While the negative results are sufficient to make her hold off treatment based on the initial diagnosis, she wants to avoid prematurely turning away

from the commonest disease, especially given its potential to kill or seriously damage the patient's health.

Due to this uncertainty, she initiates two lines of action. Firstly, although she does not have a single and simple hypothesis about what could cause the atypical presentation, she decides to repeat the tests. The fact that ACS is the most common cause of chest pain and the life-threatening character of the diagnosis makes it reasonable to repeat the tests. Secondly, her lowered confidence in the ACS diagnosis is reflected in the fact that she decides to start systematically generating alternative diagnoses in order to select a differential diagnosis.

Generation and selection of differential diagnosis: *The clinician thinks of alternative diagnoses. She asks herself a number of questions to guide her thinking: "what else could explain acute chest pain?", "which are the most common causes?", "what would be the most serious disease to miss?" "what could I test effectively and quickly?", "which conditions are effectively treatable?" She can think of a wide range of possibilities. For instance, she briefly considers Chagas disease, but decides this is too rare in the U.S.A. to merit immediate consideration. Another possibility would be referred pain e.g. from acute pancreatitis or another abdominal organ. In certain clinical circumstances this can be confirmed by re-examination after delaying an hour or more. In the present case, there is no time to delay, as the pain seems to be continuous even with morphine analgesia. She ultimately decides to focus on four alternative hypotheses in addition to acute coronary syndrome:*

1. *Acute pulmonary embolism (a blockage of a lung artery, e.g. by a blood clot).*
2. *Acute aortic syndromes, (different kinds of damage to the aorta, the main artery leading blood out of the heart and into the body).*
3. *Pericarditis (inflammation to the pericardium, the sac surrounding the heart).*
4. *Gastrointestinal reflux disease.*

Commentary: At this stage, the clinician starts systematically generating possible candidate diagnoses. She probably has already spontaneously thought of alternative possibilities, but she now tries to explicitly elicit her memory and clinical experience by asking herself a series of questions. The case presents a puzzling picture, so the clinician decides initially to cast her net widely. She first asks herself which other diagnosis could explain the symptoms. Her concern is to make sure she has not forgotten to think of a potentially dangerous condition. However, she needs to quickly limit herself to a short list of actionable diagnoses. Trying to actively consider and rule out all candidate causes of chest pain is not practically possible, especially since the patient is unstable. Thus, the physician asks herself further questions to limit and focus her search, prioritising conditions that (i) are common for this type of patient, (ii) would be dangerous if left untreated and (iii) allows of effective testing and treatment. She here tries to direct her attention towards diagnoses with characteristics which would give them a high level of pursuit worthiness.

Prioritise hypothesis for testing: *After two hours, cardiac enzymes are borderline elevated, with c-Troponin at 98.5% of the normal range. The clinician ponders if, perhaps, the origin of the troponin elevation is from the epicardium or the pleura, rather than the heart muscle, reassessing the hypotheses of enzyme origin and cause of chest* *pain. She requests a cardiologist consultation. Meanwhile, she is at the bedside. She next considers pulmonary embolism and orders a chest-computed tomogram (CT-scan) with contrast media to search for the embolism (the blockage).*

Commentary: The clinician now decides to request a second opinion from a specialist colleague. Meanwhile she prioritises the pulmonary embolism hypothesis for testing. There are several good reasons for this. First, she currently considers pulmonary embolism to be among the most likely diagnoses. Second, pulmonary embolism, as well as acute aortic syndromes, are emergency conditions and would require immediate treatment. Pericarditis is also a very serious condition but less urgent, whereas gastrointestinal reflux disease is not an immediate threat. Third, a CT-scan is a highly reliable way to detect embolism. Fourth, the chest CT-scan might also show a widened mediastinum (the area containing the heart and the pericardium), a possible sign of pericarditis. It would also show the thoracic aorta (the part of aorta situated in the chest-region), a possible site of any aortic syndrome. This last point is an example of how a test can have other epistemic consequences in addition to testing hypotheses directly, in this case by potentially providing clues for future hypothesis generation. In sum, the CT-scan would be a reliable test of one of the most likely and serious conditions, while potentially also providing information relevant for the two other serious conditions currently considered.

Further puzzling results: *The results of the CT-scan, adjusted to an early phase of contrast injection, are reported as negative for pulmonary embolism, but the ascending aorta was reported to be prominent, measuring 4.3 cm in diameter (normal ascending aorta is 3.63* *to 3.91 cm*). *The patient still complains of chest pain but feels relieved by increased dose of morphine, and breathing is improved somewhat by continuous 100% oxygen therapy.*

Consulting cardiologist arrives. He reviews the history and testing and is still convinced the patient has cardiac ischemia. Given the negative results of the CT-scan he judges that the picture is atypical but consistent with ischemia, probably caused by coronary artery disease. He also considers the other available results. Although the CT-scan has ruled out pulmonary embolism, the prominent aortic valve shadow is worrisome. He requests a transthoracic echocardiogram (TTE).

Commentary: Since the cardiologist judges pulmonary embolism to be ruled out by the CT-scan, he still considers cardiac ischemia most likely. However, the shadow on the aortic root is puzzling and he decides this merits further investigation.

Conclusion of scenario: *The TTE shows a widened mediastinum and that the aortic root is dilated to 4.5 cm. The* patient's *condition is unchanged. Because of the degree of clinical pain, the cardiologist and emergency room clinician decide that immediate coronary artery intervention (stent or bypass) is necessary. They consult with the nearest cardiac surgical unit for immediate transfer and transport helicopter arrives. The cardiologist accompanies the patient to the surgical unit.*

Upon arrival, the cardiologist is still worried about the diagnosis. He reviews the inflight recordings and TTE together with the other available clues. He returns to the patient, listens for the holosystolic murmur reported at initial examination, notes the non-stress induced pain and the dilated aortic root. He tries to think of a diagnosis which could explain these clues, thinking through different possible aortic conditions, and *realises that a dissecting thoracic aortic aneurysm*[136] *could explain all of these symptoms: the dilated root is part of the aneurysm, the dissection would cause the pain and could produce the continuous murmur. The patient is taken directly to surgery where an ascending (type A) aortic dissection is repaired. He was discharged home after one week.*

Commentary: Given the state of the patient, the physicians are forced to act even though the evidence remains puzzling. The cardiologist does not consider the diagnosis of cardiac ischemia particularly likely due to the lack of expected observations (minimal pain relief from morphine, non-diagnostic cardiac enzymes and ECG). But he currently lacks a plausible alternative.

He chooses a strategy for generating an alternative diagnosis, deciding to review the previously reported clues, including ones that initially were not considered salient (the soft murmur), to guide his search for alternative diagnoses. Like the emergency room clinician, given the negative result for embolism, he considers the most serious remaining possibility an acute aortic syndrome, which is also suggested by the dilated aortic root. Relying on his background knowledge, he considers possible aortic syndromes and quickly thinks of a possibility—a dissecting aneurysm—capable of explaining the symptoms. Once he has in mind this newly generated hypothesis he immediately recognises that it would be able to explain all of the otherwise puzzling evidence. On this basis, he judges it more likely than ACS and decides to adopt it as a basis for the surgical intervention. While one cannot guarantee such judgements to always be reliable, in this case it was correct.

[136] An aneurysm is an abnormal widening of a blood vessel. This can cause weakness in the wall of the vessel. A dissection is a rupture of the blood vessel where blood flows into layers of the wall of the vessel, forcing them apart.

6.4. Limitations of Current Accounts of Medical Diagnosis

Throughout the preceding case study, we discussed a number of decisions made by the physicians regarding the generation and pursuit of diagnostic hypotheses. In the following two sections, we will, first, highlight some limitations of the two primary frameworks used for discussing diagnostic reasoning in the medical literature: the normative, probabilistic framework associated with the threshold approach and descriptive frameworks based on cognitive psychology. We will then discuss our constructive proposal: to conceptualise the process of diagnosis in terms of strategic reasoning.

The probabilistic framework is the most popular normative framework employed in methodological discussions of diagnosis in the medical literature, especially among proponents of evidence-based medicine. This approach is typically summarised as follows (e.g. Richardson and Wilson 2015). First, physicians identify a plausible differential diagnosis for the patient and assigns an initial prior (or “pretest”) probability to each of the hypotheses in the differential diagnosis. Second, they compare the initial probabilities of the hypotheses to the probability thresholds, as determined by the decision-theoretic models of the threshold approach, in order to decide whether to test or treat for the disease. Third, as test-results become available, physicians should use Bayes’ theorem, together with information about test reliability, to update the probabilities.

While this probabilistic framework can highlight important lessons for clinical reasoning,[137] it does not provide a general framework for explicating this reasoning; the probabilistic approach presents an idealised, simplified picture of clinical decision-

[137] For instance, threshold models highlight the importance of weighing initial probability against the potential benefits and harms of testing or treating. Similarly, Bayes’ theorem can highlight important lessons about probabilistic reasoning. Thus, they may provide useful analytic frameworks for teaching clinical reasoning (Sox, Higgins and Owens 2013). We are less optimistic about proposals to reform clinical practice to conform more closely to probabilistic models (Richardson 2007); see Marewski and Gigerenzer’s (2012) critique of information-greedy procedures in clinical decision-making.

making which leaves out many important aspects of the process of diagnosis. In the case study, factors that eventually led to successful diagnoses included: (i) decisions about when to generate more diagnoses for consideration, both initially by the emergency room clinician and later by the cardiologist; (ii) choosing effective and efficient strategies for generating relevant hypotheses; (iii) recognising whether the generated hypotheses can explain the salient symptoms; (iv) recognising the importance of subtle clues, such as the dilated aortic root or the holosystolic murmur, that may initially appear puzzling or unimportant, as well as knowing which features (most of them unmentioned in the description of the case) to ignore; (v) strategic choices about pursuit, especially the choice of a test (the CT-scan) which could reveal important information for further inquiry even if it produced a negative result for the hypothesis tested.

This last point is especially important. The decision-theoretic models of the threshold approach are limited to considering the direct benefits and harms of testing or treating. They do not take into account the strategic considerations, i.e. the kinds of downstream consequences highlighted in Section 6.2.3. However, these considerations proved crucial to the successful resolution of the case: it was the choice to do a CT-scan, and subsequently an echocardiogram, which produced the crucial clue that eventually led the cardiologist on the right track. These kinds of considerations are difficult to represent directly in the probabilistic framework, since it is difficult to assign meaningful probabilities or utilities to unknown unknowns. What is the probability that a given test will produce a valuable clue for a diagnosis we have not thought of yet? What is the utility of treating this as-yet-unknown disease? Successful diagnosis depends, at least in part, on recognising and considering these possibilities. Of course, one can always add a term into the decision-theoretic calculus to represent the weight given to these considerations relative to the direct consequences of testing/treating. But this would not shed any further light on the reasoning that leads physicians to give them that weight.

Finally, probabilistic models start from the assumption that one has already formulated a diagnostic hypothesis. In their current form, they only address the question of whether the hypotheses generated satisfy the goal of being pursuit worthy. To the extent that it succeeds in the latter, it at best represents the *aim* of generative reasoning, rather than explicating this reasoning in itself.

When hypothesis generation is discussed in the medical literature, it is done primarily within the framework of cognitive psychology. For instance, while Kassirer, Wong and Kopelman (2010: ch. 13) discuss hypothesis generation in several case studies, their focus is on which structures of memory allow (or prevent) physicians from recalling the correct diagnosis—e.g. perhaps the physician's memory is structured in condition-action pairs, one of which states, e.g. that "**If** an adult has a high serum cholesterol value, **then** consider the possibility of hypothyroidism" (*ibid.*: 75, original boldface)? While much can no doubt be learned from a better understanding of the relevant psychology, these analyses currently lack a guiding normative framework. The questions they ask are about what structures of memory allow us to recall the correct diagnosis. Ultimately, this is of course what successful diagnosis requires, but "try to recall the correct diagnosis" does not exhaust the relevant considerations in generative reasoning. In cases involving non-textbook, or otherwise puzzling presentations (as in the case study), starting by generating every possible diagnosis is not feasible. As argued, the relevant question is rather: what strategies for hypothesis generation allows physicians to generate a manageable set of hypotheses that are important to consider at the given stage of inquiry?

6.5. Diagnosis as Strategic Reasoning

The framework we present in this section does not aim to rival the probabilistic approach in formal rigour. Rather, we want to outline a more flexible, general framework for thinking about diagnostic reasoning, based on the idea that choices about how to generate hypotheses and select them for pursuit can be usefully thought of as instances of strategic reasoning. We here draw on Hintikka's (1998) suggestion that abduction can be understood as in terms of strategic reasoning. Hintikka's proposal is based on an analogy between game-theoretic reasoning and scientific inquiry. Knowing how to play a game such as chess involves at least two kinds of knowledge. The *definitory rules* tells us what kinds of moves are allowed and what the consequences are of those moves. The *strategic principles*, by contrast, tell us what would be a wise or an unwise move in a given situation, i.e. whether the move is likely to help us achieve the goals of the game.

Applying this distinction to medical diagnosis, the definitory rules correspond to the physician's knowledge of how of a given diagnostic hypothesis should be evaluated based on the available evidence, what the available tests are, how the hypothesis should be re-evaluated in light of different possible test results and what the potential consequences are for the health of the patient given different outcomes. Based on this, the clinician has to adopt an overall *diagnostic strategy*. By a diagnostic strategy we mean a strategy for how to generate hypotheses, select a differential diagnosis and prioritise hypotheses for testing. As Hintikka emphasises (513), one usually has to evaluate entire strategies, rather than individual steps. This is because it is often only possible to evaluate the importance of potential consequences—e.g. providing clues for hypothesis generation—within the context of a broader strategy.

We thus propose to see diagnostic reasoning as two-tiered. Individual moves (ordering a given test, choosing to stop generating new hypotheses, etc.) are justified in terms of whether they contribute to an overall strategy. The crucial choice then concerns which strategy to pursue. Diagnostic strategies can be thought of as analogous to the strategies a seasoned chess player might employ.[138] Choosing a chess strategy depends, in part, on what kind of opponent you think you are up against, together with knowledge of the definitory rules. Similarly, the choice of a diagnostic strategy will be informed by what kinds of diseases the physician thinks are most likely to be causing the salient symptoms as well as what she judges the to be the potential consequences for the health and well-being of the patient, what can be reliably tested and so on. In some very simple cases, it may be possible to represent this in an explicit decision-theoretic model. In this sense, the threshold approach is not incompatible with the broader framework we propose here. However, in many cases and for the reasons given above, the most adequate account of a physician's reasoning regarding the best strategy in the given clinical context will not be captured by any generally applicable, formal model. This does not mean that we cannot say anything about the kinds of considerations involved in this type of reasoning.

In the case study discussed, we can identify three crucial choices of strategy. The first is the initial choice to pursue the diagnosis of ACS before systematically generating new hypotheses. The emergency room clinician is trying to achieve a quick resolution, thus sparing the patient from the potential harm of leaving the condition untreated, as well as from unnecessary testing. The choice of this strategy is in part justified by beliefs about what diagnosis is most likely. In the chess analogy, seasoned players may try to push for a quick checkmate because they think their opponent is likely not to recognise a certain trap. In our case study, the clinician immediately recalls the most common cause of the presenting symptoms (ACS) and knows of tests which, if positive, would quickly and

[138] Kassirer, Wong and Kopelman (2013: 46) also mention analogies between expertise in chess and in medical diagnosis.

conclusively verify the hypothesis. This choice of strategy in turn justified the choice not to systematically generate further hypotheses. *Given* that she was pursuing the strategy of achieving a quick resolution, it made sense to stop generating new hypotheses once she had identified what was—given the available evidence—the likeliest cause, and had realised that this could be quickly and reliably tested. Spending more time generating hypotheses in this context would have been unnecessary.

Unfortunately, many opponents—whether chess players or diseases—will not be defeated by such a direct strategy. After the initial tests fail to confirm the ACS diagnosis, putting the clinician in a more uncertain situation, she adopts a new strategy. Since she no longer has a clear view of what the likeliest diagnosis is, her priority shifts to ruling out the most serious threats, hoping in the process to discover—or better: create—a "strategic opening", that is, a clue which could lead to the correct diagnosis. Adopting this strategy, she systematically attempts to recall the most dangerous alternative possible causes of chest pain, while considering future possible moves. She focuses on hypotheses that can be reliably tested, choosing a test which is both highly reliable for ruling out the target hypothesis (pulmonary embolism) and which might enable future moves, in this case by producing information relevant to generating other possible diagnoses. Her decision to request a cardiologist consultation at this stage also makes sense in light of this strategy, since his expertise would (i) complement her ability to think of relevant hypotheses and (ii) enable him recognise the relevance of any emerging clues.

Finally, at the concluding stages of the case, the cardiologist adopts a strategy for generating hypotheses that is focused on the salient clue—a dilated aortic root—brought out by the CT-scan and the echocardiogram, as well as the puzzling holosystolic murmur. The cardiologist is not satisfied with the ACS hypothesis but lacks a plausible alternative. However, due to the worsening state of the patient, there is no time for further testing.

Whether to maintain the hypothesis of ACS or to adopt the newly generated hypothesis has to depend on his judgement of which hypothesis best ‘fits’ the clinical picture. He therefore chooses a strategy of thinking quickly through a range of hypotheses, and counts on his experience to allow him to recognise the correct hypothesis when he ‘sees’ it. Given his specialisation as a cardiologist, and the constraints of the situation, thinking through the possible aortic syndromes with a focus on explaining the dilated root and the murmur was a reasonable—and as it turned out successful—strategy.

6.6. Conclusion

In this chapter, we have outlined a general framework for analysing diagnostic reasoning. We have distinguished between reasoning concerned with generating, pursuing and accepting/rejecting diagnostic hypotheses, illustrating these throughout the case study. Finally, we argued that currently existing frameworks for conceptualising diagnostic reasoning do not present a unified, normative framework, and proposed that diagnosis can be fruitfully thought of in terms of strategic reasoning.

As illustrated in Section 6.5, the framework of strategic reasoning allows us to naturally describe the reasons underlying the diagnostic process in our case study. We do not intend this to support any strong prescriptive conclusions: our analysis should not be taken to suggest that the strategies pursued by the physicians in our case study should be used as a model for diagnostic reasoning in other clinical contexts. We have explicated reasons which in the concrete situation made the strategies adopted by the physicians reasonable, but do not claim that these represent the best possible strategies. However, we believe our framework can contribute to a better normative understanding of diagnostic reasoning as it occurs in existing clinical practice, providing a basis for future discussions of improvements of clinical practice and the teaching of diagnostic reasoning.

Conclusion

In this thesis, I have argued for a number of conclusions regarding each of the types of scientific reasoning discussed in Chapters 3-6. I will not repeat these here. Instead, I want to highlight some general questions and topics for further research which have been raised in the process.

The first of these concerns the status of the decision-theoretic models developed in Chapter 2. On the one hand, some of my arguments (especially in Chapters 3 and 4) are based on these. On the other hand, as illustrated in Chapter 6, I do not think they can account for all reasoning about generation or pursuit. As explained in Section 2.9, I do not take these models to provide a general account of pursuit worthiness, but instead regard them as a tool for thinking about some of the factors and trade-offs related to *epistemic* pursuit worthiness. As with many idealised models in science, decisions about whether they can be applied in a given context has to be made on a case-by-case basis. Further work would be needed to determine, for instance, the extent to which decision-theoretic models (possibly integrated into an AI system) can be used to improve reasoning about generation and pursuit in medical practice, or elsewhere in science.

While I have argued for a consequentialist approach to (epistemic) pursuit worthiness, the extent to which this translates into a general account of the different types of scientific reasoning discussed in later chapters depends on the specific case. In the case of the Peircean view, I have argued (Section 3.4) on the basis of a few, relatively uncontroversial assumptions that the explanatoriness of a hypothesis does in general provide reasons for pursuing it. I have also criticised arguments for explanationism, i.e. that explanatoriness provides reasons for acceptance (Sections 3.5 and 3.6). By contrast, I do not have a single account of the role of analogies in science or even of how they justify pursuit. On my view, analogies can, and do, play a number of roles and can justify pursuit for different reasons. While the billiard ball analogy for gases arguably did so by promising to unify thermodynamics and statistical mechanics (Section 4.4), the liquid drop model, by contrast, allowed physicists such as Gamow and Bohr to transfer a modelling framework and its associated understanding-with to a new domain (Section 4.5). In the case of archaeology, I have argued that analogies raise interesting questions about the similarities and differences between human culture at different times and places (Section 5.3.3). There is obviously much scope for exploring the role of analogies in further disciplines and contexts.

A particularly important further question, to my mind, concerns the ‘value’ or utility of accepting or rejecting a hypothesis or, more generally, of different epistemic states. My arguments in this thesis have committed me to several general claims about this type of epistemic value: I have argued (Section 3.4) that it consists partly in having explanations of the phenomena that interest us and that we can compare hypotheses in terms of how much explanation and understanding they give us. Furthermore, in order to fully cash in the epistemic value of an explanation, it is not enough to know that the explanatory theory or model is true; we must also have understanding-with of the theoretical or modelling framework within which the explanation is framed (Section 4.5). Finally, I have claimed (Section 5.3.3) that part of the value of archaeological inquiry is that it teaches us about how human culture varies across time and space, about the *limitations* in our knowledge about past societies (Section 5.5). Apart from brief surveys of some salient options, e.g. in Sections 2.8.2 and 3.4, I have not considered *what* this kind of epistemic value consists in in any depth. An important future extension of the research begun in this thesis will be to examine the nature of epistemic value, how it relates to non-epistemic (e.g. political or ethical) values and whether different accounts of these matters have any implications for discussions of pursuit worthiness.

As mentioned in the Introduction, the purpose of this thesis has not been to present a single, overreaching argument. Instead, my aim has been, first, to highlight the importance of distinguishing between reasoning concerned with the acceptance, generation and pursuit of scientific hypotheses, and to present a framework for approaching the latter two from a normative perspective. Second, I have tried to demonstrate the fruitfulness of this framework within debates regarding different kinds of scientific reasoning. In addition to the conclusions argued within each chapter, I have, hopefully, made a compelling case for the importance of paying attention to the generation and pursuit of hypotheses in debates about scientific reasoning.

Bibliography

Achinstein, Peter. 1993. "How to Defend a Theory Without Testing It: Niels Bohr and the "Logic of Pursuit"". *Midwest Studies in Philosophy* 18: 90-120.

———. 1992. "Inference to the Best Explanation: Or, Who Won the Mill-Whewell Debate?". *Studies in the History and Philosophy of Science* 23: 349-364.

———. 1990. "The Only Game in Town". *Philosophical Studies* 58: 179-201.

Airy, George B. 1846. "Account of Some Circumstances Historically Connected with the Discovery of the Planet Exterior to Uranus". *Monthly Notices of the Royal Astronomical Society* 7: 121-144.

Armstrong, David. 1983. *What is a Law of Nature?* Cambridge: Cambridge University Press.

Allison, Penelope. 2009. Understanding Pompeian Household Practices Through Their Material Culture. *FACTA* 3: 11-33.

———. 2001. Using the Material and Written Sources: Turn of the Millennium Approaches to Roman Domestic Space. *American Journal of Archaeology* 105: 181-208.

———. 1999. Labels for Ladels: Interpreting the Material Culture of Roman Households. In Allison, P. (ed.): *The Archaeology of Household Activities*, London: Routledge, pp. 57-77.

Andersen, Hanne. 1997. "Categorization, Anomalies and the Discovery of Nuclear Fission", *Studies in the History and Philosophy of Modern Physics* 27: 463-492.

Ascher, Robert. 1961. "Analogy in Archaeological Interpretation". *Southwestern Journal of Archaeology* 17: 317-325.

Bailey, D. M. 2012. "Classical Architecture". In C. Riggs (ed.) *The Oxford Handbook of Roman Egypt*, Oxford: Oxford University Press, pp. 189–204.

Bartha, Paul. 2013. "Analogy and Analogical Reasoning". In Zalta, E. (ed.): *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), http://plato.stanford.edu/archives/fall2013/entries/reasoning-analogy/.

———. 2010. *By Parallel Reasoning: The Construction and Evaluation of Analogical Arguments*, New York: Oxford University Press.

Baum, Richard & William Sheehan. 1997/2013. *In Search of Planet Vulcan: The Ghost in Newton's Clockwork Universe*. New York: Plenum Press. Reprinted 2013 by Springer.

Bell, James. 1994. *Reconstructing Prehistory*. Philadelphia: Temple University Press.

Binford, Lewis. 1972. *An Archaeological Perspective*, New York: Seminar Press.

———. 1967. "Smudge Pits and Hide Smoking: The Use of Analogy in Archaeological Reasoning". *American Antiquity* 32: 1-12.

Bond, William F., Linda M. Schwartz, Kevin R. Weaver, Donald Levick, Michael Giuliano, and Mark L. Graber. 2012. "Differential Diagnosis Generators: An Evaluation of Currently Available Computer Programs". *Journal of General Internal Medicine* 27: 213-9.

BonJour, Laurence. 1985. *The Structure of Empirical Knowledge*. Cambridge, MA.: Harvard University Press.

Boozer, Anna. 2015. "The Tyranny of Typologies: Evidential Reasoning in Romano-Egyptian Archaeology". In Chapman, R. and A. Wylie (eds.): *Material Evidence: Learning from Archaeological Practice*, London: Routledge, pp. 92-109.

Boyd, Richard. 1983. “On the Current Status of the Issue of Scientific Realism”. *Erkenntniss* 19: 45-90.

Buchdahl, Gerd. 1970. “History of Science and Criteria of Choice”. In Stuewer, R. (ed.): *Historical and Philosophical Perspectives of Science*, Minnesota Studies in the Philosophy of Science, Vol. 5, Minneapolis: University of Minnesota Press, pp. 204-230.

Burks, Arthur. 1946. “Peirce’s Theory of Abduction”. *Philosophy of Science* 13: 301-306.

Busch, Jacob. 2008. “No New Miracles, Same Old Tricks”. *Theoria* 74: 102-114.

Campbell, Norman. 1920. *Physics: The Elements*. Cambridge: Cambridge University Press.

Campbell, W.W. 1909. “The Closing of a Famous Astronomical Problem”. *Publications of the Astronomical Society of the Pacific* 21: 103-115.

Campos, Daniel. 2011. “On the Distinction Between Peirce’s Abduction and Lipton’s Inference to the Best Explanation”. *Synthese* 180: 419-442.

Carnap, Rudolf. 1928. *Der Logische Aufbau der Welt*. Leipzig: Felix Meiner Verlag.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie*. Oxford: Clarendon Press.

Chang, Hasok. 2012. *Is Water H_2O? Evidence, Pluralism and Realism*. New York: Springer.

———. 2003. “Preservative Realism and Its Discontents: Revisiting Caloric”. *Philosophy of Science* 30: 902-912.

Chiffi, Daniele and Renzo Zanotti. 2015. “Medical and Nursing Diagnosis: A Critical Comparison. *Journal of Evaluation in Clinical Practice* 21: 1-6.

Childe, V. Gordon. 1956. *Piecing Together the Past*. London: Routledge and Kegan Paul.

Clark, J. Grahame. 1953. “Archaeological Theories and Interpretation: Old World”. In Kroeber, A. L. (ed.): *Anthropology Today*, Chicago: University of Chicago Press, pp. 343-383.

———. 1951. “Folk-Culture and the Study of European History”. In Grimes, W. F. (ed.): *Aspects of Archaeology*, pp. 49-65.

Cramp, Lucy, Richard Evershed and Hella Eckardt. 2011. “What Was a Mortarium Used for? Organic Residues and Cultural Change in Iron Age and Roman Britain”. *Antiquity* 85: 1339-1352.

Cohen, Robert, Paul Feyerabend and Marx Wartofsky (eds.). 1976. *Essays in Memory of Imre Lakatos*. Boston Studies in the Philosophy of Science, Vol. 39. Dordrecht: Reidel.

Cool, H. E. M. 2004. “Some Notes on Spoons and Mortaria”. In Croxford, B., H. Eckardt, J. Meade & H. Weekes (eds.): *TRAC 2003. Proceedings of the Thirteenth Annual Theoretical Roman Archaeology Conference, Leicester 2003*, Oxford: Oxbow, pp. 28–36.

Curd, Martin. 1980. “The Logic of Discovery: An Analysis of Three Approaches”. In Nickles (1980a), pp. 201-220.

Curren, Cailup. 1977. “Potential Interpretations of “Stone Gorget” Function. *American Antiquity* 42: 97-101.

Dawes, Gregory. 2013. “Belief Is Not the Issue: A Defence of Inference to the Best Explanation”. *Ratio* 26: 62-78.

Daremburg, C. V. and E. Saglio. 1881-1904. *Dictionnaire des Antiquités grecques et romaines*. 3rd ed. 3 Vols. Paris: Libraire Hachette.

Day, Timothy and Harold Kincaid. 1994. “Putting Inference to the Best Explanation in Its Place”. *Synthese* 98: 271-295.

DeBoer, W.R. and D.W. Lathrap. 1979. “The making and breaking of Shipibo-Conibo ceramics”. In C. Kramer (ed.): *Ethnoarchaeology: Implications of Ethnography for Archaeology* New York: Columbia University Press, pp. 102–38.

Denny, Don. 2013. “Baptism”. In Roberts, H. E. (ed.): *Encyclopedia of Comparative Iconography*. London: Routledge.

Djulbegovic, Benjamin, Jef van den Ende, Robert M. Hamm, Thomas Mayrhofer, Iztok Hozo and Stephen G. Pauker. 2015. “When Is It Rational to Order a Diagnostic Test, or Prescribe Treatment: The Threshold Model as an Explanation of Practice Variation. *European Journal of Clinical Investigation* 45: 485-493.

Douglas, Heather. 2009. *Science, Policy and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Douven, Igor. 2011. “Abduction”. In Zalta, E. (ed.): *The Stanford Encyclopedia of Philosophy* (Spring 2011 Edition), <http://plato.stanford.edu/archives/spr2011/entries/abduction/>.

———. 2005. “Evidence, Explanation, and the Empirical Status of Scientific Realism”, *Erkenntnis* 63: 253-291.

———. 2002. “Testing Inference to the Best Explanation”. *Synthese* 130: 355-377.

Duhem, Pierre. 1954. *The Aim and Structure of Scientific Theory.* Princeton: Princeton University Press.

Elliot, Kevin and Daniel McKaughan. 2014. “Nonepistemic Values and the Multiple Goals of Science”. *Philosophy of Science* 81: 1-21.

———. 2009. “How Values in Scientific Discovery and Pursuit Alter Theory Appraisal”. *Philosophy of Science* 76: 598-611.

Fann, K. T. 1970. *Peirce's Theory of Abduction*. The Hague: Martinus Nijhoff.

Feigl, Herbert. 1971. “Research Programmes and Induction”. In Buck, R.C. and Cohen R.S. (eds.): *P.S.A. 1970*. Boston Studies in the Philosophy of Science Vol. 8. Dordrecht: Reidel, pp. 147-150.

———. 1970. “The “Orthodox” View of Theories: Remarks in Defense as well as Critique”. In Radner, M. and Winokour (eds.): *Theories and Methods of Physics and Psychology.* Minnesota Studies in Philosophy of Science, Vol. 4. Minneapolis: University of Minnesota Press, pp. 3-15.

Feyerabend, Paul. 1970. “Consolations for the Specialist”. In Lakatos and Musgrave (1970), pp. 197-230.

Fontenrose, Robert. 1973. “In Search of Vulcan”. *Journal for the History of Astronomy* 4: 145-158.

Frankfurt, Harry. 1958. “Peirce's Notion of Abduction”. *Journal of Philosophy* 55: 593-597.

Franklin, Allan. 1993a. “Discovery, Pursuit and Justification”. *Perspectives on Science* 1: 252-284.

———. 1993b. *The Rise and Fall of the Fifth Force*. New York: American Institute of Physics.

———. 1986. *The Neglect of Experiment*. Cambridge: Cambridge University Press.

Freeman, L. G. 1968. “A theoretical framework for interpreting archaeological materials”. In Lee, R. B. and I. DeVore (eds.): *Man the Hunter*. Chicago: Aldine, pp. 262–267.

French, Steven. 1995. “The Esperable Uberty of Quantum Chromodynamics”. *Studies in History and Philosophy of Modern Physics* 26: 87-105.

Fresnel, Augustin-Jean. 1866. *Oeuvres complètes*. Paris: Imprimerie Impériale.

Galle, Johan. 1846. [Letter to Le Verrier, 25 September 1846]. Cited in *Comptes rendus hebdomadaires des seánces de l'Acadamie des sciences* 23: 659.

Gero, Joan. 2007. “Honoring Ambiguity/Problematizing Certitude”. *Journal of Archaeological Theory and Method* 14: 311-327.

Giere, Ronald. 2004. How Models Are Used to Represent Reality. *Philosophy of Science* 71: 742-752.

Glymour, Clark. 1984. “Explanation and Realism”. In Leplin (1984), pp. 173-192.

Godfrey-Smith, Peter. 2003. *Theory and Reality*. Chicago: University of Chicago Press.

Gould, Benjamin A. 1850. *Report to the Smithsonian Institution of the History of the Discovery of Neptune*. Washington City: Smithsonian Institution.

Gould, Richard. 1980. *Living Archaeology*. Cambridge: Cambridge University Press.

———. 1978. “The Archaeology of Human Residues”. *American Anthropologist* 80: 815-835.

Grosser, Morton. 1962. *The Discovery of Neptune*. Cambridge, MA: Harvard University Press.

Hacking, Ian. 1984. “Experimentation and Scientific Realism”. In Leplin (1984), pp. 154-172.

Hanson, Norwood R. 1965. “The Idea of a Logic of Discovery”. *Dialogue* 4: 48-61.

———. 1962. “Leverrier: The Zenith and Nadir of Newtonian Mechanics”. *Isis* 53: 359-378.

———. 1960a. “Is There a Logic of Discovery?”. *Australasian Journal of Philosophy* 38: 91-106.

———. 1960b. “More on “The Logic of Discovery””, *Journal of Philosophy* 57: 182-188.

———. 1958. “The Logic of Discovery”. *Journal of Philosophy* 55: 1073-1089.

Harman, Gilbert. 1965. “The Inference to the Best Explanation”. *The Philosophical Review* 74: 88-95.

Harp, Randall and Kareem Khalifa. 2015. “Why Pursue Unification: A Social-Epistemological Puzzle”. *Theoria* 30: 431-477.

Hawkes, Christopher. 1954. “Archaeological Theory and Method: Some Suggestions from the Old World”. *American Anthropologist* 56: 155-168.

Hempel, Carl. 1966. *Philosophy of Natural Science*. Englewood Cliffs: Prentice Hall.

Henderson, Leah. 2014. “Bayesianism and Inference to the Best Explanation”. *British Journal for the Philosophy of Science* 65: 687-715.

Hesse, Mary. 1966. *Models and Analogies in Science*. Notre Dame: University of Notre Dame Press.

———. 1953. “Models in Physics”. *British Journal for the Philosophy of Science* 4: 198-214.

Hingley, Richard. 2000. *Roman Officers and English Gentlemen*. London: Routledge.

Hintikka, Jaakko. 1998. “What is Abduction? The Fundamental Problem of Contemporary Epistemology”. *Transactions of the Charles S. Peirce Society* 34: 503-533.

Hodder, Ian. 1982. *The Present Past: An Introduction to Anthropology for Archaeologists*. London: Batsford.

Howard, Don. 2006. “Lost Wanderers in the Forest of Knowledge: Some Thoughts on the Discovery-Justification Distinction”. In Schickore and Steinle (2006), pp. 3-22.

Hoyningen-Huene, Paul. 2006. “Context of Discovery Versus Context of Justification and Thomas Kuhn”. In Schickore and Steinle (2006), pp. 119-131.

———. 1993. *Reconstructing Scientific Revolutions: Thomas S. Kuhn’s Philosophy of Science*. Chicago: University of Chicago Press.

———. 1987. “Context of Discovery and Context of Justification”. *Studies in History and Philosophy of Science* 18: 501-515.

Iranzo, Valeriano. 2008. “Reliabilism and the Abductive Defence of Scientific Realism”. *Journal of General Philosophy of Science* 39: 115-120.

Jansson, Lina and Jonathan Tallant. *Forthcoming*. “Quantitative Parsimony: Probably for the Better”. *British Journal for the Philosophy of Science*. doi:10.1093/bjps/axv064

Joyce, James. 1999. *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

Kapitan, Tomis. 1997. “Peirce and the Structure of Abductive Inference”. In Houser, N., Don Roberts and James Van Evra (eds.): *Studies in the Logic of Charles Sanders Peirce*. Bloomington: Indiana Unversity Press, pp. 477-496.

———. 1992. “Peirce and the Autonomy of Abductive Inference”. *Erkenntnis* 37: 1-26.

Kassirer, Jerome P., John B. Wong and Richard I. Kopelman. 2010. *Learning clinical reasoning* (2nd ed.). Philadelphia: Lippincott Williams & Wilkins.

Kidd, Ian. 2015. “What’s So Great About Feyerabend? *Against Method*, Forty Years On”. *Metascience* 24: 343-349.

Kitcher, Philip. 2001a. *Science, Truth and Democracy*. Oxford: Oxford University Press.

———. 2001b. “Real Realism: The Galilean Strategy”. *The Philosophical Review* 110: 151-197.

———. 1993. *The Advancement of Science: Science without Legend, Objectivity without Illusions*. Oxford: Oxford University Press.

———. 1990. “The Division of Cognitive Labor”. *Journal of Philosophy* 87: 5-22.

———. 1989. “Explanatory Unification and the Causal Structure of the World.” In Kitcher, P. and W. Salmon (eds.): *Scientific Explanation,* Minnesota Studies in the Philosophy of Science, vol. 13, Minneapolis: University of Minnesota Press, pp. 410–505.

Kordig, Carl. 1978. “Discovery and Justification”, *Philosophy of Science* 45: 110-117.

Krieger, William. 2006. *Can There Be a Philosophy of Archaeology?* New York: Lexington Books.

Kuhn, Thomas. 1977. “Objectivity, Value Judgment and Theory Choice”. In Kuhn, T. (1977): *The Essential Tension*, Chicago: University of Chicago Press, pp. 320-339.

———. 1962/1996. *The Structure of Scientific Revolutions*. 3rd edition. Chicago: University of Chicago Press.

Kukla, André. 2010. *Extraterrestrials: A Philosophical Perspective*. Plymouth: Lexington Books.

———. 2001. “SETI: On the Prospects and Pursuitworthiness of the Search for Extraterrestrial Intelligence”. *Studies in History and Philosophy of Science* 32: 31-67.

Lakatos, Imre. 1978a. *The Methodology of Scientific Research Programmes: Philosophical Papers, Vol. 1*. Eds. J. Worrall and G. Currie. Cambridge: Cambridge University Press.

———. 1978b. *Mathematics, Science and Epistemology: Philosophical Papers, Vol. 2.* Eds. J. Worrall and G. Currie. Cambridge: Cambridge University Press.

———. 1978c. “Anomalies Versus ‘Crucial Experiments’” (A Rejoinder to Professor Grünbaum). In Lakatos (1978b), pp. 211-223.

———. 1974/1978. “Popper on Demarcation and Induction”. Originally in Schilpp (1974), Book 1, pp. 241-273. Reprinted in Lakatos (1978a), pp. 139-167.

———. 1971/1977. “History of Science and Its Rational Reconstructions”. Orginally in Buck, R.C. and Cohen R.S. (eds.): *P.S.A. 1970*. Boston Studies in the Philosophy of Science Vol. 8. Dordrecht: Reidel. Reprinted in Lakatos (1978a), pp. 102-138.

———. 1970/1978. “Falsification and the Methodology of Scientific Research Programmes”. Originally in Lakatos and Musgrave (1970), pp. 91-196). Reprinted in Lakatos (1978a), pp. 8-101.

———. 1968/1978. “Changes in the Problem of Inductive Logic”. Originally in Lakatos, I. (ed.) *The Problem of Inductive Logic*, Amsterdam: North Holland, pp. 315-417. Reprinted in Lakatos (1978b), pp. 128-200.

Lakatos, Imre and Alan Musgrave. 1970. *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press.

Laudan, Larry. 1984. *Science and Values: The Aims of Science and Their Role in Scientific Debate*. Berkeley and Los Angeles: University of California Press.

———. 1981. “A Confutation of Convergent Realism”. *Philosophy of Science* 48: 19-49.

———. 1980. “Why Was the Logic of Discovery Abandoned?”. In Nickles (1980a), pp. 173-183.

———. 1977. *Progress and Its Problems*. Berkeley and Los Angeles: University of California Press.

Leplin, Jarrett (ed.). 1984. *Scientific Realism*. Berkeley and Los Angeles: University of California Press.

Le Verrier, Urbain J. J. 1846a. “Reserches sur les mouvements d’Uranus”. *Comptes rendus hebdomadaires des séances de l’Académie des sciences* 22: 907-918.

———. 1846b. “Sur la planète qui produit les anomalies observes dans les mouvements d’Uranus – determination de sa masse, de son orbite et de sa position actuelle”. *Comptes rendus hebdomadaires des séances de l’Académie des sciences* 23: 428-438.

———.1846c. “Reserches sur les mouvements de la planète Herschel (dite Uranus)”. *Connaissance des temps pour l’année 1849*, Addition, 3-254.

———. 1845. “Première mémoire sur la théorie d’Uranus”. *Comptes rendus hebdomadaires des séances de l’Académie des sciences* 21:1050-1055.

Lewis, David. 1986. *On the Plurality of Worlds.* Oxford: Blackwell.

Lightfoot, Kent. 1995. Culture Contact Studies: Redefining the Relationship between Prehistoric and Historical Archaeology. *American Antiquity* 60: 199-217.

Lipton, Peter. 2004. *Inference to the Best Explanation* (2nd ed.). London: Routledge.

Longacre, William. 1970. “Some Aspects of Prehistoric Society in East-Central Arizona”. In Binford, S. and L. Binford (eds.): *New Perspectives in Archaeology*. Chicago: Aldine, pp. 151-160.

Lycan, William. 2012. “Explanationist Rebuttals (Coherentism Defended Again)”. *Southern Journal of Philosophy* 50: 5-20.

———. 2002. “Explanation and Epistemology”. In Moser, P. (ed.): *Oxford Hand of Epistemology*, Oxford: Oxford University Press, pp. 408-33.

———. 1985. “Epistemic Value”. *Synthese* 64: 137-164.

Lyons, Timothy. 2006. “Scientific Realism and the Strategema de Divide et Impera”. *British Journal for the Philosophy of Science* 57: 537-560.

———. 2005. “Towards a Purely Axiological Scientific Realism”. *Erkenntnis* 63: 167-204.

Magnus, P. D. 2003: “Success, Truth, and the Galilean Strategy”. *British Journal for the Philosophy of Science* 54: 465-474.

Magnus, P. D. and Craigh Callender. 2004. “Realist Ennui and the Base Rate Fallacy”. *Philosophy of Science* 71: 320-388.

Marewski, Julian N. and Gerd Gigenrenzer. 2012. “Heuristic Decision Making in Medicine”. *Dialogues in Neuroscience* 14: 77-89.

McKaughan, Daniel. 2007. *Towards a Richer Vocabulary for Epistemic Attitudes: Mapping the Cognitive Landscape*. Ph.D. Dissertation, University of Notre Dame.

———. 2008. “From Ugly Duckling to Swan: C. S. Peirce, Abduction and the Pursuit of Scientific Theories”. *Transactions of the Charles S. Peirce Society* 44: 446-468.

McLaughlin, Robert. 1982. “Invention and Appraisal”. In McLaughlin, R. (ed.): *What? Where? When? When?* Australasian Studies in History and Philosophy of Science, Vol. 1. Dordrecht: Reidel.

McMullin, Ernan. 1985. “Galilean Idealization”. *Studies in History and Philosophy of Science* 16: 247-273.

———. 1976. “The Fertility of Theory and the Unity for Appraisal in Science”. In Cohen, Feyerabend and Wartofsky (1976), pp. 395-432.

———. 1968. “What do physical models tell us?” In van Rootselaar, B. and J.F. Staal (eds.): *Logic, Methodology and Philosophy of Science III: Proceedings of the Third International Congress for Logic, Methodology and Philosophy of Science, Amsterdam 1967,* Amsterdam: North-Holland, pp. 385-396.

Mill, John Stuart. 1859/2003. *On Liberty*. New Haven: Yale University Press.

Minnameier, Gerhard. 2004. “Peirce-Suit of Truth: Why Inference to the Best Explanation and Abduction Ought Not to be Confused”. *Erkenntnis* 60: 75-105.

Morgan, Mary. 1999. “Learning from Models”. In Morgan, M. and M. Morrison (eds.): *Models as Mediators*, Cambridge: Cambridge University Press, pp. 347-388.

———. 1997. “The Technology of Analogical Models: Irving Fisher’s Monetary Worlds”. *Philosophy of Science* 64 (Proceedings): S304-S314.

Musgrave, Alan. 1976. “Method or Madness?”, In Cohen, Feyerabend and Wartofsky (1976), pp. 457-492.

Nagel, Ernest. 1936. “Critical Notice: *Wahrscheinlichkeitslehre*. By Hans Reichenbach”. *Mind* 45: 501-514.

Neurath, Otto. 1913. “Die Verirrten des Cartesius und das Auxiliarmotiv. Zur Psychologie des Entschlusses”. *Jahrbuch der Philosophischen Gesellschaft an der Universität Wien*. Leipzig: Johann Ambrosius Barth. Translated as “The Lost Wanderers of Descartes and the Auxiliary Motive (On the Psychology of Decision)”, in Neurath, Otto (1983): *Philosophical Papers, 1913-1946*. Vienna Circle Collection, Vol. 16. Ed. Cohen, R. and M. Neurath, Dordrecht: Reidel, pp. 1-12.

Nersessian, Nancy. 2008. *Creating Scientific Concepts*. Cambridge, MA: MIT Press.

———. 2002. “Maxwell and “the Method of Physical Analogy”: Model-Based Reasoning, Generic Abstraction and Conceptual Change”. In Malament, D. (ed.): *Reading Natural Philosophy: Essays in the History and Philosophy of Science and Mathematics*. Chicago and La Salla: Open Court.

———. 1988. “Reasoning from Imagery and Analogy in Scientific Concept Formation”. In Fine, A and J. Leplin (eds.), *PSA 1988, Volume One: Contributed Papers*, East Lansing: Philosophy of Science Assocation, pp. 41-47.

Nickles, Thomas. 1981. “What Is a Problem That We May Solve It?”. *Synthese* 47: 85-118.

——— (ed.). 1980a. *Scientific Discovery, Logic and Rationality*. Boston Studies in the Philosophy of Science, Vol. 56. Dordrecht: Reidel.

——— (ed.). 1980b. *Scientific Discovery: Case Studies*. Boston Studies in the Philosophy of Science, Vol. 60. Dordrecht: Reidel.

———. 1980c. “Introductory Essay: Scientific Discovery and the Future of Philosophy of Science”. In Nickles (1980a), pp. 1-60.

Niiniluoto, Ilka. 1999. “Defending Abduction”. *Philosophy of Science* 66 (Proceedings): S436-S451.

Norton, John. *ms*. “Analogy”, manuscript. Retrieved from http://www.pitt.edu/~jdnorton/papers/material_theory/4.%20Analogy.pdf, 17 March 2017.

Nozick, Robert. 1969. “Newcomb’s Problem and the Two Principles of Choice”. In Rescher, N. (ed.): *Essays in the Honor of Carl G. Hempel*. Dordrecht: Reidel, pp. 114-146.

Nyrup, Rune. 2015. “How Explanatory Reasoning Justifies Pursuit: A Peircean View of IBE”, *Philosophy of Science* 82: 749-760.

Orme, Bryony. 1981. *Anthropology for Archaeologists: An Introduction*. London: Ducksworth.

———. 1974. “Twentieth-Century Prehistorians and the Idea of Ethnographic Parallels”. *Man (New Series)* 9: 199-212.

Paavola, Sami. 2006. “Hansonian and Harmanian Abduction as Models of Discovery”. *International Studies in the Philosophy of Science*. 20: 93-108.

———. 2004. “Abduction as a Logic and Methodology of Discovery: The Importance of Strategies”. *Foundations of Science* 9: 267-283.

Parker, Wendy. 2010. “Scientific Models and Adequacy-for-Purpose”. *Modern Schoolman* 87: 285-293.

———. 2009. “Confirmation and Adequacy-for-Purpose in Climate Modelling”. *Proceedings of the Aristotelian Society Supplementary Volume* 83: 233-249.

Pauker, Stephen and Jerome Kassirer. 1980. “The Threshold Approach to Clinical Decision Making”. *The New England Journal of Medicine* 302: 1109-1117.

Peacock, Jennifer. 2016. “When Is a Mortarium Not a Mortarium? Analogies and Interpretation in Roman Cumbria”. In Erskine, G., P. Jacobsson and S. Stetkiewicz (eds.): *Proceedings of the 17th Iron Age Research Student Symposium*. Oxford: Archaeopress, pp. 20-27.

Peirce, Charles S. 1932-58. *Collected Papers of Charles Sanders Peirce*. 8 vols. Eds. P. Weiss, C. Hartshorne and A. Burks. Cambridge, MA: Harvard University Press.

Pietarinen, Ahti-Veikko and Francesco Bellucci. 2014. “New Light on Peirce’s Conceptions of Retroduction, Deduction and Scientific Reasoning”. *International Studies in the Philosophy of Science* 28: 353-373.

Popper, Karl. 1974. “Replies to My Critics”. In Schilpp (1974), Book 2, pp. 961-1197.

———. 1972. *Objective Knowledge: An Evolutionary Approach*. Oxford: Clarendon Press.

———. 1959/1992. *The Logic of Scientific Discovery*. London and New York: Routledge.

———. 1957/1972. “The Aim of Science”. *Ratio* 1: 24-35. Reprinted in Popper (1972), pp. 191-205.

———. 1935. ““Induktionslogik” und “Hypothesenwahrscheinlichkeit””, *Erkentniss* 5: 170-172.

———. 1934. *Logik der Forschung*. Wien: Springer-Verlag. Translated as Popper (1959/1992).

Psillos, Stathis. 2011a. “An Explorer Upon Untrodden Ground: Peirce on Abduction”. In Gabbay, D., S. Hartmann and J. Woods (eds.) *Handbook of the History of Logic. Volume 10: Inductive Logic*. Amsterdam: Elsevier.

———. 2011b. “The Scope and Limits of the No Miracles Argument”, in Dieks, D., G. Wenceslao, S. Hartmann, T. Uebel and M. Weber (eds.): *Explanation, Prediction, and Confirmation*. Dordrecht: Springer, pp. 23-35.

———. 2002. “Simply the Best: A Case for Abduction”. In Kakas, A. C. and F. Sadri (eds.): *Computational Logic: From Logic Programming into the Future*. Berlin and Heidelberg: Springer, pp. 605-625.

———. 1999. *Scientific Realism: How Science Tracks Truth*. London: Routledge.

Putnam, Hilary. 1975. “What is Mathematical Truth”. In Putnam, H. *Mathematics, Matter and Method*, Cambridge: Cambridge University Press, pp. 60-78.

Quine, W. V. O. 1953a. *From A Logical Point of View*. New York: Harper and Row.

———. 1953b. “On What There Is”, in Quine (1953a), pp. 1-19.

———. 1953c. “Two Dogmas of Empiricism”, in Quine (1953a), pp. 20-46.

Quinn, Philip. 1972. “Methodological Appraisal and Heuristic Advice: Problems in the Methodology of Scientific Research Programmes”. *Studies the History and Philosophy of Science* 3: 135-149.

Ravn, Mads. 2011. “Ethnographic Analogy from the Pacific: Just as Analogical as Any Other Analogy”. *World Archaeology* 43: 716-725.

Reichenbach, Hans. 1944. *Philosophic Foundations of Quantum Mechanics*. Berkeley and Los Angeles: University of California Press.

———. 1938a. “On Probability and Induction”. *Philosophy of Science* 5: 21-45.

———. 1938b. *Experience and Prediction*. Chicago: University of Chicago Press.

———. 1935a. “Wahrscheinlichkeitslogik”. *Erkenntnis* 5: 37-43.

———. 1935b. “Zur Induktions-Maschine”. *Erkenntnis* 5: 172-173.

———. 1935c. *Wahrscheinlichkeitslehre: Eine Untersuchung über die Logischen und Mathematischen Grundlagen der Wahrscheinlichkeitsrechnung*. Leiden: A. W. Sijthoff’s Uitgeversmij.

Reiss, Julian. 2015. “A Pragmatist Theory of Evidence”. *Philosophy of Science* 82: 341-362.

Richard, W. Scott. 2007. “We Should Overcome the Barriers to Evidence-Based Clinical Diagnosis!” *Journal of Clinical Epidemiology* 60: 217-227.

Richardson, W. Scott and Mark C. Wilson. 2015. “The Process of Diagnosis”. In Guyatt, G., D. Rennie, M. O. Meade and D. J. Cook (eds.): *Users’ Guides to the Medical Literature: Essentials of Evidence-Based Clinical Practice*, New York: McGraw Hill, pp. 211-222.

Richardson, W. Scott, Mark C. Wilson, Gordon H. Guyatt, Deborah J. Cook, and James Nishikawa. 1999. “Users’ Guides to the Medical Literature XV: How to Use an Article About Disease Probability for Differential Diagnosis”. *JAMA* 281: 1214-1219.

Roseveare, N.T. 1982. *Mercury's Perihelion from Le Verrier to Einstein*. Oxford: Clarendon Press.

Rueger, Alexander. 1996. “Risk and Diversification in Theory Choice”. *Synthese* 109: 263-280.

Rutherford, Ernest, Francis Aston, James Chadwick, Charles Ellis, George Gamow, Ralph Fowler, Owen Richardson and Douglas Hartree. 1929. “Discussion on the Structure of Atomic Nuclei”, *Proceedings of the Royal Society A* 123: 262–267.

Rutherford, Ernest, James Chadwick and Charles Ellis. 1930. *Radiations from Radioactive Substances*. Cambridge: Cambridge University Press.

Saatsi, Juha. 2017. “Explanation and Explanationism in Science and Metaphysics”. In Slater, M. and Z. Yudell (eds.): *Metaphysics and the Philosophy of Science: New Essays*, Oxford: Oxford University Press, pp. 163-192.

———. 2012. “Scientific Realism and Historical Evidence: Shortcomings of the Current State of Debate”. In de Regt, H., Stephan Hartmann and Samir Okasha (eds.): *EPSA Philosophy of Science: Amsterdam 2009*, Dordrecht: Springer, pp. 329-340.

Salmon, Merrilee. 1982. *Philosophy and Archaeology*. New York: Academic Press.

———. 1976. ““Deductive” and “Inductive” Archaeology”. *American Antiquity* 41: 376-381.

Salmon, Wesley. 2001. “Explanation and Confirmation: A Bayesian Critique of Inference to the Best Explanation”, in Hon, G. & S.S. Rakover (eds.): *Explanation: Theoretical Approaches and Applications*, Dordrecht: Kluwer, pp. 61-91.

———. 1990. “The Appraisal of Theories. Kuhn Meets Bayes”. Fine, A., M. Forbes, and L. Wessels, (eds.): *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1990 Volume Two: Symposia and Invited Papers (1990), pp. 325-332.

———. 1981. “Rational Prediction”. *British Journal for the Philosophy of Science* 32: 115-125.

———. 1967. *The Foundations of Scientific Inference*. Pittsburgh: University of Pittsburgh Press.

Sarkar, Husain. 1983. *A Theory of Method*. Berkeley and Los Angeles: University of California Press.

Schaffner, Kenneth 1993. *Discovery and Explanation in Biology and Medicine*. Chicago: University of Chicago Press.

——— (ed.). 1985. *Logic of Discovery and Diagnosis in Medicine*. Berkeley & Los Angeles: University of California Press.

Schickore, Jutta and Friedrich Steinle. 2006. *Revisiting Discovery and Justification*. Archimedes: New Studies in the History and Philosophy of Science and Technology, Vol. 14. Dordrecht: Springer.

Schilpp, Paul. 1974. *The Philosophy of Karl Popper*. The Library of Living Philosophers, Volume 14. 2 Books. La Salle: Open Court.

Schindler, Samuel. 2017. “Theoretical Fertility McMullin-Style”. *European Journal of Philosophy of Science* 7: 151-173.

———. 2014. “A Matter of Kuhnian Theory-Choice? The GWS Model and the Neutral Current”. *Perspectives on Science* 22: 491-522.

Scholl, Raphael. 2015. “Inference to the Best Explanation in the Catch-22: How Much Autonomy for Mill’s Method of Difference?”. *European Journal for Philosophy of Science* 5: 89-110.

Schon, Donald. 1959. “Comments on Mr. Hanson’s “The Logic of Discovery””. *Journal of Philosophy* 56: 500-503.

Seetah, Krish. 2008. “Modern Analogy, Cultural Theory and Experimental Replication: A Merging Point at the Cutting Edge of Archaeology”. *World Archaeology* 40: 135-150.

Šešelja, Dunja and Christian Straßer. 2014. “Epistemic Justification in the Context of Pursuit: A Coherentist Approach”. *Synthese* 191: 3111-3141.

———. 2013. “Kuhn and the Question of Pursuit Worthiness”. *Topoi* 32: 9-19.

Šešelja, Dunja, Laszlo Kosolosky and Christian Straßer. 2012. “The Rationality of Scientific Reasoning in the Context of Pursuit: Drawing Appropriate Distinctions”. *Philosophica* 86: 51-82.

Smith, Bruce. 1977. “Archaeological Inference and Inductive Confirmation”. *American Anthropologist* 79: 598-617.

Smith, M. A. 1955. “The limitations of inference in archaeology”. *Archaeological Newsletter* 6: 3-7.

Smith, Robert. 1989. “The Cambridge Network in Action: The Discovery of Neptune”. *Isis* 80: 395-422.

Solas, W. J. 1911. *Ancient Hunters and Their Modern Representatives*. London: MacMillan.

Sox, Harold C., Michael C. Higgins & Douglas K. Owens. 2013. *Medical Decision Making* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Stahl, Ann. 1993. “Concepts of Time and Approaches to Analogical Reasoning in Historical Perspective”. *American Antiquity* 58: 235-260.

Stanford, Kyle. 2006. *Exceeding Our Grasp: Science, History and the Problem of Unconceived Alternatives*. Oxford: Oxford University Press.

Stanley, Donald E., and Daniel G. Campos. 2016. Selecting Clinical Diagnoses: Logical Strategies Informed by Experience. *Journal of Evaluation in Clinical Practice* 22: 588-597.

———. 2013. “The Logic of Medical Diagnosis. *Perspectives in Biology and Medicine,* 56: 300-315.

Stanley, Donald E. and Rune Nyrup. *forthcoming*. “Strategies in Abduction: Generating and Selecting Diagnostic Hypotheses”, manuscript under review.

Starna, William. 1979. “A Comment on Curren’s “Potential Interpretations of ‘Stone Gorget’ Function””. *American Antiquity* 44: 337-341.

Stiles, Daniel. 1977. “Ethnoarchaeology: A Discussion of Methods and Applications”. *Man* (New Series) 12: 87-103.

Strevens, Michael. 2013. “No understanding without explanation”, *Studies in History and Philosophy of Science* 44:510-515.

Stuewer, Roger. 1994. “The Origin of the Liquid-Drop Model and the Interpretation of Nuclear Fission”, *Perspectives on Science* 2: 76-129.

Thagard, Paul. 1988. *Computational Philosophy of Science*, Cambridge, MA: MIT Press.

———. 1978. “The Best Explanation: Criteria for Theory Choice”. *Journal of Philosophy* 75: 76-82.

Trigger, B. G. 1989. *A History of Archaeological Thought*. Cambridge: Cambridge University Press.

Tulodziecki, Dana. 2013. “Shattering the Myth of Semmelweis”. *Philosophy of Science* 80: 1065-1075.

Ucko, Peter. 1969. “Ethnography and Archaeological Interpretation of Funerary Remains”. *World Archaeology* 1: 262-280.

Ucko, Peter and Andrée Rosenfeld. 1967. *Paleolithic Cave Art*. London: World University Library.

Upshur, Ross. 1997. “Certainty, Probability and Abduction: Why We Should Look to C.S. Peirce Rather Than Gödel for a Theory of Clinical Reasoning”. *Journal of Evaluation in Clinical Practice* 3: 201-206.

Van Fraassen, Bas. 1989. *Laws and Symmetry*. Oxford: Oxford University Press.

———. 1980. *The Scientific Image*. Oxford: Oxford University Press.

Vickers, Peter. 2013. “A Confrontation of Convergent Realism”. *Philosophy of Science* 80: 189-211.

Wartofsky, Marx. 1986. “Clinical Judgement, Expert Programs and Cognitive Style: A Counter-Essay in the Logic of Diagnosis”. *Journal of Medicine and Philosophy* 11: 81-92.

Watson, P. J.. 1979. “The Idea of Ethnoarchaeology: Notes and Comments”. In Kramer (ed.): *Ethnoarchaeology: Implications of Ethnography in Archaeology*, New York: Columbia University Press, pp. 277-287.

Weinberg, Steven. 1992. *Dreams of a Final Theory: The Search for the Fundamental Laws of Nature*. New York: Pantheon Books.

Weitzenfeld, Julian. 1984. “Valid Reasoning by Analogy”. *Philosophy of Science* 51: 137-149.

White, Roger. 2005. “Explanation as a Guide to Induction”. *Philosophers’ Imprint* 5: 1-29.

Whitt, Laurie. 1992. “Indices of Theory Promise”. *Philosophy of Science* 59: 612-634.

———. 1990. “Theory Pursuit: Between Discovery and Acceptance.” In Fine, A., M. Forbes, and L. Wessels, (eds.): *PSA: Proceedings of the Biennial Meetings of the Philosophy of Science Association*, 1990 Volume One: Contributed Papers. East Lansing, MI: Philosophy of Science Association, pp. 467-483.

Woody, Andrea. 2015. “Re-orienting Discussions of Scientific Explanation: A Functional Perspective”. *Studies in History and Philosophy of Science* 52: 79-87.

———. 2004. “Telltale Signs: What Common Explanatory Strategies in Chemistry Reveal about Explanation Itself”. *Foundations of Chemistry* 6: 13-43.

Wylie, Allison. 2002. *Thinking from Things: Essays in the Philosophy of Archaeology*. Berkeley: University of California Press.

———. 1988. “‘Simple’ Analogy and the Role of Relevance Assumptions: Implications of Archaeological Practice”. *International Studies in the Philosophy of Science* 2: 134-150.

———. 1985. “The reaction against analogy”. *Advances in Archaeological Method and Theory* 8: 63-111.

———. 1982. “An Analogy by Any Other Name Is Just As Analogical: A Commentary on the Gould-Watson Dialogue”. *Journal of Anthropological Archaeology* 1: 382–401.