First published 1991 by Routledge

Second edition published 2004
by Routledge
11 New Fetter Lane, London EC4P 4EE

Simultaneously published in the USA and Canada
by Routledge
29 West 35th Street, New York, NY 10001

Routledge is an imprint of the Taylor & Francis Group

Typeset in Times by
MHL Production Services Limited, Coventry
Printed and bound in Great Britain by
MPG Books Ltd, Bodmin

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data

Lipton, Peter, 1954–
Inference to the best explanation / Peter Lipton. – 2nd ed.
p. cm. – (International library of philosophy)
Includes bibliographical references and index.
1. Science–Philosophy. 2. Science–Methodology. 3. Inference. 4. Explanation. I. Title. II. Series.
Q175.L556 2004
501–dc22 2003018554

ISBN 0-415-24202-9 (hbk)
ISBN 0-415-24203-7 (pbk)

Preface to the second edition

I have resisted the temptation to rewrite this book completely; but this new edition includes substantial enlargement and reworking. All of chapter 7 is new, as are over half of chapters 8 and 9 and significant stretches of chapter 3. Most of the other chapters have enjoyed some alteration. But the overall project and the tone remain the same. My primary interest is to explore our actual inferential practices by articulating and defending the idea that explanatory considerations are an important guide to inference, that we work out what to infer from our evidence by thinking about what would explain that evidence.

Some of the changes in this edition are attempts to fill in blanks I was aware of when I wrote the first edition. The most conspicuous example of this is a discussion of the relationship between explanationism and Bayesianism, an important topic I ducked entirely the first time round. Others include probabilistic approaches to explanation, and the application of some work in cognitive psychology, especially from the 'heuristics and biases' program of Daniel Kahneman and Amos Tversky. Many of the other changes are additions or modifications prompted by reactions to the first edition or by self-criticism. Among the most prominent of these changes is an extension and additional defense of my account of contrastive explanation, further development of the idea that explanatory considerations are actually guiding inferences, and further responses to the objection that Inference to the Best Explanation would make it a miracle that our inferences tend to take us towards the truth. I hope the result is a clear improvement (more truths! fewer falsehoods!), but I remain impressed by how much work remains to be done. I wish that my account were at once more precise and more comprehensive. There is surely room for substantial improvement in the account of Inference to the Best Explanation and for work on aspects of inference that explanationism cannot address. Given the richness of our inferential practices, the analytic project may generate better sketches but never a complete or completely adequate account.

I am very grateful to all those who have helped my thinking about Inference to the Best Explanation since the first edition, through publication,

correspondence and discussion. In this regard I would especially like to thank Alexander Bird, George Botterill, Alex Broadbent, Jeremy Butterfield, John Carroll, Nancy Cartwright, Anjan Chakravartty, Steve Clarke, Chris Daly, Mark Day, Herman de Regt, Ton Derksen, Marina Frasca-Spada, Stephen Grimm, Jason Grossman, Dan Hausman, Katherine Hawley, Dan Heard, Chris Hitchcock, Dien Ho, Giora Hon, Colin Howson, Nick Jardine, Martin Kusch, Tim Lewens, Wendy Lipworth, Timothy Lyons, Hugh Mellor, Tim McGrew, Christina McLeish, Samir Okasha, Arash Pessian, Stathis Psillos, Steven Rappaport, Michael Redhead, Michael Siegal, Paul Thagard, Bas van Fraassen, Jonathan Vogel and Tim Williamson. I am particularly grateful to Eric Barnes for extensive correspondence on the topics of this book and for his incisive published critique (1995). And I have an exceptional intellectual debt to the late Wes Salmon, who wrote a paper critical of Inference to the Best Explanation (2001a), to which I replied (2001) and then had the privilege of his response (2001b). That dialectic involved a sustained email relationship that was for me a profound intellectual experience.

Some of the additions to this edition draw on work of mine previously published elsewhere. I thank Kluwer Academic Publishers for permission to use material from 'Is Explanation a Guide to Inference?' (2001: 92–120), and the Aristotelian Society for permission to use material from 'Is the Best Good Enough?' (1993b).

At about the same time as the first edition of this book appeared, I appeared at Cambridge University. The Department of History and Philosophy of Science, King's College and the University generally have provided an exceptionally congenial setting for my academic life. I am very fortunate, and I especially want to thank everyone who works in the Department for making it such a wildly stimulating place and for showing by example how historians and philosophers can learn from each other. My time here has among many other beneficial things increased my appreciation of the diversity and contingency of scientific practices through their histories; it is my hope that the general aspects of inference that I promote in this book are compatible with that complicated reality.

Finally I express my gratitude to my beloved family, for everything.

Peter Lipton
Cambridge, England

Preface to the first edition

It was David Hume's argument against induction that hooked me on philosophy. Surely we have good reason to believe that the sun will rise tomorrow, even though it is possible that it won't; yet Hume provided an apparently unanswerable argument that we have no way to show this. Our inductive practices have been reliable in the past, or we would not be here now to think about them, but an appeal to their past successes to underwrite their future prospects assumes the very practices we are supposed to be justifying. This skeptical tension between my unshakable confidence in the reliability of many of my inferences and the persuasive power of Hume's argument against the possibility of showing any such thing has continued to focus much of my philosophical thinking to this day.

Somewhat later in my education, I was introduced to the problem of description. Even if we cannot see how to justify our inductive practices, surely we can describe them. But I discovered that it is amazingly difficult to give a principled description of the way we weigh evidence. We may be very good at doing it, but we are miserable at describing how it is done, even in broad outline. This book is primarily an investigation of one popular solution to this problem of description, though I also have something to say about its bearing on the problem of justification. It is a solution that looks to explanation as a key to inference, and suggests that we find out what by asking why.

I owe a great debt to the teachers who are my models of how philosophy is to be done and ought to be taught, especially Freddie Ayer, Rom Harré, Peter Harvey, Bill Newton-Smith, and Louis Mink. I have indicated specific obligations to the literature in the body of the text, but there are a number of philosophers who, through their writing, have pervasively influenced my thinking about inference, explanation and the relations between them. On the general problem of describing our inductive practices, I owe most to John Stuart Mill, Carl Hempel and Thomas Kuhn; on the nature of explanation, to Hempel again, to Alan Garfinkel, and Michael Friedman; and on Inference to the Best Explanation, to Gilbert Harman.

I am also very pleased to be able to acknowledge the help that many colleagues and friends have given me with the material in this book. In particular, I would like to thank Ken April, Philip Clayton, Matt Ginsberg, Hyman Gross, Jim Hopkins, Colin Howson, Trevor Hussey, Norbert Kremer, Ken Levy, Stephen Marsh, Hugh Mellor, David Papineau, Philip Pettit, Michael Redhead, David Ruben, Mark Sainsbury, Morton Schapiro, Dick Sclove, Dan Shartin, Elliott Sober, Fred Sommers, Richard Sorabji, Ed Stein, Nick Thompson, Laszlo Versenyi, Jonathan Vogel, David Weissbord, Alan White, Jim Woodward, John Worrall, Eddy Zemach and especially Tim Williamson. All of these people have made this a better book and their philosophical company has been one of the main sources of the pleasure I take in my intellectual life.

I am also grateful to the National Endowment for the Humanities for a grant under which some of the research for this book was completed, and to Williams College, for a leave during which the final version was written. Several parts of the book are based on previously or soon to be published work, and I thank the editors and publishers in question for permission to use this. Chapter 3 includes material from 'A Real Contrast' (1987) and from 'Contrastive Explanation' (1991). Chapter 8 includes material from 'Prediction and Prejudice' (1990).

Finally, I would like to thank Diana, my wife. Without her, a less readable version of my book would still have been possible, but my life would have been immeasurably poorer.

Peter Lipton
Williamstown, Massachusetts
February 1990

Introduction

We are forever inferring and explaining, forming new beliefs about the way things are and explaining why things are as we have found them to be. These two activities are central to our cognitive lives, and we usually perform them remarkably well. But it is one thing to be good at doing something, quite another to understand how it is done or why it is done so well. It is easy to ride a bicycle, but hard to describe how it is done; it is easy to distinguish between grammatical and ungrammatical strings of words in one's native tongue, but hard to describe the principles that underlie those judgments. In the cases of inference and explanation, the contrast between what we can do and what we can describe is stark, for we are remarkably bad at principled description. We seem to have been designed to perform the activities, but not to analyze or to defend them. Still, epistemologists do the best they can with their limited cognitive endowment, trying to describe and justify our inferential and explanatory practices.

This book is an essay on one popular attempt to understand how we go about weighing evidence and making inferences. According to the model of Inference to the Best Explanation, our explanatory considerations guide our inferences. Beginning with the evidence available to us, we infer what would, if true, provide the best explanation of that evidence. This cannot be the whole story about inference: any sensible version of Inference to the Best Explanation should acknowledge that there are aspects of inference that cannot be captured in these terms. But many of our inferences, both in science and in ordinary life, appear to follow this explanationist pattern. Faced with tracks in the snow of a certain peculiar shape, I infer that a person on snowshoes has recently passed this way. There are other possibilities, but I make this inference because it provides the best explanation of what I see. Watching me pull my hand away from the stove, you infer that I am in pain, because this is the best explanation of my excited behavior. Having observed the motion of Uranus, the scientist infers that there is another hitherto unobserved planet with a particular mass and orbit, since that is the best explanation of Uranus's path.

Inference to the Best Explanation is a popular account, though it also has notable critics. It is widely supposed to provide an accurate description of a

central mechanism governing our inferential practices and also a way to show why these practices are reliable. In spite of this, the model has not been much developed. It is more a slogan than an articulated philosophical theory. There has been some discussion of whether this or that inference can be described as to the best explanation, but little investigation into even the most basic structural features of the model. So it is time to try to flesh out the slogan and to give the model the detailed assessment it deserves. That is the purpose of this book.

One reason Inference to the Best Explanation has been so little developed, in spite of its popularity, is clear. The model is an attempt to account for inference in terms of explanation, but our understanding of explanation is so patchy that the model seems to account for the obscure in terms of the equally obscure. It might be correct, yet unilluminating. We do not yet have an account that provides the correct demarcation between what explains a phenomenon and what does not; we are even further from an account of what makes one explanation better than another. So the natural idea of articulating Inference to the Best Explanation by plugging in one of the standard theories of explanation yields disappointing results. For example, if we were to insert the familiar deductive-nomological model of explanation, Inference to the Best Explanation would reduce to a variant of the equally familiar hypothetico-deductive model of confirmation. This would not give us a new theory of inference, but only a restatement of an old one that is known to have many weaknesses.

Nevertheless, the situation is far from hopeless. First of all, there are a number of elementary distinctions that can be made which add structure to the account without presupposing any specific and controversial account of explanation. Secondly, there has been some important work on causal explanation and the 'interest relativity' of explanation that can be developed and extended in a way that casts light on the nature of Inference to the Best Explanation and its prospects. Or so I shall try to show.

The book falls into three parts. The first part, chapters 1 through 3, introduces some of the central problems in understanding the nature of inference and of explanation. I distinguish the problems of describing these practices from the problems of justifying them, and consider some of the standard solutions to both. In the third chapter, I focus on contrastive explanations, explanations that answer questions of the form 'Why this *rather than* that?', and attempt an improved account of the way these explanations work. The second part of the book, chapters 4 through 8, considers the prospects of Inference to the Best Explanation as a partial solution to the problem of describing our inductive practices. Chapter 4 develops some of the basic distinctions the model requires, especially the distinction between actual and potential explanation, and between the explanation that is most warranted and the explanation that would, if true, provide the most understanding, the distinction between the 'likeliest' and

the 'loveliest' explanation. This chapter also flags some of the prima facie strengths and weaknesses of the model. Chapters 5 and 6 are an attempt to use the analysis of contrastive explanation from chapter 3 to defend the model and in particular to show that it marks an improvement on the hypothetico-deductive model of confirmation. Chapter 7 focuses on the relationship between Inference to the Best Explanation and Bayesian approaches to inference and develops a compatibilist position, on the grounds that explanationist thinking can be seen in part as a way cognitive agents 'realize' the probabilistic Bayesian calculation that reflects the bearing of evidence on hypothesis. Chapter 8 develops arguments for saying that it really is diverse explanatory considerations that are guiding inferences. The third part of the book, chapters 9 through 11, switches from issues of description to issues of justification. Chapter 9 answers the accusation that Inference to the Best Explanation would undermine the idea that our inferences take us towards the truth. Chapter 10 considers the use of an explanatory inference to justify the common but controversial view that the successful predictions that a scientific theory makes provide stronger support for it than data that were known before the theory was generated and which the theory was designed to accommodate. The last chapter evaluates another well-known application of Inference to the Best Explanation, as an argument for scientific realism, the view that science is in the truth business, where the truth of predictively successful theories is claimed to provide the best explanation of that success. This chapter ends with a brief sketch of some of the prospects for exploiting the explanationist structure of scientific inferences in other arguments for realism and against instrumentalist interpretations of scientific inference.

The topics of inference and explanation are vast, so there is much that this book assumes and much that it neglects. For example, I help myself to the concept of causation, without offering an analysis of it. My hope is that whatever the correct analysis of that concept turns out to be, it can be plugged into the many claims I make about causal explanation and causal inference without making them false. I also assume throughout that inferred claims, especially inferred theories, are to be construed literally and not, say, by means of some operationalist reduction. For most of the book, I also assume that when a claim is inferred, what is inferred is that the claim is true, or at least approximately true, though this becomes an issue in the final chapter. I do not attempt the difficult task of providing an adequate analysis of the notion of approximate truth or verisimilitude. (For animadversions on this notion and its application, see Laudan 1984: 228–30, and Fine 1984.) I have also neglected the various approaches that workers in artificial intelligence have taken to describe inference. These are important matters and ought to be addressed in relation to Inference to the Best Explanation, but I leave them for another time, if not to another person.

I do not count myself a stylish writer, but I can be clear and accessible, at least by the generally low standards of the philosophical literature. So I have tried to write a book that, while not introductory, would be of some use to a dedicated undergraduate unfamiliar with the enormous literature on inference and explanation. As a consequence, some of the material to follow, especially in the first two chapters, will be familiar to *aficionados*. I have also attempted to write so that each chapter stands as much on its own as is compatible with a progressive argument. Consequently, even if you are not particularly interested in Inference to the Best Explanation per se, you may find certain chapters worth your time. For example, if you are interested in contrastive explanation, you might just read chapter 3, and if you are interested in the prediction/accommodation issue, just chapter 10.

Most philosophers, today and throughout the subject's history, adopt the rhetoric of certainty. They write as if the correctness of their views has been demonstrated beyond reasonable doubt. This sometimes makes for stimulating reading, but it is either disingenuous or naive. In philosophy, if a position is interesting and important, it is almost always also controversial and dubitable. I think Inference to the Best Explanation is both interesting and important, and I have tried not to express more confidence in my claims than the arguments warrant, without at the same time being annoyingly vague or tentative. Since I probably get carried away at points, in spite of my best intentions, let me repeat now that it seems obvious to me that Inference to the Best Explanation cannot be the whole story about inference: at most, it can be an illuminating chapter. And while this book has turned out to take the form of a defense of this model of inference, it should be read rather as a preliminary exploration. Even if it wins no converts, but encourages others to provide more probing criticisms of Inference to the Best Explanation or to generate better alternatives, I will be well satisfied.

Chapter 1

Induction

Underdetermination

We infer some claims on the basis of other claims: we move from premises to a conclusion. Some inferences are deductive: it is impossible for the premises to be true but the conclusion false. All other inferences I call 'inductive', using that term in the broad sense of non-demonstrative reasons. Inductive inference is thus a matter of weighing evidence and judging probability, not of proof. How do we go about making these judgments, and why should we believe they are reliable? Both the question of description and the question of justification arise from underdetermination. To say that an outcome is underdetermined is to say that some information about initial conditions and rules or principles does not guarantee a unique solution. The information that Tom spent five dollars on apples and oranges and that apples are fifty cents a pound and oranges a dollar a pound underdetermines how much fruit Tom bought, given only the rules of deduction. Similarly, those rules and a finite number of points on a curve underdetermine the curve, since there are many curves that would pass through those points.

Underdetermination may also arise in our description of the way a person learns or makes inferences. A description of the evidence, along with a certain set of rules, not necessarily just those of deduction, may underdetermine what is learned or inferred. Insofar as we have described all the evidence and the person is not behaving erratically, this shows that there are hidden rules. We can then study the patterns of learning or inference to try to discover them. Noam Chomsky's argument from 'the poverty of the stimulus' is a good example of how underdetermination can be used to disclose the existence of additional rules (1965: ch. 1, sec. 8, esp. 58–9). Children learn the language of their elders, an ability that enables them to understand an indefinite number of sentences on first acquaintance. The talk young children hear, however, along with rules of deduction and any plausible general rules of induction, grossly underdetermine the language they learn. What they hear is limited and includes many ungrammatical sentences, and the little they hear that is well formed is compatible with

many possible languages other than the one they learn. Therefore, Chomsky argues, in addition to any general principles of deduction and induction, children must be born with strong linguistic rules or principles that further restrict the class of languages they will learn, so that the actual words they hear are now sufficient to determine a unique language. Moreover, since a child will learn whatever language he is brought up in, these principles cannot be peculiar to a particular human language; instead, they must specify something that is common to all of them. For Chomsky, determining the structure of these universal principles and the way they work is the central task of modern linguistics.

Thomas Kuhn provides another well-known example of using underdetermination as a tool to investigate cognitive principles. He begins from an argument about scientific research strikingly similar to Chomsky's argument about language acquisition (1970; 1977, esp. ch. 12). In most periods in the history of a developed scientific specialty, scientists are in broad agreement about which problems to work on, how to attack them and what counts as solving them. But the explicit beliefs and rules scientists share, especially their theories, data, general rules of deduction and induction, and any explicit methodological rules, underdetermine these shared judgments. Many possible judgments are compatible with these beliefs and rules other than the ones the scientists make. So Kuhn argues that there must be additional field-specific principles that determine the actual judgments. Unlike Chomsky, Kuhn does not argue for principles that are either innate or in the form of rules, narrowly construed. Instead, scientists acquire through their education a stock of exemplars – concrete problem solutions in their specialty – and use them to guide their research. They pick new problems that look similar to an exemplar problem, they try techniques that are similar to those that worked in that exemplar, and they assess their success by reference to the standards of solution that the exemplars illustrate. Thus the exemplars set up a web of 'perceived similarity relations' that guide future research, and the shared judgments are explained by the shared exemplars. These similarities are not created or governed by rules, but they result in a pattern of research that mimics one that is rule governed. Just how exemplars do this work, and what happens when they stop working, provide the focus of Kuhn's account of science.

As I see it, Chomsky and Kuhn are both arguing for unacknowledged principles of induction, even though the inferences in the one case concern grammaticality rather than the world around us and even though the principles governing the inference in the other case are determined by exemplars rather than by rules (cf. Curd and Cover 1998: 497). In both cases, inferences are drawn that are not entailed by the available evidence. However, here underdetermination is taken to be a symptom of the existence of highly specialized principles, whether of language acquisition or of scientific research in a particular field at a particular time, since the

underdetermination is claimed to remain even if we include general principles of induction among our rules. But it is natural to suppose that there are some general principles, and the same pattern of argument applies there. If an inference is inductive, then by definition it is underdetermined by the evidence and the rules of deduction. Insofar as our inductive practices are systematic, we must use additional principles of inference, and we may study the patterns of our inferences in an attempt to discover what those principles are and to determine what they are worth.

Justification

The two central questions about our general principles of induction concern description and justification. What principles do we actually use? Are these good principles to use? The question of description seems at first to take priority. How can we even attempt to justify our principles until we know what they are? Historically, however, the justification question came first. One reason for this is that the question of justification gets its grip from skeptical arguments that seem to apply to any principles that could account for the way we fill the gap between the evidence we have and the inference we make. It is the need for such principles rather than the particular form they take that creates the skeptical trouble.

The problem of justification is to show that our inferential methods are good methods, fit for purpose. The natural way to understand this is in terms of truth. We want our methods of inference to be 'truth-tropic', to take us towards the truth. For deduction, a good argument is one that is valid, a perfect truth conduit, where if the premises are true, the conclusion must be true as well. The problem of justification here would be to show that arguments we judge valid are in fact so. For induction, such perfect reliability is out of the question. By definition, even a good inductive argument is one where it is possible for there to be true premises but a false conclusion. Moreover, it is clear that the reasonable inductive inferences we make are not entirely reliable even in this world, since they sometimes sadly take us from truth to falsehood. Nevertheless, it remains natural to construe the task of justification as that of showing truth-tropism. We would like to show that those inductive inferences we judge worth making are ones that tend to take us from true premises to true conclusions.

A skeptical argument that makes the problem of justification pressing has two components, underdetermination and circularity. The first is an argument that the inferences in question are underdetermined, given only our premises and the rules of deduction; that the premises and those rules are compatible not just with the inferences we make, but also with other, incompatible inferences. This shows that the inferences in question really are inductive and, by showing that there are possible worlds where the principles we use take us from true premises to false conclusions, it also shows that

there are worlds where our principles would fail us. Revealing this underdetermination, however, does not yet generate a skeptical argument, since we might have good reason to believe that the actual world is one where our principles are at least moderately reliable. So the skeptical argument requires a second component, an argument for circularity, which attempts to show that we cannot rule out the possibility of massive unreliability that underdetermination raises without employing the very principles that are under investigation, and so begging the question.

Although it is not traditionally seen as raising the problem of induction, Descartes's 'First Meditation' (1641) is a classic illustration of this technique. Descartes's goal is to cast doubt on the 'testimony of the senses', which leads us to infer that there is, say, a mountain in the distance, because that is what it looks like. He begins by arguing that we ought not to trust the senses completely, since we know that they do sometimes mislead us, 'when it is a question of very small and distant things'. This argument relies on underdetermination, on the fact that the way things appear does not entail the way they are; but it does not yet have the circularity component, since we can corroborate our inferences about small and distant things without circularity by taking a closer look (Williams 1978: 51–2). But Descartes immediately moves on from the small and the distant to the large and near. No matter how clearly we seem to see something, it may only be a dream, or a misleading experience induced by an evil demon. These arguments describe possible situations where even the most compelling sensory testimony is misleading. Moreover, unlike the worry about small and distant things, these arguments also have a circle component. There is apparently no way to test whether a demon is misleading us with a particular experience, since any test would itself rely on experiences that the demon might have induced. The senses may be liars, giving us false testimony, and we should not find any comfort if they also report that they are telling us the truth.

The demon argument begins with the underdetermination of observational belief by observational experience, construes the missing principle of inference on the model of inference from testimony, and then suggests that the reliability of this principle could only be shown by assuming it. Perhaps one of the reasons Descartes's arguments are not traditionally seen as raising the problem of justifying induction is that his response to his own skepticism is to reject the underdetermination upon which it rests. Descartes argues that, since inferences from the senses must be inductive and so raise a skeptical problem, our knowledge must instead have a different sort of foundation for which the problem of underdetermination does not arise. The *cogito* and the principles of clarity and distinctness that it exemplifies are supposed to provide the non-inductive alternative. Circularity is also avoided since the senses do not have to justify themselves, even if the threat of circularity notoriously reappears elsewhere, in the attempt to justify the principles of clarity and distinctness by appeal to an argument for the existence of God

that is to be accepted because it itself satisfies those principles. Another reason why Descartes is not credited with the problem of induction may be that he does not focus directly on the principles governing inferences from experience, but rather on the fallibility of the conclusions they yield.

The moral Descartes draws from underdetermination and circularity is not that our principles of induction require some different sort of defence or must be accepted without justification, but that we must use different premises and principles, for which the skeptical problem does not arise. Thus he attempts to wean us from the senses. For a skeptical argument about induction that does not lead to the rejection of induction, we must turn to its traditional home, in the arguments of David Hume.

Hume also begins with underdetermination, in this case that our observations do not entail our predictions (1748: sec. IV). He then suggests that the governing principle of all our inductive inferences is that nature is uniform, that the unobserved (but observable) world is much like what we have observed. The question of justification is then the question of showing that nature is indeed uniform. This cannot be deduced from what we have observed, since the claim of uniformity itself incorporates a massive prediction. But the only other way to argue for uniformity is to use an inductive argument, which would rely on the principle of uniformity, leaving the question begged. According to Hume, we are addicted to the practice of induction, but it is a practice that cannot be justified.

To illustrate the problem, suppose our fundamental principle of inductive inference is 'More of the Same'. We believe that strong inductive arguments are those whose conclusions predict the continuation of a pattern described in the premises. Applying this principle of conservative induction, we would infer that the sun will rise tomorrow, since it has always risen in the past; and we would judge worthless the argument that the sun will not rise tomorrow since it has always risen in the past. One can, however, come up with a factitious principle to underwrite the latter argument. According to the principle of revolutionary induction, 'It's Time for a Change', and this sanctions the dark inference. Hume's argument is that we have no way to show that conservative induction, the principle he claims we actually use for our inferences, will do any better than intuitively wild principles like the principles of revolutionary induction. Of course conservative induction has had the more impressive track record. Most of the inferences from true premises that it has sanctioned have also had true conclusions. Revolutionary induction, by contrast, has been conspicuous in failure, or would have been, had anyone relied on it. The question of justification, however, does not ask which method of inference has been successful; it asks which one will be successful.

Still, the track record of conservative induction appears to be a reason to trust it. That record is imperfect (we are not aspiring to deduction), but very impressive, particularly as compared with revolutionary induction and its ilk.

In short, induction will work because it has worked. This seems the only justification our inductive ways could ever have or require. Hume's disturbing observation was that this justification appears circular, no better than trying to convince someone that you are honest by saying that you are. Much as Descartes argued that we should not be moved if the senses give testimony on their own behalf, so Hume argued that we cannot appeal to the history of induction to certify induction. The trouble is that the argument that conservative inductions will work because they have worked is itself an induction. The past success is not supposed to prove future success, only make it very likely. But then we must decide which standards to use to evaluate this argument. It has the form 'More of the Same', so conservatives will give it high marks, but since its conclusion is just to underwrite conservatism, this begs the question. If we apply the revolutionary principle, it counts as a very weak argument. Worse still, by revolutionary standards, conservative induction is likely to fail precisely because it has succeeded in the past, and the past failures of revolutionary induction augur well for its future success (Skyrms 1986: ch. II). The justification of revolutionary induction seems no worse than the justification of conservative induction, which is to say that the justification of conservative induction looks very bad indeed.

The problem of justifying induction does not show that there are other inductive principles better than our own. Instead it argues for a deep symmetry: many sets of principles, most of them wildly different from our own and incompatible with each other, are yet completely on a par from a justificatory point of view. This is why the problem of justification can be posed before we have solved the problem of description. Whatever inductive principles we use, the fact that they are inductive seems enough for the skeptic to show that they defy justification. We fill the gap of under-determination between observation and prediction in one way, but it could be filled in many other ways that would have led to entirely different predictions. We have no way of showing that our way is any better than any of the other ways that would certify their own reliability. Each is on a par in the sense that it can only argue for its principles by appeal to those very principles. And it is not just that the revolutionaries will not be convinced by the justificatory arguments of the conservatives: the conservatives should not accept their own defense either, since among their standards is one which says that a circular argument is a bad argument, even if it is in one's own aid. Even if I am honest, I ought to admit that the fact that I say so ought not to carry any weight. We have a psychological compulsion to favor our own inductive principles but, if Hume is right, we should see that we cannot even provide a cogent rationalization of our behavior.

It seems to me that we do not yet have a satisfying solution to Hume's challenge and that the prospects for one are bleak. (Though some seem unable to give up trying. See e.g. Lipton 2000 and forthcoming.) There are,

however, other problems of justification that are more tractable. The peculiar difficulty of meeting Hume's skeptical argument against induction arises because he casts doubt on our inductive principles as a whole, and so any recourse to induction to justify induction appears hopeless. But one can also ask for the justification of particular inductive principles and, as Descartes's example of small and distant things suggests, this leaves open the possibility of appeal to other principles without begging the question. For example, among our principles of inference is one that makes us more likely to infer a theory if it is supported by a variety of evidence than if it is supported by a similar amount of homogenous data. This is the sort of principle that might be justified in terms of a more basic inductive principle, say that we have better reason to infer a theory when all the reasonable competitors have been refuted, or that a theory is only worth inferring when each of its major components has been separately tested. Another, more controversial, example of a special principle that might be justified without circularity is that, all else being equal, a theory deserves more credit from its successful predictions than it does from data that the theory was constructed to fit. This appears to be an inductive preference most of us have, but the case is controversial because it is not at all obvious that it is rational. On the one hand, many people feel that only a prediction can be a real test, since a theory cannot possibly be refuted by data it is built to accommodate; on the other, that logical relations between theory and data upon which inductive support exclusively depends cannot be affected by the merely historical fact that the data were available before or only after the theory was proposed. In any event, this is an issue of inductive principle that is susceptible to non-circular evaluation, as we will see in chapter 10. Finally, though a really satisfying solution to Hume's problem would have to be an argument for the reliability of our principles that had force against the inductive skeptic, there may be arguments for reliability that do not meet this condition yet still have probative value for those of us who already accept some forms of induction. We will consider some candidates in chapter 11.

Description

We can now see why the problem of justification, the problem of showing that our inductive principles are reliable, did not have to wait for a detailed description of those principles. The problem of justifying our principles gets its bite from skeptical arguments, and these appear to depend only on the fact that these principles are principles of induction, not on the particular form they take. The crucial argument is that the only way to justify our principles would be to use an argument that relies on the very same principles, which is illegitimate; an argument that seems alas to work whatever the details of our inferences. The irrelevance of the details comes out in the symmetry of Hume's argument: just as the future success of conservative induction gains

no plausibility from its past success, so the future success of revolutionary induction gains nothing from its past failures. As the practice varies, so does the justificatory argument, preserving the pernicious circularity. Thus the question of justification has had a life of its own: it has not waited for a detailed description of the practice whose warrant it calls into doubt.

By the same token, the question of description has fortunately not waited for an answer to the skeptical arguments. Even if our inferences were unjustifiable, one still might be interested in saying how they work. The problem of description is not to show that our inferential practices are reliable; it is just to describe them as they stand. One might have thought that this would be a relatively simple problem. First of all, there are no powerful reasons for thinking that the problem of description is insoluble, as there are for the problem of justification. There is no great skeptical argument against the possibility of description. It is true that any account of our principles will itself require inductive support, since we must see whether it jibes with our observed inductive practice. This, however, raises no general problem of circularity now that a general justification of induction is not the issue. Using induction to investigate the actual structure of our inductive practices is no more suspect than using observation to study the structure and function of the eye. Secondly, it is not just that a solution to the problem of describing our inductive principles should be possible, but that it should be fairly easy. After all, they are our principles, and we use them constantly. It thus comes as something of a shock to discover how extraordinarily difficult the problem of description has turned out to be. It is not merely that ordinary reasoners are unable to describe what they are doing: years of focused effort by epistemologists and philosophers of science have yielded little better. Again, it is not merely that we have yet to capture all the details, but that the most popular accounts of the gross structure of induction are wildly at variance with our actual practice.

Why is description so hard? One reason is a quite general gap between what we can do and what we can describe. You may know how to do something without knowing how you do it; indeed, this is the usual situation. It is one thing to know how to tie one's shoes or to ride a bike; it is quite another thing to be able to give a principled description of what it is that one knows. Chomsky's work on principles of language acquisition and Kuhn's work on scientific method are good cognitive examples. Their investigations would not be so important and controversial if ordinary speakers knew how they distinguished grammatical from ungrammatical sentences or normal scientists knew how they made their methodological judgments. Speakers and scientists employ diverse principles, but they are not conscious of them. The situation is similar in the case of inductive inference generally. Although we may partially articulate some of our inferences if, for example, we are called upon to defend them, we are not conscious of the diverse principles of inductive inference we constantly use.

Since our principles of induction are neither available to introspection, nor otherwise observable, the evidence for their structure must be indirect. The project of description is one of black box inference, where we try to reconstruct the underlying mechanism on the basis of the superficial patterns of evidence and inference we observe in ourselves. This is no trivial problem. Part of the difficulty is simply the fact of underdetermination. As the examples of Chomsky and Kuhn show, underdetermination can be a symptom of missing principles and a clue to their nature, but it is one that does not itself determine a unique answer. In other words, where the evidence and the rules of deduction underdetermine inference, that information also underdetermines the missing principles. There will always be many different possible mechanisms that would produce the same patterns, so how can one decide which one is actually operating? In practice, however, as epistemologists we usually have the opposite problem: we can not even come up with a single description that would yield the patterns we observe. The situation is the same in scientific theorizing generally. There is always more than one account of the unobserved and often unobservable world that would account for what we observe, but scientists' actual difficulty is often to come up with even one theory that fits the observed facts. On reflection, then, it should not surprise us that the problem of description has turned out to be so difficult. Why should we suppose that the project of describing our inductive principles is going to be easier than it would be, say, to give a detailed account of the working of a computer on the basis of the correlations between keys pressed and images on the screen?

Now that we are prepared for the worst, we may turn to some of the popular attempts at description. In my discussion of the problem of justification, I suggested that, following Hume's idea of induction as habit formation, we describe our pattern of inference as 'More of the Same'. This is pleasingly simple, but the conservative principle is at best a caricature of our actual practice. We sometimes do not infer that things will remain the same and we sometimes infer that things are going to change. When my mechanic tells me that my brakes are about to fail, I do not suppose that he is therefore a revolutionary inductivist. Again, we often make inductive inferences from something we observe to something invisible, such as from people's behavior to their beliefs or from the scientific evidence to unobservable entities and processes, and this does not fit into the conservative mold. 'More of the Same' might enable me to predict what you will do on the basis of what you have done (if you are a creature of habit), but it will not tell me what you are or will be thinking.

Faced with the difficulty of providing a general description, a reasonable strategy is to begin by trying to describe one part of our inductive practice. This is a risky procedure, since the part one picks may not really be describable in isolation, but there are sometimes reasons to believe that a particular part is independent enough to permit a useful separation. Chomsky

must believe this about our principles of linguistic inference. Similarly, one might plausibly hold that, while simple habit formation cannot be the whole of our inductive practice, it is a core mechanism that can be treated in isolation. Thus one might try to salvage the intuition behind conservative principle by giving a more precise account of the cases where we are willing to project a pattern into the future, leaving to one side the apparently more difficult problems of accounting for predictions of change and inferences to the unobservable. What we may call the instantial model of inductive confirmation may be seen in this spirit. According to it, a hypothesis of the form 'All As are B' is supported by its positive instances, by observed As that are also B (Hempel 1965: ch. 1). This is not, strictly speaking, an account of inductive *inference*, since it does not say either how we come up with the hypothesis in the first place or how many supporting instances are required before we actually infer it, but this switching of the problem from inference to support may also be taken as a strategic simplification. In any event, the underlying idea is that if enough positive instances and no refuting instances (As that are not B) are observed, we will infer the hypothesis, from which we may then deduce the prediction that the next A we observe will be B.

This model could only be a very partial description of our inductive principles but, within its restricted range, it strikes many people initially as a truism, and one that captures Hume's point about our propensity to extend observed patterns. Observed positive instances are not necessary for inductive support, as inferences to the unobserved and to change show, but they might seem at least sufficient. But the instantial model has been shown to be wildly over-permissive. Some hypotheses are supported by their positive instances, but many are not. Observing only black ravens may lead one to believe that all ravens are black, but observing only bearded philosophers would probably not lead one to infer that all philosophers are bearded. Nelson Goodman has generalized this problem, by showing how the instantial model sanctions any prediction at all if there is no restriction on the hypotheses to which it can be applied (Goodman 1983: ch. III). His technique is to construct hypotheses with factitious predicates. Black ravens provide no reason to believe that the next swan we see will be white, but they do provide positive instances of the artificial hypothesis that 'All raveswans are blight', where something is a raveswan just in case it is either observed before today and a raven, or not so observed and a swan, and where something is blight just in case it is either observed before today and black, or not so observed and white. But the hypothesis that all raveswans are blight entails that the next observed raveswan will be blight which, given the definitions, is just to say that the next swan will be white.

The other famous difficulty facing the instantial model arises for hypotheses that do seem to be supported by their instances. Black ravens support the hypothesis that all ravens are black. This hypothesis is logically

equivalent to the contrapositive hypothesis that all non-black things are non-ravens: there is no possible situation where one hypothesis would be true but the other false. According to the instantial model, the contrapositive hypothesis is supported by non-black, non-ravens, such as green leaves. The rub comes with the observation that whatever supports a hypothesis also supports anything logically equivalent to it. This is very plausible, since support provides a reason to believe true, and we know that if a hypothesis is true, then so must be anything logically equivalent to it. But then the instantial model once again makes inductive support far too easy, counting green leaves as evidence that all ravens are black (Hempel 1965: ch. 1). We will discuss this paradox of the ravens in chapter 6. Something like conservative induction must play a role in both everyday and scientific inferences (cf. Achinstein 1992), but it can at best be only part of the story, and it has turned out to be tantalizingly difficult to articulate in a way that could even give it a limited role.

Another famous account of inductive support is the hypothetico-deductive model (Hempel 1966: chs 2, 3). On this view, a hypothesis or theory is supported when it, along with various other statements, deductively entails a datum. Thus a theory is supported by its successful predictions. This account has a number of attractions. First, although it leaves to one side the important question of the source of hypotheses, it has much wider scope than the instantial model, since it allows for the support of hypotheses that appeal to unobservable entities and processes. The big bang theory of the origin of the universe obviously cannot be directly supported; but along with other statements it entails that we ought to find ourselves today traveling through a uniform background radiation, like the ripples left by a rock falling into a pond. The fact that we do now observe this radiation (or effects of it) provides some reason to believe the big bang theory. Thus, even if a hypothesis cannot be supported by its instances, because its instances are not observable, it can be supported by its observable logical consequences. Secondly, the model enables us to co-opt our accounts of deduction for an account of induction, an attractive possibility since our understanding of deductive principles is so much better than our understanding of inductive principles. Lastly, the hypothetico-deductive model seems genuinely to reflect scientific practice, which is perhaps why it has become the scientists' philosophy of science.

In spite of all its attractions, our criticism of the hypothetico-deductive model here can be brief, since it inherits all the over-permissiveness of the instantial model. Any case of support by positive instances will also be a case of support by consequences. The hypothesis that all As are B, along with the premise that an individual is A, entails that it will also be B, so the thing observed to be B supports the hypothesis, according to the hypothetico-deductive model. That is, any case of instantial support is also a case of hypothetico-deductive support, so the model has to face the problem of

insupportable hypotheses and the raven paradox. Moreover, the hypothetico-deductive model is similarly over-permissive in the case of vertical inferences to hypotheses about unobservables, a problem that the instantial model avoided by ignoring such inferences altogether. The difficulty is structurally similar to Goodman's problem of factitious predicates. Consider the conjunction of the hypotheses that all ravens are black and that all swans are white. This conjunction, along with premises concerning the identity of various ravens, entails that they will be black. According to the model, the conjunction is supported by black ravens, and it entails its own conjunct about swans. The model thus appears to sanction the inference from black ravens to white swans (cf. Goodman 1983: 67–8). Similarly, the hypothesis that all swans are white taken alone entails the inclusive disjunction that either all swans are white or there is a black raven, a disjunction we could establish by seeing a black raven, again giving illicit hypothetico-deductive support to the swan hypothesis. These maneuvers are obviously artificial, but nobody has managed to show how the model can be modified to avoid them without also eliminating most genuine cases of inductive support (cf. Glymour 1980b: ch. II). Finally, in addition to being too permissive, finding support where none exists, the model is also too strict, since data may support a hypothesis which does not, along with reasonable auxiliary premises, entail them. We will investigate this problem in chapter 5 and return to the problem of over-permissiveness in chapter 6.

We believe some things more strongly than others, and our next approach to the descriptive problem represents degrees of belief in terms of probabilities and so is able to exploit the probability calculus for an account of when evidence inductively confirms a hypothesis. The probabilities one assigns to some statements – which will all fall between zero for the impossible and one for the certain – will constrain the probabilities one assigns to others, in order to preserve something analogous to deductive consistency (Howson 2000). Thus, if one statement entails another, the probability given to the conclusion must be at least as great as the probability of the premise, since if the premise is true the conclusion must be true as well. One consequence of the probability calculus is Bayes's theorem, which forms the basis for this approach to confirmation. In its near-simplest form, the theorem gives the probability of hypothesis H given evidence E in terms of the probability of E given H and the prior probabilities of H and of E:

$$P(H/E) = P(E/H) \times P(H)/P(E)$$

The Bayesian approach to confirmation exploits this formula by imagining the scientist's situation just before E is observed. The probabilities of H and of E are equated with her degree of belief in those two statements – their 'prior' probability – and the probability of H given E is then taken to be the probability she should assign to H after observing E – the 'posterior' probability of H. The basic Bayesian view is that E confirms H just in case

the posterior probability of H is higher than the prior probability of H. This is a natural thing to say, since this is the condition under which observing E raises the probability of H.

To use the Bayesian equation to calculate the posterior of H requires not just the priors of H and E; it also requires the probability of E given H. But in a situation where H entails E, then E must be true if H is, so the probability of E given H is one and posterior probability becomes a simple ratio:

$$P(H/E) = P(H)/P(E)$$

If H entails E, and the priors of H and E are neither zero nor one, the posterior of H must be greater than the prior of H, since we are dividing the prior by something less than one. Thus E will count as confirming H whenever E is a consequence of H. Moreover, this simple form of the equation makes clear that the degree of confirmation – the degree to which observing E raises the probability of H – will be greater as the prior probability of E is lower. In English, this is to say that there is high confirmation when your hypothesis entails an unlikely prediction that turns out to be correct, a very plausible claim.

The Bayesian account inevitably faces its own share of objections. These may be that the account is too permissive, too strict, or that it fails to describe a mechanism that could represent the way scientists actually determine the bearing of data on theory (Earman 1992; Howson and Urbach 1989). For example, since the Bayesian account has it that a hypothesis is confirmed by any of its logical consequences (so long as all the probabilities lie between zero and one), it seems to inherit the over-permissiveness of the hypothetico-deductive model. The account also threatens to be too strict, because of the problem of 'old evidence'. There is considerable dispute over whether evidence available before a hypothesis is formulated provides as strong confirmation as evidence only gathered afterwards to test a prediction. This is a dispute we enter in chapter 10; but it is agreed by almost all that old evidence can provide *some* confirmation. On its face, however, the Bayesian account does not allow for this, since old evidence will have a prior probability of one, and so have no effect on the posterior probability of the hypothesis. Finally, there are a series of objections to the basic ingredients of the Bayesian scheme, that beliefs do not in fact come in degrees well represented by probabilities, that there is no proper source for the values of the priors, and that in the realistic cases where the hypothesis in question does not on its own entail the evidence, there is no plausible way that the scientist has to determine the probability of the evidence given the hypothesis. We will investigate some of these difficulties in chapter 7.

We have now briefly canvassed four attempts to tackle the descriptive problem: 'More of the Same', the instantial model, the hypothetico-deductive model and the Bayesian approach. At least on this first pass, all four appear both too permissive and too strict, finding inductive support where there is

none and overlooking cases of genuine support. They do not give enough structure to the black box of our inductive principles to determine the inferences and judgments we actually make. This is not to say that these accounts describe mechanisms that would yield too many inferences: they would probably yield too few. A 'hypothetico-deductive box', for example, would probably have little or no inferential output, given the plausible additional principle that we will not make inferences we know to be contradictory. For every hypothesis that we would be inclined to infer on the basis of the deductive support it enjoys, there will be an incompatible hypothesis that is similarly supported, and the result is no inference at all, so long as both hypotheses are considered and their incompatibility recognized.

A fifth account of induction, the last I will consider in this section, focuses on causal inference. It is a striking fact about our inductive practice, both lay and scientific, that so many of our inferences depend on inferring from effects to their probable causes. This is something that Hume himself emphasized (Hume 1748: sec. IV). Causal inferences are legion, such as the doctor's inference from symptom to disease, the detective's inference from evidence to crook, the mechanic's inference from the engine noises to what is broken, and many scientific inferences from data to theoretical explanation. Moreover, it is striking that we often make a causal inference even when our main interest is in prediction. Indeed, the detour through causal theory on the route from data to prediction seems to be at the centre of many of the dramatic successes of scientific prediction. All this suggests that we might do well to consider an account of the way causal inference works as a central component of a description of our inductive practice.

The best known account of causal inference is John Stuart Mill's discussion of the 'methods of experimental inquiry' (Mill 1904: bk III, ch. VIII; cf. Hume 1739: bk I, pt 3, sec. 15). The two central methods are the Method of Agreement and especially the Method of Difference. According to the Method of Agreement, in idealized form, when we find that there is only one antecedent that is shared by all the observed instances of an effect, we infer that it is a cause (bk III, ch. VIII, sec. 1; hereafter as 'III.VIII.1'). This is how we come to believe that hangovers are caused by heavy drinking. According to the Method of Difference, when we find that there is only one prior difference between a situation where the effect occurs and an otherwise similar situation where it does not, we infer that the antecedent that is only present in the case of the effect is a cause (III.VIII.2). If we add sodium to a blue flame, and the flame turns yellow, we infer that the presence of sodium is a cause of the new color, since that is the only difference between the flame before and after the sodium was added. If we once successfully follow a recipe for baking bread, but fail another time when we have left out the yeast and the bread does not rise, we would infer that the yeast is a cause of the rising in the first case. Both methods work by a combination of retention and variation. When we apply the Method of Agreement, we hold the effect

constant, vary the background, and see what stays the same; when we apply the Method of Difference, we vary the effect, hold the background constant, and see what changes.

Mill's methods have a number of attractive features. Many of our inferences are causal inferences, and Mill's methods give a natural account of these. In science, for example, the controlled experiment is a particularly common and self-conscious application of the Method of Difference. The Millian structure of causal inference is also particularly clear in cases of inferential dispute. When you dispute my claim that C is the cause of E, you will often make your case by pointing out that the conditions for Mill's methods are not met; that is, by pointing out C is not the only antecedent common to all cases of E, or that the presence of C is not the only salient difference between a case where E occurs and a similar case where it does not. Mill's methods may also avoid some of the over-permissiveness of other accounts, because of the strong constraints that the requirements of varied or shared backgrounds place on their application. These requirements suggest how our background beliefs influence our inferences, something a good account of inference must do. The methods also help to bring out the roles in inference of competing hypotheses and negative evidence, as we will see in chapter 5, and the role of background knowledge, as we will see in chapter 8. Of course, Mill's methods have their share of liabilities, of which I will mention just two. First, they do not themselves apply to unobservable causes or to any causal inferences where the cause's existence, and not just its causal status, is inferred. Secondly, if the methods are to apply at all, the requirement that there be only a single agreement or difference in antecedents must be seen as an idealization, since this condition is never met in real life. We need principles for selecting from among multiple agreements or similarities those that are likely to be causes, but these are principles Mill does not himself supply. As we will see in later chapters, however, Mill's method can be modified and expanded in a way that may avoid these and other liabilities it faces in its simple form.

This chapter has set part of the stage for an investigation of our inductive practices. I have suggested that many of the problems those practices raised can be set out in a natural way in terms of the underdetermination that is characteristic of inductive inference. The underdetermination of our inferences by our evidence provides the skeptics with their lever, and so poses the problem of justification. It also elucidates the structure of the descriptive problem, and the black box inferences it will take to solve it. I have canvassed several solutions to the problem of description, partly to give a sense of some of our options and partly to suggest just how difficult the problem is. But at least one solution to the descriptive problem was conspicuous by its absence, the solution that gives this book its title and which will be at the center of attention from chapter 4 onwards. According to Inference to the Best Explanation, we infer what would, if true, be the best

explanation of our evidence. On this view, explanatory considerations are a guide to inference. So to develop and assess this view, we need first to look at another sector of our cognitive economy, our explanatory practices. This is the subject of the next two chapters.

Chapter 2

Explanation

Understanding explanation

Once we have made an inference, what do we do with it? Our inferred beliefs are guides to action that help us to get what we want and to avoid trouble. Less practically, we also sometimes infer simply because we want to learn more about the way the world is. Often, however, we are not satisfied to discover that something is the case: we want to know *why*. Thus our inferences may be used to provide explanations, and they may themselves be explained. The central question about our explanatory practices can be construed in several ways. We may ask what principles we use to distinguish between a good explanation, a bad explanation and no explanation at all. Or we may ask what relation is required between two things for one to count as an explanation of the other. We can also formulate the question in terms of the relationship between knowledge and understanding. Typically, someone who asks why something is the case already knows that it is the case. The person who asks why the sky is blue knows that it is blue, but does not yet understand why. The question about explanation can then be put this way: What has to be added to knowledge to yield understanding?

As in the case of inference, explanation raises problems both of justification and of description. The problem of justification can be understood in various ways. It may be seen as the problem of showing whether things we take to be good explanations really are, whether they really provide understanding. The issue here, to distinguish it from the case of inference, is not whether there is any reason to believe that our putative explanations are themselves true, but whether, granting that they are true, they really explain. When we reject the explanation that the sky is blue because the sea is blue, we do not deny the blueness of the sea. There is no argument against the possibility of explanation on a par with Hume's argument against induction. The closest thing is the why-regress. This is a feature of the logic of explanation many of us discovered as children, to our parents' cost. I vividly recall the moment it dawned on me that, whatever my mother's answer to my latest why-question, I could simply retort by asking

'Why?' of the answer itself, until even my mother ran out of answers or patience. But if only something that is itself understood can be used to explain, and understanding only comes through being explained by something else, then the infinite chain of whys makes explanation impossible. Sooner or later, we get back to something unexplained, which ruins all the attempts to explain that are built upon it (cf. Friedman 1974: 18–19).

This skeptical argument is not very troubling. One way to stop the regress is to argue for phenomena that are self-explanatory or that can be understood without explanation. But while there may be such phenomena, this reply concedes too much to the skeptical argument. A better reply is that explanations need not themselves be understood. A drought may explain a poor crop, even if we don't understand why there was a drought; I understand why you didn't come to the party if you tell me that you had a bad headache, even if I have no idea why you had a headache; the big bang explains the background radiation, even if the big bang is itself inexplicable, and so on. Understanding is not like a substance that the explanation has to possess in order to pass it on to the phenomenon to be explained. Rather than show that explanation is impossible, the regress argument brings out the important facts that explanations can be chained and that what explains need not itself be understood, and so provides useful constraints on a proper account of the nature of understanding. Any model that does not allow for a benign why-regress is suspect.

It is fairly clear why there is no skeptical argument against explanation on a par with Hume's argument against induction. Hume's argument, like all the great skeptical arguments, depends on our ability to see how our methods of acquiring beliefs could lead us into massive error. There are possible worlds where our methods persistently mislead us, and Hume exploits these possibilities by arguing that we have no non-circular way of showing that the actual world is not one of these misleading worlds. In the case of explanation, by contrast, the skeptic does not have this handle, since the issue is not whether the explanation is true, but whether the truth really explains. We do not appear to know how to make the contrast between understanding and merely seeming to understand in a way that would make sense of the possibility that most of the things that meet all our standards for explanation might nonetheless not really explain. To put the matter another way, we do not see a gap between meeting our standards for the explanation and actually understanding in the way we easily see a gap between meeting our inductive standards and making an inference that is actually correct.

It is not clear whether this is good or bad news for explanation. On the one hand, in the absence of a powerful skeptical argument, we feel less pressure to justify our practice. On the other, the absence seems to show that our grasp on explanation is even worse than our grasp on inference. We know that inferences are supposed to take us to truths and, as Hume's argument

illustrates, we at least have some appreciation of the nature of these ends independently of the means we use to try to reach them. The situation is quite different for explanation. We may say that understanding is the goal of explanation, but we do not have a clear conception of understanding apart from whatever it is our explanations provide. If this is right, the absence of powerful skeptical arguments against explanation does not show that we are in better shape here than we are in the case of inference. Perhaps things are even worse for explanation: here we may not even know what we are *trying* to do. Once we know that something is the case, what is the point of asking why?

Reason, familiarity, deduction, unification, necessity

Explanation also raises the problem of description. Whatever the point of explanation or the true nature of understanding, we have a practice of giving and judging explanations, and the problem of description is to give an account of how we do this. As with the problem of justification, the central issue here is not how we judge whether what we claim to be an explanation is true, but whether, granting that it is true, it really does explain what it purports to explain. Like our inductive practices, our explanatory practices display the gap between doing and describing. We discriminate between things we understand and things we do not, and between good explanations and bad explanations, but we are strikingly poor at giving any sort of principled account of how we do this. As before, the best way to make this claim convincing is to canvass some of the objections to various popular accounts of explanation. I will consider briefly five accounts.

According to the reason model of explanation, to explain a phenomenon is to give a reason to believe that the phenomenon occurs (Hempel 1965: 337, 364–76). On this view, an engineer's explanation of the collapse of a bridge succeeds by appealing to theories of loading, stress and fatigue which, along with various particular facts, show that the collapse was likely. There is a germ of truth in this view, since explanations do quite often make the phenomenon likely and give us a reason to believe it occurs. A particularly satisfying type of explanation takes a phenomenon that looks accidental and shows how, given the conditions, it was really inevitable, and these deterministic explanations do seem to provide strong reasons for belief. Moreover, the reason model suggests a natural connection between the explanatory and predictive uses of scientific theories since, in both cases, the theory would work by providing grounds for belief.

On balance, however, the reason model is extremely implausible. It does not account for the central difference between knowing that a phenomenon occurs and understanding why it occurs. The model claims that understanding why the phenomenon occurs is having a reason to believe that it occurs, but typically we already have this when we know that it occurs. We

already have a reason to believe the bridge collapsed when we ask why it did, so we can not simply be asking for a reason for belief when we ask for an explanation. Explanations may provide reasons for belief, but that is not enough. It is also often too much: many explanations do not provide any actual reason to believe that the phenomenon occurs. Suppose you ask me why there are certain peculiar tracks in the snow in front of my house. Looking at the tracks, I explain to you that a person on snowshoes recently passed this way. This is a perfectly good explanation, even if I did not see the person and so an essential part of my reason for believing my explanation are the very tracks whose existence I am explaining. Similarly, an astronomer may explain why the characteristic spectrum of a particular galaxy is shifted towards the red by giving its velocity of recession, even if an essential part of the evidence for saying that the galaxy is indeed moving away from us at that speed is the very red-shift that is being explained. These 'self-evidencing explanations' (Hempel 1965: 370–4) have a distinctive circularity: the person passing on snowshoes explains the tracks and the tracks provide the evidence for the passing. What is significant is that the circularity is benign: it spoils neither the explanation of the tracks nor the justification for the belief that someone did pass on snowshoes, neither the explanation of the red-shift nor the justification for the claim that the galaxy moves with that velocity. Self-evidencing explanations do, however, show that the reason model of explanation is untenable, since to take the explanation to provide a reason to believe the phenomenon after the phenomenon has been used as a reason to believe the explanation would be vicious. In other words, if the reason model were correct, self-evidencing explanations would be illicit, but self-evidencing explanations may be perfectly acceptable and are indeed ubiquitous, as we will see in chapter 4, so the reason model is wrong. Providing reasons for belief is neither necessary nor sufficient for explanation.

Another answer to the descriptive question is the familiarity model. On this view, unfamiliar phenomena call for explanation, and good explanations somehow make them familiar (Hempel 1965: 430–3; Friedman 1974: 9–11). On one version of this view, there are certain familiar phenomena and processes that we do not regard as in need of explanation, and a good explanation of an unfamiliar phenomenon consists in showing it to be the outcome of a process that is analogous to the processes that yield the familiar phenomena. The kinetic theory of gases explains various phenomena of heat by showing that gases behave like collections of tiny billiard balls; Darwin's theory of natural selection explains the traits of plants and animals by describing a mechanism similar to the mechanism of artificial selection employed by animal breeders; and an electronic theory explains by showing that the flow of current through a wire is like the flow of water through pipes. But this is not a particularly attractive version of the familiarity view, in part because it does not help us to understand why certain phenomena are

familiar in the first place, and because not all good explanations rely on analogies.

A more promising version of the familiarity view begins with the idea that a phenomenon is unfamiliar when, although we may know that it occurs, it remains surprising because it is in tension with other beliefs we hold. A good explanation shows how the phenomenon arises in a way that eliminates the tension and so the surprise (Hempel 1965: 428–30). We may know that bats navigate with great accuracy in complete darkness, yet find this very surprising, since it seems in tension with our belief that vision is impossible in the dark. Finding out about echolocation shows that there is no real tension, and we are no longer surprised. The magician tells me the number I was thinking of, to my great surprise; a good explanation of the trick ruins it by making it unsurprising. This version of the familiarity view has the virtue of capturing the fact that it is very often surprise that prompts the search for explanations. It also calls our attention to the process of 'defamiliarization', which is often the precursor to asking why various common phenomena occur. In one sense, the fact that the sky is blue is a paradigm of familiarity, but we become interested in explaining this when we stop to consider how odd it is that the sky should have any color at all. Again, the fact that the same side of the moon always faces us at first seems not to call for any interesting explanation, since it just seems to show that the moon is not spinning on its own axis. It is only when we realize that the moon must actually spin in order to keep the same side towards us, and moreover with a period that is exactly the same as the period of its orbit around the earth, that this phenomenon cries out for explanation. The transformation of an apparently familiar phenomenon into what seems an extraordinary coincidence prompts the search for an adequate explanation. The surprise version of the familiarity model also suggests that a good explanation of a phenomenon will sometimes show that beliefs that made the phenomenon surprising are themselves in error, and this is an important part of our explanatory practice. If I am surprised to see a friend at the supermarket because I expected him to be away on vacation, he will not satisfy my curiosity simply by telling me that he needed some milk, but must also say something about why my belief about his travel plans was mistaken.

There are three objections that are commonly made to familiarity models of explanation. One is that familiarity is too subjective to yield a suitably objective account of explanation. The surprise version of the familiarity model does make explanation audience-relative, since what counts as a good explanation will depend on prior expectations, which vary from person to person. But it is not clear that this is a weakness of the model. Nobody would argue that an account of inference is unacceptable because it makes warranted inductions vary with prior belief. Similarly, an account that makes explanation 'interest-relative' does not thereby make explanation perniciously subjective. (We will return to the interest relativity of explanation

in the next chapter.) In particular, the familiarity model does not collapse the distinction between understanding a phenomenon and mistakenly thinking one does. The explanation must itself be true, and it must say something about how the phenomenon actually came about, but just what it says may legitimately depend on the audience's interests and expectations.

More telling are explanation by the unfamiliar and explanation of the familiar. Good explanations often themselves appeal to unfamiliar events and processes, especially in the sciences. And we often explain familiar phenomena. An appeal to the process of defamiliarization only partially meets this objection. The rattle in my car is painfully familiar, and consistent with everything else I believe, but while I am sure there is a good explanation for it, I don't have any idea what it is. Nor do you have to convince me that, in fact, it is somehow surprising that there should be a rattle, in order to get me interested in explaining it. Surprise is often a precursor to the search for explanation, but it is not the only motivation. A reasonable version of the familiarity theory has more going for it than some of its critics suppose, but does not by itself provide an adequate description of our explanatory practices.

The third and best known account of explanation is the deductive-nomological model, according to which we explain a phenomenon by deducing it from a set of premises that includes at least one law that is necessary to the deduction (Hempel 1965: 335–76). The case of the galaxy illustrates this. We can explain why a particular galaxy has its characteristic spectrum shifted towards the red by deducing this shift from the speed at which the galaxy is receding from us, and the Doppler law that links recession and red-shift. (This Doppler effect is similar to the change in pitch of a train whistle as the train passes by.) Of course many scientific explanations and most lay explanations fail the strict requirements of the model, since they either do not contain exceptionless laws or do not strictly entail the phenomena, but they can be seen as 'explanation sketches' (Hempel 1965: 423–4) that provide better or worse approximations to a full deductive-nomological explanation.

The deductive-nomological model is closely related to the reason model, since the premises that enable us to deduce the phenomenon often also provide us with a reason to believe that the phenomenon occurs, but it avoids some of the weaknesses of the reason model. For example, unlike the reason model, the deductive-nomological model allows for self-evidencing explanations. As we have seen, the explanation of red-shift in terms of recession satisfies the deductive-nomological model even though, as it happens, the red-shift is itself crucial evidence for the speed of recession that explains it. That is, the Doppler explanation is self-evidencing. The deductive-nomological model also does better than the reason model in accounting for the difference between knowing and understanding. When we know that a phenomenon occurs but do not understand why, we usually do

not know laws and supporting premises that entail the phenomenon. So at least a deductive-nomological argument usually gives us something new. Also to the model's credit, it does seem that when theories are used to give scientific explanations, these explanations often do aspire to deductive-nomological form. Moreover, the model avoids the main objection to the familiarity model, since a phenomenon may be common and unsurprising, but still await a deductive-nomological explanation.

The model nevertheless faces debilitating objections. It is almost certainly too strong: very few explanations fully meet the requirements of the model and, while some scientific explanations at least aspire to deductive-nomological status, many ordinary explanations include no laws and allow no deduction, yet are not incomplete or mere sketches. The model is also too weak. Perhaps the best known objection is that it does not account for the asymmetries of explanation (Bromberger 1966; Friedman 1974: 8; van Fraassen 1980: 112). Consider again the Doppler explanation of the red-shift. In that explanation, the law is used to deduce the shift from the recession, but the law is such that we could equally well use it to deduce the recession from the shift. Indeed this is how we figure out what the recession is. What follows is that, according to the deductive-nomological model, we can explain why the galaxy is receding by appeal to its red-shift. But this is wrong: the shift no more explains the recession than a wet sidewalk explains why it is raining. The model does not account for the many cases where there is explanatory asymmetry but deductive symmetry.

The deductive-nomological model should produce a sense of *déjà vu*, since it is isomorphic to the hypothetico-deductive model of confirmation, which we considered in the last chapter. In one case we explain a phenomenon by deducing it from a law; in the other we show that the evidence confirms a hypothesis by deducing the evidence from the hypothesis. That deduction should play a role both in explanation and in inductive support is not itself suspicious, but the isomorphism of the models suggests that weaknesses of one may also count against the other, and this turns out to be the case. As we saw, the main weakness of the hypothetico-deductive model is that it is over-permissive, counting almost any datum as evidence for almost any hypothesis. The deductive-nomological model similarly makes it far too easy to explain (Hempel 1965: 273, n. 33, 293–4). This comes out most clearly if we consider the explanation of a general phenomenon, which is itself described by a law. Suppose, for example, that we wish to explain why the planets move in ellipses. According to the deductive-nomological model, we can 'explain' the ellipse law by deducing it from the conjunction of itself and any law you please, say a law in economics. The model also suffers from a problem analogous to the raven paradox. We may explain why an object was warmed by pointing out that it was in the sun and everything warms when in the sun, but we cannot explain why an object was not in the sun by pointing out that it was not warmed. Just

as the hypothetico-deductive model leaves the class of hypotheses confirmed by the available data dramatically underdetermined, so the deductive-nomological model underdetermines the class of acceptable explanations for a given phenomenon.

I will be brief with the fourth and fifth models of explanation, though both could easily support an extended discussion. According to the unification model, we come to understand a phenomenon when we see how it fits together with other phenomena into a unified whole (Friedman 1974; Kitcher 1989). This conception chimes with the ancient idea that to understand the world is to see unity that underlies the apparent diversity of the phenomena. The unification conception allows for both the gap between knowledge and understanding and the legitimacy of self-evidencing explanations without difficulty. We can know that something is the case without yet being able to fit it together appropriately with other things we know, so there can be knowledge without understanding. Self-evidencing explanations are also accounted for, since a piece of a pattern may provide evidence for the pattern as a whole, while the description of the whole pattern places the piece in a unifying framework.

One salient difficulty for the unification model is that the notion of unification turns out to be surprisingly difficult to analyze. Certainly our analytic grip on this notion turns out to be far weaker than our grip on deduction. One source of the difficulty is that while it is usually supposed that unification is a metaphysical notion, depending on how things in the world actually fit together, it is difficult to come up with an analysis that avoids making unification a function of the vocabulary with which we describe the world. A second possible objection to the unification model is that it does not allow sufficiently for the why-regress. Presumably a unifying explanation is itself unified, so there seems to be no room for explanations that we do not already understand. But this is not clear. For one might say that to explain a phenomenon is to embed it appropriately into a *wider* pattern. In this case a hypothesis might suitably embed some evidence, even though we have no wider pattern in which to embed the hypothesis, and the requirements of the why-regress would be satisfied. A third objection to the unification model is that it does not seem to account for what is perhaps the most common sort of explanation, when a singular effect is explained by saying something about its causes, since such singular causal explanations do not seem to provide any strong form of unification. Unification is an explanatory virtue, a point we will return to in chapter 8, but it does not promise an adequate general model of explanation.

The final model to be canvassed in this chapter appeals to necessity. According to the necessity model, an explanation shows that the phenomenon in question *had* to occur (Glymour 1980a). This conception of understanding acknowledges the gap between knowing that and understanding why, since one may know that something did in fact occur

without knowing that it had to occur. The necessity model also appears to allow for self-evidencing explanations, since there seems to be no vicious circularity involved in supposing that one thing shows another to be in some sense necessary while the latter gives a reason for believing the former. More generally, many explanations do appear to work by showing some kind of necessity or eliminating some kind of apparent contingency. Both the appeal to law and to the relation of entailment in the deductive-nomological model can be seen as attempts to capture the role of necessity in explanation, but given the weaknesses of that model it may well be better to appeal to necessity more directly.

On the debit side, the necessity model faces a number of objections similar to those that face the appeal to unification. Our grasp on the notion of logical necessity is pretty good, as philosophical grips go, but this notion of necessity is too strong for a general model of explanation. Outside explanations in pure mathematics, it is at least very rare for explanations to show that it is logically impossible for the phenomenon in question to have been otherwise. And weaker notions of necessity resist analysis. It is also unclear whether the necessity model passes the test of the why-regress. It fails if only what is itself necessary can confer necessity, or if only what is already known to be necessary can be used to show that something else is necessary too. Finally, many mundane causal explanations seem not to depend on displaying even a relatively weak necessity. Why did we abandon last week's Sunday morning football game? Because nobody remembered to bring a ball. Even if we live in a strongly deterministic world where everything really happens by some serious necessity, many of the explanations we actually give and accept do not need to reveal it.

I conclude that none of the five models we have briefly considered gives an adequate general description of our explanatory practices. Each of them captures distinctive features of certain explanations: some explanations provide reasons for belief, some make the initially unfamiliar or surprising familiar, some are deductions from laws, some show how things fit together, and some show that things had to happen the way they did. But there are explanations that have none of these features, and something can have all of these features without being an acceptable explanation. I want now to consider a sixth model of explanation, the causal model. This model has its share of difficulties, but I believe it is more promising than the other five. I also can offer a modest development of the model that improves its descriptive adequacy and that will make it an essential tool for the discussion of Inference to the Best Explanation to follow. For these reasons, the causal model deserves a chapter of its own.

Chapter 3

The causal model

Fact and foil

According to the causal model of explanation, to explain a phenomenon is simply to give information about its causal history (Lewis 1986) or, where the phenomenon is itself a causal regularity, to explain it is to give information about the mechanism linking cause and effect. If we explain why smoking causes cancer, we do not give a cause of this causal connection, but we do give information about the causal mechanism that makes it. Not only is the causal model of explanation natural and plausible, but it avoids many of the problems that beset the other views we have canvassed. It provides a clear distinction between understanding why a phenomenon occurs and merely knowing that it does, since it is possible to know that a phenomenon occurs without knowing what caused it. Moreover, the model draws this distinction in a way that makes understanding unmysterious and objective. Understanding is not some sort of super-knowledge, but simply more knowledge: knowledge of causes.

Unlike the unification and necessity models of explanation, the causal model makes it clear how something can explain without itself being explained or already understood, and so has no difficulty accounting for the possibility of explanation in the face of the regress of whys. One can know a phenomenon's cause without knowing the cause of that cause. And unlike the reason model, which requires that an explanation provide a reason to believe the phenomenon occurs, the causal model accounts for the legitimacy of self-evidencing explanations, where the phenomenon being explained is also an essential part of the evidence for the explanation. The causal model also avoids the most serious objection to the familiarity model, since a phenomenon can be common and unsurprising, even though we do not know its cause. Finally, it avoids many of the objections to the deductive-nomological model. Ordinary explanations do not have to meet the requirements of that model, because one need not give a law to give a cause, and one need not know a law to have good reason to believe that a cause is a cause. As for the over-permissiveness of the deductive-

nomological model, the reason recession explains red-shift but not conversely is that causes explain effects and not conversely; the reason a conjunction does not explain its conjuncts is that conjunctions do not cause their conjuncts; and the reason the sun explains the warmth, while not being warmed does not explain not being in the sun, is that the sun causes an object to warm, but not being warmed does not cause something to be in the shade.

There are three natural objections to the causal model of explanation. The first is that we do not have a fully adequate analysis of causation, and not through want of trying (cf. Sosa and Tooley 1993). This, however, is no reason to abjure the model. The notion of causation is indispensable to philosophy, ordinary life and much of science, we know a good deal about the relation without a full philosophical account, and if we wait for a fully adequate analysis of causation before we use it to analyze other things we will probably wait forever. I will not, in this book, say anything on the large topic of the nature of causation, but trust that what I do say about the role of causation in explanation and inference holds for the causal relation as it is. Applying the moral of the why-regress reflexively, we do not need fully to explain causation in order to use causation to explain other things, in this case, the nature of explanation itself.

The second objection to the causal model of explanation is simply that there are non-causal explanations. Mathematicians and philosophers, for example, give explanations, but mathematical explanations are never causal, and philosophical explanations seldom are. A mathematician may explain why Godel's Theorem is true, and a philosopher may explain why there can be no inductive justification of induction, but these are not explanations that cite causes. (Some philosophical explanations are, however, broadly causal, such as the explanations of inferential and explanatory practices that we are considering in this book.) There are even physical explanations that seem non-causal. I am particularly fond of two examples. First, suppose that a bunch of sticks are thrown into the air with a lot of spin, so that they separate and tumble about as they fall. Now freeze the scene at a moment during the sticks' descent. Why are appreciably more of them near the horizontal axis than near the vertical, rather than in more or less equal numbers near each orientation as one might have expected? The answer, roughly speaking, is that there are many more ways for a stick to be near the horizontal than near the vertical. To see this, consider purely horizontal and vertical orientations for a single stick with a fixed midpoint. There are indefinitely many horizontal orientations, but only two vertical orientations. Or think of the shell that the ends of that stick trace as it takes every possible orientation. The areas that correspond to near the vertical are caps centered on the north and south poles formed when the stick is forty-five degrees or less off the vertical, and this area is substantially less than half the surface area of the entire sphere. Another way of putting it is that the explanation why more sticks are near the horizontal than near the vertical is that there are two

horizontal dimensions but only one vertical one. This is a lovely explanation, but apparently not a causal one, since geometrical facts cannot be causes.

My second example of a lovely non-causal explanation concerns reward and punishment, and is based on the influential work in cognitive psychology by Daniel Kahneman and Amos Tversky (Kahneman *et al.* 1982: 66–8), work we will return to in chapters 7 and 8. Flight instructors in the Israeli air force had a policy of strongly praising trainee pilots after an unusually good performance and strongly criticizing them after an unusually weak performance. What they found is that trainees tended to improve after a poor performance and criticism; but they actually tended to do worse after good performance and praise. What explains this pattern? Perhaps it is that criticism is much more effective than praise. That would be a causal explanation. But this pattern is also what one should expect if neither praise nor criticism had any effect. It may just be regression to the mean: extreme performances tend to be followed by less extreme performances. If this is what is going on, we can have a lovely explanation of the observed pattern by appeal to chance (or the absence of causal influence) rather than any cause. (This example ought to give pause to parents who are too quick to infer that punishing children for bad behavior is more effective than rewarding them for good behavior.)

The existence of non-causal explanations show that a causal model of explanation cannot be complete. One reaction to this would be to attempt to expand the notion of causation to some broader notion of 'determination' that would encompass the non-causal cases (Ruben 1990: 230–3). This approach has merit, but it will be difficult to come up with such a notion that we understand even as well as causation, without falling into the relation of deductive determination, which will expose the model to many of the objections to the deductive-nomological model. For the time being at least, I believe that the causal view is still our best bet, because of the backward state of alternate views of explanation, and the overwhelming preponderance of causal explanations among all explanations. Nor does it seem ad hoc to limit our attention to causal explanations. The causal view does not simply pick out a feature that certain explanations happen to have: causal explanations are explanatory *because* they are causal.

The third objection is that the causal model is too weak or permissive, that it underdetermines our explanatory practices. Let us focus on the causal explanation of particular events. We may explain an event by giving some information about its causal history, but causal histories are long and wide, and most causal information does not provide a good explanation. The big bang is part of the causal history of every event, but explains only a few. The spark and the oxygen are both part of the causal history that led up to the fire, but only one of them explains it. In a particular context, most information about the causal history of a phenomenon is explanatorily irrelevant, so explaining cannot simply be giving such information. This is an important

objection, but I prefer to see it as a challenge. How can the causal model be developed to account for the causal selectivity of our explanatory practices? The rest of this chapter is a partial answer to this question. The answer is interesting in its own right, and it will also turn out to be a crucial tool for developing and assessing an account of Inference to the Best Explanation, the central project of this book.

What makes one piece of information about the causal history of an event explanatory and another not? The short answer is that the causes that explain depend on our interests. But this does not yield a very informative model of explanation unless we can go some way towards spelling out how explanatory interests determine explanatory causes. One natural way to show how interests help us to select from among causes is to reveal additional structure in the phenomenon to be explained, structure that varies with interest and that points to particular causes. The idea here is that we can account for the specificity of explanatory answer by revealing the specificity in the explanatory question, where a difference in interest is an interest in explaining different things. Suppose we started by construing a phenomenon to be explained simply as a concrete event, say a particular eclipse. The number of causal factors is enormous. As Carl Hempel has observed, however, we do not explain events, only aspects of events (1965: 421–3). We do not explain the eclipse *tout court*, but only why it lasted as long as it did, or why it was partial, or why it was not visible from a certain place. Which aspect we ask about depends on our interests, and reduces the number of causal factors we need consider for any particular phenomenon, since there will be many causes of the eclipse that are not, for example, causes of its duration. More recently, it has been argued that the interest relativity of explanation can be accounted for with a contrastive analysis of the phenomenon to be explained. What gets explained is not simply 'Why this?', but 'Why this *rather than* that?' (Garfinkel 1981: 28–41; van Fraassen 1980: 126–9). A contrastive phenomenon consists of a fact and a foil, and the same fact may have several different foils. We may not explain why the leaves turn yellow in November *simpliciter*, but only for example why they turn yellow in November rather than in January, or why they turn yellow in November rather than turn blue.

The contrastive analysis of explanation is extremely natural. We often pose our why-questions explicitly in contrastive form and it is not difficult to come up with examples where different people select different foils, requiring different explanations. When I asked my, then, 3-year old son why he threw his food on the floor, he told me that he was full. This may explain why he threw it on the floor rather than eating it, but I wanted to know why he threw it rather than leaving it on his plate. An explanation of why I went to see *Jumpers* rather than *Candide* will probably not explain why I went to see *Jumpers* rather than staying at home, an explanation of why Able rather than Baker got the philosophy job may not explain why Able rather than

Charles got the job, and an explanation of why the mercury in a thermometer rose rather than fell may not explain why it rose rather than breaking the glass. The proposal that phenomena to be explained have a complex fact–foil structure can be seen as another step along Hempel's path of focusing explanation by adding structure to the why-question. A fact is often not specific enough: we also need to specify a foil. Since the causes that explain a fact relative to one foil will not generally explain it relative to another, the contrastive question provides a further restriction on explanatory causes.

The role of contrasts in explanation will not account for all the factors that determine which cause is explanatory. For one thing, I do not assume that all why-questions are contrastive. For another, even in the cases of contrastive questions, the choice of foil is not, as we will see, the only relevant factor. Nevertheless, it does provide a central mechanism, so I want to try to show in some detail how contrastive questions help select explanatory causes. My discussion will fall into three parts. First, I will make three general observations about contrastive explanation. Then, I will use these observations to show why contrastive questions resist reduction to non-contrastive form. Finally, I will describe the mechanism of 'causal triangulation' by which the choice of foils in contrastive questions helps to select explanatory causes.

When we ask a contrastive why-question – 'Why the fact rather than the foil?' – we presuppose that the fact occurred and that the foil did not. Often we also suppose that the fact and the foil are in some sense incompatible. When we ask why Kate rather than Frank won the Philosophy Department Prize, we suppose that they could not both have won. Similarly, when we asked about leaves, we supposed that if they turn yellow in November, they cannot turn yellow in January, and if they turn yellow in November they cannot also turn blue then. Indeed, it is widely supposed that fact and foil are always incompatible (Garfinkel 1981: 40; Temple 1988: 144; Ruben 1987). My first observation is that this is false: many contrasts are compatible. We often ask a contrastive question when we do not understand why two apparently similar situations turned out differently. In such a case, far from supposing any incompatibility between fact and foil, we ask the question just because we expected them to turn out the same. By the time we ask the question, we realize that our expectation was disappointed, but this does not normally lead us to believe that the fact precluded the foil, and the explanation for the contrast will usually not show that it did. Consider the much discussed example of syphilis and paresis (Scriven 1959: 480; Hempel 1965: 369–70; van Fraassen 1980: 128). Few with syphilis contract paresis, but we can still explain why Jones rather than Smith contracted paresis by pointing out that only Jones had syphilis. In this case, there is no incompatibility. Only Jones contracted paresis, but they both could have: Jones's affliction did not protect Smith. Of course, not every pair of compatible fact and foil would yield a sensible why-question but, as we will

see, it is not necessary to restrict contrastive why-questions to incompatible contrasts to distinguish sensible questions from silly ones.

The existence of compatible contrasts is somewhat surprising, since the 'rather than' construction certainly appears to suggest some sort of incompatibility (Carroll 1997, 1999). I think there are a number of reasons why we find the 'rather than' construction natural even when the P and Q in 'Why P rather than Q?' are compatible. As I have already mentioned, we often ask a contrastive question when cases we expected to turn out the same in fact turn out differently. In this case, though fact and foil are compatible, we are also interested in the contrast between turning out the same and turning out differently, and *this* contrast is obviously incompatible. In other cases, when we ask 'Why P rather than Q?' we may also be interested in why things turned out one way rather than the other way around – the contrast between (P & not-Q) and (Q & not-P) – which is again an incompatible contrast, even though P and Q are compatible (Jones 1994). A third underlying incompatible contrast in cases of compatible fact and foil is between the fact and its negation: the foil Q is a device for asking a certain kind of question about the contrast between P and not-P. In ways we will investigate, the choice of foil serves to focus on which aspect of the fact we are interested in explaining, in effect by specifying which way of the fact not occurring is of interest. The foil provides, in Alan Garfinkel's words, a 'limited negation' (1981: 30).

So we see there are a number of reasons why the suggestion of incompatibility carried by 'rather than' is apt even when fact and foil are compatible, since there may be incompatible contrasts at play beneath the surface. Now in such cases one could take the terms of those underlying contrasts to be the 'real' facts and foils, so making the questions about an incompatible contrast after all, but I prefer to hold on to P and Q in the surface structure of the question as fact and foil, and these will often be compatible. This way of proceeding gives us a univocal structure for contrastive questions and will make my subsequent analysis more perspicuous, but this is not to deny that the underlying contrasts are real as well.

So that is my first observation: fact and foil may be compatible. My second and third observations concern the relationship between an explanation of the contrast between a fact and foil and the explanation of the fact alone. I do not have a general account of what it takes to explain a fact on its own. As we will see, this is not necessary to give an account of what it takes to explain a contrast; indeed, this is one of the advantages of a contrastive analysis. Yet, based on our intuitive judgments of what is and what is not an acceptable or decent explanation of a fact alone, we can see that the requirements for explaining a fact are different from the requirements for explaining a contrast. Of course intuitions about particular non-contrastive explanations may vary, so that for example what to one

person seems no explanation at all will to another seem a genuine though very weak explanation. My hope is that this will not matter in what follows. A good explanation of P rather than Q that is no explanation at all of P alone is a particularly dramatic difference, but for my purposes it is probably enough if you find that the one explanation is considerably better than the other, even if in my presentation I speak in starker terms.

My second observation, then, is that explaining a contrast is sometimes easier than explaining the fact alone, in the sense that an explanation of 'P rather than Q' is not always an explanation of P (cf. Garfinkel 1981: 30). This is particularly clear in examples of compatible contrasts. Jones's syphilis does not, it seems to me, explain why he got paresis, since the vast majority of people who get syphilis do not get paresis, but it does explain why Jones rather than Smith got paresis, since Smith did not have syphilis. (For a different view, see Carroll 1999.) The relative ease with which we explain some contrasts also applies to many cases where there is an incompatibility between fact and foil. My preference for contemporary plays may not explain why I went to see *Jumpers* last night, since it does not explain why I went out, but it does explain why I went to see *Jumpers* rather than *Candide*. A particularly striking example of the relative ease with which some contrasts can be explained is the explanation that I chose A rather than B because I did not realize that B was an option. If you ask me why I ordered eggplant rather than sea bass (a 'daily special'), I may give the perfectly good answer that I did not know there were any specials; but this would not be a very good answer to the simple question, 'Why did you order eggplant?' (even if one does not hold that only sufficient causes explain non-contrastive facts (cf. Hitchcock 1999: 603–4)). One reason we can sometimes explain a contrast without explaining the fact alone seems to be that contrastive questions incorporate a presupposition that makes explanation easier. To explain 'P rather than Q' is to give a certain type of explanation of P, *given* 'P or Q', and an explanation that succeeds with the presupposition will not generally succeed without it.

My final observation is that explaining a contrast is also sometimes harder than explaining the fact alone. An explanation of P is not always an explanation of 'P rather than Q'. This is obvious in the case of compatible contrasts: we cannot explain why Jones rather than Smith contracted paresis without saying something about Smith. But it also applies to incompatible contrasts. To explain why I went to *Jumpers* rather than *Candide*, it is not enough for me to say that I was in the mood for a philosophical play. To explain why Kate rather than Frank won the prize, it is not enough that she wrote a good essay; it must have been better than Frank's. One reason that explaining a contrast is sometimes harder than explaining the fact alone is that explaining a contrast requires giving causal information that distinguishes the fact from the foil, and information that we accept as an

explanation of the fact alone may not do this, since it may not include information about the foil.

Failed reductions and false differences

There have been a number of attempts to reduce contrastive questions to non-contrastive and generally truth-functional form. One motivation for this is to bring contrastive explanations into the fold of the deductive-nomological model since, without some reduction, it is not clear what the conclusion of a deductive explanation of 'P rather than Q' ought to be. Armed with our three observations – that contrasts may be compatible, and that explaining a contrast is both easier and harder than explaining the fact alone – we can show that contrastive questions resist a reduction to non-contrastive form. We have already seen that the contrastive question 'Why P rather than Q?' is not equivalent to the simple question 'Why P?', where two why-questions are explanatorily equivalent just in case any adequate answer to one is an adequate answer to the other. One of the questions may be easier or harder to answer than the other. Still, a proponent of the deductive-nomological model of explanation may be tempted to say that, for incompatible contrasts, the question 'Why P rather than Q?' is equivalent to 'Why P?'. But it is not plausible to say that a deductive-nomological explanation of P is generally necessary to explain 'P rather than Q'. More interestingly, a deductive-nomological explanation of P is not always sufficient to explain 'P rather than Q', for any incompatible Q. Imagine a typical deductive explanation for the rise of mercury in a thermometer. Such an explanation would explain various contrasts, for example why the mercury rose rather than fell. It may not, however, explain why the mercury rose rather than breaking the glass. A full deductive-nomological explanation of the rise will have to include a premise saying that the glass does not break, but it does not need to explain this.

Another natural suggestion is that the contrastive question 'Why P rather than Q?' is equivalent to the conjunctive question 'Why P and not-Q?'. On this view, explaining a contrast between fact and foil is tantamount to explaining the conjunction of the fact and the negation of the foil (Temple 1988; Carroll 1997, 1999). In ordinary language, a contrastive question is often equivalent to its corresponding conjunction, simply because the 'and not' construction is often used contrastively. Instead of asking, 'Why was the prize won by Kate rather than by Frank?', the same question could be posed by asking 'Why was the prize won by Kate and not by Frank?'. But this colloquial equivalence does not seem to capture the point of the conjunctive view. To do so, the conjunctive view should be taken to entail that explaining a conjunction at least requires explaining each conjunct; that an explanation of 'P and not-Q' must also provide an explanation of P and an explanation of not-Q. Thus, on the conjunctive view, to explain why Kate

rather than Frank won the prize at least requires an explanation of why Kate won it and an explanation of why Frank did not.

The conjunctive view falls to the observation that explaining a contrast is sometimes easier than explaining the fact alone, since explaining P and explaining not-Q is at least as difficult as explaining P. If your horse is lame and mine isn't, that explains why my horse won rather than yours, but it does not explain why my horse won, there being many other fit horses in the race. The conjunctive view makes contrastive explanation too hard. Somewhat surprisingly, it also makes it too easy, on any model of explanation that is deductively closed. A model is deductively closed if it entails that an explanation of P will also explain any logical consequence of P. (The deductive-nomological model is nearly but not entirely closed, since it requires that the law premise be essential to the deduction and this condition will not be satisfied by every consequence of P.) Consider cases where the fact is logically incompatible with the foil. Here P entails not-Q, so the conjunction 'P and not-Q' is logically equivalent to P alone. Furthermore, all conjunctions whose first conjunct is P and whose second conjunct is logically incompatible with P will be equivalent to each other, since they are all logically equivalent to P. Hence, for a deductively closed model of explanation, explaining 'P and not-Q' is tantamount to explaining P, whatever Q may be, so long as it is incompatible with P. We have seen, however, that explaining 'P rather than Q' is not generally tantamount to explaining P, and that an explanation of P relative to one contrast is not in general an explanation of P relative to another. The conjunction in these cases is explanatorily equivalent to P, and the contrast is not, so the conjunction is not equivalent to the contrast.

The failure to represent a contrastive phenomenon by the fact alone or by the conjunction of the fact and the negation of the foil suggests that, if we want a non-contrastive paraphrase, we ought instead to try something logically weaker than the fact. In some cases it does seem that an explanation of the contrast is really an explanation of a logical consequence of the fact. This is closely related to what Hempel has to say about 'partial explanation' (1965: 415–18). He gives the example of Freud's explanation of a particular slip of the pen that resulted in writing down the wrong date. Freud explains the slip with his theory of wish-fulfillment, but Hempel objects that the explanation does not really show why that particular slip took place, but at best only why there was some wish-fulfilling slip or other. Freud gave a partial explanation of the particular slip, since he gave a full explanation of the weaker claim that there was some slip. Hempel's point fits naturally into contrastive language: Freud did not explain why it was this slip rather than another wish-fulfilling slip, though he did explain why it was this slip rather than no slip at all. And it seems natural to analyze 'Why this slip rather than no slip at all?' as 'Why some slip?'. In general, however, we cannot paraphrase contrastive questions with consequences of their facts. We

cannot, for example, say that to explain why the leaves turn yellow in November rather than in January is just to explain why the leaves turn (some color or other) in November. This attempted paraphrase fails to discriminate between the intended contrastive question and the question, 'Why do the leaves turn in November rather than fall right off?'. Similarly, we cannot capture the question, 'Why did Jones rather than Smith get paresis?', by asking about some consequence of Jones's condition, such as why he contracted a disease.

A general problem with finding a paraphrase entailed by the fact P is that, as we have seen, explaining a contrast is sometimes harder than explaining P alone. There are also problems peculiar to the obvious candidates. The disjunction 'P or Q' will not do: explaining why I went to *Jumpers* rather than *Candide* is not the same as explaining why I went to either. Indeed, this proposal gets things almost backwards: the disjunction is what the contrastive question assumes, not what calls for explanation. This suggests, instead, that the contrast is equivalent to the conditional, 'if P or Q, then P' or, what comes to the same thing if the conditional is truth-functional, to explaining P on the assumption of 'P or Q'. Of all the reductions we have considered, this proposal is the most promising, but I do not think it will do. On a deductive model of explanation it would entail that any explanation of not-Q is also an explanation of the contrast, which is incorrect. We cannot explain why Jones rather than Smith has paresis by explaining why Smith did not get it. It would also wrongly entail that any explanation of P is an explanation of the contrast, since P entails the conditional.

By asking a contrastive question, we can achieve a specificity that we do not seem to be able to capture either with a non-contrastive sentence that entails the fact or with one that the fact entails. But how then does a contrastive question specify the sort of information that will provide an adequate answer? It now appears that looking for a non-contrastive reduction of 'P rather than Q' is not a useful way to proceed. The contrastive claim may entail no more than 'P and not-Q' or perhaps better, 'P but not-Q', but explaining the contrast is not the same as explaining these conjuncts. We will do better to leave the analysis of the contrastive question to one side, and instead consider directly what it takes to provide an adequate answer. Intuitively, it seems that to explain a contrast requires citing a cause of the fact that marks a difference between fact and foil. But how is this difference to be analyzed? In the remainder of this section we consider two approaches that do not seem quite right; in the next section I shall try to do better.

David Lewis has given an interesting account of contrastive explanation that does not depend on paraphrasing the contrastive question and that does give one sense of a cause marking a difference between fact and foil. According to him, we explain why event P occurred rather than event Q by giving information about the causal history of P that would not have applied

to the history of Q, if Q had occurred (Lewis 1986: 229–30). Roughly, we cite a cause of P that would not have been a cause of Q. In Lewis's example, we can explain why he went to Monash rather than to Oxford in 1979 by pointing out that only Monash invited him, because the invitation to Monash was a cause of his trip, and that invitation would not have been a cause of a trip to Oxford, if he had taken one. On the other hand, Lewis's desire to go to places where he has good friends would not explain why he went to Monash rather than Oxford, since he has friends in both places and so the desire would have been part of either causal history.

Lewis's counterfactual account, however, is too weak: it allows for unexplanatory causes. Suppose that both Oxford and Monash had invited him, but he went to Monash anyway. On Lewis's account, we can still explain this by pointing out that Monash invited him, since *that* invitation still would not have been a cause of a trip to Oxford. Yet the fact that he received an invitation from Monash clearly does not explain why he went there rather than to Oxford in this case, since Oxford invited him too. Similarly, Jones's syphilis satisfies Lewis's requirement even if Smith has syphilis too, since Jones's syphilis would not have been a cause of Smith's paresis, had Smith contracted paresis, yet in this case Jones's syphilis would not explain why he rather than Smith contracted paresis.

It might be thought that Lewis's account could be saved by construing the causes more broadly, as types rather than tokens. In the case of the trip to Monash, we might take the cause to be receiving an invitation rather than the particular invitation to Monash he received. If we do this, we can correctly rule out the attempt to explain the trip by appeal to an invitation if Oxford also invited since, in this case, receiving an invitation would also have been a cause of going to Oxford. This, however, will not do, for two reasons. First, it does not capture Lewis's intent: he is interested in particular elements of a particular causal history, not general causal features. Secondly, and more importantly, the suggestion throws out the baby with the bath water. Now we have also ruled out the perfectly good explanation by invitation in some cases where only Monash invites. To see this, suppose that Lewis is the sort of person who only goes where he is invited. In this case, an invitation would have been part of a trip to Oxford, if he had gone there.

A second plausible attempt to say what contrastive explanation requires in terms of a cause that marks a difference between fact and foil appeals to a probabilistic notion of favoring. Such an account could take various forms, but a simple version would say that an explanation of why P rather than Q must cite a cause of P that raises the probability of P without raising the probability of Q, where the probabilities are construed as physical chances (van Fraassen 1980: 146–51; Hitchcock 1999: 597–608). (Construing the probabilities instead as degrees of belief is another option, but we then face something analogous to the old evidence problem for Bayesian accounts of confirmation, since P is typically already known when the why-question is

posed.) This favoring criterion rightly counts Lewis's invitation to Monash as an explanation of why he went there rather than to Oxford in a situation where only Monash invites, since the invitation to Monash increased the probability of his going there but did not increase the probability of his going to Oxford. Similarly, the favoring criterion rightly excludes the explanation by appeal to Lewis's desire to go to a place where he has good friends, since while that raises the probability of his going to Monash, it also raises the probability of his going to Oxford.

An account of contrastive explanation in terms of shifting probabilities appears naturally to capture a notion of a cause that marks a difference. Moreover, as Christopher Hitchcock (1999) observes, it also provides an account that appears to capture the independently plausible link I flagged above between contrast and presupposition, namely that to explain P rather than Q is to explain why P, given P or Q. Nevertheless, the favoring account seems too permissive. Suppose we ask not why Jones rather than Smith contracted paresis, but why Jones contracted paresis rather than remaining relatively healthy. I take it that the fact that Jones had syphilis does not explain this contrast concerning Jones alone, since he might well have had syphilis even if he had been relatively healthy, given that so few with syphilis go on to contract paresis. Yet Jones's syphilis meets the favoring condition in this case, since it raises the probability of Jones with paresis without raising the probability of healthy Jones. (Lewis's counterfactual account gives the same wrong answer here, since the syphilis would not have been a cause of Jones remaining healthy.) Moreover, a favoring account faces the same difficulty that Lewis has already faced: it wrongly rules in the use of the invitation to Monash to explain why Lewis went to Monash rather than to Oxford, even when Oxford also invites. For even in such a case, the invitation to Monash raises the probability of going to Monash without raising the probability of going to Oxford. (Eric Barnes has made this objection to Hitchcock's version of a favoring account in correspondence; for Hitchcock's reply see his 1999: 605–6.) What seems to be wrong with requiring only that the cause of P would not have been a cause of Q, or with requiring only that the cause of P not raise the probability of Q is that neither of these accounts captures the need in contrastive explanation not just for the presence, for example, of an invitation to Monash, but also the need for the absence of an invitation to Oxford. To explain why P rather than Q, we seem to need not just a cause of P, but also the *absence* of a corresponding event.

Causal triangulation

In an attempt to improve on the counterfactual and favoring approaches to contrastive explanation, consider John Stuart Mill's Method of Difference, his version of the controlled experiment, which we discussed in chapter 1 (Mill 1904: III.VIII.2). Mill's method rests on the principle that a cause must

lie among the antecedent differences between a case where the effect occurs and an otherwise similar case where it does not. The difference in effect points back to a difference that locates a cause. Thus we might infer that contracting syphilis is a cause of paresis, since it is one of the ways Smith and Jones differed. The cause that the Method of Difference isolates depends on which control we use. If, instead of Smith, we have Doe, who does not have paresis but did contract syphilis and had it treated, we would be led to say that a cause of paresis is not syphilis, but the failure to treat it. The Method of Difference also applies to incompatible as well as to compatible contrasts. As Mill observes, the method often works particularly well with diachronic (before and after) contrasts, since these give us histories of fact and foil that are largely shared, making it easier to isolate a difference. If we want to determine the cause of a person's death, we naturally ask why he died when he did rather than at another time, and this yields an incompatible contrast, since you can only die once.

The Method of Difference concerns the discovery of causes rather than the explanation of effects, but the similarity to contrastive explanation is striking (Garfinkel 1981: 40). Accordingly, I propose that, for the causal explanations of events, explanatory contrasts select causes by means of the Difference Condition. *To explain why P rather than Q, we must cite a causal difference between P and not-Q, consisting of a cause of P and the absence of a corresponding event in the case of not-Q.* Instead of pointing to a counterfactual difference, a particular cause of P that would not have been a cause of Q, as Lewis suggests, or a single cause with differential probabilistic effect, as a favoring account suggests, contrastive questions select as explanatory an actual causal difference between P and not-Q, consisting of both a presence and an absence. Lewis's invitation to Monash does not explain why he went there rather than to Oxford if he was invited to both places because, while there is an invitation in the history of his trip to Monash, there is also an invitation in the history that led him to forgo Oxford. Similarly, the Difference Condition correctly entails that Jones's syphilis does not explain why he rather than Smith contracted paresis if Smith had syphilis too, and that Kate's submitting an essay does not explain why she rather than Frank won the prize. Consider now some of the examples of successful contrastive explanation. If only Jones had syphilis, that explains why he rather than Smith has paresis, since having syphilis is a condition whose presence was a cause of Jones's paresis and a condition that does not appear in Smith's medical history. Writing the best essay explains why Kate rather than Frank won the prize, since that is a causal difference between the two of them. Lastly, the fact that *Jumpers* is a contemporary play and *Candide* is not caused me both to go to one and to avoid the other. As most of these examples illustrate, the required absence in the case of not-Q is typically an absence from the causal history of not-Q, but this is not always the case. Where both Jones and Doe have syphilis, but only Jones

also has paresis, Jones's syphilis clearly does not explain why he rather than Doe has paresis; nevertheless, Doe's syphilis is not a cause of Doe not having paresis. The Difference Condition must thus be read as requiring not just the absence of the corresponding event from the causal history of not-Q, but its absence, *tout court*. (The need to rule out cases where the corresponding event is present but not part of the causal history was made clear to me by Michael Gaylard and Tom Grimes.)

The application of the Difference Condition is easiest to see in cases of compatible contrasts, since here the causal histories of P and of not-Q are generally distinct, but the condition applies to incompatible contrasts too. In cases of choice, for example, the causal histories are usually the same: the causes of my going to *Jumpers* are the same as the causes of my not going to *Candide*. The Difference Condition may nevertheless be satisfied if my belief that *Jumpers* is a contemporary play is a cause of going, and I do not believe that *Candide* is a contemporary play. That is why my preference for contemporary plays explains my choice. Similarly, the invitation from Monash explains why Lewis went there rather than to Oxford and satisfies the Difference Condition, so long as Oxford did not invite. The condition does not require that the same event be present in the history of P but absent in the history of not-Q, a condition that could never be satisfied when the two histories are the same, but only that the cited cause of P find no corresponding event in the case of not-Q where, roughly speaking, a corresponding event is something that would bear the same relation to Q as the cause of P bears to P.

The application of the Difference Condition is perhaps most difficult to see in cases where the contrastive question does not supply two distinct instances, like Jones and Smith. For example, we may explain why a particle was deflected upward rather than moving in a straight line by observing that the particle passed through a particular field: this field is a causal presence that explains the contrast, but it is not clear in such a case what the corresponding absence might be. (I owe this point and example to Jonathan Vogel.) This sort of case is not unusual, since we often ask why a change occurred rather than the status quo. Similarly, instead of asking why Jones rather than Smith contracted paresis, a case to which the Difference Condition is easy to apply, we might ask a similar question about Jones alone, that is why he contracted paresis rather than staying healthy. As in the particle case, we seem here to have only one instance, and that seems at first to block the application of the notion of a corresponding event.

In fact I think the Difference Condition does apply in single instance cases, including the case of the wayward particle, though this is perhaps particularly difficult to see at first because what the Condition requires in that particular case is the absence of an absence. To see how the Condition works in such cases, it is helpful to work up to it in stages (which will also give us another good illustration of how explanation is sensitive to contrast).

Suppose first that there are two particles, one with a field and deflection upward and the other with no field and no deflection. This is an easy case for the Difference Condition: the presence of the field in the one case and its absence in the other explains why one particle rather than the other was deflected, because the field in the one case is a cause of the deflection, and the corresponding event in the other case would be a similar field there, which is duly absent. And even in a single particle case, there need be no difficulty. Thus, if the question is why the particle deflected upward rather than downward, we can explain this in terms of the presence of a field with a particular orientation, since the corresponding event would be a field oriented in the opposite direction, again duly absent. Now we can return to the original example, where we ask why the particle was deflected upward rather than moving in a straight line. The Difference Condition again applies after all, since we have the presence of the field, where in this case the corresponding event would be the absence of any field, and this (absence) is absent, there being a field present. Although the application of the Difference Condition is easiest to see in cases like those where the Method of Difference applies, with two quite distinct instances in one of which the effect occurs and in the other of which it does not, the Condition applies to single instance contrasts as well.

I hope this helps to show how to apply the notion of a corresponding event in cases where there is only one instance. There is, however, another related challenge to the notion of a corresponding event: not that it is sometimes inapplicable, but that it is vague. I have no full response to this difficulty, but will worry it a bit here. As a first approximation, I suggested above that a corresponding event is something that would bear the same relation to Q as the cause of P bears to P. But what does this mean? Clearly it cannot mean simply anything that would have caused Q (Achinstein 1992: 353). If Lewis was invited to both places, the invitation to Monash does not explain why he went there rather than to Oxford, even though of course he was not abducted to Oxford. We would do better to think of the Difference Condition as requiring the presence of one token of a type and the absence of another token of the same type. But not any type will do. Thus, if both Monash and Oxford invited Lewis, the invitation to Monash will not explain the contrast, even if that invitation falls under the type 'invitation printed on pink paper' and the invitation to Oxford is not of that type. Perhaps this difficulty can be met by requiring that the type be pitched at a level of causal efficacy. Thus the type 'invitation' is appropriate because it is in virtue of being an invitation that the token caused Lewis to go to Monash. So it is only the absence of an invitation to Oxford, not merely the absence of an invitation on pink paper, that would allow the invitation to Monash to explain the contrast. (On the other hand, if Lewis had been the sort of person more likely to be swayed by invitations on pink paper, then that type would be explanatorily relevant.)

The idea of the presence of one token and the absence of another is, however, still only an approximation to the requirements of the Difference Condition. For one thing, the absence needs to be tied to the foil. (The absence of an invitation to Cambridge will obviously not help us to explain why Lewis went to Monash rather than to Oxford.) For another, although many explanatory presence–absence pairs are naturally seen as two tokens of the same type, not all are. Thus if you ask why the mercury in the thermometer rose rather than fell, one token is of increasing temperature, while the other is of decreasing temperature. I have attempted to capture both of these points in the gloss I gave above on the corresponding event for the P/Q contrast – 'something that would bear the same relation to Q as the cause of P bears to P' – but the account does remain somewhat vague, and it is not clear how to make it more precise. It is for example now tempting to return to the condition that the corresponding event must be something that would have been a cause of Q, if Q had occurred, only now as an additional requirement alongside something like the token-type condition. But while satisfying this further condition may yield particularly satisfying contrastive explanations, it is not necessary. For we can explain why Kate rather than Frank won the prize by pointing out that she wrote the better essay, even though had Frank's essay been better than Kate's, that still would not have assured him of the prize, since a third party might have written something better still.

I have been unable to give a precise account of the notion of a corresponding event. By way of mitigation, I would say that any unclarity in the notion of a corresponding event is one that we clearly negotiate successfully in practice, since it is a notion we deploy frequently and uncontroversially in inferential contexts when we apply Mill's Method of Difference to infer causes from effects and their absence. As Mill puts it:

> If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ is the effect, or the cause, or an indispensable part of the cause, of the phenomenon. (1904: III.VIII.2)

Here 'the circumstance in which ... the two instances differ' is tantamount to the presence–absence pair in contrastive explanation, and faces the same vagueness problems that I have been worrying in the notion of corresponding event. How do we tell whether a circumstance attaches to an instance? And do two instances differ if they both for example contain invitations but only one is issued on pink paper? A fuller analysis of the notion of differing circumstances would be welcome, but the application and development of the Method of Difference does not wait on this, and I suggest that the

situation is similar with respect to the notion of corresponding event and the Difference Condition on contrastive explanation.

One of the merits of the Difference Condition is that it brings out the way the incompatibility of fact and foil, when it obtains, is not sufficient to transform an explanation of the fact into an explanation of the contrast, even if the cause of the fact is also a cause of the foil not obtaining. Perhaps we could explain why Able got the philosophy job by pointing out that Quine wrote him a strong letter of recommendation, but this will only explain why Able rather than Baker got the job if Quine did not also write a similar letter for Baker. If he did, Quine's letter for Able does not alone explain the contrast, even though that letter is a cause of both Able's success and Baker's failure, and the former entails the latter. The letter may be a partial explanation of why Able got the job, but it does not explain why Able rather than Baker got the job. In the case where they both have strong letters from Quine, a good explanation of the contrast will have to find an actual difference, say that Baker's dossier was weaker than Able's in some other respect, or that Able's specialties were more useful to the department. There are some cases of contrastive explanation that do seem to rely on the way the fact precludes the foil, but I think these can be handled by the Difference Condition. For example, suppose we explain why a bomb went off prematurely at noon rather than in the evening by saying that the door hooked up to the trigger was opened at noon (I owe this example to Eddy Zemach). Here it may appear that the Difference Condition is not in play, since the explanation would stand even if the door was also opened in the evening. But the Difference Condition is met, if we take the cause not simply to be the opening of the door, but the opening of the door when it is rigged to an armed bomb.

My goal in this chapter is to show how the choice of contrast helps to determine an explanatory cause, not to show why we choose one contrast rather than another. The latter question is not part of providing a model of explanation, as that task has been traditionally construed. It is no criticism of the deductive-nomological model that it does not tell us which phenomena we care to explain, so long as it tells us what counts as an adequate explanation of the phenomena we select; similarly, it is no criticism of my account of contrastive explanation that it does not tell us why we are interested in explaining some contrasts rather than others. Still, an account of the considerations that govern our choice of why-questions ought to form a part of a full model of our explanatory practices, and it is to the credit of the contrastive analysis that it lends itself to this. As we will see in later chapters, our choice of why-questions is often governed by our *inferential* interests, so that we choose contrasts that help us to determine which of competing explanatory hypotheses is correct. For now, however, we may just note that not all contrasts make for sensible contrastive questions. It does not make sense, for example, to ask why Lewis went to

Monash rather than Baker getting the philosophy job. One might have thought that a sensible contrast must be one where fact and foil are incompatible, but we have seen that this is not necessary, since there are many sensible compatible contrasts. There are also incompatible contrasts that do not yield reasonable contrastive questions, such as why someone died when she did rather than never having been born. The Difference Condition suggests instead that the central requirement for a sensible contrastive question is that the fact and the foil have a largely similar history, against which the differences stand out (cf. Barnes 1994). When the histories are too disparate, we do not know where to begin to answer the question. There are, of course, other considerations that help to determine the contrasts we actually choose. For example, in the case of incompatible contrasts, we often pick as foil the outcome we expected; in the case of compatible contrasts, as I have already mentioned, we often pick as foil a case we expected to turn out the same way as the fact. The condition of a similar history also helps to determine what will count as a corresponding event. If we were to ask why Lewis went to Monash rather than Baker getting the job, it would be difficult to see what in the history of Baker's failure would correspond to Lewis's invitation, but when we ask why Able rather than Baker got the job, the notion of a corresponding event is relatively clear.

I will now consider three further issues connected with my analysis of contrastive explanation: the need for further principles for distinguishing explanatory from unexplanatory causes, the prospects for treating all why-questions as contrastive, and a more detailed comparison of my analysis with the deductive-nomological model. When we ask contrastive why-questions, we choose our foils to point towards the sorts of causes that interest us. As we have just seen, when we ask about a surprising event, we often make the foil the thing we expected. This focuses our inquiry on causes that will illuminate the reason our expectation went wrong. Failed expectations are not, however, the only things that prompt us to ask why-questions. If a doctor is interested in the internal etiology of a disease, he will ask why the afflicted have it rather than other people in similar circumstances, even though the shared circumstances may be causally relevant to the disease. Again, if a machine malfunctions, the natural diagnostic contrast is its correct behavior, since that directs our attention to the causes that we want to change. But the contrasts we construct will almost always leave multiple differences that meet the Difference Condition. More than one of these may be explanatory: my account does not entail that there is only one way to explain a contrast. At the same time, however, some causally relevant differences will not be explanatory in a particular context, so while the Difference Condition may be necessary for the causal contrastive explanations of particular events, it is not generally sufficient. For that we need further principles of causal selection.

The considerations that govern selection from among causally relevant differences are numerous and diverse; the best I can do here is to mention what a few of them are. An obvious pragmatic consideration is that someone who asks a contrastive question may already know about some causal differences, in which case a good explanation will have to tell her something new. If she asks why Kate rather than Frank won the prize, she may assume that it was because Kate wrote the better essay, in which case we will have to tell her more about the differences between the essays that made Kate's better. A second consideration, and one that I have already mentioned, is that when they are available we usually prefer explanations where the foil would have occurred if the corresponding event had occurred. Suppose that only Able had a letter from Quine, but even a strong letter from Quine would not have helped Baker much, since his specialties do not fit the department's needs. Suppose also that, had Baker's specialties been appropriate, he would have gotten the job, even without a letter from Quine. In this case, the difference in specialties is a better explanation than the difference in letters. As we have seen, however, an explanation that does not meet this condition of counterfactual sufficiency for the occurrence of the foil may be perfectly acceptable, if we do not know of a sufficient difference. To give another example, the explanation of why Jones rather than Smith contracted paresis is an example of this: even if Smith had syphilis in his medical history, he probably would not have contracted paresis (but cf. Carroll 1999). Moreover, even in cases where a set of known causes does supply a counterfactually sufficient condition, the inquirer may be much more interested in some than in others. The doctor may be particularly interested in causes he can control, the lawyer in causes that are connected with legal liability and the accused in causes that cannot be held against him.

We also prefer differences where the cause is causally necessary for the fact in the circumstances. Consider a case of overdetermination. Suppose that you ask me why I ordered eggplant rather than beef, when I was in the mood for eggplant and not for beef, and I am a vegetarian. My mood and my convictions are separate causes of my choice, each causally sufficient in the circumstance and neither necessary. In this case, it would be better to give both differences than just one. The Difference Condition could easily be modified to require necessary causes, but I think this would make the Condition too strong. One problem would be cases of 'failsafe' over-determination. Suppose we change the restaurant example so that my vegetarian convictions were not a cause of the particular choice I made: that time, it was simply my mood that was relevant. Nevertheless, even if I had been in the mood for beef, I would have resisted, because of my convictions. In this case, my explanation does not have to include my convictions, even though my mood was not a necessary cause of my choice. Again, we sometimes don't know whether a cause is necessary for the effect, and in such cases the cause still seems explanatory. But when there are differences

that supply a necessary cause, and we know that they do, we tend to prefer them.

Another reason satisfying the Difference Condition is not always sufficient for a good contrastive explanation is that a difference favoring the fact may be balanced against another favoring the foil. If I tell you that Lewis went to Monash rather than Oxford because only Monash invited him, you might reply, 'Yes, but Oxford has much better book shops, and Lewis loves book shops.' In such a case, I will have to supplement my original explanation by showing, or at least claiming, that the actual cause of the fact trumped or outweighed the potential cause of the foil. Thus I might claim that his preference for places that invite him was stronger than his preference for places with outstanding book shops. Of course this might not be true: the difference I originally cite may not by itself be stronger than the countervailing force you mention. In this case, I must find other or additional differences that are. Here the hardware of a probabilistic favoring approach has a natural application. There are doubtless other principles that also play a role in determining which differences yield the best explanation in a particular context. So there is more to contrastive explanation than the Difference Condition describes, but that Condition does seem to describe the central mechanism of causal selection.

Since contrastive questions are so common and foils play such an important role in determining explanatory causes, it is natural to wonder whether all why-questions are not at least implicitly contrastive. Often the contrast is so obvious that it is not worth mentioning. If you ask me why I was late for our appointment, the question is why I was late rather than on time, not why I was late rather than not showing up at all. Moreover, in cases where there is no specific contrast, stated or implied, we might construe 'Why P?' as 'Why P rather than not-P?', thus subsuming all causal why-questions under the contrastive analysis.

How does the Difference Condition behave for these 'global' contrasts? I once thought the answer was 'pathologically', since the Condition would impossibly require that we find a cause of P that is at once present and absent; but I now see (thanks to Naomi Rosenberg) that things are not so bad. When the question is, 'Why P rather than not-P?', what the Difference Condition requires is the absence of something that bears the same relation to *not*-P that the cited cause bears to P. If C is the cause of P, then what would bear the same relation to not-P is presumably not C itself, but something else. But what would it be? The difficulty in answering the question arises because 'not-P' is not a limited negation, but encompasses all the different specific ways P might not have been the case. The way to construe the Difference Condition as it applies to the limiting case of the contrast, P rather than not-P, is that we must find a difference for events logically or causally incompatible with P, not for a single event, 'not-P'. Suppose that we ask why Jones has paresis, with no implied contrast. This

would require a difference for foils where he does not have paresis. Saying that he had syphilis differentiates between the fact and the foil of a thoroughly healthy Jones, but this is not enough, since it does not differentiate between the fact and the foil of Jones with syphilis but without paresis. Excluding many incompatible foils will push us towards a sufficient cause of Jones's syphilis, since it is only by giving such a 'full cause' that we can be sure that some bit of it will be missing from the history of all the foils.

To explain P rather than not-P we do not, however, need to explain every incompatible contrast. We do not, for example, need to explain why Jones contracted paresis rather than being long dead or never being born. The most we can require is that we exclude all incompatible foils with histories similar to the history of the fact. I nevertheless remain unsure whether every apparently non-contrastive question should be analyzed in contrastive form, so I am an agnostic about the existence of non-contrastive why-questions.

Finally, let us compare my analysis of contrastive explanation to the deductive-nomological model. First, as we have already seen, a causal view of explanation has the merit of avoiding all the counterexamples to the deductive-nomological model where causes are deduced from effects. It also avoids the unhappy consequence of counting almost every explanation we give as a mere sketch, since one can give a cause of P that meets the Difference Condition for various foils without having the laws and singular premises necessary to deduce P. Many explanations that the deductive model counts as only very partial explanations of P are in fact reasonably complete explanations of P rather than Q. The excessive demands of the deductive model are particularly striking for cases of compatible contrasts, at least if the deductive-nomologist requires that an explanation of P rather than Q provide an explanation of P and an explanation of not-Q. In this case, the model makes explaining the contrast substantially harder than providing a deductive explanation of P, when in fact it is often substantially easier. Our inability to find a non-contrastive reduction of contrastive questions is, among other things, a symptom of the inability of the deductive-nomological model to give an accurate account of this common type of explanation.

There are at least two other conspicuous advantages of a causal contrastive view of explanation over the deductive-nomological model. One odd feature of the model is that it entails that an explanation cannot be ruined by adding true premises, so long as the additional premises do not render the law superfluous to the deduction by entailing the conclusion outright (assuming they are not themselves laws). This consequence follows from the elementary logical point that additional premises can never convert a valid argument into an invalid one. In fact, however, irrelevant additions can spoil an explanation. If I say that Jones rather than Smith contracted paresis because only Jones had syphilis and only Smith was a regular church-goer, I have not simply said more than I need to, I have given an incorrect explanation, since going to church is not a prophylactic. By requiring that

explanatory information be causally relevant, the contrastive model avoids this problem. Another related and unhappy feature of the deductive-nomological model is that, as we have seen, it entails that explanations are virtually deductively closed: an explanation of P will also be an explanation of any logical consequence of P, so long as the consequence is not directly entailed by the initial conditions alone. (For an example of the slight non-closure in the model, notice that a deductive-nomological explanation of P will not also be a deductive-nomological explanation of the disjunction of P and one of the initial conditions of the explanation.) In practice, however, explanation seems to involve a much stronger form of non-closure. I might explain why all the men in the restaurant are wearing paisley ties by appealing to the fashion of the times for ties to be paisley, but this might not explain why they are all wearing ties, which is because of a rule of the restaurant. (I owe this example to Tim Williamson.) The contrastive view gives a natural account of this sort of non-closure. When we ask about paisley ties, the implied foil is other sorts of tie; but when we ask simply about ties, the foil is not wearing ties. The fashion marks a difference in one case, but not in the other.

A defender of the deductive-nomological model might respond to some of these points by arguing that, whatever the merits of a contrastive analysis of lay explanation, the deductive model (perhaps with an additional restriction blocking 'explanations' of causes by effects) gives a better account of scientific explanation. For example, it has been claimed that scientific explanations, unlike ordinary explanations, do not exhibit the interest relativity of foil variation that a contrastive analysis exploits, so a contrastive analysis does not apply to scientific explanation (Worrall 1984: 76–7). It is, however, a mistake to suppose that all scientific explanations even aspire to deductive-nomological status. The explanation of why Jones rather than Smith contracted paresis is presumably scientific, but it is not a deduction *manqué*. Moreover, as the example of the thermometer shows, even a full deductive-nomological explanation may exhibit interest relativity: it may explain the fact relative to some foils but not relative to others. A typical deductive-nomological explanation of the rise of mercury in a thermometer will simply assume that the glass does not break and so while it will explain, for example, why the mercury rose rather than fell, it will not explain why it rose rather than breaking the thermometer. Quite generally, a deductive-nomological explanation of a fact will not explain that fact relative to any foils that are themselves logically inconsistent with one of the premises of the explanation. Again, a Newtonian explanation of the earth's orbit (ignoring the influence of the other planets) will explain why the earth has its actual orbit rather than some other orbit, but it will not explain why the earth does not have any of the other orbits that are compatible with Newton's theory. The explanation must assume information about the earth's position and velocity at some time that will rule out the other Newtonian orbits, but it

will not explain why the earth does not travel in those paths. To explain this would require quite different information about the early history of the earth. Similarly, an adaptionist explanation of why members of a species possess a certain trait may explain why they have that trait rather than various maladaptive traits, but it may not explain why they have that trait rather than other traits that would perform the same functions equally well. To explain why an animal has one trait rather than another functionally equivalent trait requires instead an appeal to the evolutionary history of the species, insofar as it can be explained at all.

With rather more justice, a deductive-nomologist might object that scientific explanations do very often essentially involve laws and theories, and that the contrastive view does not seem to account for this. For even if the fact to be explained carries no restricting contrast, the contrastive view, if it is extended to this case by analyzing 'Why P?' as 'Why P rather than not-P', requires at most that we cite a condition that is causally sufficient for the fact, not that we actually give any laws. In reply, one might mention first that laws may nevertheless be part of a correct analysis of the causal relation itself, and that knowledge of laws is sometimes essential evidence for causal claims. Moreover, the contrastive view can help to account for the explicit role of laws in many scientific explanations. To see this, notice that scientists are often and perhaps primarily interested in explaining regularities, rather than particular events (Friedman 1974: 5; though explaining particular events is also important when, for example, scientists test their theories, since observations are of particular events). Something like the Difference Condition applies to many explanations of regularities, but to give a contrastive explanation of a regularity will require citing a law, or at least a generalization, since here we need some general cause (cf. Lewis 1986: 225–6). To explain, say, why people feel the heat more when the humidity is high, we must find some general causal difference between cases where the humidity is high and cases where it is not, such as the fact that the evaporation of perspiration, upon which our cooling system depends, slows as the humidity rises. So the contrastive view, in an expanded version that applies to general facts as well as to events (a version I do not here provide), should be able to account for the role of laws in scientific explanations as a consequence of the scientific interest in general why-questions. Similarly, although the contrastive view does not require deduction for explanation, it is not mysterious that scientists should often look for explanations that do entail the phenomenon to be explained. This may not have to do with the requirements of explanation per se, but rather with the uses to which explanations are put. Scientists often want explanations that can be used for accurate prediction, and this requires deduction. Again, the construction of an explanation is a way to test a theory, and some tests require deduction.

Another way of seeing the compatibility between the scientific emphasis on theory and the contrastive view of explanation is by observing that

scientists are not just interested in this or that explanation, but in a unified explanatory scheme. Scientists want theories, in part, because they want engines that will provide many explanations. The contrastive view does not entail that a theory is necessary for any particular explanation, but a good theory is the best way to provide the many and diverse contrastive explanations that the scientist is after. This also helps to account for the familiar point that scientists are often interested in discovering causal mechanisms. The contrastive view will not require a mechanism to explain why one input into a black box causes one output, but it pushes us to specify more and more of the detailed workings of the box as we try to explain its full behavior under diverse conditions. So I conclude that the contrastive view of explanation does not fly in the face of scientific practice.

The Difference Condition shows how contrastive questions about particular events help to determine an explanatory cause by a kind of *causal triangulation*. This contrastive model of causal explanation cannot be the whole story about explanation since, among other things, not all explanations are causal and since the choice of foil is not the only factor that affects the appropriate choice of cause. The model does, however, give a natural account of much of what is going on in many explanations, and it captures some of the merits of competing accounts while avoiding some of their weaknesses. We have just seen this in some detail for the case of the deductive-nomological model. It also applies to the familiarity model. When an event surprises us, a natural foil is the outcome we had expected, and meeting the Difference Condition for this contrast will help to show us why our expectation went wrong. The mechanism of causal triangulation also accounts for the way a change in foil can lead to a change in explanatory cause, since a difference for one foil will not in general be a difference for another. It also shows why explaining 'P rather than Q' is sometimes harder and sometimes easier than explaining P alone. It may be harder, because it requires the absence of a corresponding cause in the history of not-Q, and this is something that will not generally follow from the presence of the cause of P. Explaining the contrast may be easier, because the cause of P need not be even close to being sufficient for P, so long as it is part of a causal difference between P and not-Q. Causal triangulation also elucidates the interest relativity of explanation. We express some of our interests through our choice of foils and, by construing the phenomenon to be explained as a contrast rather than the fact alone, the interest relativity of explanations reduces to the important but unsurprising point that different people are interested in explaining different phenomena. Moreover, the Difference Condition shows that different interests do not require incompatible explanations to satisfy them, only different but compatible causes. The mechanism of causal triangulation also helps to account for the failure of various attempts to reduce contrastive questions to some non-contrastive form. None of these bring out the way a foil serves to select a

location on the causal history leading up to the fact. Causal triangulation is the central feature of contrastive explanation that non-contrastive paraphrases suppress. Lastly, we will find that the structure of contrastive explanations helps us with the problem of describing our inferential practices, a problem whose difficulties we met in chapter 1, when it is wed to Inference to the Best Explanation, an account to which we now finally turn.

Chapter 4

Inference to the Best Explanation

Spelling out the slogan

Our initial survey of the problems of induction and explanation is now complete. We have considered some of the forms these problems take, some of the reasons they are so difficult to solve, and some of the weaknesses that various attempts to solve them suffer. In the last chapter, I also offered something more constructive, by attempting an improved version of the causal model of explanation. Up to now, however, we have treated inference and explanation in near mutual isolation, a separation that reflects most of the literature on these subjects. Although the discussion of inference in chapter 1 construed the task of describing our practices as itself an explanatory inquiry, the attempt to specify the black box mechanism that takes us from evidence to inference and so explains why we make the inferences we do, none of the models of inference we considered explicitly invoked explanatory relations between evidence and conclusion. Similarly, in the discussion of explanation in chapters 2 and 3, inferential considerations played a role in only one of the models, the reason model. That model uses an inferential notion to account for explanation, by claiming that we explain a phenomenon by giving some reason to believe that the phenomenon occurs, and it was found to be unacceptable for inferential reasons, since it does not allow for self-evidencing explanations, such as the explanation of the tracks in the snow or of the red-shift of the galaxy, virtuous explanations where the phenomenon that is explained nevertheless provides an essential part of the reason for believing that the explanation is correct.

In this chapter, the relationship between the practices of explanation and of inference will take center stage, where it will remain for the rest of this book. Let us begin with a simple view of that relationship. First we make our inferences; then, when we want to explain a phenomenon, we draw upon our pool of beliefs for an explanation, a pool filled primarily by those prior inferences. This, however, must be too simple, since our pool may not contain the explanation we seek. So a slightly less simple view is that, if we

do not find an explanation in our pool, we search for a warranted inference that will explain, a process that may also require further observation. Explanatory considerations thus have some bearing on inference, since they may focus our inquiry but, on this view, inference still comes before explanation. After all, the most basic requirement of an explanation is that the explanatory information be correct, so how can we be in a position to use that information for an explanation unless we first know that it is indeed correct?

This picture of inference first, explanation second, however, seriously underestimates the role of explanatory considerations in inference. Those considerations tell us not only what to look for, but also whether we have found it. Take the cases of self-evidencing explanations. The tracks in the snow are the evidence for what explains them, that a person passed by on snowshoes; the red-shift of the galaxy is an essential part of the reason we believe the explanation, that it has a certain velocity of recession. In these cases, it is not simply that the phenomena to be explained provide reasons for inferring the explanations: we infer the explanations precisely because they would, if true, explain the phenomena. Of course, there is always more than one possible explanation for any phenomenon – the tracks might have instead been caused by a trained monkey on snowshoes, or by the elaborate etchings of an environmental artist – so we cannot infer something simply because it is a possible explanation. It must somehow be the best of competing explanations.

These sorts of explanatory inferences are extremely common. The sleuth infers that the butler did it, since this is the best explanation of the evidence before him. The doctor infers that his patient has measles, since this is the best explanation of the symptoms. 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. Kuhn infers that normal science is governed by exemplars, since they provide the best explanation for the observed dynamics of research. This suggests a new model of induction, one that binds explanation and inference in an intimate and exciting way. According to Inference to the Best Explanation, our inferential practices are governed by explanatory considerations. Given our data and our background beliefs, we infer what would, if true, provide the best of the competing explanations we can generate of those data (so long as the best is good enough for us to make any inference at all). Far from explanation only coming on the scene after the inferential work is done, the core idea of Inference to the Best Explanation is that explanatory considerations are a guide to inference.

Inference to the Best Explanation has become extremely popular in philosophical circles, discussed by many and endorsed without discussion by many more (for discussions see, e.g., Peirce 1931: 5.180–5.212, esp. 5.189;

Hanson 1972: ch. IV; Harman 1965; Brody 1970; Thagard 1978; Cartwright 1983, essay 5; Ben-Menahem 1990; Vogel 1990; Day and Kincaid 1994; Barnes 1995; Rappaport 1996; Bird 1998; Psillos 2002). Yet it still remains more of a slogan than an articulated account of induction. In the balance of this section, I will take some first steps towards improving this situation. In the next section, we will consider the initial attractions of the view, as well as some apparent liabilities. The balance of this book is devoted to the questions of whether Inference to the Best Explanation really will provide an illuminating model of our inductive practices and whether it is an improvement over the other accounts we have considered.

The obvious way to flesh out Inference to the Best Explanation would be to insert one of the standard models of explanation. This, however, yields disappointing results, because of the backward state of those models. For example, we would not get very far if we inserted the deductive-nomological model, since this would just collapse Inference to the Best Explanation into a version of the hypothetico-deductive model of confirmation. Indeed one suitable acid test for Inference to the Best Explanation is that it mark an improvement over the hypothetico-deductive model. As we saw in chapter 1, the deductive-nomological model of explanation has many unattractive features; it also provides almost no resources for saying when one explanation is better than another. We will do better with the causal model of contrastive explanation I developed in the last chapter, as we will see in chapters 5 and 6, but for now we are better off not burdening Inference to the Best Explanation with the details of any specific model of explanation, trying instead to stick to the actual explanatory relation itself, whatever its correct description turns out to be. Let us begin to flesh out the account by developing two signal distinctions that do not depend on the details of explanation. The first of these is the distinction between actual and potential explanations. The second is the distinction between the explanation best supported by the evidence, and the explanation that would provide the most understanding: in short, between the likeliest and the loveliest explanation.

Our discussion of inference, explanation and the connection between the two is being conducted under the assumption of inferential and explanatory realism, an assumption we will not seriously investigate until the final chapter of this book. Until then, I am assuming that a goal of inference is truth, that our actual inferential practices are truth-tropic, i.e. that they generally take us towards this goal, and that for something to be an actual explanation, it must be (at least approximately) true. But Inference to the Best Explanation cannot then be understood as inference to the best of the *actual* explanations. Such a model would make us too good at inference, since it would make all our inferences true. Our inductive practice is fallible: we sometimes reasonably infer falsehoods. This model would also fail to account for the role of competing explanations in inference. These competitors are typically incompatible and so cannot all be true, so we

cannot represent them as competing actual explanations. The final and most important reason why Inference to the Best Actual Explanation could not describe our inductive practices is that it would not characterize the process of inference in a way we could follow, since we can only tell whether something is an actual explanation *after* we have settled the inferential question. It does not give us what we want, which is an account of the way explanatory considerations could serve as a guide to the truth. Telling someone to infer actual explanations is like a dessert recipe that says start with a soufflé. We are trying to describe the way we go from evidence to inference, but Inference to the Best Actual Explanation would require us to have already arrived in order to get there. The model would not be epistemically effective.

The obvious solution is to distinguish actual from *potential* explanation, and to construe Inference to the Best Explanation as Inference to the Best Potential Explanation. We have to produce a pool of potential explanations, from which we infer the best one. Although our discussion of explanation in the last two chapters considered only actual explanations, the distinction between actual and potential explanations is familiar in the literature on explanation. The standard version of the deductive-nomological model gives an account of potential explanation: there is no requirement that the explanation be true, only that it include a general hypothesis and entail the phenomenon. If we then add a truth requirement, we get an account of actual explanation (Hempel 1965: 338). Similarly, on a causal model of explanation, a causal story is a potential explanation, and a true causal story is an actual explanation. By shaving the truth requirement off explanation, we get a notion suitable for Inference to the Best Explanation; one that allows for the distinction between warranted and successful inferences, permits the competition between explanations to take place among incompatible hypotheses, and gives an account that is epistemically effective. According to Inference to the Best Explanation, then, we do not infer the best actual explanation; rather we infer that the best of the available potential explanations is an actual explanation.

The intuitive idea of a potential explanation is of something that satisfies all the conditions of an actual explanation, except possibly that of truth (Hempel 1965: 338). This characterization may, however, be somewhat misleading, since it seems to entail that all true potential explanations are actual explanations, which is probably false. It may not be the case for explanations that fit the deductive-nomological model, since on some views a lawlike statement could hold in one possible world as a law, but in another as a mere coincidence. Even more clearly, it does not hold in the context of a causal model. A potential cause may exist yet not be an actual cause, say because some other cause pre-empted it. Of course one could construct a technical notion of potential explanation that satisfied the equality between true potential explanation and actual explanation, but this would not be a

suitable notion for Inference to the Best Explanation. As the literature on Gettier cases shows, we often infer potential causes that exist but are not actual causes. (Gilbert Harman's two-candle case is a good example of this (1973: 22–3).)

So we may need to do more work to characterize the notion of potential explanation that is suitable for Inference to the Best Explanation. One issue is how large we should make the pool. We might say that a potential explanation is any account that is logically compatible with all our observations (or almost all of them) and that is a possible explanation of the relevant phenomena. In other words, the potential explanations of some phenomena are those that do explain them in a possible world where our observations hold. This pool is very large, including all sorts of crazy explanations nobody would seriously consider. On the other hand, we might define the pool more narrowly, so that the potential explanations are only the 'live options': the serious candidates for an actual explanation. The advantage of the second characterization is that it seems to offer a better account of our actual procedure. When we decide which explanation to infer, we often start from a group of plausible candidates, and then consider which of these is the best, rather than selecting directly from the vast pool of possible explanations. But it is important to notice that the live options version of potential explanation already assumes an epistemic 'filter' that limits the pool of potential explanations to plausible candidates. This version of Inference to the Best Explanation thus includes two filters, one that selects the plausible candidates, and a second that selects from among them. This view has considerable verisimilitude, but a strong version of Inference to the Best Explanation will not take the first filter as an unanalyzed mechanism, since epistemic filters are precisely the mechanisms that Inference to the Best Explanation is supposed to illuminate. We will return to this issue in chapter 9.

Let us turn now to the second distinction. It is important to distinguish two senses in which something may be the best of competing potential explanations. We may characterize it as the explanation that is most warranted: the 'likeliest' or most probable explanation. On the other hand, we may characterize the best explanation as the one which would, if correct, be the most explanatory or provide the most understanding: the 'loveliest' explanation. The criteria of likeliness and loveliness may well pick out the same explanation in a particular competition, but they are clearly different sorts of standard. Likeliness speaks of truth; loveliness of potential understanding. Moreover, the criteria do sometimes pick out different explanations. Sometimes the likeliest explanation is not very enlightening. It is extremely likely that smoking opium puts people to sleep because of its dormative powers (though not quite certain: it might be the oxygen that the smoker inhales with the opium, or even the depressing atmosphere of the opium den), but this is the very model of an unlovely explanation. An

explanation can also be lovely without being likely. Perhaps some conspiracy theories provide examples of this. By showing that many apparently unrelated events flow from a single source and many apparent coincidences are really related, such a theory may have considerable explanatory power. If only it were true, it would provide a very good explanation. That is, it is lovely. At the same time, such an explanation may be very unlikely, accepted only by those whose ability to weigh evidence has been compromised by paranoia.

One of the reasons likeliness and loveliness sometimes diverge is that likeliness is relative to the total available evidence, while loveliness is not, or at least not in the same way. We may have an explanation that is both lovely and likely given certain evidence, unlikely given additional evidence, yet still a lovely explanation of the original evidence. Newtonian mechanics is one of the loveliest explanations in science and, at one time, it was also very likely. More recently, with the advent of special relativity and the new data that support it, Newtonian mechanics has become less likely, but it remains as lovely an explanation of the old data as it ever was. Another reason for the divergence is that the two criteria are differently affected by additional competition. A new competitor may decrease the likeliness of an old hypothesis, but it will usually not change its loveliness. Even without the evidence that favored special relativity, the production of the theory probably made Newtonian mechanics less likely but probably not less lovely.

This gives us two more versions of Inference to the Best Explanation to consider: Inference to the Likeliest Potential Explanation and Inference to the Loveliest Potential Explanation. Which should we choose? There is a natural temptation to plump for likeliness. After all, Inference to the Best Explanation is supposed to describe strong inductive arguments, and a strong inductive argument is one where the premises make the conclusion likely. But in fact this connection is too close and, as a consequence, choosing likeliness would push Inference to the Best Explanation towards triviality. We want a model of inductive inference to describe what principles we use to judge one inference more likely than another, so to say that we infer the likeliest explanation is not helpful. To put the point another way, we want our account of inference to give the *symptoms* of likeliness, the features an argument has that lead us to say that the premises make the conclusion likely. A model of Inference to the Likeliest Explanation begs these questions. It would still have some content, since it suggests that inference is a matter of selection from among competitors and that inference is often inference to a cause. But for Inference to the Best Explanation to provide an illuminating account, it must say more than that we infer the likeliest cause (cf. Cartwright 1983: 6). This gives us a second useful acid test for Inference to the Best Explanation, alongside the requirement that it do better than the hypothetico-deductive model. Inference to the Best Explanation is an advance only if it reveals more about inference than that it is often inference

to the likeliest cause. It should show how judgments of likeliness are determined, at least in part, by explanatory considerations.

So the version of Inference to the Best Explanation we should consider is Inference to the Loveliest Potential Explanation. Here at least we have an attempt to account for epistemic value in terms of explanatory virtue. This version claims that the explanation that would, if true, provide the deepest understanding is the explanation that is likeliest to be true. Such an account suggests a really lovely explanation of our inferential practice itself, one that links the search for truth and the search for understanding in a fundamental way. Similar remarks apply to the notion of potential explanation, if we opt for the narrower live option characterization I favor. We want to give an account of the plausibility filter that determines the pool of potential explanations, and a deep version of Inference to the Best Explanation will give this characterization in explanatory terms: it will show how explanatory considerations determine plausibility.

The distinction between likeliness and loveliness is, I hope, reasonably clear. Nevertheless, it is easy to see why some philosophers may have conflated them. After all, if Inference to the Loveliest Explanation is a reasonable account, loveliness and likeliness will tend to go together, and indeed loveliness will be a guide to likeliness. Moreover, given the opacity of our 'inference box', we may be aware only of inferring what seems likeliest even if the mechanism actually works by assessing loveliness. Our awareness of what we are doing may not suggest the correct description. In any event, if there is a tendency to conflate the distinction, this helps to explain why Inference to the Best Explanation enjoys more popularity among philosophers than is justified by the arguments given to date in its favor. By implicitly construing the slogan simply as Inference to the Likeliest Explanation, it is rightly felt to apply to a wide range of inferences; by failing to notice the difference between this and the deep account, the triviality is suppressed. At the same time, the distinction between likeliness and loveliness, or one like it, is one that most people who were seriously tempted to develop the account would make, and this may help to explain why the temptation has been so widely resisted. Once one realizes that an interesting version requires an account of explanatory loveliness that is conceptually independent of likeliness, the weakness of our grasp on what makes one explanation lovelier than another is discouraging.

In practice, these two versions of Inference to the Best Explanation are probably ideal cases: a defensible version may well need to combine elements of each, accounting for likeliness only partially in explanatory terms. For example, one might construct a version where a non-explanatory notion of likeliness plays a role in restricting membership in the initial set of potential explanations, but where considerations of loveliness govern the choice from among the members of that set. Again, we may have to say that considerations of likeliness having nothing to do with explanation will, under

various conditions, defeat a preference for loveliness. This may be the only way to account for the full impact of disconfirming evidence. So the distinction between likeliness and loveliness leaves us with considerable flexibility. But I think we may take it as a rule of thumb that the more we must appeal to likeliness analyzed in non-explanatory terms to produce a defensible version of Inference to the Best Explanation, the less interesting that model is. Conversely, the more use we can make of the explanatory virtues, the closer we will come to fulfilling the exciting promise of Inference to the Best Explanation, of showing how explanatory considerations are our guide to the truth.

It bears emphasizing that my aspirations for Inference to the Best Explanation are thus modest in one way and bold in another. On the modest side, it is no part of my brief to defend the view that Inference to the Best Explanation gives a complete account of scientific inferences, much less of scientific practices generally, or that it describes the fundamental form of inference to which everything else somehow reduces. Inference to the Best Explanation can only ever be a part of a very complicated story. On the bold side, however, I want to insist that the account makes out explanatory considerations to be an important guide to judgments of likeliness, that Inference to the Best Explanation not reduce to the true but very weak claim that scientists are in the habit of explaining the phenomena they observe. The central idea that explanatory considerations are an important guide to inference is thus meant to steer a middle course between triviality and manifest falsity.

We have now gone some way towards spelling out the slogan, by making the distinctions between potential and actual explanation and between the likeliest and the loveliest explanation. By seeing how easy it is to slide from loveliness to likeliness, we have also sensitized ourselves to the risk of trivializing the model by making it so flexible that it can be used to describe almost any form of inference. But there are also various respects in which the scope of Inference to the Best Explanation is greater than may initially appear. Two apparent and damaging consequences of Inference to the Best Explanation are that only one explanation can be inferred from any set of data and that the only data that are relevant to a hypothesis are data the hypothesis explains. Both of these are, however, merely apparent consequences, on a reasonable version of Inference to the Best Explanation. The first is easily disposed of. Inference to the Best Explanation does not require that we infer only one explanation of the data, but that we infer only one of *competing* explanations. The data from a flight recorder recovered from the wreckage of an airplane crash may at once warrant explanatory inferences about the motion of the plane, atmospheric conditions at the time of the accident, malfunctions of equipment in the airplane and the performance of the pilot, and not simply because different bits of information from the recorder will warrant different inferences, but because

the same bits may be explained in many different but compatible ways. When I notice that my front door has been forced open, I may infer both that I have been robbed and that my deadbolt is not as force-resistant as the locksmith claimed. Thus, in spite of the suggestion of uniqueness that the word 'best' carries, Inference to the Best Explanation should be construed so as to allow multiple explanations.

So the account allows us to infer more than one explanation. It also allows us to infer fewer. As I have already suggested in passing, although 'best' suggests existence, the account clearly should be construed so as to allow that we infer none of the potential explanations we have come up with, since they may all be too weak. The best explanation must be good enough to merit inference: Inference to the Best Explanation must allow for agnosticism. Indeed we could, I think, maintain the spirit of the account without construing it as the basis for theory *acceptance* at all. We could say instead that what we have is an account of theory 'confirmation' (the misleading term philosophers of science use for the notion of inductive support), according to which evidence confirms the theory that best explains it. As we will see in chapter 7, this notion might be cashed out in terms of probabilities construed as degrees of belief that fall short of full acceptance. The crucial issue, so far as I am concerned, is whether explanatory considerations are a guide to the bearing of evidence on theory. Nevertheless, for the time being I will usually write as if we are talking about a mechanism of acceptance.

Inference to the Best Explanation can also account for some of the ways evidence may be relevant to a hypothesis that does not explain it. The most obvious mechanism for this depends on a deductive consequence condition on inference. If I am entitled to infer a theory, I am also entitled to infer whatever follows deductively from that theory, or from that theory along with other things I reasonably believe (Hempel 1965: 31–2). This is at least highly plausible: it would be a poor joke to say one is entitled to believe a theory but not its consequences. (Although the consequence condition is more controversial as a condition on confirmation, it is near irresistible as a condition on acceptance.) Suppose now that I use Inference to the Best Explanation to infer from the data to a high level theory, and then use the consequence condition to deduce a lower level hypothesis from it. There is now no reason to suppose that the lower level theory will explain all of the original data that indirectly support it. Newton was entitled to infer his dynamical theory in part because it explained the result of various terrestrial experiments. This theory in turn entails laws of planetary orbit. Inference to the Best Explanation with its consequence condition counts those laws as supported by the terrestrial evidence, even though the laws do not explain that evidence. It is enough that the higher level theory does so. The clearest cases of the consequence condition, however, are deduced predictions. What will happen in the future does not

explain what happened in the past, but a theory that entails the prediction may.

Since Inference to the Best Explanation will sometimes underwrite inferences to high level theories, rich in deductive consequences, the consequence condition substantially increases the scope of the model. Even so, we may wish to broaden the scope of the condition to include 'explanatory consequences' as well as strictly deductive ones. Seeing the distinctive flash of light, I infer that I will hear thunder. The thunder does not explain the flash, but the electrical discharge does and would also explain why I will hear thunder. But the electrical discharge does not itself entail that I will hear thunder. It is not merely that there is more to the story, but that there is always the possibility of interference. I might go temporarily deaf, the lightning may be too far away, I might sneeze at just the wrong moment, and there are always other possibilities that I do not know about. Someone who favors deductive models will try to handle these possibilities by including extra premises, but this will almost certainly require an unspecifiable *ceteris paribus* (all else being equal) clause. So in at least many cases, it may be more natural to allow 'Inference from the Best Explanation' (Harman 1986: 68–70). Noticing that it is extraordinarily cold this morning, I infer that my car will not start. The failure of my car to start would not explain the weather, but my inference is naturally described by saying that I infer that it will not start because the weather would provide a good explanation of this, even though it does not entail it. (The risk of interference also helps to explain why we are often more confident of inferences to an explanation than inferences from an explanation. When we start from the effect, we know that there was no effective interference.)

Attractions and repulsions

We have now said enough to give some content to the idea of Inference to the Best Explanation. What the account now needs is some specific argument on its behalf, which I will begin to provide in the next chapter. First, however, it will be useful to compile a brief list of its general and prima facie advantages and disadvantages, some of which have already been mentioned, to prepare the ground for a more detailed assessment. First of all, Inference to the Best Explanation seems itself to be a relatively lovely explanation of our inductive practices. It gives a natural description of familiar aspects of our inferential procedures. The simplest reason for this is that we are often aware that we are inferring an explanation of the evidence; but there is more to it. We are also often aware of making an inferential choice between competing explanations, and this typically works by means of the two-filter process my favored version of Inference to the Best Explanation describes. We begin by considering plausible candidate explanations, and then try to find data that discriminate between them.

The account reflects the fact that a hypothesis that is a reasonable inference in one competitive milieu may not be in another. An inference may be defeated when someone suggests a better alternative explanation, even though the evidence does not change. Inference to the Best Explanation also suggests that we assess candidate inferences by asking a subjunctive question: we ask how good the explanation *would* be, if it were true. There seems to be no reason why an inferential engine has to work in this way. If induction really did work exclusively by simple extrapolation, it would not involve subjunctive assessment. We can also imagine an inductive technique that included selection from among competitors but did not involve the subjunctive process. We might simply select on the basis of some feature of the hypotheses that we directly assess. In fact, however, we do often make the inductive decision whether something is true by asking what would be the case if it were, rather than simply deciding which is the likeliest possibility. We construct various causal scenarios and consider what they would explain and how well. Why is my refrigerator not running? Perhaps the fuse has blown. Suppose it has; but then the kitchen clock should not run either, since the clock and refrigerator are on the same circuit. Is the clock running? By supposing for the moment that a candidate explanation is correct, we can work out what further evidence is relevant to our inference. The role of subjunctive reasoning is partially captured by the familiar observation about the 'priority of theory over data'. Induction does not, in general, work by first gathering all the relevant data and only then considering the hypotheses to which they apply, since we often need to entertain a hypothesis first in order to determine what evidence is relevant to it (Hempel 1966: 12–13). But the point about subjunctive evaluation is not only that explanatory hypotheses are needed to determine evidential relevance, but also a partial description of how that determination is made. (One of the attractions of the hypothetico-deductive model is that it also captures this subjunctive aspect of our methods for assessing relevant evidence, since we determine what a hypothesis entails by asking what would have to be the case if the hypothesis were true.)

Although we often infer an explanation just because that is where our interests lie, Inference to the Best Explanation correctly suggests that explanatory inferences should be common even in cases where explaining is not our primary purpose. Even when our main interest is in accurate prediction or effective control, it is a striking feature of our inferential practice that we often make an 'explanatory detour'. If I want to know whether my car will start tomorrow, my best bet is to try to figure out why it sometimes failed to start in the past. When Ignaz Semmelweis wanted to control the outbreak of childbed fever in one of the maternity wards in a Vienna hospital, he proceeded by trying to explain why the women in that ward were contracting the disease, and especially the contrast between the women in that ward and the women in another ward in the same hospital,

who rarely contracted it. (We will consider this case at length in the next chapter.) The method of explanatory detour seems to be one of the sources of the great predictive and manipulative successes of many areas of science. In science, the detour often requires 'vertical' inference to explanations in terms of unobserved and often unobservable entities and processes, and Inference to the Best Explanation seems particularly well equipped to account for this process.

In addition to giving a natural description of these various features of our inferential practice, Inference to the Best Explanation has a number of more abstract attractions. The notion of explanatory loveliness, upon which an interesting version of Inference to the Best Explanation relies, should help to make sense of the common observation of scientists that broadly aesthetic considerations of theoretical elegance, simplicity and unification are a guide to inference. More generally, as I have already mentioned, the account describes a deep connection between our inferential and explanatory behavior, one that accounts for the prevalence of explanatory inferences even in cases where our main interests lie elsewhere. As such, it also helps with one of the problems of justifying our explanatory practices, since it suggests that one of the points of our obsessive search for explanations is that this is a peculiarly effective way of discovering the structure of the world. The explicit point of explaining is to understand *why* something is the case but, if Inference to the Best Explanation is correct, it is also an important tool for discovering *what* is the case.

Another sort of advantage to the view that induction is Inference to the Best Explanation is that it avoids some of the objections to competing models of inductive inference or confirmation that we discussed in chapter 1. One of the weaknesses of the simple Humean extrapolation model ('More of the Same') is that we are not always willing to extrapolate and, when we are, the account does not explain which of many possible extrapolations we actually choose. Inference to the Best Explanation does not always sanction extrapolation, since the best explanation for an observed pattern is not always that it is representative (Harman 1965: 90–1). Given my background knowledge, the hypothesis that I always win is not a good explanation of my early successes at the roulette wheel. Similarly, given a finite number of points on a graph marking the observed relations between two quantities, not every curve through those points is an equally good explanation of the data. One of the severe limitations of both the extrapolation view and the instantial model of confirmation is that they do not cover vertical inferences, where we infer from what we observe to something at a different level that is often unobservable. As we have seen, Inference to the Best Explanation does not have this limitation: it appears to give a univocal account of horizontal and vertical inferences, of inferences to what is observable and to the unobservable. The instantial model is also too permissive, since it generates the raven paradox. In

chapter 6, I will attempt to show how Inference to the Best Explanation helps to solve it.

Inference to the Best Explanation also seems a significant advance over the hypothetico-deductive model. First, while that model has very little to say about the 'context of discovery', the mechanisms by which we generate candidate hypotheses, the two-filter version of Inference to the Best Explanation suggests that explanatory considerations may apply to both the generation of candidates and the selection from among them. Secondly, since the deductive model is an account of confirmation rather than inference, it does not say when a hypothesis may actually be inferred. Inference to the Best Explanation does better here, since it brings in competition selection. Thirdly, while the hypothetico-deductive model allows for vertical inference, it does not say much about how 'high' the inference may legitimately go. It allows the evidence to confirm a hypothesis however distant from it, so long as auxiliary premises can be found linking the two. In the next chapter, I will argue that Inference to the Best Explanation rightly focuses the impact of evidence more selectively, so that only some hypotheses that can be made to entail the evidence are supported by it. Fourthly, if the model limits auxiliary hypotheses to independently known truths, it is too restrictive, since evidence may support a hypothesis even though the evidence is not entailed by the hypothesis and those auxiliaries, and it may disconfirm a hypothesis without contradicting it. Inference to the Best Explanation allows for this sort of evidence, since explanation does not require deduction. Finally, Inference to the Best Explanation avoids several of the sources of over-permissiveness that are endemic to the deductive model. In addition to avoiding the raven paradox, Inference to the Best Explanation blocks confirmation by various non-explanatory deductions. One example is the confirmation of an arbitrary conjunction by a conjunct, since a conjunction does not explain its conjuncts. For these as well as for some of the other liabilities of the hypothetico-deductive model, a symptom of the relative advantages of Inference to the Best Explanation is that many of the problems of the hypothetico-deductive model of confirmation and of the deductive-nomological model of explanation 'cancel out': many of the counter-examples of the one are also counterexamples of the other. This suggests that the actual explanatory relation offers an improved guide to inference.

We have so far canvassed two sorts of advantage of Inference to the Best Explanation. The first is that it is itself a lovely explanation of various aspects of inference; the second is that it is better than the competition. The third and final sort of advantage I will mention is that, in addition to accounting for scientific and everyday inference, Inference to the Best Explanation has a number of distinctively philosophical applications. The first is that it accounts for its own discovery. In chapter 1, I suggested that the task of describing our inductive behavior is itself a broadly inductive project, one of going from what we observe about our inferential practice to

the mechanism that governs it. If this is right, a model of induction ought to apply to itself. Clearly the extrapolation and the instantial models do not do well on this criterion, since the inference is to the contents of a black box, the sort of vertical inference those models do not sanction. Nor does the hypothetico-deductive model do much better, since it does not entail many observed features of our practice. It does not, for example, entail any of the inferences we actually make. Inference to the Best Explanation does much better on this score. The inference to an account of induction is an explanatory inference: we want to explain why we make the inferences we do. Our procedure has been to begin with a pool of plausible candidate explanations – the various models of induction we have canvassed – and then select the best. This is a process of competition selection which works in part by asking the subjunctive question of what sorts of inferences we would make, if we used the various models. Moreover, if we do end up selecting Inference to the Best Explanation, it will not simply be because it seems the likeliest explanation, but because it has the features of unification, elegance, and simplicity that make it the loveliest explanation of our inductive behavior.

Another philosophical application of Inference to the Best Explanation is to the local justification of some of our inferential practices. For example, it is widely supposed that a theory is more strongly supported by successful predictions than by data that were known before the theory was constructed and which the theory was designed to accommodate. At the same time, the putative advantage of prediction over accommodation is controversial and puzzling, because the logical relationships between theory and data upon which inductive support is supposed to depend seem unaffected by the merely historical fact of when the data were observed. But there is a natural philosophical inference to the best explanation that seems to defend the epistemic distinction. When data are predicted, the best explanation for the fit between theory and data, it is claimed, is that the theory is true. When the data are accommodated, however, there is an alternative explanation of the fit, namely that the theory was designed just for that purpose. This explanation, which only applies in the case of accommodation, is better than the truth explanation, and so Inference to the Best Explanation shows why prediction is better than accommodation (cf. Horwich 1982: 111). We will assess this argument in chapter 10.

Another example of the application of Inference to the Best Explanation to local philosophical justification is in connection with Thomas Kuhn's provocative discussion of 'incommensurability' (Kuhn 1970, esp. chs IX–X). According to him, there is no straightforward way of resolving scientific debates during times of 'scientific revolutions', because the disputants disagree about almost everything, including the evidence. This seems to block resolution by any appeal to a crucial experiment. On the traditional view of such experiments, they resolve theoretical disputes by providing

evidence that simultaneously refutes one theory while supporting the other. Competing theories are found to make conflicting predictions about the outcome of some experiment; the experiment is performed and the winner is determined. But this account seems to depend on agreement about the outcomes of experiment, which Kuhn denies. These experiments can, however, be redescribed in terms of Inference to the Best Explanation in a way that does not assume shared observations. A crucial experiment now becomes two experiments, one for each theory. The outcome of the first experiment is explained by its theory, whereas the outcome of the second is not explained by the other theory, so we have some basis for a preference. Shared standards of explanation may thus compensate for observational disagreement: scientists should prefer the theory that best explains its proper data. There is, however, more to Kuhn's notion of incommensurability than disagreement over the data; in particular, there is also tacit disagreement over explanatory standards. But this may turn out to be another advantage of Inference to the Best Explanation. Insofar as Kuhn is right here, Inference to the Best Explanation will capture the resulting indeterminacy of scientific debate that is an actual feature of our inferential practices.

Another well-known philosophical application of Inference to the Best Explanation is to argue for various forms of realism. For example, as part of an answer to the Cartesian skeptic who asks how we can know that the world is not just a dream or that we are not just brains in vats, the realist may argue that we are entitled to believe in the external world since hypotheses that presuppose it provide the best explanation of our experiences. It is possible that it is all a dream, or that we are really brains in vats, but these are less good explanations of the course of our experiences that the ones we all believe, so we are rationally entitled to our belief in the external world. There is also a popular application of Inference to the Best Explanation to realism in the philosophy of science, which we have already briefly mentioned. The issue here is whether scientific theories, particularly those that appeal to unobservables, are getting at the truth, whether they are providing an increasingly accurate representation of the world and its contents. There is an inference to the best explanation for this conclusion. In brief, later theories tend to have greater predictive success than those they replace, and the best explanation for this is that later theories are better descriptions of the world than earlier ones. We ought to infer scientific realism, because it is the best explanation of predictive progress. We will assess this argument in chapter 11.

Let us conclude this chapter with some of the challenges to Inference to the Best Explanation we will consider in detail in chapters to come. First, several of the philosophical applications of Inference to the Best Explanation can be questioned. In the case of the argument for the advantages of prediction over accommodation, one may ask whether the 'accommodation explanation' really competes with the truth explanation (Horwich 1982:

112–16). If not, then as we saw in the last section, Inference to the Best Explanation does not require that we choose between them. Moreover, the assumption that they do compete, that explaining the fit between theory and accommodated data by appeal to the act of accommodation pre-empts explaining the fit by appeal to the truth of the theory, seems just to assume that accommodation does not provide support, and so to beg the question. (See chapter 10.) As for the argument for realism about the external world, do our beliefs about the world really provide a better explanation than the dream hypothesis, or is it simply that this is the explanation we happen to prefer? Again, doesn't the inference to scientific realism as the best explanation for predictive success simply assume that Inferences to the Best Explanation are guides to the truth about unobservables, which is just what an opponent of scientific realism would deny? (See chapter 11.) A second sort of liability is the suspicion that Inference to the Best Explanation is still nothing more than Inference to the Likeliest Cause in fancy dress, and so fails to account for the symptoms of likeliness. (See chapters 7 and 8.) Third, insofar as there is a concept of explanatory loveliness that is conceptually distinct from likeliness, one may question whether this is a suitable criterion of inference. On the one hand, there is what we may call 'Hungerford's objection', in honor of the author of the line 'Beauty is in the eye of the beholder'. Perhaps explanatory loveliness is too subjective and interest relative to give an account of inference that reflects the objective features of inductive warrant. On the other hand, supposing that loveliness is as objective as inference, we have 'Voltaire's objection'. What reason is there to believe that the explanation that would be loveliest, if it were true, is also the explanation that is most likely to be true? Why should we believe that we inhabit the loveliest of all possible worlds? And why should we suppose that *any* of the potential explanations of a given set of data we happen to think of is likely to be true? (See chapter 9.) As we saw in the last section, Inference to the Best Explanation requires that we work with a notion of potential explanation that does not carry a truth requirement. Once we have removed truth from explanation, however, it is not clear how we get it back again (Cartwright 1983: 89–91). Lastly, and perhaps most importantly, it will be claimed that Inference to the Best Explanation is only as good as our account of explanatory loveliness, and this account is non-existent. In the next chapter, I begin to meet this objection.

Chapter 5

Contrastive inference

A case study

In this chapter and the next four, I will consider some of the prospects of Inference to the Best Explanation as a solution to the descriptive problem of inductive inference. We want to determine how illuminating that account is as a partial description of the mechanism inside the cognitive black box that governs our inductive practices. To do this, we need to show how explanatory considerations are a guide to inference or confirmation; how loveliness helps to determine likeliness. In particular, we want to see whether the model can meet the two central challenges from the last chapter, to show that inferences to the best explanation are more than inferences to the likeliest cause, and to show that Inference to the Best Explanation marks an advance over the simple hypothetico-deductive model.

As I have stressed, a major challenge facing this project is our poor understanding of what makes one explanation lovelier than another. Little has been written on this subject, perhaps because it has proven so difficult even to say what makes something an explanation. How can we hope to determine what makes one explanation better than another, if we cannot even agree about what distinguishes explanations of any quality from something that is not an explanation at all? Moreover, most of what has been written about explanatory loveliness has focused on the interest relativity of explanation, which seems to bring out pragmatic and subjective factors that are too variable to provide a suitably objective measure of inductive warrant.

Yet the situation is not hopeless. My analysis of contrastive explanation in chapter 3 will help. There I argued that phenomena we explain often have a contrastive fact–foil structure, and that the foil helps to select the part of the causal history of the fact that provides a good explanation by means of a mechanism of causal triangulation. According to my Difference Condition, to explain why P rather than Q, we need a causal difference between P and not-Q, consisting of a cause of P and the absence of a corresponding event in the case of not-Q. Thus we can explain why Jones rather than Smith contracted paresis by pointing out that only Jones had syphilis, since this is

to point out a causal difference between the two men, even though most people with syphilis do not get paresis. This account of contrastive explanation shows how what counts as a good explanation depends on interests, since interests determine the choice of foil, and a cause that marks a difference for one foil will not generally do so for another. Jones's syphilis would not explain why he rather than Doe (who also had syphilis) contracted paresis; here the explanation might be instead that only Jones left his syphilis untreated. In this respect, then, what counts as a lovely explanation of P depends on one's interests, but this cashes out into the question of whether the cited cause provides any explanation at all of the contrast that expresses a particular interest.

The sensitivity of explanation to choice of foils captures much of what has been said about interest relativity, and it also shows that these factors are not strongly subjective in a way that would make them irrelevant to inference. An account of the interest relativity of explanation would be strongly subjective if it showed that what counts as a good explanation depends on the tastes of the audience rather than the causal structure of the world. Examples that would threaten the idea of Inference to the Best Explanation would be cases where people favor incompatible explanations of the same phenomenon, even though their evidence and their inferential inclinations are the same. It is no threat to the objectivity of explanation that different people should be interested in explaining different phenomena, and it is obvious that a good explanation of one phenomenon is not usually a good explanation of another. A contrastive analysis of explanation supports only this innocuous form of relativity, if we construe the phenomena themselves as contrastive, so that a change in foil yields a different phenomenon. Moreover, my analysis of contrastive explanation shows that a change in foil helps to select a different part of the same causal history. Differences in interest require different but compatible explanations, which does not bring in strong subjectivity. And this much interest relativity is also something any reasonable account of inference must acknowledge: different people may all reasonably infer different things from shared evidence, depending on their inferential interests, when the inferences are compatible.

So my account of contrastive explanation helps to defuse the objection to Inference to the Loveliest Explanation that loveliness is hopelessly subjective. (We will return to this issue in later chapters.) It also provides the core of a positive account of one way that explanatory considerations can serve as a guide to inference. The reason for this is the structural similarity between the Difference Condition and Mill's Method of Difference. According to Mill, we find the cause of a fact in some prior difference between a case where the fact occurs and an otherwise similar case where it does not. Mill's central mechanism for inferring the likeliest cause is almost the same as the mechanism of causal triangulation that helps to determine the loveliest explanation. This near-isomorphism

provides an important argument in favor of Inference to the Best Explanation, since it shows that a criterion we use to evaluate the quality of potential explanations is the same as one we use to infer causes. By inferring something that would provide a good explanation of the contrast if it were a cause, we are led to infer something that is likely to be a cause. Returning to poor Jones, we may find that his condition, taken alone, points to no particular explanation. But if we try instead to explain why Jones rather than Smith contracted paresis, we will be led, by means of the Difference Condition, to look for some possibly relevant difference in the medical histories of the two men. Thus we may infer that Jones's syphilis was a cause of his paresis, since this is an explanatory difference. And this is just where Mill's method would take us, if syphilis was the only possibly relevant difference. Moreover, our explanation and our inference will both change if we change the foil. If we ask why Jones rather than Doe contracted paresis, we will be led to explain this contrast by appeal to Doe's treatment. By varying the foil, we change the best explanation, and this leads us to different but compatible inductive inferences, taking us to different stages of Jones's medical history.

By considering inferences to contrastive explanations, we go some way towards meeting the challenge that Inference to the Best Explanation is nothing more than Inference to the Likeliest Cause, where likeliness is judged on some basis entirely independent of explanatory considerations. Since looking for residual differences in similar histories of fact and foil is a good way of determining a likely cause, as Mill taught us, and contrastive explanation depends on just such differences, looking for potential contrastive explanations can be a guide to causal inference. Given contrastive data, the search for explanation is an effective way of determining just what sort of causal hypotheses the evidence supports. This procedure focuses our inferences, by eliminating putative causes that are in the shared part of antecedents of fact and foil. These antecedents may well be causally relevant, but the fact that they would not explain the contrast shows that the contrast does not (at least by itself) provide evidence that they are causes. This version of Inference to the Best Explanation thus sheds some light on the context of discovery, since the requirement that a potential explanation cite a difference severely restricts the class of candidate hypotheses. It also brings out one role of background knowledge in inference in a natural way, since our judgment of which antecedents are shared, a judgment essential to the application of the method, will depend on such knowledge.

I also want to argue, in this chapter and the next, that Inference to the Best Contrastive Explanation helps to meet the second challenge, to show that the model is better than simple hypothetico-deductivism. It marks an improvement both where the deductive model is too strict, neglecting evidential relevance in cases where there is no appropriate deductive

connection between hypothesis and data, and where it is too lenient, registering support where there is none to be had as, for example, revealed by the raven paradox. Inference to the Best Explanation does better in the first case because, as the analysis of contrastive explanation shows, explanatory causes need not be sufficient for their effects, so the fact that a hypothesis would explain a contrast may provide some reason to believe the hypothesis, even though the hypothesis does not entail the data. It does better in the second case because, while some contrapositive instances (e.g. non-black non-ravens) do support a hypothesis, not all do, and the requirement of shared antecedents helps to determine which do and which do not. The structural similarity between the Method of Difference and contrastive explanation that I will exploit in these chapters will also eventually raise the question of why Inference to the Best Explanation is an improvement on Mill's methods, a question I will address in chapter 8.

To develop these arguments and, more generally, to show just how inferences to contrastive explanations work, it is useful to consider a simple but actual scientific example in some detail. The example I have chosen is Ignaz Semmelweis's research from 1844–8 on childbed fever, inspired by Carl Hempel's well-known and characteristically clear discussion (1966: 3–8) and Semmelweis's own account of his work (1860). Semmelweis wanted to find the cause of this often fatal disease, which was contracted by many of the women who gave birth in the Viennese hospital in which he did his research. Semmelweis's central datum was that a much higher percentage of the women in the First Maternity Division of the hospital contracted the disease than in the adjacent Second Division, and Semmelweis sought to explain this difference. The hypotheses he considered fell into three types. In the first were hypotheses that did not mark differences between the divisions, and so were rejected. Thus, the theory of 'epidemic influences' descending over entire districts did not explain why more women should die in one division than another; nor did it explain why the mortality among Viennese women who gave birth at home or on the way to the hospital was lower than in the First Division. Similarly, the hypotheses that the fever is caused by overcrowding, by diet or by general care were rejected because these factors did not mark a difference between the divisions.

One striking difference between the two divisions was that medical students only used the First Division for their obstetrical training, while midwives received their training in the Second Division. This suggested the hypothesis that the high rate of fever in the First Division was caused by injuries due to rough examination by the medical students. Semmelweis rejected the rough examination hypothesis on the grounds that midwives performed their examinations in more or less the same way, and that the injuries due to childbirth are in any case greater than those due to rough examination.

The second type of hypotheses were those that did mark a difference between the divisions, but where eliminating the difference in putative cause did not affect the difference in mortality. A priest delivering last sacrament to a dying woman had to pass through the First Division to get to the sickroom where dying women were kept, but not through the Second Division. This suggested that the psychological influence of seeing the priest might explain the difference, but Semmelweis ruled this out by arranging for the priest not to be seen by the women in the First Division either and finding that this did not affect the mortality rates. Again, women in the First Division delivered lying on their backs, while women in the Second delivered on their sides, but when Semmelweis arranged for all women to deliver on their sides, the mortality remained the same.

The last type of hypothesis that Semmelweis considered is one that marked a difference between the divisions, and where eliminating this difference also eliminated the difference in mortality. Kolletschka, one of Semmelweis's colleagues, received a puncture wound in his finger during an autopsy, and died from an illness with symptoms like those of childbed fever. This led Semmelweis to infer that Kolletschka's death was due to the 'cadaveric matter' that the wound introduced into his blood stream, and Semmelweis then hypothesized that the same explanation might account for the deaths in the First Division, since medical students performed their examinations directly after performing autopsies, and midwives did not perform autopsies at all. Similarly, the cadaveric hypothesis would explain why women who delivered outside the hospital had a lower mortality from childbed fever, since they were not examined. Semmelweis had the medical students disinfect their hands before examination, and the mortality rate in the First Division went down to the same low level as that in the Second Division. Here at last was a difference that made a difference, and Semmelweis inferred the cadaveric hypothesis.

This case is a gold mine for inferences to the best contrastive explanation. Let us begin by considering Semmelweis's strategy for each of the three groups of hypotheses: those of no difference, of irrelevant differences and of relevant differences. Semmelweis's rejection of the hypotheses in the first group – epidemic influences, overcrowding, general care, diet and rough examination – show how Inference to the Best Explanation can account for negative evidence. These hypotheses are rejected on the grounds that, though they are compatible with the evidence, they would not explain the contrast between the divisions. Epidemic influences, for example, still might possibly be part of the causal history of the deaths in the First Division, say because the presence of these influences is a necessary condition for any case of childbed fever. And nobody who endorsed the epidemic hypothesis would have claimed that the influences are sufficient for the fever, since it was common knowledge that not all mothers in the district contracted childbed fever. Still, Semmelweis took the fact that the hypotheses in the first group

would not explain the contrast between the divisions or the contrast between the First Division and mothers who gave birth outside the hospital to be evidence against them.

Semmelweis also used a complementary technique for discrediting the explanations in the first group that is naturally described in terms of Inference to the Best Explanation, when he argued against the epidemic hypothesis on the grounds that the mortality rate for births outside the hospital was lower than in the First Division. What he has done is to change the foil, and point out that the hypothesis also fails to explain this new contrast. It explains neither why mothers get fever in the First Division rather than in the Second, nor why mothers get fever in the First Division rather than outside the hospital. Similarly, when Semmelweis argued against the rough examination hypothesis on the grounds that childbirth is rougher on the mother than any examination, he pointed out not only that it fails to explain why there is fever in the First Division rather than in the Second, but also why there is fever in the First Division rather than among other mothers generally. New foils provide new evidence, in these cases additional evidence against the putative explanations.

The mere fact that the hypotheses in the first group did not explain some evidence can not, however, account for Semmelweis's negative judgment. No hypothesis explains every observation, and most evidence that is not explained by a hypothesis is simply irrelevant to it. But Semmelweis's observation that the hypotheses do not explain the contrast in mortality between the divisions seems to count against those hypotheses in a way that, say, the observation that those hypotheses would not explain why the women in the First Division were wealthier than those in the Second Division (if they were) would not. Of course, since Semmelweis was interested in reducing the incidence of childbed fever, he was naturally more interested in an explanation of the contrast in mortality than in an explanation of the contrast in wealth, but this does not show why the failure of the hypotheses to explain the first contrast counts against them. This poses a general puzzle for Inference to the Best Explanation: how can that account distinguish negative evidence from irrelevant evidence, when the evidence is logically consistent with the hypothesis?

One straightforward mechanism is rival support. In some cases, evidence counts against one hypothesis by improving the explanatory power of a competitor. The fact that the mortality in the First Division went down when the medical students disinfected their hands before examination supports the cadaveric matter hypothesis, and so indirectly counts against all the hypotheses inconsistent with it that cannot explain this contrast. But this mechanism of disconfirming an explanation by supporting a rival does not seem to account for Semmelweis's rejection of the hypotheses in the first group, since at that stage of his inquiry he had not yet produced an alternative account.

Part of the answer to this puzzle about the difference in the epistemic relevance of a contrast in mortality and a contrast in wealth is that the rejected hypotheses would have enjoyed some support from the fact of mortality but not from the fact of wealth. The epidemic hypothesis, for example, was not Semmelweis's invention, but a popular explanation at the time of his research. Its acceptance presumably depended on the fact that it seemed to provide an explanation, if a weak one, for the non-contrastive observations of the occurrence of childbed fever. In the absence of a stronger and competing explanation, this support might have seemed good enough to justify the inference. But by pointing out that the hypothesis does not explain the contrast between the divisions, Semmelweis undermines this support. On the other hand, the epidemic hypothesis never explained and so was never supported by observations about the wealth of the victims of childbed fever, so its failure to explain why the women in the First Division were wealthier than those in the Second Division would not take away any support it had hitherto enjoyed.

On this view, the observation that the hypotheses in the first group do not explain the contrast in mortality and the observation that they do not explain the contrast in wealth are alike in that they both show that these data do not support the hypothesis. The difference in impact only appears when we take into account that only evidence about mortality had been supposed to support the hypothesis, so only in this case is there a net loss of support. This view seems to me to be correct as far as it goes, but it leaves a difficult question. Why, exactly, does the failure to explain the contrast in mortality undermine prior support for hypotheses in the first group? Those hypotheses would still give some sort of explanation for the cases of the fever in the hospital, even if they would not explain the contrast between the divisions. Consider a different example. Suppose that we had two wards of patients who suffer from syphilis and discovered that many more of them in one ward contracted paresis than in the other. The hypothesis that syphilis is a necessary cause of paresis would not explain this contrast, but this would not, I think, lead us to abandon the hypothesis on the grounds that its support had been undermined. Instead, we would continue to accept it and look for some further and complementary explanation for the difference between the wards, say in terms of a difference in the treatments provided. Why, then, is Semmelweis's case any different?

The difference must lie in the relative weakness of the initial evidence in support of the hypotheses in the first group. If the only evidence in favor of the epidemic hypothesis is the presence of childbed fever, the contrast in mortality does undermine the hypothesis, because it suggests that the correct explanation of the contrast will show that epidemic influences have nothing to do with fever. If, on the other hand, the epidemic hypothesis would also explain why there were outbreaks of fever at some times rather than others, or in some hospitals rather than others, even though these cases seemed

similar in all other plausibly relevant respects, then we would be inclined to hold on to that hypothesis and look for a complementary explanation of the contrast between the divisions. In the case of syphilis and paresis, we presumably have extensive evidence that there are no known cases of paresis not preceded by syphilis. The syphilis hypothesis not only would explain why those with paresis have it, but also the many contrasts between people with paresis and those without it. This leads us to say that the correct explanation of the contrast between the wards is more likely to complement the syphilis hypothesis than to replace it.

If this account is along the right lines, then the strength of the disconfirmation provided by the failure to explain a contrast depends on how likely it seems that the correct explanation of the contrast will pre-empt the original hypothesis. This explains our different reaction to the wealth case. We may have no idea why the women in the First Division are wealthier than those in the second, but it seems most unlikely that the reason for this will pre-empt the hypotheses of the first group. When we judge that pre-emption is likely, we are in effect betting that the best explanation of the contrast will either contradict the original hypothesis or show it to be unnecessary, and so that the evidence that originally supported it will instead support a competitor. So the mechanism here turns out to be an attenuated version of disconfirmation by rival support after all. The inability of the hypotheses in the first group to explain the contrast between the divisions and the contrast between the First Division and births outside the hospital disconfirms those hypotheses because, although the contrastive data do not yet support a competing explanation, since none has yet been formulated, Semmelweis judged that the best explanation of those contrasts would turn out to be a competing rather than a complementary account. This judgment can itself be construed as an overarching inference to the best explanation. If we reject the hypotheses in the first group because they fail to explain the contrasts, this is because we regard the conjecture that the hypotheses are wrong to be a better explanation of the failures than that they are merely incomplete. Judgments of this sort are speculative, and we may in the end find ourselves inferring an explanation of the contrasts that is compatible with the hypotheses in the first group, but insofar as we do take their explanatory failures to count against them, I think it must be because we do make these judgments.

On this view, given a hypothesis about the etiology of a fact, and faced with the failure of that hypothesis to explain a contrast between that fact and a similar foil, the scientist must choose between the overarching explanations that the failure is due to incompleteness or that it is due to incorrectness. Semmelweis's rejections of the hypotheses in the first group are examples of choosing the incorrectness explanation. It is further corroboration of the claim that these choices must be made that we cannot make sense of Semmelweis's research without supposing that he also

sometimes inferred incompleteness. For while the cadaveric hypothesis had conspicuous success in explaining the contrast between the divisions, it fails to explain other contrasts that formed part of Semmelweis's evidence. For example, it does not explain why some women in the Second Division contracted childbed fever while others in that division did not, since none of the midwives who performed the deliveries in that division performed autopsies. Similarly, the cadaveric hypothesis does not explain why some women who had street births on their way to the hospital contracted the fever, since those women were rarely examined by either medics or midwives after they arrived. Consequently, if we take it that Semmelweis nevertheless had good reason to believe that infection by cadaveric matter was a cause of childbed fever, it can only be because he reasonably inferred that the best explanation of these explanatory failures was only that the cadaveric hypothesis is incomplete, not the only cause of the fever, rather than that it is incorrect. These cases also show that we cannot in general avoid the speculative judgment by waiting until we actually produce an explanation for all the known relevant contrasts, since in many cases this would postpone inference indefinitely.

Let us turn now to the two hypotheses of the second group, concerning the priest and delivery position. Unlike the hypotheses of the first group, these did mark differences between the divisions and so might explain the contrast in mortality. The priest bearing the last sacrament only passed through the First Division, and only in that division did mothers deliver on their backs. Since these factors were under Semmelweis's control, he tested these hypotheses in the obvious way, by seeing whether the contrast in mortality between the divisions remained when these differences were eliminated. Since that contrast remained, even when the priest was removed from the scene and when the mothers in both divisions delivered on their sides, these hypotheses could no longer be held to explain the original contrast.

This technique of testing a putative cause by seeing whether the effect remains when it is removed is widely employed. Semmelweis could have used it even without the contrast between the divisions, and it is worth seeing how a contrastive analysis could account for this. Suppose that all the mothers in the hospital delivered on their backs, and Semmelweis tested the hypothesis that this delivery position is a cause of childbed fever by switching positions. He might have only done this for some of the women, using the remainder as a control. In this case, the two groups would have provided a potential contrast. If a smaller percentage of the women who delivered on their sides contracted childbed fever, the delivery hypothesis would have explained and so been supported by this contrast. And even if Semmelweis had switched all the mothers, he would have had a potential diachronic contrast, by comparing the incidence of fever before and after the switch. In either case, a contrast would have supported the explanatory inference. In fact, however, these procedures would not have produced a

contrast, since delivery position is irrelevant to childbed fever. This absence of contrast would not disprove the delivery hypothesis. Delivering on the back might still be a cause of fever, but there might be some obscure alternate cause that came into play when the delivery position was switched. But the absence of the contrast certainly would disconfirm the delivery hypothesis. The reason for this is the same as in the case of the epidemic hypothesis: the likeliness of a better, pre-emptive explanation. Even if Semmelweis did not have an alternative explanation for the cases of fever, there must be another explanation in the cases of side delivery, and it is likely that this explanation will show that back delivery is irrelevant even when it occurs. As in the case of the hypotheses in the first group, when we take an explanatory failure to count against a hypothesis, even when we do not have an alternative explanation, this is because we infer that the falsity of the hypothesis is a better explanation for its explanatory failure than its incompleteness.

This leaves us with Semmelweis's final hypothesis, that the difference in mortality is explained by the cadaveric matter that the medical students introduced into the First Division. Here too we have an overarching explanation in play. Semmelweis had already conjectured that the difference in mortality was somehow explained by the fact that mothers were attended by medical students in the First Division and by midwives in the Second Division. This had initially suggested the hypothesis that the rough examinations given by the medical students was the cause, but this neither explained the contrast between the divisions nor the contrast between the mothers in the First Division and mothers generally, who suffer more from labor and childbirth than from any examination. The cadaveric hypothesis is another attempt to explain the difference between the divisions under the overarching hypothesis that the contrast is due to the difference between medical students and midwives. In addition to explaining the difference between divisions, this hypothesis would explain Kolletschka's illness, as well as the difference between the First Division and births outside the hospital.

Finally, Semmelweis tested this explanation by eliminating the cadaveric matter with disinfectant and finding that this eliminated the difference in the mortality between the divisions. This too can be seen as the inference to a contrastive explanation for a new contrast, where now the difference that is explained is not the simple difference in mortality between the divisions, but the diachronic contrast between the initial presence of that difference and its subsequent absence. The best explanation of the fact that removing the cadaveric matter is followed by the elimination of the difference in mortality is that it was the cadaveric matter that was responsible for that difference. By construing Semmelweis's evidence as a diachronic contrast, we bring out the important point that the comparative data have a special probative force that we would miss if we simply treated them as two separate confirmations of Semmelweis's hypothesis.

Semmelweis's research into the causes of childbed fever brings out many of the virtues of Inference to the Best Explanation when that account is tied to a model of contrastive explanation. In particular, it shows how explanatory considerations focus and direct inquiry. Semmelweis's work shows how the strategy of considering potential contrastive explanations focuses inquiry, even when the ultimate goal is not simply an explanation. Semmelweis's primary interest was to eliminate or at least reduce the cases of childbed fever, but he nevertheless posed an explanatory question: Why do women contract childbed fever? His initial ignorance was such, however, that simply asking why those with the fever have it did not generate a useful set of hypotheses. So he focused his inquiry by asking contrastive why-questions. His choice of the Second Division as foil was natural because it provided a case where the effect is absent yet the causal histories are very similar. By asking why the contrast obtains, Semmelweis focused his search for explanatory hypotheses on the remaining differences. This strategy is widely applicable. If we want to find out why some phenomenon occurs, the class of possible causes is often too big for the process of Inference to the Best Explanation to get a handle on. If, however, we are lucky or clever enough to find or produce a contrast where fact and foil have similar histories, most potential explanations are immediately 'cancelled out' and we have a manageable and directed research program. The contrast will be particularly useful if, as in Semmelweis's case, in addition to meeting the requirement of shared history, it is also a contrast that various available hypotheses will not explain. Usually, this will still leave more than one hypothesis in the field, but then further observation and experiment may produce new contrasts that leave only one explanation. This shows how the interest relativity of explanation is at the service of inference. By tailoring his explanatory interests (and his observational and experimental procedures) to contrasts that would help to discriminate between competing hypotheses, Semmelweis was able to judge which hypothesis would provide the best overall explanation of the wide variety of contrasts (and absences of contrast) he observed, and so to judge which hypothesis he ought to infer. Semmelweis's inferential interests determined his explanatory interests, and the best explanation then determined his inference.

Before assessing the prospects for a hypothetico-deductive analysis of Semmelweis's research, it is worth mentioning that while I have followed Hempel's sensible selection from among the arguments, Semmelweis's own presentation contains many more. Indeed he provides a kind of orgy of arguments from explanatory power that can only warm the heart of defenders of Inference to the Best Explanation. As Semmelweis himself put it, 'As soon as one knows that childbed fever rises from decaying matter which is conveyed from external sources, explanations are easy' (1860: 156). In his introduction to the translation of Semmelweis's work, K. Codell Carter gives a good sense of this remarkable explanatory range:

> Semmelweis drew from his account explanations for dozens of facts that had been recorded but never explained. To choose only a few examples, Semmelweis explained why infants never died from [childbed] fever while their mothers remained healthy, why the mortality rates of infants changed in certain ways, why women who delivered on the way to hospital or who delivered prematurely had a lower mortality rate, why the disease often appeared in particular patterns among patients, why the mortality rate was different in the two clinics and why it had changed in certain ways through history, why infections were rare during pregnancy or after delivery, why the disease appeared to be contagious, why it exhibited seasonal patterns, why the disease was concentrated in teaching hospitals, why some non-teaching hospitals had much lower mortality rate that others, and why the disease appeared with different frequencies in different countries and in different historical periods. (Semmelweis 1860: 39–40)

Semmelweis's hypothesis would explain all these things and his account of his work is a sustained argument that we should accept that hypothesis precisely because of this explanatory power and because of the failure of competing hypotheses to provide equally good explanations.

Explanation and deduction

Semmelweis's research is a striking illustration of inferences to the best explanation in action, and of the way they often exploit contrastive data. It is also Hempel's paradigm of the hypothetico-deductive method. So this case is particularly well suited for a comparison of the virtues of Inference to the Best Explanation and the deductive model. It shows, I will suggest, that Inference to the Best Explanation is better than hypothetico-deductivism.

Consider first the context of discovery. Semmelweis's use of contrasts and prior differences to help generate a list of candidate hypotheses illustrates one of the ways Inference to the Best Explanation elucidates the context of discovery, a central feature of our inductive practice neglected by the hypothetico-deductive model. The main reason for this neglect is easy to see. Hypothetico-deductivists emphasize the hopelessness of narrow inductivism, the view that scientists ought to proceed by first gathering all the relevant data without theoretical preconception and then using some inductive algorithm to infer from those data to the hypothesis they best support. Scientists never have all the relevant data, they often cannot tell whether or not a datum is relevant without theoretical guidance, and there is no general algorithm that could take them from data to a hypothesis that refers to entities and processes not mentioned in the data (Hempel 1966: 10–18). The hypothetico-deductive alternative is that, while scientists never have all the data, they can at least determine relevance if the hypothesis comes first.

Given a conjectural hypothesis, they know to look for data that either can be deduced from it or would contradict it. The cost of this account is that we are left in the dark about the source of the hypotheses themselves. According to Hempel, scientists need to be familiar with the current state of research, and the hypotheses they generate should be consistent with the available evidence but, in the end, generating good hypotheses is a matter of 'happy guesses' (1966: 15).

The hypothetico-deductivist must be right in claiming that there are no universally shared mechanical rules that generate a unique hypothesis from any given pool of data since, among other things, different scientists generate different hypotheses, even when they are working with the same data. Nevertheless, this 'narrow hypothetico-deductivist' conception of inquiry badly distorts the process of scientific invention. Most hypotheses consistent with the data are non-starters, and the use of contrastive evidence and explanatory inference is one way the field is narrowed. In advance of an explanation for some effect, we know to look for a foil with a similar history. If we find one, this sharply constrains the class of hypotheses that are worth testing. A reasonable conjecture must provide a potential explanation of the contrast, and most hypotheses that are consistent with the data will not provide this. (For hypotheses that traffic in unobservables, the restriction to potential contrastive explanations still leaves a lot of play: we will consider further ways the class of candidate hypotheses is restricted in chapters 8 and 9.)

The slogan 'Inference to the Best Explanation' may itself bring to mind an excessively passive picture of scientific inquiry, suggesting perhaps that we simply infer whatever seems the best explanation of the data we happen to have. But the Semmelweis example shows that the account, properly construed, allows for the feedback between the processes of hypothesis formation and data acquisition that characterizes actual inquiry. Contrastive data suggest explanatory hypotheses, and these hypotheses in turn suggest manipulations and controlled experiments that may reveal new contrasts that help to determine which of the candidates is the best explanation. This is one of the reasons the subjunctive element in Inference to the Best Explanation is important. By considering what sort of explanation the hypothesis would provide, if it were true, we assess not only how good an explanation it would be, but also what as yet unobserved contrasts it would explain, and this directs future observation and experiment. Semmelweis's research also shows that Inference to the Best Explanation is well suited to describe the role of overarching hypotheses in directing inquiry. Semmelweis's path to his cadaveric hypothesis is guided by his prior conjecture that the contrast in mortalities between the divisions is somehow due to the fact that deliveries are performed by medical students in the First Division, but by midwives in the Second Division. He then searches for ways of fleshing out this explanation and for the data that would test various proposals. Again, I have

suggested that we can understand Semmelweis's rejection of the priest and the birth position hypotheses in terms of an inference to a negative explanation. The best explanation for the observed fact that eliminating these differences between the divisions did not affect mortality is that the mortality had a different cause. In both cases, the intermediate explanations focus research, either by marking the causal region within which the final explanation is likely to be found, or by showing that a certain region is unlikely to include the cause Semmelweis is trying to find.

The hypothetico-deductive model emphasizes the priority of theory as a guide to observation and experiment, at the cost of neglecting the sources of theory. I want now to argue that the model also fails to give a good account of the way scientists decide which observations and experiments are worth making. According to the deductive model, scientists should check the observable consequences of their theoretical conjectures, or of their theoretical systems, consisting of the conjunction of theories and suitable auxiliary statements. As we will see below, this account is too restrictive, since there are relevant data not entailed by the theoretical system. As we have already seen, it is also too permissive, since most consequences are not worth checking. Any hypothesis entails the disjunction of itself and any observational claim whatever, but establishing the truth of such a disjunction by checking the observable disjunct rarely has any bearing on the truth of the hypothesis. The contrastive account of Inference to the Best Explanation is more informative, since it suggests Semmelweis's strategy of looking for observable contrasts that distinguish one causal hypothesis from competing explanations.

Even if we take both Semmelweis's hypotheses and his data as given, the hypothetico-deductive model gives a relatively poor account of their relevance to each other. This is particularly clear in the case of negative evidence. According to the deductive model, evidence disconfirms a hypothesis just in case the evidence either contradicts the hypothesis or contradicts the conjunction of the hypothesis and suitable auxiliary statements. None of the hypotheses Semmelweis rejects contradicts his data outright. For example, the epidemic hypothesis does not contradict the observed contrast in mortality between the divisions. Proponents of the epidemic hypothesis would have acknowledged that, like any other epidemic, not everyone who is exposed to the influence succumbs to the fever. They realized that not all mothers contract childbed fever, but rightly held that this did not refute their hypothesis, which was that the epidemic influence was a cause of the fever in those mothers that did contract it. So the hypothesis does not entail that the mortality in the two divisions is the same. Similarly, the delivery position hypothesis does not entail that the mortality in the two divisions is different when the birth positions are different; nor does it entail that the mortality will be the same when the positions are the same. Even if back delivery is a cause of childbed fever, the mortality in the

Second Division could have been as high as in the First, because the fever might have had other causes there. Similarly, the possibility of additional causes shows that back delivery could be a cause of fever even though the mortality in the First Division is lower than in the Second Division where all the mothers deliver on their sides. The situation is the same for all the other hypotheses Semmelweis rejects.

What does Hempel say about this? He finds a logical conflict, but in the wrong place. According to him, the hypotheses that appealed to overcrowding, diet or general care were rejected because the claims that the difference in mortality between the divisions was due to such differences 'conflict with readily observable facts' (1966: 6). The claim that, for example, the difference in mortality is due to a difference in diet is incompatible with the observation that there is no difference in diet. These are clearly cases of logical incompatibility, but they are not the ones Hempel needs: the claims that are incompatible with observation are not the general hypotheses Semmelweis rejects. Like the cadaveric hypothesis he eventually accepts, the hypotheses of overcrowding, diet and care are surely general conjectures about causes of childbed fever, not specific claims about the differences between the divisions. But the hypotheses that overcrowding, diet or general care is a cause of childbed fever is logically compatible with everything Semmelweis observes.

The hypothetico-deductivist must claim that hypotheses are rejected because, although they are compatible with the data, each of them, when conjoined with suitable auxiliary statements, is not. But what could such statements be? Each hypothesis must have a set of auxiliaries that allows the deduction that the mortality in the divisions is the same, which contradicts the data. The auxiliaries need not be known to be true, but they need to be specified. This, however, cannot be done. The proponent of the epidemic hypothesis, for example, does not know what additional factors determine just who succumbs to the influence, so he cannot say how the divisions must be similar in order for it to follow that the mortality should be the same. Similarly, Semmelweis knew from the beginning that back delivery cannot be necessary for childbed fever, since there are cases of fever in the Second Division where all the women delivered on their sides, but he cannot specify what all the other relevant factors ought to be. The best the hypothetico-deductivist can do, then, is to rely on *ceteris paribus* auxiliaries. If fever is caused by epidemic influence, or by back delivery, and everything else 'is equal', the mortality in the divisions ought to be the same. This, however, does not provide a useful analysis of the situation. Any proponent of the rejected hypotheses will reasonably claim that precisely what the contrast between the divisions shows is that not everything is equal. This shows that there is more to be said about the etiology of childbed fever, but it does not show why we should reject any of the hypotheses that Semmelweis does reject. Semmelweis's observations show that none of these hypotheses

would explain the contrasts, but they do not show that the hypotheses are false, on hypothetico-deductive grounds.

There is another objection to the *ceteris paribus* approach, and indeed to any other scheme that would generate the auxiliaries the deductive model requires to account for negative evidence. It would disprove too much. Recall that the cadaveric hypothesis does not itself explain all the relevant contrasts, such as why some women in the Second Division contracted childbed fever while others in that division did not, or why some women who had 'street births' on their way to the hospital contracted the fever while others did not. If the other hypotheses were rejected because they, along with *ceteris paribus* clauses, entail that there ought to be no difference in mortality between the divisions, then the model does not help us to understand why similar clauses did not lead Semmelweis to reject the cadaveric hypothesis as well.

From the point of view of Inference to the Best Explanation, we can see that there are several general and related reasons why the hypothetico-deductive model does not give a good description of the way causal hypotheses are disconfirmed by contrastive data. The most important is that the model does not account for the negative impact of explanatory failure. Semmelweis rejected hypotheses because they failed to explain contrasts, not because they were logically incompatible with them. Even on a deductive-nomological account of explanation, the failure to explain is not tantamount to a contradiction. In order to register the negative impact of these failures, the hypothetico-deductive model must place them on the Procrustean bed of logical incompatibility, which requires auxiliary statements that are not used by scientists and not usually available even if they were wanted. Second, the hypothetico-deductive model misconstrues the nature of explanatory failure, in the case of contrastive explanations. As we saw in chapter 3, to explain a contrast is not to deduce the conjunction of the fact and the negation of the foil, but to find some causal difference. The hypotheses Semmelweis rejects do not fail to explain because they do not entail the contrast between the divisions: the cadaveric hypothesis does not entail this either. They fail because they do not mark a difference between the divisions, either initially or after manipulation. Third, the model does not reflect the need, in the case of explanatory failure, to judge whether this is due to incompleteness or error. In the model, this decision becomes the one of whether we should reject the hypothesis or the auxiliaries in a case where their conjunction contradicts the evidence. This, however, is not the decision Semmelweis has to make. When he had all the mothers in both divisions deliver on their sides, and found that this did not affect the contrast in mortality, he did not have to choose between saying that the hypothesis that delivery position is a cause of fever is false and saying that the claim that everything else was equal is false. After his experiment, he knew that not everything else was equal,

but this left him with the question of whether he ought to reject the delivery hypothesis or just judge it to be incomplete.

The failures of the hypothetico-deductive model to capture the force of disconfirmation through explanatory failure also clearly count against Karl Popper's account of theory testing through falsification (1959). Although he is wrong to suppose that we can give an adequate account of science without relying on some notion of positive inductive support, Popper is right to suppose that much scientific research consists in the attempt to select from among competing conjectures by disconfirming all but one of them. Popper's mistake here is to hold that disconfirmation and elimination work exclusively through refutation. As the Semmelweis example shows, scientists also reject theories as false because, while they are not refuted by the evidence, they fail to explain the salient contrasts. Moreover, if my account of the way this sort of negative evidence operates is along the right lines, this is a form of disconfirmation that Popper's account cannot be modified to capture without abandoning his central proscription on positive support, since it requires that we make a positive judgment about whether the explanatory failure is more likely to be due to incompleteness or error, a judgment that depends on inductive considerations.

The hypothetico-deductive model appears to do a better job of accounting for Semmelweis's main positive argument for his cadaveric hypothesis, that disinfection eliminated the contrast in mortality between the divisions. Suppose we take it that the cadaveric hypothesis says that infection with cadaveric matter is a necessary cause of childbed fever, that everyone who contracts the fever was so infected. In this case, the hypothesis entails that where there is no infection, there is no fever which, along with plausible auxiliaries about the influence of the disinfectant, entails that there should be no fever in the First Division after disinfection. But this analysis does not do justice to the experiment, for three reasons. First of all, the claim that cadaveric infection is strictly necessary for fever, which is needed for the deduction, is not strictly a tenable form of the cadaveric hypothesis, since Semmelweis knew of some cases of fever, such as those in the Second Division and those among street births, where there was no cadaveric infection. Similarly, given that disinfection is completely effective, this version of the hypothesis entails that there should be no cases of fever in the First Division after disinfection, which is not what Semmelweis observed. What he found was rather that the mortality in the First Division went down to the same low level (just over 1 percent) as in the Second Division. As Hempel himself observes, Semmelweis eventually went on to 'broaden' his hypothesis, by allowing that childbed fever could also be caused by 'putrid matter derived from living organisms' (1966: 6); but if this is to count as broadening the hypothesis, rather than rejecting it, the original cadaveric hypothesis cannot have been that cadaveric infection is a necessary cause of the fever.

The second reason the deductive analysis of the disinfection experiment does not do justice to it is that the analysis does not bring out the special probative force of the contrastive experiment. Even if we suppose that cadaveric infection is necessary for fever, the hypothesis does not entail the *change* in mortality, but only that there should be no fever where there is disinfection, since it does not entail that there should be fever where there is no disinfection. But it is precisely this contrast that makes the experiment persuasive. What the hypothetico-deductivist could say here, I suppose, is that the change is entailed if we use the observed prior mortality as a premise in the argument. If cadaveric infection is necessary for fever, and if there was fever and infection but then the infection is removed, it follows that the fever will disappear as well. Even this, however, leaves out an essential feature of the experiment, which was the knowledge that, apart from disinfection, all the antecedents of the diachronic fact and foil were held constant. Finally, what makes the cadaveric experiment so telling is not only that it provides evidence that is well explained by the cadaveric hypothesis, but that the evidence simultaneously disconfirms the competitors. None of the other hypotheses can explain the temporal difference, since they all appeal to factors that were unchanged in this experiment. As we have seen in our discussion of negative evidence, however, the deductive model does not account for this process of disconfirmation through explanatory failure, and so it does not account for the way the evidence makes the cadaveric hypothesis the best explanation by simultaneously strengthening it and weakening its rivals.

I conclude that Inference to the Best Explanation, linked to an account of contrastive explanation that provides an alternative to the deductive-nomological model, is an improvement over the hypothetico-deductive model in its account of the context of discovery, the determination of relevant evidence, the nature of disconfirmation and the special positive support that certain contrastive experiments provide. In particular, Inference to the Best Explanation is an improvement because it allows for evidential relevance in the absence of plausible deductive connections, since contrastive explanations need not entail what they explain. If Inference to the Best Explanation is to come out as a suitable replacement for the hypothetico-deductive model, however, it is important to see that it does not conflict with the obvious fact that scientific research is shot through with deductive inferences. To deny that all scientific explanations can be cast in deductive form is not to deny that some of them can, or that deduction often plays an essential role in those that cannot be so cast. Semmelweis certainly relied on deductive inferences, many of them elementary. For example, he needed to use deductive calculations to determine the relative frequencies of fever mortalities for the two divisions and for street births. Moreover, in many cases of causal scientific explanation, deduction is required to see whether a putative cause would explain a particular contrast. One reason for

this is that an effect may be due to many causes, some of which are already known, and calculation is required to determine whether an additional putative cause would explain the residual effect. Consider, for example, the inference from the perturbation in the orbit of Uranus to the existence of Neptune. In order to determine whether Neptune would explain this perturbation, Adams and Leverrier had first to calculate the influence of the sun and known planets on Uranus, in order to work out what the perturbation was, and then had to do further calculations to determine what sort of planet and orbit would account for it (cf. Grossner 1970). As Mill points out, this 'Method of Residues' is an elaboration of the Method of Difference where the negative instance is 'not the direct result of observation and experiment, but has been arrived at by deduction' (1904: III.VIII.5). Through deduction, Adams and Leverrier determined that Neptune would explain why Uranus had a perturbed orbit rather than the one it would have had if only the sun and known planets were influencing its motion. This example also illustrates other roles for deduction, since calculation was required to solve Newton's equations even for the sole influence of the sun, and to go from the subsequent observations of Neptune to Neptune's mass and orbit. This particular inference to the best contrastive explanation would not have been possible without deduction.

Let us return now to the two challenges for Inference to the Best Explanation, that it mark an improvement over the hypothetico-deductive model, and that it tell us more than that inductive inference is often inference to the likeliest cause. I have argued that the Semmelweis case shows that Inference to the Best Explanation passes the first test. It helps to show how the account passes the second test, by illustrating some of the ways explanatory considerations guide inference and judgments of likeliness. Although Semmelweis's overriding interest was in control rather than in understanding, he focused his inquiry by asking a contrastive explanatory question. Faced with the brute fact that many women were dying of childbed fever, and the many competing explanations for this, Semmelweis did not simply consider which explanation seemed the most plausible. Instead, he followed an organized research program based on evidential contrasts. By means of a combination of conjecture, observation, and manipulation, Semmelweis tried to show that the cadaveric hypothesis is the only available hypothesis that adequately explains his central contrast in mortality between the divisions. This entire process is governed by explanatory considerations that are not simply reducible to independent judgments of likeliness. By asking why the mortality in the two divisions was different, Semmelweis was able to generate a pool of candidate hypotheses, which he then evaluated by appeal to what they could and could not explain, and Semmelweis's experimental procedure was governed by the need to find contrasts that would distinguish between them on explanatory grounds. When Semmelweis inferred the cadaveric hypothesis, it was not simply that what turned out to

be the likeliest hypothesis also seemed the best explanation: Semmelweis judged that the likeliest cause of most of the cases of childbed fever in his hospital was infection by cadaveric matter *because* this was the best explanation of his evidence.

The picture of Inference to the Best Explanation that has emerged from the example of Semmelweis's research is, I think, somewhat different from the one that the slogan initially suggests in two important respects. The slogan calls to mind a fairly passive process, where we take whatever data happen to be to hand and infer an explanation, and where the central judgment we must make in this process is which of a battery of explanations of the same data would, if true, provide the loveliest explanation. But as the example shows, Inference to the Best Explanation supports a picture of research that is at once more active and realistic, where explanatory considerations guide the program of observation and experiment, as well as of conjecture. The upshot of this program is an inference to the loveliest explanation but the technique is eliminative. Through the use of judiciously chosen experiments, Semmelweis determined the loveliest explanation by a process of manipulation and elimination that left only a single explanation of the salient contrasts. In effect, Semmelweis converted the question of the loveliest explanation of non-contrastive facts into the question of the only explanation of various contrasts. Research programs that make this conversion are common in science, and it is one of the merits of Inference to the Best Explanation that it elucidates this strategy. And it is because Semmelweis successfully pursues it that we have been able to say something substantial about how explanatory considerations can be a guide to inference without getting bogged down in the daunting question of comparative loveliness where two hypotheses do both explain the same data. At the same time, this question cannot be avoided in a full assessment of Inference to the Best Explanation, since scientists are not always as fortunate as Semmelweis in finding contrasts that discriminate between all the competitors. Accordingly, I will attempt partial answers in later chapters. First, however, I will consider in the next chapter the resources of Inference to the Best Explanation to avoid some of the over-permissiveness of the hypothetico-deductive model.

Chapter 6

The raven paradox

Unsuitable contrasts

The hypothetico-deductive model has three weaknesses. It neglects the context of discovery; it is too strict, discounting all relevant evidence that is compatible with the hypothesis but not entailed by it; and it is over-permissive, counting some irrelevant data as relevant. In the last chapter, I argued that Inference to the Best Explanation does better in the first two respects. In this chapter, I argue that it also does better in the third. The problem of over-permissiveness arises because there are consequences of a hypothetical system that do not support the hypothesis, and there are several related ways to generate such consequences. One is by strengthening the premise set. An observed consequence of a hypothesis about childbed fever may support that hypothesis, but it will not support the conjunctive hypothesis consisting of the fever hypothesis and some unrelated hypothesis, even though the conjunction will of course also entail the observation (Goodman 1983: 67–8). Similar problems arise with the indiscriminate use of auxiliary statements. At the limit, any hypothesis is supported by any datum consistent with it, if we use a conditional auxiliary whose antecedent is the hypothesis and whose consequent is the irrelevant datum. This places the hypothetico-deductivist in a bind: to meet the objections in the last chapter that some supporting evidence is not entailed by the hypothesis, he will have to be extremely permissive about the sorts of auxiliaries he allows, but what he then gains in coverage he pays for by also admitting irrelevant evidence. A second way to generate irrelevant consequences is to weaken the conclusion. For any consequence of a hypothesis that supports the hypothesis, there are also innumerable disjunctions of that consequence and the description of an irrelevant observation, and the truth of the disjunction can then be established by the irrelevant observation. 'All sodium burns yellow' entails that either this piece of sodium will burn yellow or there is a pen on my table, but the observation of the pen is irrelevant to the hypothesis. Third, as Goodman has shown, by using factitious predicates we can construct hypotheses that are not lawlike, which

is to say that they are not supported by the instances they entail (1983: 72–5). A grue emerald observed today does not support the hypothesis that all emeralds are grue, where something is grue just in case it is either observed before midnight tonight and green or not so observed and blue. Finally, we may generate apparently irrelevant consequences through contraposition: this is Hempel's raven paradox. The hypothesis that all Fs are G entails that this F is G, a consequence that seems to support the hypothesis. Since, however, this hypothesis is logically equivalent to 'All non-Gs are non-F', it also entails that this non-G is non-F, a consequence that does not seem to support the original hypothesis. 'All ravens are black' is logically equivalent to 'All non-black things are non-ravens', so the raven hypothesis entails that this white object (my shoe) is not a raven, but observations of my shoe seem irrelevant to the hypothesis (Hempel 1965: 14–15).

Inference to the Best Explanation has the resources to avoid some of these irrelevant consequences. One way of avoiding the problem of arbitrary conjuncts is to observe that a causal account of explanation, unlike the deductive-nomological model, does not allow such arbitrary strengthening. Adding irrelevant information to a good explanation can spoil it. On a causal model of explanation, when we say that something happened *because* of X, we are claiming that X is a cause, and this claim will be false if X includes causally irrelevant material. Alternatively, one might argue that logically strengthening the explanation by adding a gratuitous conjunct worsens the explanation, even if it does not spoil it altogether, so we ought only to infer the weaker premise. Similarly, Inference to the Best Explanation can rule out some weakened consequences, since the sodium hypothesis does not explain the disjunction of itself and an arbitrary statement. It is less clear what to say in the case of the disjunction of a relevant instance and an irrelevant one. Perhaps arbitrary disjunctions are not suitable objects of causal explanation, since it is unclear whether it makes sense to ask for the cause of such disjunctions. As for Goodman's new riddle of induction, it may be argued that while 'All emeralds are green' enables us to give an obliquely causal explanation of why this object is green, 'All emeralds are grue' does not explain why an object is grue, because 'grue' does not pick out the sort of feature that can be an effect. Finally, explanatory considerations seem to help with the raven paradox since, while the raven hypothesis may provide some sort of explanation for the blackness of a particular raven, neither the original hypothesis nor its contrapositive explain why this shoe is white. Inductive irrelevance and explanatory irrelevance appear largely to coincide.

All of these suggestions would require development and defense, but in what follows I will concentrate only on the case of the raven paradox, since it is here that my account of contrastive inference gets the best purchase. My main conclusion will be that the raven paradox does not arise for contrastive inferences. Before we see how that might be, it is worth emphasizing the generality of the problem. It is faced by any account of induction that

maintains the principle of instance confirmation, that any statement of the form 'All Rs are B' is confirmed by something that is R and B, and the equivalence condition, that whatever confirms a statement also confirms any other statement logically equivalent to it. By the principle of instance confirmation, the statement 'All non-Bs are non-R' is confirmed by something that is both non-B and non-R. And since 'All non-Bs are non-R' is logically equivalent to 'All Rs are B', by the equivalence condition that non-B non-R instance will also confirm the statement 'All Rs are B'. Thus a white shoe confirms the hypothesis that all non-black things are non-ravens, and so that all ravens are black. The paradox is tantalizing, because both the principle of instance confirmation and the equivalence condition are so plausible, yet the consequence that observing a white shoe provides some reason to believe that all ravens are black is so implausible. Surely this is not a good description of our actual inductive practices.

Three well-known solutions to the raven paradox will provide foils to my approach. The first is Hempel's own strategy, which is to bite the bullet. He frames the question of whether observing a non-black non-raven supports the raven hypothesis with the 'methodological fiction' that this observation is the only available information that might be relevant to the hypothesis; that is, that there is no relevant background information or other relevant evidence. He then claims that, under this idealizing assumption, a non-black non-raven does indeed provide some support for the hypothesis that all ravens are black, and that any intuitions to the contrary are due to the failure to respect the idealization (Hempel 1965: 18–20). This solution is consistent with the hypothetico-deductive model, but it is also unsatisfying, for three reasons. First of all, even if white shoes do support the raven hypothesis under the idealization, this leaves the interesting question unanswered, which is why in methodological fact we do not look to non-black non-ravens for support of the raven hypothesis. Secondly, the methodological fiction takes us so far from our actual inductive practice that there may be no way to determine what the right answer in the fictional case ought to be. Perhaps it is instead that even black ravens would not support the hypothesis under such extreme ignorance since, if we really knew nothing else relevant to the hypothesis, we would have no reason to believe that it is lawlike, that is, confirmable even by its direct instances. Not all generalizations are instance confirmable so, if we do not know anything relevant to the raven hypothesis, except that there exists one black raven, or one non-black non-raven, we do not know whether the hypothesis is confirmable or not. Finally, it is not clear that the idealization is sensible, because it is not clear that the question of support is coherent under those conditions. It makes sense to ask what the period of a pendulum is under the idealization of no friction, but not under the idealization of no length since, without length, you have no pendulum (cf. Cummins 1989: 78). Similarly, there may be no such thing as inductive support without

background information, just as there is no such thing as support without a hypothesis to be supported.

Another well-known attempt to solve the problem is due to Quine who, drawing on Goodman's discussion of the new riddle of induction, suggests that only instances of 'projectible' predicates provide inductive support and that the complement of a projectible predicate is not projectible. On this view, black ravens support both the raven hypothesis and its contrapositive, but non-black non-ravens support neither (Quine 1969: 114–16). This solution is not compatible with the simple hypothetico-deductive model, since it makes many consequences of a hypothesis inductively irrelevant to it. The main objection to this proposal is that some complements of projectible predicates are projectible. For example, some things that are neither rubbed nor heated do support the hypothesis that friction causes heat. As we will shortly see, this is far from an isolated case.

The last solution I will mention is Goodman's own. He claims that inductive support can be analyzed in terms of 'selective confirmation', which requires that confirming evidence provide both an instance of the hypothesis and a counterexample to its contrary. The contrary of 'All ravens are black' is 'No ravens are black', which is incompatible with 'This raven is black', but not with 'This shoe is white' (Goodman 1983: 70–1). Like Quine's proposal, this is incompatible with simple hypothetico-deductivism. The most serious objection to Goodman's solution is that, unlike Quine's proposal, it violates the very plausible equivalence condition on inductive support, that whatever supports a hypothesis also supports any logically equivalent hypothesis. A white shoe does not selectively confirm the raven hypothesis, but it does selectively confirm the logically equivalent hypothesis that all non-black things are non-ravens, since it is incompatible with the contrary hypothesis that no non-black things are non-ravens. It hardly removes the sense of paradox to be told that, while white shoes do not support the raven hypothesis, they do support a logical equivalent.

Bearing these proposals and their drawbacks in mind, I want now to argue that contrastive inference avoids the paradox. To do so, it is best to leave the particular example of the ravens to one side for the moment and to consider instead a more straightforwardly causal example. We will then return to those puzzling birds. In spite of its apparent simplicity, the raven hypothesis raises special problems, and an excessive focus on this example has made the general problem seem harder than it is. Semmelweis's cadaveric hypothesis, which so occupied our attention in the last chapter, is a more suitable case. The hypothesis is that childbed fever is caused by cadaveric infection. Simplifying somewhat, let us suppose that Semmelweis's hypothesis has the same form as the raven hypothesis, that all cases of mothers with cadaveric infection are cases of mothers with childbed fever.

Semmelweis's evidence for the cadaveric hypothesis consisted primarily in two contrasts, one synchronic and one diachronic. He noticed that the

hypothesis would explain why the incidence of the fever was much higher in one maternity division of the hospital than in the other at the same time, since it was only in the First Division that mothers were examined by people who had handled cadaveric matter. It would also explain why the incidence of fever in the First Division went down to the same low level as the Second Division after he had the medics who performed the examinations in the First Division disinfect their hands. In each case, Semmelweis relied *both* on instances of infection and fever and on instances of non-fever and non-infection, that is, both on instances of the direct hypothesis that all infections are fevers and on instances of the equivalent contrapositive hypothesis that all non-fevers are non-infections.

Contrastive inferences place the raven paradox in a new light. The problem is no longer to show why contrapositive instances do not provide support, or to explain away the intuition that they do not. In contrastive inferences, they provide strong support and this is consonant with our intuitions about these cases. Contrapositive instances are essential to contrastive inference (or to the Method of Difference, upon which much of my account of contrastive inference has so far been based), itself an indispensible vehicle of inductive support. The real problem is rather to show why some contrapositive instances support while others do not. The mothers who were not infected and did not contract the fever were contrapositive instances that provided an essential part of Semmelweis's evidence for his hypothesis, but Semmelweis's observation that his shoe was neither infected nor feverish would not have supported his hypothesis.

The structure of contrastive inference suggests a natural solution to this problem. The uninfected mothers supported Semmelweis's hypothesis because they provided a suitable foil to the direct instances of the infected mothers with fever. As we have seen, the main condition for the suitability of a foil is shared history with its fact. The ideal case is one where the only difference in the history of fact and foil is the presence of the putative cause. This ideal can never be strictly achieved, but the investigator will look for contrasts that appear to differ in no respects that are likely to be causally relevant, or at least for contrasts that share other suspect causes. One reason Semmelweis's shoe did not support his hypothesis is that it provided a contrast that was too different from infected mothers, so there was no reason to say that those mothers contracted the fever while the shoe did not because only the mothers were infected by cadaveric matter. Infection cannot be inferred as the best explanation of this contrast, because there are too many other differences.

The requirement that a supporting contrapositive instance be known to share most of its history with an observed direct instance shows how background knowledge is essential to contrastive support. Unless we have this knowledge, we cannot make the inference. Nobody would suggest that we consider applications of the Method of Difference under the

methodological fiction that we know nothing about the antecedents of fact and foil, since the method simply would not apply to such a case. Background knowledge may rule out a contrapositive instance either because we do not know enough history or because we do know that the histories of fact and foil differ in many respects. Another way background knowledge may rule out contrastive inference is if we already know that the absence of the effect in the case of the foil is not due to the absence of the putative cause. In the case of the shoe, Semmelweis already knew that, even if cadaveric infection was indeed a cause of childbed fever, the reason his shoe did not have fever was not because it was uninfected. He knew that only living organisms can contract fever, and this pre-empts the explanation by appeal to lack of infection. A foil is unsuitable if it is already known that it would not have manifested the effect, even if it had the putative cause. In sum, then, a contrapositive instance only provides support if it is known to have a similar history to a direct instance and if it is not already known that the absence of the effect is due to some pre-empting cause. It is only if both of these conditions are satisfied that the hypothesis is a candidate for the best explanation of the contrast.

The method of contrastive inference avoids the raven paradox, because the restrictions on suitable contrasts rule out irrelevant instances. This is a solution to the paradox that differs from those of Hempel, Quine and Goodman. Unlike Hempel's solution, it allows for a distinction between relevant and irrelevant instances, though this does not line up with the distinction between direct and contrapositive instances, and the contrastive solution does not require that we fly in the face of our intuitions about inductive relevance. It also rejects the terms of methodological fiction Hempel uses to frame his position, since the question of whether a contrapositive instance is a suitable foil can only be answered by appeal to background knowledge. Finally, since the contrastive inference does not count all contrapositive instances as providing support it is, unlike Hempel's solution, incompatible with the hypothetico-deductive model (and with the narrower principle of instance confirmation). The contrastive solution is also different from Quine's appeal to projectibility, since it shows that the complement of a projectible predicate may itself be projectible, and indeed must be in any situation where contrastive inferences can be made. Lastly, the contrastive solution differs from Goodman's use of selective confirmation, since it does not violate the equivalence condition. A hypothesis and its contrapositive are supported by just the same contrastive data. Whichever form of hypothesis we use, we need the same pairs of instances subject to the same constraints of similar histories and the absence of known pre-empting explanations.

This completes my central argument for the claim that Inference to the Best Explanation avoids some of the over-permissiveness of the deductive-nomological model. Not all contrapositive instances can be used to infer a

contrastive explanation, and the model helps to say which can and which cannot. But I suppose that no discussion of the raven paradox can ignore the ravens themselves, so we will now return to them, though what I will have to say about them is not essential to my general claim. Contrastive inference does not suffer from the raven paradox, but neither does it provide much help with the particular example of the raven hypothesis, that all ravens are black. The trouble is that this hypothesis does not find support in contrastive evidence at all. To find such evidence, we might look for birds similar to ravens that are not black that might be paired up with observed black ravens, but in fact this is not something worth doing. The problem is not that there are no such birds, but that finding them would not support the raven hypothesis. If we found those birds, they might even tend to disconfirm the raven hypothesis: if there are non-black birds very much like ravens in other respects, this may increase the likelihood that there are some non-black ravens we have yet to observe.

Why can the raven hypothesis not find contrastive support? One natural suspicion is that this is because it is not a causal hypothesis. This, however, is not the problem. First, the hypothesis is broadly causal, though the cause is only obliquely described. Here is one way of analyzing the situation. When we consider the hypothesis that all ravens are black, what we are considering is that there is something in ravens, a gene perhaps, that makes them black. Moreover, the hypothesis implies that this cause, whatever it is, is part of the essence of ravens (or at least nomologically linked to their essence). Roughly speaking, the hypothesis implies that birds similar to ravens except with respect to this cause would not interbreed with ravens. The hypothesis thus implicitly makes a claim that could be false even as it applies to ravens that are in fact black, since the cause of blackness could be merely accidental to those birds, a feature they could lack without forfeiting their species membership. The raven hypothesis is causal, but the cause is only characterized as something in ravens that is essential to them. This characterization is very broad, but it is necessary if we are to make sense of the way we would support the hypothesis. For if we supposed instead that the hypothesis is entirely neutral with respect to the essentiality of the cause, it would not be lawlike; that is, it would not be supported even by black ravens. If we thought it was merely some contingent feature of the black ravens we observed that made them black, then even observing many black ravens should give us no confidence that all ravens, past, present and future, are black. This raises the question of how we could experimentally distinguish between an essential and an accidental cause, a question we will address in the next section.

The reason the raven hypothesis is not susceptible to contrastive support cannot be that it is non-causal, if in fact it is causal. But even if its causal status is questionable, there is reason to believe that this is not the source of the problem, since contrastive inference is applicable to other hypotheses

that have no stronger claim to causal status. Consider the hypothesis that all sodium burns yellow. Being sodium does not cause something to have the dispositional property of burning yellow any more or less than being a raven causes it to be black. As in the case of the raven hypothesis, I would say that the sodium hypothesis claims that there is something in sodium that causes it to burn yellow, and that this feature is essential to sodium. So the sodium hypothesis does not have a stronger claim to causal status than the raven hypothesis, yet it is contrastively supported. One way we convince ourselves of its truth is by producing a flame that has no sodium in it and is not burning yellow, and then adding sodium and noticing the immediate change to a yellow flame. This diachronic contrast clearly supports the hypothesis.

What then is the difference between the sodium hypothesis and the raven hypothesis that explains why one enjoys contrastive support and the other does not? It is easy to see one reason why the raven hypothesis does not lend itself to diachronic contrast. We cannot transform a non-black non-raven into a raven to see whether we get a simultaneous transformation from non-black to black, in the way that we can transform a flame without sodium into a flame with sodium. Perhaps if such a transformation were possible, we could usefully apply the Method of Difference to the raven hypothesis. There is probably another reason why diachronic contrast is useful for sodium, but not for ravens. In the case of yellow flames that contain sodium, we may wonder whether it is the sodium or some other factor that is responsible for the color. The diachronic contrast eliminates other suspects. In the case of the ravens, on the other hand, we already presume that the color is caused by something in the ravens rather than, say, the lighting conditions under which ravens are generally observed, so we do not look to experiments that would support this. For the raven hypothesis, only the essentiality claim and not the causal claim are in question. For both these reasons, then, it is no mystery that the raven hypothesis is not susceptible to diachronic contrastive support. What about the synchronic case? Here the sodium and the ravens are similar: neither finds support. We have already seen this for the ravens. Similarly, even if there are elements like sodium that do not burn yellow, observing this contrast would not increase our confidence that all sodium burns yellow.

Neither hypothesis lends itself to synchronic contrastive support, as a matter of principle, and the raven hypothesis does not lend itself to a diachronic application, as a matter of fact, since we do not know how to transform non-ravens into ravens and we do already know that the ravens themselves are responsible for their color. The remaining question, then, is why there is not synchronic application in either case. The answer, I think, lies in the extreme obliqueness of the causal descriptions implicit in the two hypotheses. This makes it impossible for us to know, in the synchronic case, whether we have satisfied the shared background condition. Faced with a raven and generally similar but non-black bird, we know that the cause of the blackness of the raven must lie among the differences, but we are none the

wiser about what the relevant difference is, and whether it is essential or accidental to ravens, the crucial question. Similar remarks apply to a contrast between a sample of sodium and an apparently similar substance that does not burn yellow.

My view that it is the obliqueness of the causal description that disqualifies the raven hypothesis from synchronic contrastive support is corroborated by comparing this case to a more directly causal hypothesis about the coloration of birds. Suppose we wanted to test the hypothesis that gene B causes blackness, where we have some independent way to identify the presence of that gene. In this case, finding that all the black birds we have observed have gene B, and that otherwise genetically similar birds without gene B are uniformly non-black, would support the genetic hypothesis. Without something like the genetic hypothesis, however, we simply do not know what should count as a relevantly similar bird. Similar considerations apply to the case of sodium. If we had a hypothesis about just what feature of sodium causes it to burn yellow, and we found an element that was just like sodium except with respect to this feature and burning color, we would have a useful synchronic contrast. Without this, we do not know what a relevantly similar element would be. In the case of sodium, we can avoid this problem by using a diachronic contrast but, as we have seen, this is not an option for ravens.

The Method of Agreement

Contrastive inference avoids the raven paradox, but it does not account for the way we support the raven hypothesis itself. Since contrastive inference is modeled on Mill's Method of Difference, the natural place to look for a kind of causal inference that would cover the ravens is in Mill's other main method, the Method of Agreement. In the Method of Agreement, we look for two instances of the effect whose histories ideally have nothing in common but the putative cause (Mill 1904: III.VIII.1). Here we hold the effect constant and vary the history to see what stays the same, while in contrastive inference or the Method of Difference, we hold the history constant and vary the effect. The inference to the cadaveric hypothesis from the evidence of infected mothers who contracted fever, even though they differed in birth position, exposure to the priest, diet and so on, would be an application of the Method of Agreement.

Like the Method of Difference, applications of the Method of Agreement are naturally construed as inferences to the best explanation, where what is explained is a similarity rather than a contrast. To explain a similarity, we must cite a cause of the effect shared by the instances. For example, to explain why all the members of the club have red hair, we might point out that red hair is a requirement for admission. This explains why all of them have red hair, even though it does not explain why each of them does (cf.

Nozick 1974: 22), so explanations of similarity exhibit a divergence between explaining P and explaining 'P and Q', similar to the divergence we found in chapter 3 between explaining P and explaining 'P rather than Q'. Similarly, the cause that we cite to explain P in one milieu of similar effect may differ as we change the milieu, since a common cause in one case may not be in another. In the context of inference, the evidence of agreement focuses our inference on these shared elements that would explain the agreement, much as the contrastive evidence focuses our inference on explanatory differences.

The primary restriction on supporting instances of agreement is varying history. Ideally the pair of instances share only effect and putative cause; in practice, what is required is that the instances must at least differ in the presence of something that might otherwise have been thought to be a cause of the effect. The Method of Agreement requires a further restriction, because of the risk of what Mill calls 'the plurality of causes'. As he observes, the method may give the wrong result in cases where the effect in question can be caused in more than one way (III.X.2). Suppose that two people die, one from poisoning, the other from a knife wound. The Method of Agreement will discount both causes, since they are part of the variation between the instances. Worse, the method may count as cause an incidental similarity between the two, say that they both smoked. Mill suggests two ways of handling this problem. The first is to also apply the Method of Difference, where this is possible. The second is to gather additional and varied instances for the Method of Agreement: if many instances share only one suspect, either this is a cause or there are as many causes as there are instances, an unlikely result. There is, however, a third obvious way to handle the problem of the plurality of causes, and this imposes a further restriction on the instances suitable to the method. We will only use pairs of instances that our background knowledge leads us to believe have the same kind of cause, even if we do not yet know what it is.

These two restrictions of varied history and common etiology seem to protect the Method of Agreement from the raven paradox. Unlike contrastive inference, which always joins a direct-instance and a contrapositive-instance, the Method of Agreement requires different pairs for a hypothesis and its contrapositive. For the hypothesis that all infected mothers contract fever, we would look at a pair of infected mothers with fever; for the contrapositive that everything that does not contract childbed fever is not an infected mother, we might look instead at a pair of uninfected and healthy mothers. As in the case of contrastive inference, the Method of Agreement correctly allows that some contrapositive instances provide support, such as a pair of uninfected mothers without fever who had different delivery positions. Similarly, if we wanted to know whether a vitamin C deficiency causes scurvy, we would find support among healthy people without the deficiency, as well as among sufferers of scurvy with it. The method also correctly excludes a pair of shoes in the case of the cadaveric hypothesis, since these

shoes do not differ in any respect that we think might be causally relevant to contracting childbed fever. But what about a pair consisting of a healthy mother without fever and an uninfected shoe? The histories of the mother and the shoe vary a great deal, yet we do not want to say that they support the cadaveric hypothesis. I think we can exclude cases of this sort by appeal to the common etiology condition. We do not believe the shoe is free from childbed fever because it has not been infected, even if infection is the cause of fever; the reason shoes do not have fever is that they are not living organisms.

Now we can return to the ravens themselves. To determine whether a pair of agreeing instances would support the raven hypothesis, we must consider what the alternative causes might be. This is not obvious, because the raven hypothesis is only obliquely causal. Following my suggestion in the last section, let us take the raven hypothesis to imply that the cause of blackness in ravens is some feature of ravens that is essential to them, and the other suspects to be features of ravens that are non-essential. How can we discriminate empirically between these possibilities? The best we can do at the level of the raven hypothesis is to employ a direct application of the Method of Agreement. By finding only black ravens in varied locations and environments, we eliminate various accidental suspects, namely those that vary, and this provides some support for the hypothesis.

Why will we not bother to look for two very different non-black non-ravens? Some contrapositive instances are clearly irrelevant to the raven hypothesis, because they violate the common etiology restriction: the reason my shoe is white is not that it lacks some feature essential to ravens that makes them black. But why will we not bother to look at non-black birds? The reason is again the obliqueness of our description of the cause of blackness in ravens. If we observe one non-black bird, we know that it lacks whatever it is that makes black ravens black. But we are none the wiser about whether this missing factor is essential or accidental to ravens, which is the point at issue. If we have a pair of very different non-black birds, we may infer that the missing factor is among those that they share. This narrows down the suspects, but we still have no way of saying whether the remaining candidates are essential or accidental to ravens. Indeed, even if we discovered that these birds are not black because they do not have gene B, the 'blackness gene', we are in no better shape, since this information does not help us to determine whether the presence of this gene is essential to ravens. This, I think, is the solution to the raven paradox as it applies to the peculiar example that gives it its name. When we look at various black ravens, at least we are ruling out some alternative causes, since we know that every respect in which these ravens differ is non-essential to them, but when we look at non-black birds, the information that we may gain about the etiology of coloration does us no good, since it does not discriminate between the raven hypothesis and its competitors.

In the last chapter, I argued that contrastive inference is a common evidential procedure that brings out many of the virtues of Inference to the Best Explanation and shows some of the ways it differs from the hypothetico-deductive model. In this chapter, I have continued that argument by suggesting why contrastive inference does not have to swallow the raven paradox. Since the hypothetico-deductive model does, this marks another difference that is to the credit of Inference to the Best Explanation. At the same time, the brief discussion of the Method of Agreement in this section underlines the fact that contrastive inference cannot be the whole of induction though, as I suggested, the Method of Agreement also seems amenable to explanatory analysis. I have chosen to focus in this book on the contrastive case, not because I think that it exhausts Inference to the Best Explanation, but because it is a particularly important form of inference and one that links up with a form of explanation that I have been able to articulate in some detail. What made this articulation possible is the close analogy between contrastive explanation and the Method of Difference. The analogy is so close that some readers, convinced that Inference to the Best Explanation really is different from hypothetico-deductivism, may now wonder whether it is not just a catchy name for Mill's method itself, rather than a distinctive account. This raises a third challenge to Inference to the Best Explanation. Even if the account is more than Inference to the Likeliest Cause, and improves on the hypothetico-deductive model, is it really any better than Mill's methods? I will consider this challenge in chapter 8. Before that, however, I want to compare Inference to the Best Explanation to yet another account of induction, one so far conspicuous by its absence. This is the Bayesian approach, which I consider in the next chapter.

Chapter 7

Bayesian abduction

The Bayesian approach

Bayesians hold that belief is a matter of degree and can be represented in terms of probabilities. Thus p(E) is the probability I give to the statement E, which may range from 0, if I am certain E is false, to 1, if I am certain E is true. By representing beliefs as probabilities, it is possible to use the mathematical theory of probability to give an account of the dynamics of belief, and in particular an account of inductive confirmation. The natural thought is that evidence E supports hypothesis H just in case the discovery of E causes (or ought to cause) me to raise my degree of belief in H. To put the point in terms of probabilities, E supports H just in case the probability of H after E is known is higher than the probability of H beforehand. In the jargon, what is required is that the posterior probability of H be greater than its prior probability.

What makes Bayesianism exciting is that the standard axioms of probability theory yield an equation that appears to tell us just when this condition of confirmation is satisfied, and so to give us a precise theory of induction. That equation is Bayes's theorem, which in its near simplest form looks like this:

$$p(H|E) = p(E|H)p(H)/p(E)$$

On the left-hand side, we have the conditional probability of H given E. Bayesians treat this as the posterior probability of H, so the figure on the left-hand side represents the degree of belief you should have after the evidence E is in. The right-hand side contains three probabilities, which together determine the posterior. The first of these – p(E|H) – is the probability of E given H, known as the 'likelihood' of the evidence E, because it represents how likely H would make E. The other two probabilities on the right-hand side – p(H) and p(E) – are the priors of H and E respectively. They represent degree of belief in hypothesis H before the evidence described by E is in, and degree of belief in E itself before the relevant observation is made. This process of moving from prior probabilities and likelihood to posterior probability by moving from right to left in Bayes's theorem is known as

conditionalizing and is claimed by the Bayesian to characterize the dynamic of degrees of belief and so the structure of inference.

To get the flavor of the way Bayes's theorem works, consider a hypothetico-deductive case; that is, a case where H deductively entails E. Here the likelihood p(E|H) is simply 1, since if H entails E then E must be true given H. Under this happy circumstance, Bayes's theorem becomes even simpler: the posterior of H is simply the prior of the hypothesis divided by the prior of the evidence: p(H)/p(E). Bayes's theorem thus tells us that a successful prediction of H confirms H, so long as the correctness of the prediction was neither certainly true or false before it was checked. For so long as the prior of E is less than 1 but more than 0 (something I will assume is always the case), p(H)/p(E) must be greater than p(H), and this is to say that the posterior of H will be greater than the prior of H, which is just when the Bayesian claims there is inductive confirmation. The theorem as we are using it also tells us that the successful prediction will provide greater confirmation the lower the prior of E, since the lower the prior of E, the more it boosts the posterior of H. Successful surprising predictions will count as providing stronger support for H than do predictions that were to be expected to be correct whether or not H was true. Bayes's theorem also tells us when evidence disconfirms a hypothesis. That is a situation where the posterior probability of H is less than its prior probability, and the theorem tells us that this will occur just in case the likelihood of E is lower than its prior probability: p(E|H) is lower than p(E). In other words, a hypothesis is disconfirmed when it would make the evidence more surprising. These are the sorts of things Bayes's theorem tells us, and pretty plausible things they are. Considerable work has now been done in this spirit, showing how Bayes's theorem can be made to capture diverse aspects of scientific judgment of the bearing of evidence on hypothesis, that is, how it can provide a good answer to the descriptive problems of induction (e.g. Horwich 1982; Howson and Urbach 1989; Earman 1992).

Bayesianism has been taken to pose a serious threat to Inference to the Best Explanation (van Fraassen 1989: ch. 7; Salmon 2001a). In its simplest form, the threatening argument says that Bayesianism is right, so Inference to the Best Explanation must be wrong. Here 'right' means a good description of our inductive practices, but this objection may also have a normative edge, since there are so-called 'dutch book' arguments which are supposed to show that anyone whose dynamic of belief failed to obey Bayes's theorem would be irrational. For example, someone who gives lovely explanations a higher posterior than Bayes's theorem would sanction is in trouble. Such a person would end up with an incoherent distribution of states of belief which, it is claimed, is the probabilistic analog to holding contradictory beliefs, and such a person would be exposed to a set of wagers at particular odds that he would both be disposed to accept and that would yield a certain monetary loss.

One response to the challenge is to argue directly that Bayesianism is not right: it does not give a good account of our inductive practices. Here there is already a substantial literature of objections to Bayesianism, mostly developed independently of Inference to the Best Explanation (e.g. Glymour 1980b: ch. III; Chihara 1987; Earman 1992). For example, one can question whether beliefs come in degrees as measured by the probability calculus, and whether the degree of belief in a hypothesis after the evidence is in should really be identified with the conditional probability of that hypothesis on the evidence. There is also the problem of old evidence. Evidence available before a hypothesis is formulated presumably may confirm it, yet in this case it will already have affected the prior probability of the hypothesis and so its impact on the posterior probability does not seem registered by the Bayesian formula. (We will consider whether the question of whether evidence is available before or after a hypothesis is formulated is at all relevant to inductive support in chapter 10.) Relatedly, there is a problem of new theories. The probability you assign to H may go up because you think of a new way of using it to explain your evidence, but this will not occur through the application of Bayes's theorem, since there is no new evidence for the theorem to register. And while the Bayesian scheme has the merit, unlike the hypothetico-deductive model, of allowing for confirmation without entailment (that is, with likelihoods less than 1), it shares with the hypothetico-deductive model the consequence that every hypothesis is confirmed by every consequence. Thus the Bayesian account I have sketched has the consequence that an arbitrary conjunction 'A and B' is confirmed by A, and that A is confirmed by an arbitrary disjunction 'A or B'.

Those are some of the objections that have been used to argue that Bayesianism fails to capture various aspects of our good inferential practices, the kinds of objections that philosophical opponents of Bayesianism tend to make. But some cognitive psychologists have also objected to Bayesianism on the interestingly different grounds that it fails to capture our *bad* inferential practices, and so fails to capture the way we actually reason (e.g. Kahneman *et al.* 1982). What they have done is to construct inferential situations where a Bayesian calculation would yield what is clearly the correct answer, yet most people give a different and incorrect answer. For example, Bayes's theorem makes it clear that where the two priors are the same (so p(H)/p(E) equals 1), the likelihood and the posterior must also be the same, yet it is apparently easy to construct examples where people will both judge the priors to be the same yet the likelihood and the posterior to be different. This pulls apart the normative and descriptive issues. It is here assumed that Bayesianism dictates the inferences one ought to make, but argued on empirical grounds that it does not describe the inferences that people actually make. It is thus inadequate as a solution to the descriptive problem of induction, and the door is opened to the claim that Inference to the Best Explanation does better, descriptively if not normatively.

A third response to the Bayesian challenge is to argue that Bayes's theorem is in fact compatible with Inference to the Best Explanation, because the constraint that Bayes's theorem imposes is much weaker than has been supposed. The probability calculus in general, and Bayes's theorem in particular, do place constraints on permissible combinations of degree of belief, much as the rules of deduction place the constraint of consistency on permissible combinations of belief (Howson 2000: ch. 7). But these sorts of formal or structural constraints do not rule out or rule in any particular belief, so long as it is neither self-contradictory nor a tautology. The rules of deduction may seem to impose a dynamic of belief, such that we should believe whatever follows deductively from what we already believe, but this is an illusion, since we may always turn the argument around by rejecting a consequence and restoring consistency by also revising some of our previous beliefs. Logic doesn't anchor those premises, and so does not compel belief in the conclusion. Similarly, Bayes's theorem may seem to impose a dynamic on degrees of belief, such that we must change the probability of H from its prior to its posterior upon observing E, but this is an illusion, since we may instead choose to alter the priors, so as to retain probabilistic consistency. Probability theory doesn't anchor the priors, and it does not entail the temporal dynamic that standard applications of Bayes's theorem apply to it. On this view, Bayes's theorem tells us something correct about what consistency in degree of belief requires, but it is no threat to Inference to the Best Explanation, since explanatory considerations may be a guide to inference without breaking any probabilistic rule. A good explanation may be given a high posterior for that reason, but this may be squared with Bayes's theorem by also giving it a high prior, say because it is simpler than other explanations (Harman 1999: ch. 4). As I am interpreting this third response, it is explanatory considerations rather than Bayesian conditionalization that is driving our inferential practices, though no theorem of the probabilistic calculus need be flouted in the process.

That gives us three broad responses to the Bayesian challenge. One can argue that Bayesianism is not to be preferred to Inference to the Best Explanation because it is in various ways incorrect or incomplete, or because while normatively correct it does not accurately describe the way people actually reason, or because it does not in fact conflict with Inference to the Best Explanation. I have considerable sympathy with all of these responses. But in the balance of this chapter I would like to consider the prospects of a fourth response, one that draws to some extent on the other three. Like the last of these, it is irenic. My objection to the argument that Inference to the Best Explanation is wrong because Bayesianism is right will not be that the premise is false, but that the argument is a non-sequitur, because Bayesianism and Inference to the Best Explanation are broadly compatible. It goes beyond the third response, however, in suggesting not only that Bayes's theorem and explanationism are compatible, but that they are

complementary. Bayesian conditionalization can indeed be an engine of inference, but it is run in part on explanationist tracks. That is, explanatory considerations may play an important role in the actual mechanism by which inquirers 'realize' Bayesian reasoning. As we will see, explanatory considerations may help inquirers to determine prior probabilities, to move from prior to posterior probabilities, and to determine which data are relevant to the hypothesis under investigation. Explanatory considerations may also play a crucial role in scientists' expression of their preference for hypotheses that promise to be fertile, that is to explain phenomena in addition to those directly under scrutiny at the time of inference. One way of putting this fourth response to Bayesianism is that explanatory considerations provide a central heuristic we use to follow the process of conditionalization, a heuristic we need because we are not very good at making the probabilistic calculations directly. We will see that this simple realization model is probably too simple, but I aim to show that something like this role for inferences to the best explanation is possible and even plausible, and in so doing promote the philosophical project of properly articulating the relationship between these two approaches to inference.

The Bayesian and the explanationist should be friends

There is no difficulty in showing that Bayesianism is compatible with the idea that scientists often infer explanations of their evidence. Just let H be such an explanation, and Bayes's theorem will then tell you whether the evidence confirms it. That is, Bayesianism is clearly compatible with something like 'Inference to the Likeliest Explanation', so long as likeliness is posterior probability as determined by the theorem. As we saw in chapter 4, however, this is a very thin notion of Inference to the Best Explanation, because here it is the Bayesianism and not explanatory notions that seems to be supplying most of the substance. Perhaps not quite all of it, since Bayesianism says little about the context of discovery and we could here take Inference to the Likeliest Explanation to help, by placing a constraint on the hypotheses we consider, namely that they should be potential explanations of the evidence. Once the hypotheses are selected, however, the Bayesian formula would take over.

The use of explanatory considerations in the context of discovery is an important contribution that explanationist thinking can make to Bayesianism, and one to which we shall return, but my immediate aim is different. I want to suggest that Bayesianism is compatible with the thicker and more ambitious notion of 'Inference to the Loveliest Explanation'. That is, Bayesianism is compatible with the governing idea behind Inference to the Best Explanation as I have been developing that account, the idea that explanatory considerations are a guide to likeliness. In Bayesian terms, this is to say that explanatory considerations help to determine posterior

probability. This may sound as though explanatory considerations are somehow to modify the Bayesian formula, say by giving some hypothesis a posterior 'bonus' beyond what Bayes's theorem itself would grant in cases where the hypothesis bears the recommended explanatory relationship to the data or is otherwise sufficiently 'lovely'. This is where the spectre of dutch book irrationality appears (van Fraassen 1989: 160–70), but it is not what I propose. Instead I want to suggest some of the ways in which explanatory considerations may be our way of running or realizing the mechanism of Bayesian conditionalization – the movement from prior to posterior probability – and our way of handling certain aspects of inference that conditionalizing does not address. (Okasha (2000) takes the same irenic approach; for a different angle on the relationship between Inference to the Best Explanation and Bayesianism, see Psillos (2003).) If these suggestions are along the right lines, then arguing that Inference to the Best Explanation is wrong because Bayesianism is right is like arguing that thinking about technique cannot help my squash game because the motion of the ball is governed by the laws of mechanics. Even if Bayesianism gave the mechanics of belief revision, Inference to the Best Explanation might yet illuminate its psychology.

Research in cognitive psychology suggests that we are sometimes remarkably bad at probabilistic reasoning: here the work of Daniel Kahneman and Amos Tversky has been particularly influential (Kahneman *et al.* 1982). I here retail a few of their most striking results. In chapter 3 in the context of my discussion of non-causal explanations we already had the case of praise and criticism from the air force instructors. Noticing that performance typically improves after punishing unusually poor performance but gets worse after rewarding unusually good behavior, we may erroneously infer that punishment is more effective than reward, because we ignore the ubiquitous probabilistic effect of regression to the mean (Kahneman *et al.* 1982: 66–8; page references in the balance of this section refer to this book unless otherwise noted). Extreme behavior tends to be followed by less extreme behavior, for statistical reasons alone, but people commonly ignore this and instead make an unwarranted causal inference.

Equally striking is research that shows how easy it is to get people to violate one of the simplest rules of probability, by judging the probability of a conjunction to be higher than the probability of one of its conjuncts, as in the notorious case of Linda the bank teller. Having been given a brief description of Linda which characterizes her as politically active on the left, 85 percent of the respondents judged that it is less likely that Linda is a bank teller than that she is a bank teller and active in the feminist movement (92–3). It does not seem that this effect can be explained away and the rationality of respondents saved by supposing that they took 'bank teller' to mean 'bank teller but not active in the feminist movement'. The way the question was posed did nothing to encourage such an interpretation, and

follow-on experiments helped further to rule it out. For example, when another group were asked to predict how Bjorn Borg would fare in the 1981 Wimbledon finals, 72 percent judged it less likely that he would lose the first set than that he would lose the first set but then go on to win the match (96), even though the way the question was presented made it very unlikely that respondents took 'he will lose the first set' to entail that he would lose the match (see also Tversky and Kahneman 1984).

A third notorious casualty of probabilistic reasoning that Kahneman and Tversky discuss concerns base rates. Suppose that one in a thousand people contract a disease for which there is a diagnostic test. That test suffers no false negatives but has a false positive rate of 5 percent. That is, 5 percent of those without the disease will nevertheless test positive for it. When 60 students and staff at Harvard Medical School were asked what the probability is that their patient has the disease if the test comes out positive, almost half said 95 percent (154; the original work is in Casscells *et al.* 1978: 999). In this they were wildly mistaken: the correct answer is just under 2 percent. For suppose we have a thousand healthy people and the one poor fellow with the disease. He will test positive, but so will 50 of the thousand healthy people, so of the 51 who test positive, only one will have the disease.

A final example from Kahneman and Tversky involves conditional probabilities. As Bayes's theorem shows, if p(A) and p(B) are the same, then p(A|B) must be the same as p(B|A). Yet people often judge those two probabilities to be different. For example most people judge the probability higher that a competitor won the first event in the decathlon if he won the decathlon than they judge the probability that he won the decathlon if he won the first event (120–1). In fact the two conditional probabilities must be the same, since the prior of a competitor winning the first event is the same as the prior of winning the decathlon, namely 1 divided by the number of competitors.

What is the bearing of these striking psychological results and many others like them on the relationship between Bayesianism and Inference to the Best Explanation? They might be taken to suggest that my irenic strategy is misguided: explanationism should be defended from the Bayesian menace not by arguing for compatibility but by arguing that Bayesianism is a fundamentally inadequate answer to the descriptive problem of induction. This is the second response I mentioned above. Thus one might choose to side with Kahneman and Tversky, who maintain: 'In his evaluation of evidence, man is apparently not a conservative Bayesian: he is not a Bayesian at all' (46).

In addition to providing a critique of the descriptive adequacy of Bayesianism, some of these psychological results strongly suggest the presence of an underlying proclivity for explanationist thinking. Arguing negatively, one could claim that the failure to take sufficient account of various probabilistic facts may be accounted for in part by their lack of a

causal-explanatory role. Thus underlying base rates do not help to explain causally why a particular individual has a particular test result or medical condition, and while they do help to explain statistical outcomes, for example why most people who test positive do not have the disease, even here the explanation is not causal. And arguing positively, several of the cases suggest that causal-explanatory considerations are helping to drive the inferences. In the case of the flight instructors, it is natural to account for the hasty inference that punishment is more effective than reward in terms of an inclination to infer causal explanations of striking patterns in one's evidence. The contrastive evidence of the apparent efficacy of punishment but not reward is like the paradigm of contrastive explanatory inference we explored in chapter 3. We have an inveterate habit of inferring causal explanations of contrasts in our evidence, and this spills over into cases where the inference is ill advised because the actual explanation is instead statistical regression. As Kahneman and Tversky put it, 'when people observe regression, they typically invent spurious dynamic explanations for it' (66).

Explanatory considerations may also be at work in the case of Linda. Kahneman and Tversky suggest that the reason people tend to say it is more likely that she is a bank teller active in the feminist movement than that she is a bank teller is because the conjunction is closer than is the conjunct to the stereotype of the kinds of activities that someone with Linda's interests goes on to pursue (90). We can make a similar point in terms of explanatory connection. The personality type suggested by the description of Linda would provide no explanation of her being a bank teller, but it provides a partial explanation of the conjunction, because it would help to explain why she was active in the feminist movement. More generally, if we have a tendency to prefer better explanations, as the explanationist asserts, it is not surprising that we will often prefer more detailed hypotheses to sparser consequences alone even though they are more probable, since most of the logical consequences of an explanation of E will not also explain E; and even if a consequence does explain, it may not be as good an explanation as the stronger hypothesis that entails it. As Kahneman and Tversky say, 'a good story is often less probable than a less satisfactory one' (98). Nor is this preference generally irrational, even though it may have led people astray in the case of Linda. As we saw in chapter 4, according to Inference to the Best Explanation our aim in inferring an explanation is not to infer the most probable claim, but rather to infer the most probable of competing explanations. No statement can be more probable than any of its logical consequences, but there is nothing irrational about preferring a good story to limiting oneself to a less good consequence of that story. If all we wanted was to maximize probability, we should never venture beyond our data.

All this adds up to a pretty strong case for the second response to the Bayesian challenge, because it adds up to a pretty strong case for saying that Bayesianism and explanationism are incompatible, with Bayesianism

descriptively wrong and explanationism descriptively right. But while I certainly want to capitalize on the way the psychological results suggest both that we find probabilistic calculation difficult and explanatory thinking natural, I think that the second response makes too much of the incompatibility. I accept that the psychological research shows that there are cases where Bayesianism would dictate one inference but people make a different and incompatible inference, but I do not take these cases to be typical. It would seem rather that people's actual inferences will more often track the Bayesian result, or else the extent to which we are able to cope with our environment becomes a mystery. Another reason not to take the psychological results as showing that we are simply not Bayesians is that those who are caught out by the cases we have discussed are also people who can be brought to acknowledge the correctness of the probabilistic considerations they neglected. This is the way they would like to reason and, I would suggest, this is the way they often do reason. What makes the psychological results so striking is that we both find the mistake natural to make and that we can come to see that the reasoning is flawed. This suggests that we do have something like the constraints of probability theory embedded somewhere in our epistemic competence, however it may be masked in awkward cases. The probabilistic constraints that Bayesianism exploits can naturally be seen as analogous to the requirement of deductive consistency (Howson 2000) and so to have a normative force that we do recognize if imperfectly in our actual inferential practices.

So on my reading Kahneman and Tversky's results show us how much help we need with Bayesianism. In one respect they may nevertheless understate that need. For the power of their cases depends in part on how simple they are, from a Bayesian point of view, since it is this that enables us to see so clearly the divergence between the right answer and the answer given. And the fact that there is this divergence even in such simple cases does underline our probabilistic disabilities. But most real life cases are much more complex, and the Bayesian calculation is hence much more difficult. So the need for some help in realizing the requirements of Bayesianism is all the greater. The great simplicity of cases may also lead us to exaggerate the extent of systematic irrationality they expose. For it is a natural enough thought that if we do this poorly on simple problems, we will be even more pathetic in complicated real life cases. But this thought may be mistaken, for two different reasons. First, although it might be more difficult to get the right answer in complex cases, in a sense it is easier to avoid blatant irrationality in such cases, precisely because it is so unobvious what the correct answer is. So while we might be less reliable in complex cases, there is no reason to suppose that we will be less rational. Secondly, it might even turn out that, surprisingly enough, we are sometimes more reliable in complex cases, because we use inferential techniques that are better suited to the kind of complexity we typically encounter in the real world than the

bespoke simplicity of the cases Kahneman and Tversky discuss. For example, while the false positive case might suggest that many doctors are completely hopeless at estimating the probability that you have a given disease after a test, their estimates in real life situations may be far more accurate than the divergence between 95 percent and 2 percent suggests, because in real life it is unusual to give a test to a person at random and to have no other relevant information to hand. The estimates the doctors make under more realistic and complex circumstances may well be more reliable than would be any result based on an attempt directly to apply Bayes's theorem.

Although they claim that we are not Bayesians at all, Kahneman and Tversky's positive view is that we reason by means of heuristics that often but not always yield the normatively correct result:

> people rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations. In general, these heuristics are quite useful, but sometimes they lead to severe and systematic errors. (3)

Even though Kahneman and Tversky take their work to show that people are not Bayesians, I take it that their position commits them to saying that people very often reason in a way that is consistent with Bayesian constraints, since they take those constraints to be normatively binding and what they take their research to show is not that we are generally bad at reasoning, but rather that we use heuristics that very often work well but sometimes fail in striking ways. The point of dwelling on those striking failures is not to show a general cognitive disability, but to reveal general heuristics, because the heuristics reveal themselves most clearly when they let us down. (Gerd Gigerenzer and his colleagues have argued that remarkably simple heuristics can be very effective, because of a co-operative environment. Much of what they say is compatible with Kahneman and Tversky's work, although Gigerenzer also argues that Kahneman and Tversky's heuristics are vague and that their conception of rationality is too narrowly probabilistic. See e.g. Gigerenzer *et al.* (2000).)

We are not good at probabilistic reasoning, so we use other methods or heuristics. With this I agree, only where Kahneman and Tversky take these heuristics to replace Bayesian reasoning, I am suggesting that it may be possible to see at least one heuristic, Inference to the Best Explanation, in part as a way of helping us to respect the constraints of Bayes's theorem, in spite of our low aptitude for abstract probabilistic thought. That is, we can continue to investigate the proposal that explanatory considerations are a way of realizing the Bayesian mechanism, while acknowledging that there will be cases, sometimes striking, where explanatory and Bayesian considerations pull in different directions. To do this, we should consider

how and how extensively the components of the explanationist heuristic map onto the components of the Bayesian system.

Bayesianism provides an account of the evolution of degrees of belief, whereas Inference to the Best Explanation as I characterized it in chapter 4 is rather an account of hypothesis acceptance. This contrast raises interesting questions, but I do not consider them here. Instead, I will leave the notion of acceptance to one side to focus on the question of the relationship between explanatory considerations and probabilities. A natural first thought is that the distinction between the loveliness and the likeliness of an explanation corresponds to the Bayesian distinction between prior and posterior probability. Things are not that neat, however, since although likeliness corresponds to posterior probability, loveliness can not be equated with the hypothesis's prior. Perhaps the easiest way of seeing this is to note the relational character of loveliness. A hypothesis is only a good or bad explanation relative to the specific phenomenon explained. Contrastive explanations make the point vividly, since a good explanation of P rather than Q may not be a good explanation of P rather than R, but the point applies to non-contrastive cases as well, since clearly a good explanation of P will not in general be a good explanation of S. Prior probability is also a relative notion – it is relative to previous evidence and background belief – but it is not relative to the new evidence E on which the Bayesian would have the inquirer conditionalize in order to move from prior to posterior. Loveliness, by contrast, is relative to that new evidence.

Another tempting connection would be to link loveliness not to the prior but to the Bayesian notion of likelihood – to the probability of E given H. The identification of loveliness with likelihood is a step in the right direction, since both loveliness and likelihood are relative to E, the new evidence. But I am not sure that this is quite correct either, since H may give E high probability without explaining E. Indeed H may entail E yet not explain it, as some of the counterexamples to the deductive-nomological model of explanation show. Nevertheless, as Samir Okasha has observed, it may be that whenever H1 is a lovelier explanation of E than H2, the likelihood of H1 is greater than the likelihood of H2 (2000: 705).

It appears that loveliness does not map neatly onto any one component of the Bayesian scheme. Some aspects of loveliness, some explanatory virtues – including scope, unification and simplicity – are related to prior probability (Harman 1999: 110); others seem rather to do with the transition from prior to posterior. But what does this mean? My thought is this. In many real life situations, the calculation that the Bayesian formula would have us make does not, in its bare form, meet the general requirement of epistemic effectiveness: it is not a recipe we can readily follow. We do not always know how to work out the probabilities that are required in order to move from prior to posterior probability simply on the basis of a (presumably tacit) grasp of the abstract principles of the probability calculus. My suggestion is

that explanatory considerations of the sort to which Inference to the Best Explanation appeals are often more accessible than those probabilistic principles to the inquirer on the street or in the laboratory, and provide an effective surrogate for certain components of the Bayesian calculation. On this proposal, the resulting transition of probabilities in the face of new evidence might well be just as the Bayesian says, but the process that actually brings about the change is explanationist.

I want briefly to suggest how explanatory considerations might help to lubricate the Bayesian mechanism, in three ways. The first role for explanatory considerations is in the determination of likelihood, which is needed for the transition from prior to posterior probability. The second is with the determination of the priors, the input to conditionalizing. The third concerns the determination of relevant evidence.

One way in which explanatory considerations might be part of the actual mechanism by which inquirers move from prior to posterior probabilities is by helping inquirers to assess likelihoods, an assessment essential to Bayesian conditionalizing. For although likelihood is not to be equated with loveliness, it might yet be that one way we judge how likely E is on H is by considering how well H would explain E. This would hardly be necessary in cases where H entails E, but in real life inference this is rarely the case and, where H does not entail E, it is not so clear how in fact we do work out how likely H makes E, and how likely not-H makes E. (As Jason Grossman has pointed out to me, this difficulty will not however arise in situations where H is a probabilistic hypothesis *about* likelihoods.) One of the signal advantages of Bayesianism over hypothetico-deductivism is that it allows for confirmation without entailment; but it is just such cases where the Bayesian formula can be difficult to implement. Explanatory considerations might help here, if in fact loveliness is reasonably well correlated with likelihood, and we find it easier in practice to judge loveliness than likelihood. What would be required, I think, is that lovelier explanations tend to make what they explain likelier (even if high likelihood is no guarantee of good explanation), and that we sometimes exploit this connection by using judgments of loveliness as a barometer of likelihood.

For example, when we consider the loveliness of a potential causal explanation, we may consider how the mechanism linking cause and effect might run, and in so doing we are helped in forming a judgment of how likely the cause would make the effect and how unlikely the effect would be without the cause. This mechanism may also be at work in the context of contrastive explanation. When Semmelweis was investigating the causes of childbed fever, he repeatedly considered to what extent various hypotheses explained the contrasts in his data, such as contrasts between rates of the fever in different wards and within a single ward under different conditions. The suggestion is that Semmelweis was aided in coming to a view of likelihoods by considering how well those data would be explained by the

competing hypotheses. The case would have to be made out; here I only wish to make the suggestion clear. Inference to the Best Explanation proposes that loveliness is a guide to likeliness (a.k.a. posterior probability); the present proposal is that the mechanism by which this works may be understood in part by seeing the process as operating in two stages. Explanatory loveliness is used as a symptom of likelihood (the probability of E given H), and likelihoods help to determine likeliness or posterior probability. This is one way Inference to the Best Explanation and Bayesianism may be brought together.

Another obvious place to look for a way explanatory considerations might in practice play an important role in a Bayesian calculation is in the determination of prior probabilities. I begin with a general observation about the role of priors in Inference to the Best Explanation. Choices between competing potential explanations of some phenomenon are often driven by judgments of which of the explanations has the highest prior. This is one important source of suspicion about Inference to the Best Explanation: the choices here seem actually to be based on judgments of which is the likeliest explanation, judgments which in many cases depend on which potential explanation is judged to have the highest prior, not on which would be the loveliest explanation (Salmon 2001a: 83–4). My reply is to agree about the crucial role that priors play in this way, but to deny that this is in tension with Inference to the Best Explanation. Consider what the Bayesian himself says about the priors. He of course does not take their crucial role to undermine the importance of the Bayesian formula, in large part because today's priors are usually yesterday's posteriors. That is, the Bayesian claims that today's priors are generally themselves the result of prior conditionalizing. Similarly, the defender of Inference to the Best Explanation should not deny that inference is mightily influenced by the priors assigned to competing explanations, but she will claim that those priors were themselves generated in part with the help of explanatory considerations.

This means that, insofar as my suggestion that explanatory considerations play a role in conditionalizing has merit, explanatory considerations also have a role to play in the determination of priors, since priors are partially determined by earlier conditionalization. Explanatory considerations may also enter into the determination of priors in other ways. This is where considerations of unification, simplicity and their ilk would naturally come into play. The Bayesian is happy to acknowledge the role that these sorts of factors may play in fixing prior probabilities for hypotheses, and the prospects seem promising for showing that some of these may in practice be determined by considering explanatory quality. And explanatory considerations may also play a role in the determination of the priors of the evidence, in part because the status of E before observation is in many respects that of another hypothesis, whose probability will be influenced by prior conditionalization and by considerations of simplicity *et al.* More

directly, the prior probability of E is tantamount to how surprising it would be to observe E, and this will be determined in part by how good an explanation of E my current beliefs would supply.

Given a specified H and E, the process of conditionalization requires the determination of likelihoods and of priors, and I have suggested how explanationist considerations may help with both. But H and E are not simply given, and Inference to the Best Explanation helps to account for their source. Let us begin with the evidence. Bayes's theorem describes the transition from prior to posterior, in the face of specified evidence. It does not, however, say *which* evidence one ought to conditionalize on. In principle perhaps, non-demonstrative inference should be based on 'total evidence', indeed on everything that is believed. In practice, however, investigators must think about which bits of what they know really bear on their question, and they need also to decide which further observations would be particularly relevant. So it seems that a Bayesian view of inference needs some account of how the evidential input into the conditionalizing process is selected, and this seems yet another area where the explanationist may contribute. To give just one example of how this might work, consider how we sometimes discover supporting evidence for a hypothesis by seeing what it would explain. My suggestion is that we sometimes come to see that a datum is epistemically relevant to a hypothesis precisely by seeing that the hypothesis would explain it. (Arthur Conan Doyle often exploited this phenomenon to dramatic effect: in 'Silver Blaze', the fact that the dog did not bark would have seemed quite irrelevant, had not Sherlock Holmes observed that the hypothesis that a particular individual was on the scene would explain this, since that person was familiar to the dog.)

The other obvious application of explanationist thinking to the process of conditionalization is one that I already mentioned earlier in this chapter: the context of discovery. Bayes's theorem says nothing about where H comes from. Inference to the Best Explanation helps here since, as we have seen, asking what would explain the available evidence is an aid to hypothesis construction. This is so for a number of reasons. Most obviously, our aim is often to explain some phenomenon, so that will constrain the process of hypothesis generation. Secondly, we wish to generate hypotheses that have a reasonably high prior, and as we have already seen explanatory considerations may be a guide here. But thirdly, and pulling in the opposite direction, explanationist considerations might be particularly well suited to account for scientists' preference for fertile hypotheses with high content (Salmon 2001b: 121). As we have already noted, high probability is not the only aim of inference. Scientists also have a preference for theories with great content, even though that is in tension with high probability, since the more one says the more likely it is that what one says is false (James 1897; Popper 1959: sec. 83). This interest in scope and fertility is a promising area in which explanationist considerations may operate, since scientists may

judge theoretical fertility or promise by assessing the explanatory potential of the hypotheses they are evaluating. By requiring that H explain E, and even more by requiring that it provide a lovely explanation of E – where one dimension of loveliness is how much H explains – explanationist considerations keep H from coming too close to E, and so from wrongly sacrificing content for the sake of high probability. At the same time, the preference for fertile hypotheses is more than just a blind preference for greater content – something that could be satisfied by arbitrary conjunction – but is a preference for hypotheses that hold out the promise of unifying hitherto unconnected phenomena, and this too is a plausible dimension of explanatory loveliness.

Contrastive inference revisited

If my approach to the relationship between Bayesianism and Inference to the Best Explanation is to serve the explanationist cause, it is not enough that it show that the two approaches are compatible. It must suggest that explanatory considerations capture aspects of our inferential practices about which Bayesianism is silent and that they actually help us to realize or run the Bayesian mechanism. The possibilities for this that I have briefly canvassed certainly do not establish that this is the case; but I hope that they make the claim plausible, particularly in concert with the psychological research that suggests both that we need help with probabilistic calculations and that explanatory considerations figure in the heuristics we use. I will end this chapter by considering a particular type of inference which I think strengthens the case. It is our old friend from chapter 5, contrastive inference, or inference from a controlled experiment.

We can think of an inference from a controlled experiment as being based on two pieces of evidence, the positive instance and the contrapositive or control instance. Thus in the case of Semmelweis's hypothesis of cadaveric infection, we have two observations, one of an infected woman without childbed fever, and one of an uninfected woman with childbed fever. We could imagine conditionalizing on each of these observations. The salient point, however, is that the net confirmatory power of the pair of observations is substantially greater than the sum of the confirmation provided by each taken on its own. How does the Bayesian account for this?

It seems the Bayesian must claim that the conjunction of the observations is less probable than the sum of their probabilities taken singly. So the one observation must lower the probability of the other: the combination must be particularly surprising. At first this seems odd. After all, why should observing a black raven make it less likely that I will observe a white shoe? Indeed insofar as observing a black raven confirms the hypothesis that all ravens are black, one might expect it to increase the probability that the next non-black thing I encounter will also be a non-raven. But this idea of the one

bit of evidence lowering the probability of the other becomes less odd if, in a Millian spirit, we focus on the requirement for a good controlled experiment (or application of the method of difference) that the two instances have largely shared histories. If I observe one person with lung cancer, that will presumably raise the probability that another person with a very similar medical history will have the cancer as well. So finding that such a person does not have cancer is made more surprising. That is the situation in a controlled experiment. We have one group of people with cancer, and then another very similar group without it: a surprising combination. And it will remain surprising even if the first group are smokers and the second are not, if we have not already given a high probability to the hypothesis that smoking causes cancer. This composite evidence, however, will then provide strong support to that hypothesis, since the hypothesis gives a high likelihood to the observations, and the observations have a low prior.

Peter Urbach (1985: 267–71) has suggested another way of thinking about the special power of controlled experiment from a Bayesian perspective. According to him, the crucial factor in this case is the way the combined observations disconfirm rival hypotheses, and so boost the probability of H. And this occurs because the likelihoods the rivals confer on the composite evidence will be low, lower than the prior of the evidence. Thus the hypothesis that obesity causes lung cancer will be disconfirmed if the people in both groups are obese, since the probability of that evidence given the obesity hypothesis is low. Like the first Bayesian account in terms of a surprising conjunction, Urbach's account depends on the shared factors between the two cases. It is because the people in both groups are obese that the obesity hypothesis does not make it likely that one group would have lung cancer while the other doesn't.

Bayesians thus have at least two ways to account for the special probative force of a well-controlled experiment. And while I am here pressing against the edge of the envelope of my probabilistic competence, I suppose that these two accounts are compatible. In any event, both accounts reflect aspects of Semmelweis's research, as discussed in chapter 5. The low composite probability corresponds to the surprise Semmelweis must have felt that the two maternity wards in his hospital should have such different rates of childbed fever. And the disconfirmation of alternative explanations corresponds to the way Semmelweis repeatedly tested and dismissed competing hypotheses on the grounds of their failure to explain contrasts that he found or created. Indeed it may well be correct to say that Semmelweis was in effect performing a complex series of Bayesian calculations. But the central suggestion of this chapter has been that investigators use explanatory considerations to perform these calculations, and this appears to apply to Semmelweis's case. As he presents his work, he does not consider directly the prior probability of his evidential contrasts. Nor does he reject the competing hypotheses directly on the grounds that they generate low

likelihoods. For example, he does not seem directly to consider the question of whether the prior of the composite evidence is higher than the likelihoods the competitors would give it. What counts for Semmelweis is rather what the various hypotheses would and would not explain. His procedure seems entirely defensible from a Bayesian perspective, dominated though his thinking is by explanatory considerations.

Some of the hypotheses Semmelweis considered included no articulated mechanism, as in the cases of those that appealed to 'epidemic influences' or to delivery position. Nevertheless, Semmelweis's causal-explanatory perspective enabled him to think about his research in concrete physical terms. Most of us find this mode of thought far more congenial than the more abstract process of calculation that Bayesianism taken straight would require. The point is not that we are always more reliable in concrete judgment. Although there are many reasoning tasks that we do much better at when they come in concrete form, we sometimes get the abstract principle right while failing to apply it in the concrete. Thus most of those who violate the conjunction rule of probability in the concrete case of the Linda example nevertheless endorse the correct abstract principle that a conjunction can never be more probable than one of its conjuncts (Kahneman *et al.* 1982: 97). The point is descriptive, not normative: it is that we tend to think in concrete terms, even when we would do better with a little abstraction. In that sense, the case of Linda is the exception that proves the rule. But I do suppose that it is an exception, and that generally speaking the explanationist heuristic does a good job of enabling us effectively to perform the Bayesian calculation, or at least to end up in pretty much the same cognitive place.

This chapter has provided a brief exploration of the prospects for a compatibilist view of the relationship between Inference to the Best Explanation and Bayesianism. The relation that has motivated my discussion has been one of realization. Perhaps the simplest version of this view would make explanatory considerations a heuristic employed to make the likelihood judgments that the Bayesian process of conditionalization requires. Prior assessment of the quality of the explanation would be a way to fix on a likelihood, according to the rule of thumb that the better the explanation, the higher the likelihood. I have suggested that this may indeed be part of the story, but we have seen the relationship between Inference to the Best Explanation and Bayesianism is considerably more complicated and extensive than this, for a number of reasons. We have found that explanatory loveliness does not map simply onto likelihood, but may also play a role in assessing the priors. We have also found that explanatory considerations may play diverse roles in addition to the substantial jobs of helping to judge priors and likelihoods, such as determining relevant evidence and guiding hypothesis construction. We have also considered some research in cognitive psychology which suggests that there are certain cases where Bayesianism and heuristics like Inference to the Best Explanation would yield different

and incompatible inferences, with Bayesianism giving the right normative result but the heuristic providing the description of the process most people actually employ.

Nevertheless, a broadly compatibilist picture remains attractive. The psychological research supports the idea that we are not very good at abstract probabilistic calculation and that we need to use heuristics, including explanatory heuristics. And although that research focuses on cases where the heuristics we use yield different results than would a straight Bayesian calculation, I have suggested that it is compatible with the view that in most situations our heuristics can be seen as a way of at least approximating the Bayesian result. Indeed I think it is compatible with the view that explanatory considerations help us to perform what is in effect a Bayesian calculation. The upshot is that Bayesianism poses no particular threat to Inference to the Best Explanation. Bayes's theorem provides a constraint on the rational distribution of degrees of belief, but this is compatible with the view that explanatory considerations play a crucial role in the evolution of those beliefs, and indeed a crucial role in the mechanism by which we attempt, with considerable but not complete success, to meet that constraint. That is why the Bayesian and the explanationist should be friends.

Chapter 8

Explanation as a guide to inference

The guiding claim

An articulation and defense of Inference to the Best Explanation might proceed in three stages: identification, matching and guiding. First we identify both the inferential and the explanatory virtues. We specify what increases the probability of a hypothesis and what makes it a better potential explanation; that is, what makes a hypothesis likelier and what makes it lovelier. Second, we show that these virtues match: that the lovelier explanation is the likelier explanation, and vice versa. Third, we show that loveliness is the inquirer's guide to likeliness, that we judge the probability of a hypothesis on the basis of how good an explanation it would provide.

To be sure, this battle plan is both too optimistic and too simple. The first stage – the project of identifying the inferential and explanatory virtues – is enormously difficult, as we have seen, and our achievements to date are surprisingly limited. And this of course places severe constraints on our ability to carry out the second stage of the campaign, which is to show that the inferential and explanatory virtues match. On the other hand, my initial characterization of the second stage also overstates what one should aspire to show. In my view an interesting account of Inference to the Best Explanation should be bold in one respect but relatively modest in another. The boldness consists in not settling for the claim of inference to the likeliest explanation, the claim simply that we often infer to an explanation that we judge to be more probable than its competitors, but in insisting on the claim of Inference to the Loveliest Explanation, the claim that explanatory loveliness is a guide to judgments of likeliness. The sensible modesty consists in making no claim that Inference to the Best Explanation is the foundation of every aspect of non-demonstrative inference (cf. Day and Kincaid 1994). It is glory enough to show that explanatory considerations are an important guide to inference. Consequently, there is no need to argue heroically for a perfect match between the explanatory and the inferential virtues. Similarly, in the third stage there is no need to argue that explanatory considerations are our only guide to inference, just that they are a significant guide, an important heuristic.

Shaded in these sorts of ways, the three-stage picture is a helpful representation of what needs to be done to defend an interesting version of Inference to the Best Explanation. In this book so far we have done some work in the first, identification stage. Standard accounts of inference and confirmation, such as hypothetico-deductivism, Mill's methods and Bayesianism, alongside a discussion of their limitations, have brought out some of the inferential virtues. On the side of the explanatory virtues I have been able to be somewhat more original, albeit on a limited front, by developing an account of contrastive explanation, which shows why some potential causes are more explanatory than others. This has also enabled me to do some work on the second stage, because of the structural similarity between the Method of Difference and the conditions on lovely contrastive explanation. And in our discussion of Bayesianism in the last chapter we have been able to see further possibilities of matching, for example between the conditional probability of the evidence given the hypothesis, which contributes to the posterior probability of the hypothesis, and the quality of the explanation that hypothesis provides.

There is plenty of room for further productive work on the first two stages, because there are a number of other virtues that appear to be at once explanatory and inferential. To provide a proper account of these virtues is a very difficult project that lies well beyond the scope of this book. But there is considerable agreement over the identity if not the analysis of many inferential virtues, and it is striking how many of these appear also to be explanatory virtues, features that make one explanation lovelier than another (cf. Thagard 1978). Thus among the inferential virtues commonly cited are mechanism, precision, scope, simplicity, fertility or fruitfulness, and fit with background belief (cf. e.g. Hempel 1966: ch. 4; Kuhn 1977: 321–2; Quine and Ullian 1978: chs VI, VIII; Newton-Smith 1981: 226–32). All of these are also plausibly seen as explanatory virtues. We understand a phenomenon better when we know not just what caused it, but how the cause operated. And we understand more when we can explain the quantitative features of a phenomenon, and not just its qualitative ones. An explanation that explains more phenomena is for that reason a lovelier explanation. Again, simplicity, in its various guises, contributes to loveliness. For example, some forms of simplicity enable us to achieve one of the cardinal goals of understanding, namely to reveal the unity that underlies the apparent diversity of the phenomena (Friedman 1974). Fruitfulness, the power of a theory to, in Kuhn's words, 'disclose new phenomena or previously unnoted relationships among those already known' (1977: 322), is linked to scope and simplicity, and is again an explanatory virtue. And fit with background is also an inferential factor that has an explanatory aspect. One reason this is so is because background beliefs may include beliefs about what sorts of accounts are genuinely explanatory. For example, at given stages of science no appeal to action at a distance or to an irreducibly chance mechanism could count as

an adequate explanation, whatever its empirical adequacy. The role of background belief in determining the quality of an explanation shows how explanatory virtue is 'contextual', since the same hypothesis may provide a lovely explanation in one theoretical milieu but not be explanatory in another (Day and Kincaid 1994: 285).

All this bodes very well for the matching claim, but still one might question whether these inferential virtues are also really explanatory virtues, because of a general skepticism about the very notion of explanatory loveliness. Thus one might say that it is whatever is true that explains, however the cards fall, and it is meaningless to say that one hypothesis is lovelier than a competitor, since it is only the one that is correct that offers any understanding at all. But this extreme position is not attractive, because truth is obviously not sufficient for explanation. On any tenable view of explanation, there are further conditions that a hypothesis must meet if it is to provide an explanation, whether one analyzes these conditions in causal, logical or structural terms. And meeting these conditions could be a matter of degree, supporting the notion that some explanations are lovelier than others. There are, however, less extreme forms of skepticism about loveliness that would still cast doubt on the explanatory value of the standard theoretical virtues. Thus, in a sensitive discussion of the first edition of this book, Eric Barnes has suggested that if one endorses a causal conception of explanation, then the loveliness of a potential explanation is exhausted by the extent to which it would, if true, give the relevant causal history of the effect to be explained. This would presumably allow for a limited notion of explanatory loveliness, because some causes might be more explanatory than others. But in Barnes's view unification, for example, does not in itself contribute to loveliness, and 'should the ultimate causes of the world be highly fragmented, it would be a fragmented theory, and not a unified one, that would maximise scientific understanding' (1994: 272).

In the context of this discussion I accept that only actual (i.e. true) explanations provide actual understanding. (Thus I leave to one side the understanding provided by hypotheses that are only approximately true and by falsehoods that nevertheless enlighten by showing how a phenomenon could have come about.) So if the world is a fragmented place, then no hypothesis that attributed a false unity would provide genuine understanding. But it seems to me that there is more to degree of understanding or loveliness than amount or relevance of causal history. Even if we limit our attention to true claims, some of these will provide much better explanations than others along dimensions that do not seem exhausted by the extent of the causal history they describe. For example, how explanatory a causal account is may depend on how the causes are described, not just on how much of the relevant history is cited. We see this for example in the context of discussions of reductionism, where the same causes described at one level may provide lovelier explanations than those causes described at another

level; more generally, many aspects of what is known as the interest relativity of explanation suggest that there are many dimensions to explanatory loveliness (Putnam 1978; Garfinkel 1981). We also see this in the history of science, where we find that what counts as an explanation at all changes over time. And if the world is a chaotic, disunified place, then I would say it is less comprehensible than if it is simple and unified. Some possible worlds make more sense than others.

Although the view is obviously not unanimous, my impression is that many will be happy to accept the matching claim, to accept that many of the inferential virtues are also explanatory virtues. This does not, however, mean that Inference to the Best Explanation is out of the woods. For correlation does not entail causation: the central explanationist claim that explanatory considerations are a guide to inference is not the only view compatible with the overlap between inferential and explanatory virtues. Perhaps it is just a coincidence, or perhaps it is the reverse of the story the explanationist would tell. That is, maybe it is likeliness that influences our judgments of loveliness; hypotheses that strike us as more probable for that reason also strike us as lovelier potential explanations. Or maybe both kinds of judgments have a common cause. These are challenges to the guiding claim.

The guiding claim is not entailed by the matching claim: it is possible that many inferential virtues are also explanatory virtues, but we do not use explanatory considerations as a guide to inference. The situation is even worse. For while the matching and guiding claims are clearly compatible, there is an annoying way in which arguments for the matching claim interfere with arguments for the guiding claim, generating a kind of catch-22. The obvious way to argue for the matching claim is to give some independent account of inference and then to show that it fits with some account of explanation. Thus one may try to show that inferential virtues, as revealed by the hypothetico-deductive model, Mill's methods, Bayesianism or what have you correspond to various explanatory virtues. This makes the matching claim plausible. But by invoking these independent characterizations of the inferential virtues and then showing a match with explanatory virtues, one provides ammunition to the enemies of the guiding claim, who may then argue that it is the factors cited in the independent characterization that are doing the inferential work, rather than explanatory considerations. Here is the most obvious example of this difficulty in the context of my discussion in this book. Having devoted considerable energy to defending the matching claim by arguing for the structural similarity between Mill's Method of Difference and the conditions on a good contrastive explanation, I open myself to the objection that what this really shows is not that Inference to the Best Explanation is ubiquitous, but rather that much of what passes as Inference to the Best Explanation is really just Mill's methods (Rappaport 1996; Barnes 1994: 255).

The general form of the objection, and it is surely legitimate, is to argue for a reductive claim: whenever it appears the explanatory considerations are guiding inference, it is really something else, that has nothing particular to do with explanation. What is slightly irritating is the way arguing for Inference to the Best Explanation by defending the matching claim gives one's reductive opponents their ammunition in their attack on the guiding claim. You already have another account of inference you quite like, and indeed it is quite a good account, as these things go; but I want to win you over to Inference to the Best Explanation. In order to do this, I need to convince you of both the matching claim and the guiding claim. Now I will either convince you that Inference to the Best Explanation is roughly co-extensive with your account or I will fail in this. If I fail, you will not buy the matching claim; but if I succeed, you will not buy the guiding claim, since you will maintain that it is your account that describes what is doing the real inferential work, without any appeal to explanatory virtues. So either way I lose: that is the catch-22.

This difficulty is real, but I trust that I have been exaggerating its extent. For those not already deeply invested in another account of inference, the strong case for the matching claim may happily encourage acceptance of Inference to the Best Explanation. And I have already in this book been arguing for the guiding claim. Thus my account of Semmelweis's research into childbed fever was supposed to be a particularly vivid example of inference being strongly guided by explanatory considerations. And in the last chapter I attempted to avoid the catch-22 by suggesting that even committed Bayesians could endorse Inference to the Best Explanation, by seeing explanatory considerations as an important part of the mechanism by which we realize Bayesian calculations. Nevertheless, since the guiding claim is at the heart of the argument of this book, I want to stay with it, and in the balance of this chapter to bring together a variety of considerations in its favor, especially but not exclusively as the guiding claim arises in the context of contrastive inference and the Method of Difference. None of these arguments to come will conclusively establish the guiding claim, but my hope is that jointly they amount to a strong case. The arguments fall into three parts. In the first, I consider ways in which an appeal to explanatory considerations gives a better description of our inductive practices than other accounts. In the next part I suggest that the guiding claim is itself the best explanation for the otherwise disproportionate role that explanatory thinking plays in our cognitive economy and of our tendency to take account in inference of features that fit into an explanatory story and to ignore those that do not. In the last part, I suggest why our inclination to solve inferential problems by constructing causal models leads us to proceed in explanatory terms.

Improved coverage

The obvious way to avoid the catch-22 and to defend the guiding claim is to argue that Inference to the Best Explanation gives a better description of our inductive practices than other accounts on offer, either because it avoids some of the misdescriptions those other accounts give, or because it goes beyond those accounts, correctly describing aspects of our inferential life about which they are silent. This is a strategy I have adopted repeatedly in this book. We have seen that Inference to the Best Explanation does better than an instantial model of confirmation because it has broader scope, and better than the hypothetico-deductive model because it has both broader scope, rightly allowing for confirmation where there is no deduction, and also narrower scope, since some deductions are not explanatory. This narrowing is, I suggested, also to the credit of Inference to the Best Explanation, since non-explanatory deductions (such as the deduction of a conjunct from a conjunction) appear also to be cases where the conclusion does not provide genuine inductive support to the premises.

In the last chapter I argued that Inference to the Best Explanation may give a better account than bare Bayesianism, not because Bayesianism gives the wrong answers, but because it is incomplete, in two ways. First, we need help in making the calculation that Bayes's theorem enjoins, and explanatory considerations can help with this. Secondly, there are aspects of inference involving especially the generation of hypotheses and the determination of what would count as relevant evidence about which Bayesianism has little to say, but which explanationism can address.

I want now to suggest that this strategy of arguing for Inference to the Best Explanation by showing that it has better coverage than the other accounts of inference also applies in the case of Mill's Method of Difference which figured so prominently in chapter 5. There I seem to have talked myself into a particularly serious version of the catch-22, by stressing the isomorphism between that method and the conditions on contrastive explanation. To talk myself out of this situation, the first thing to say is that the Method of Difference and even the full four 'experimental methods' (Difference, Agreement, Residues and Concomitant Variation) are agreed on all sides to be an incomplete account of inductive inference. Thus Mill allows that his methods must be supplemented by the Method of Hypothesis, which looks a great deal like a method of explanatory inference (1904: III.XIV.4–5). Explanationism has here the virtue of giving a unitary account of both aspects of Mill's discussion of induction. To have this scope, we have to go beyond the simple account of contrastive inference, to include other explanatory virtues. But even that simple account has a scope broader than the Method of Difference, as we shall now see.

The Method of Difference has two serious limitations not shared by Inference to the Best Explanation. The first is that it does not account for

inferred differences. The Method of Difference sanctions the inference that the only difference between the antecedents of a case where the effect occurs and one where it does not marks a cause of the effect. Here the contrastive evidence is not evidence for the *existence* of the prior difference, but only for its causal role. The method says nothing about the discovery of differences, only about the inference from sole difference to cause. So it does not describe the workings of the many contrastive inferences where the existence of the difference must be inferred, either because it is unobservable or because is it observable but not observed. Many of these cases are naturally described by Inference to the Best Explanation, since the difference is inferred precisely because it would explain the contrast.

Why does Inference to the Best Explanation account for inferred differences while the Method of Difference does not? The crucial point is that only Inference to the Best Explanation has us consider the connection between the putative cause and the evidence as a part of the process of inference. In an application of the Method of Difference, having found some contrastive evidence, we look at the antecedents or histories of the two cases. What governs our inference is not any relation between the histories and the evidence, but only a question about the histories themselves – whether they differ in only one respect. Inference to the Best Explanation works differently. Here we consider the potential explanatory connection between a difference and the contrastive evidence in order to determine what we should infer. We are to infer that a difference marks a cause just in case the difference would provide the best explanation of the contrast. Because of this subjunctive process, absent from the Method of Difference, we may judge that the difference that would best explain the evidence is one we do not observe, in which case Inference to the Best Explanation sanctions an inference to the existence of the difference, as well as to its causal role. Semmelweis did not observe that only women in the First Division were infected by cadaveric matter, so the Method of Difference does not show why he inferred the existence of this difference. Inference to the Best Explanation does show this, since the difference would explain the contrast in mortality between the divisions. (As Steven Rappaport (1996) observes, Mill's Method of Residues does, however, appear to support an inference to the existence of a cause, as in the case where Neptune was discovered on the basis of observed perturbations in the orbit of Uranus. Even here, however, I think that Mill may have thought of the Method of Residues as being fully applied only after the cause is observed; in any event, it does not seem to apply to the inference of unobservable causes.)

So Inference to the Best Explanation in the context of contrastive inference extends the Method of Difference to inferred differences. Given how common such inferences are, this is a substantial advance; but it may still understate the case for the superiority of Inference to the Best Explanation. For if the Method of Difference does not account for inferred

differences, it is unclear how it can account for any contrastive inferences at all. The trouble lies in the requirement that we know that there is only one difference. Even if only one difference is observed, the method also requires that we judge that there are no unobserved differences, but it gives no account of the way this judgment is made.

The second weakness of the Method of Difference is that it does not fully account for the way we select from among multiple differences. Although Mill's strict statement of the Method of Difference sanctions an inference only when we know that there is a sole difference in the histories of fact and foil, Mill recognizes that this is an idealization. However similar the fact and foil, there will always be more than one difference between their antecedents. Some of these will be causally relevant, but others not. The problem of multiple differences is the problem of making this discrimination. Mill proposes a double response. First, we may ignore differences 'such as are already known to be immaterial to the result' (III.VIII.3). Secondly, while passive observations will seldom satisfy the requirements of the method, Mill claims that carefully controlled experiments where a precise change is introduced into a system will often leave that change as the only possibly material difference between the situation before the change and the situation afterwards.

These are sensible suggestions, but they leave work to be done. Our ability to use differences to infer causes does often depend both on background causal knowledge and on careful experiment, but the Method of Difference does not itself tell us where this knowledge comes from, and even a careful experiment will seldom if ever leave us with only one possible cause, once we allow for the possibility of unobserved and indeed unobservable causes. (Indeed, as Trevor Hussey has pointed out to me, under Mill's own deterministic assumptions there can never be just one prior material difference, since every difference has to be preceded by another.) This background knowledge, and the means of selecting from among the possible differences that even a careful experiment will leave open requires broader inductive principles. In a Millian context, this means that we probably cannot see the method of hypothesis as an extension beyond the relatively secure four methods, but something required even to apply the Method of Difference. And Inference to the Best Explanation helps to provide this broader account.

Explanatory obsessions

A different way to argue for the guiding claim is by appeal to the remarkably large role that causal explanation plays in our cognitive economy. We can think about this from an evolutionary perspective. We are members of a species obsessed with making inferences and giving explanations. That we should devote so much of our cognitive energy to inference is no surprise, on

evolutionary grounds. Much knowledge is good for survival. That we should also be so concerned with understanding is more surprising, at least if explaining is just something we do after we have finished making our inferences. Yet we are constantly generating and assessing explanations: one empirical study suggests that around 15 percent of all conversation concerns causal explanation (Nisbett and Ross 1980: 184). What is the point of all this activity? If explanation and explanatory judgments are only consequent to inference, the answer is unclear, since explanation would then be epiphenomenal from the point of view of inference and so also it would appear from the point of view of coping with our environments. If however the guiding claim is correct then explanation has a central evolutionary point, since one of its functions is as an aid to inference.

The fact that Inference to the Best Explanation can account for the point of explanation in adaptive terms while the epiphenomenal view cannot seems a reason to favor it. This style of argument does not depend on the claim that we have a predisposition to explain, though I find this likely, just as it seems likely that, whatever the actual nature of our inductive practices, they have some innate basis. But in the end this argument for the guiding claim does not depend on armchair evolutionary biology, since the same sort of consideration applies to our learned behaviors. Inference to the Best Explanation gives the practice of seeking and assessing explanations a central role in our own well-being, and so shows why we have all been so keen to learn how to do it.

I do not apologize for defending the guiding claim and hence Inference to the Best Explanation on the grounds that it gives a good explanation of some phenomena – in this case the match between inferential and explanatory virtues and our intense interest in explanation. But while I have no objection to its form, the substance of the argument by itself is hardly conclusive. There are perhaps other useful but obscure functions which all that explanatory activity might perform, or perhaps it is all a cognitive spandrel of some sort, a by-product of other adaptive pressures. Fortunately, the evolutionary explanation for our explanatory obsession can be supplemented by other features of that obsession that suggest more directly that the explanatory activity we engage in is doing inferential work. Here we are helped by research in cognitive psychology, some of which already made an appearance in the last chapter.

This work suggests that we are in fact so keen to use explanation as a guide to inference that we tend to impose causal explanations on the patterns we find in our data even when such inferences seem unwarranted. Thus as we saw in the last chapter in the case of the flight instructors, we find it all too easy to infer that punishment is more effective than reward, when in fact improvement of exceptionally poor performance and deterioration of exceptionally strong performance needs no such explanation, since it may simply be a case of regression to the mean (Nisbett and Ross 1980: 161–3).

And we infer causal-explanatory connections even when we ought to know better. Thus people think that being able to choose their own lottery number substantially increases their chances of winning, believing there to be a causal link where they have every reason not to believe this (Langer 1975; Nisbett and Ross 1980: 134).

The flip side of our obsession with causal explanation is our tendency to ignore or undervalue information that does not fit naturally into an explanationist scheme. Ignoring regression to the mean 'effects' is one example of this. Another striking sort of case concerns base rates, a topic we also touched upon in the last chapter. People obviously do not always ignore base rate information, particularly if there is no other obviously relevant information available (Nisbett and Ross 1980: 156–7). Thus we do judge that the odds of picking the ace of spades out of a full and well-shuffled deck is 1 in 52. But people can be remarkably insensitive to base rate information, especially if there is other information available. This is what we saw in the case of the medical test. Told that a test for a disease has a false positive rate of 5 percent – that 5 percent of those without the disease will test positive – many medics will conclude that someone who tests positive has a 95 percent chance of having the disease, even if they are also told that the base rate of the disease in the population is one in a thousand, making the correct answer around 2 percent (Kahneman *et al.* 1982: 154). The point I wish to flag here is that while there is a clear explanatory connection between having the disease and testing positive, it is not obvious how to factor the base rate information into the causal story. This, I suggest, is a reason it tends to be ignored.

The point comes out particularly clearly in the context of another of Kahneman and Tversky's famous experiments, this time concerning a taxi accident (Kahneman *et al.* 1982: 156–8). People were told that a taxi was involved in a hit and run accident at night. They were also given the following information. First, 85 percent of the cabs in the city are green and 15 percent are blue (the base rate); second, there was a witness who reported that the taxi was blue; and third, the witness was found to be able to distinguish the colors at night correctly 80 percent of the time. Given this information, Bayes's theorem would show the base rate to dominate over the witness: the taxi was more likely to have been green than blue. Nevertheless, as in the case of the disease and the test, many people simply ignore the base rate here, judging that there is an 80 percent chance that the taxi was blue.

What is particularly interesting about the case of the taxi for our purposes is the way the results changed when the example was modified. Instead of being told that 85 percent of taxis are green and 15 percent are blue, subjects were now told that although there are roughly the same number of green and blue taxis, 85 percent of taxi accidents involve green taxis and only 15 percent involve blue taxis. From a Bayesian point of view, this change makes no difference; but subjects were now considerably more inclined to take

some account of the base rate information. The reason seems to be that here, unlike in the original case, the base rate can be made explanatorily relevant. As Kahneman and Tversky suggest, it may prompt the inference that drivers of the green taxis are less safe than drivers of the blue taxis.

In another experiment, people were asked how likely it was that a person had passed a particular exam (Ajzen 1977; Nisbett and Ross 1980: 158). If they were told that the person was chosen at random from a pool of students constructed to include only a quarter who passed, subjects tended to ignore that base rate information in favor of information about the particular student. But if subjects were told instead that a quarter of the students who took the test passed it, which suggests that it was a tough test, then the information had a substantial impact. Richard Nisbett and Lee Ross conclude their discussion of the cases of the taxi and the test as follows: 'if the base rate could be interpreted as reflecting causal influence ... it was utilized very heavily; if the base rate reflected only "arbitrary" group composition, it was utilized only slightly' (1980: 158).

These interesting results about people's uptake of base rate information strengthen the case for Inference to the Best Explanation. We have three situations. The first is one where there is base rate information that does not obviously fit into a causal explanatory scheme, but is all the information we have, as in dealing a card from a shuffled deck. Here we tend to use the information. The second is one where there is also evidence that can form the basis for an explanationist inference, such as a witness's report, the content of which would be explained by the fact that the witness saw what she reported. Here, the information that fits into an explanatory scheme trumps the base rate information. The third situation is one where both kinds of information are available, but the base rate information does fit into a causal narrative, because it can itself be causally explained. Here, the base rate information is used alongside the other information. The evidential structure of these cases is complex, and the final inferences – that it was a blue taxi or that the student failed – are not what explain the base rate data, but these results nevertheless seem strongly to support the claim that explanatory considerations influence inference.

In the last section, I attempted to avoid the catch-22 that threatens when one tries to make out both the matching and the guiding claims by suggesting ways in which the coverage of Inference to the Best Explanation improves on standard accounts. The cases discussed in the present section – of hasty explanatory inference and of a tendency to ignore information when it does not fit into an explanationist scheme but to use the same information when it does – can be seen as an application of the same general strategy. These phenomena, in part because many of them seem to involve inferential errors, are not covered by Bayesianism or any of the other accounts of induction we have been using as foils to Inference to the Best Explanation. And since our main goal is descriptive rather than normative, this is to the explanationist's

credit. Of course the moral I draw from this story is not that the explanationist impulses here revealed are something that only apply to our irrational moments, but that they are constantly in play but brought out with particular clarity in some cases where they may lead us astray. Evidence that we go with explanation even when normative considerations should stay our hand provides particularly strong evidence that explanationist thinking is a deep-seated aspect of our cognitive economy.

These cases support the guiding claim not just because they are instances the explanationist account covers while the other accounts do not – not just because they show Inference to the Best Explanation to have better coverage than other accounts – but also because they are cases that suggest more directly that explanatory considerations are guiding inference, for better and occasionally for worse. One of the particularly attractive features of research in the Kahneman and Tversky tradition (at least to a philosopher, if not to a psychologist) is the extent to which one can empathize with the poor subjects' inferential weaknesses, the extent to which their mistakes strike one as quite natural. And in thinking about these cases it thus also seems to me that one finds the guiding claim increasingly plausible and intuitive as a description of an aspect of one's own thinking.

From cause to explanation

I turn now to a different kind of argument for Inference to the Best Explanation. It is based on the observation that we often find it easier to reason in physical than in logical terms, in terms of causes rather than in terms of logical relations. This is one of the reasons we find it difficult to take Bayes's theorem neat, as we saw in the last chapter. And one reason we find physical thinking congenial is that our inductive reasoning is often abetted by processes of simulation, where we run causal scenarios in order to determine what inferences to draw. Inference to the Best Explanation suits this way of thinking. We are constantly asking why things are as we find them to be, and we answer our own questions by inventing explanatory stories and thus coming to believe some of them, based on how good are the explanations they would provide.

I find this picture compelling, but it can be resisted. In particular, while one may agree with Mill about 'the notion of cause being the root of the whole theory of induction' (1904: III.V.2), it does not follow from this alone that we must be explanationists in the strong sense I here promote. To be sure, talk explicitly only about causes may in fact concern explanation. Thus, to ask for the cause of an effect is often an explanation request, a way of asking why. Nevertheless, I take it that we may think about causes without thinking especially about explanations, and so we might judge likeliest cause without considering loveliest explanation. It is that gap that I want to focus on here.

The problem of the gap between causation and explanation is vivid in the case of Mill's methods. As we have seen, there is in particular a striking isomorphism between the structure of the Method of Difference in particular and contrastive explanation, and this made it possible for me to argue in some detail in the context of Semmelweis's research into childbed fever how explanatory considerations can be a guide to inference. But this argument faces the catch-22. For as Eric Barnes has argued (1995: 255–6), we can see the Method of Difference as resting simply on the deterministic principle 'same cause, same effect', thus requiring no appeal to explanatory considerations. If a cause of a certain type will always produce its characteristic effect, then where there is only one difference between a case where the effect occurs and one where it does not, that difference must mark part of the cause. Otherwise, the cause would lie entirely in what is shared between the two cases, even though the effect only occurs in one, violating the principle. If we are willing to accept the deterministic principle then, we seem to need no recourse to further explanatory considerations to motivate the Method of Difference. Those differences we take to be causes may well also explain the contrast between the two cases, but that fact need not appear on our inferential path. Perhaps, as Barnes suggests, the explanation for the isomorphism between inference and explanation in the contrastive case is not that inference is guided by explanation, but that both are determined by the principle of same cause, same effect. (One might also argue that the Method of Difference rests on a counterfactual conception of causation, a conception that makes causes necessary but not sufficient conditions for their effects. For the contrastive case could be taken as evidence that, had the putative cause not occurred, the effect would not have occurred either.)

As in the rest of life, so in philosophy: it is not always easy to distinguish a case of causation – here explanatory considerations causing inference – from a case of common cause – both inference and explanation caused by belief in the deterministic principle. And in the case of the Method of Difference, I do not even want to claim that the inference is *always* run on explanatory grounds. One reason for my modesty here is that thinking in terms of explanation, though ubiquitous, is a kind of conscious, articulated thought process, while applications of the Method of Difference are sometimes near-automatic. Take for example what is probably our most secure basis for causal inference: manipulation. I am completely convinced that flicking the switch causes the light to go on, because I find I can control the light with the switch; similarly, I am completely convinced that the motion of the mouse causes the movement of the cursor on the screen, no matter how little I understand of how this could be, because I find I can use the mouse to manipulate the cursor. These sorts of inference are tantamount to applications of the Method of Difference, because they generate sets of cases where contrasts in effect are matched with contrasts in (here very immediate) history. And they are particularly convincing applications of the

method because they provide so many contrasts (rather than a single pair) and because by introducing and then retracting the antecedent difference ourselves we may be particularly confident that there is no other hidden difference that is doing the real causal work. (Mill recognizes this latter advantage in the context of 'artificial experiment'; see III.VIII.3.) These particularly confident applications of the Method of Difference are not always plausibly described as inferences to the best explanation, because we may make the inference without needing to think about it. I became confident that the movement of the mouse was causing the movement of the cursor as I found I could control the one with the other. I did not have to consider whether the movement of the one explains the movement of the other, much less whether mouse movement is a better explanation for cursor movement than some other hypothesis. Here the Method of Difference operates as it were on its own, no more requiring an explanatory guide than the even simpler application of the method given by Mill (III.IX.6), namely our discovery that fire burns.

So, just as it is no part of my general case to claim that explanatory considerations are the sole guides to inference, it is no part of my particular case about contrastive inference that every inference that can be described as an application of the Method of Difference is in fact an Inference to the Best Explanation. Nevertheless, I do of course want to argue that there are many such cases, and there seem to be two broad strategies for doing this. The first is to argue that there are signs even in some of the purest applications of the Method of Difference of explanatory considerations at work. The second is to argue that while such applications taken alone are compatible with the common-cause view that gives no inferential role to explanation, once we take into account the roles for explanation in capturing richer inferential contexts – such as those of unification and mechanism – then we have reason to read back an explanationist gloss in the simpler cases of the Method of Difference as well. I want briefly to explore both of these strategies.

To begin with the first strategy, we are here considering simple cases which can be described as flowing from the deterministic principle directly, so there is no attempt to show that the use of explanatory considerations is inevitable. The question is rather whether even in such a simple context there are any reasons for supposing that explanatory considerations are often in inferential play. I think there are. My position here is very close to the view of the relationship between explanationism and Bayesianism I presented in the last chapter. Explanatory considerations can be seen primarily not as in competition with Bayesian calculation, but as a way of realizing that calculation. Similarly, here I want to suggest not that explanatory inference somehow refutes the deterministic principle, but rather that we often find ourselves reasoning through that principle in explanatory terms. But how can I say this? Bayes's theorem is difficult to think through, so it is plausible enough to say that we need the explanationist crutch. But what could be

simpler than 'same cause, same effect'? How could one claim that we need to lean on explanatory considerations in this case? Well, my claim is not that we need to, just that we do, and there are several reasons one might give for this. One sort of reason might be considered phenomenological. It is simply that there seem to be many cases of contrastive inference where one's inference does not seem to travel through the deterministic principle but does seem to follow explanationist lines.

One reason for saying this is that the course of inquiry is often initiated by a why-question. Thus, in trying to find the cause of childbed fever, Semmelweis focused on the question of why the mortality rate was so different in the two maternity divisions of his hospital. And having framed his inquiry in explanatory terms, it is not surprising that he went on to assess his hypotheses in explanatory terms as well, by considering whether the hypotheses would indeed explain the diverse contrasts in his evidence. Semmelweis's primary aim was not intellectual understanding but rather practical control; nevertheless, he conceptualized the question of the cause of the fever as a search for an explanation and, having done this, he generated answers that were themselves in the form of explanations.

One may go on to wonder why it is so natural for us to conceive of requests for causes as requests for explanations. My present argument is strictly independent of this issue. That argument is that causal inquiries deploying the Method of Difference are often initiated by why-questions, and this leads us to assess the answers in explanatory terms. But the skeptic might question whether our initial question is really an explanation request. Is there really a significant difference between asking 'What causes E?' and 'Why E?'? As I have already mentioned, I take it that many questions of the first form are in fact explanation requests, but there may be a relevant difference between the two forms in the particular context of contrastive questions. For while it was perfectly in order for Semmelweis to ask for an explanation for the difference in mortality rates, this is not quite the same as asking what caused the difference. Indeed to my ear it is not quite right to say in these sorts of contrastive cases that one is looking for a cause of the difference, since that suggests some kind of interaction between the two cases. But an explanation of the difference is exactly what we are looking for. And this asymmetry seems preserved when we consider answers to our contrastive questions. It is not so natural to ask 'Does C cause the difference?', but entirely natural to ask 'Does C explain the difference?'

So, looking for a cause we often ask a why-question about our contrastive evidence, and this leads us to think of potential causes as potential explanations. And as we saw in chapter 5, the next step is often to create or attempt to create more evidential contrasts to narrow down to a single candidate. Thus the difference in birth position between the two wards was a potential cause of the fever, but was ruled out because changing to a uniform position did not make a difference, and the infection hypothesis was ruled in

when it was found that introducing disinfectant did make a difference. It is also easy to find examples where a hasty inference could have been avoided with additional data that would have ruled out an irrelevant difference, on explanatory grounds. In the seventeenth century, Sir Kenelm Digby, a founding member of the Royal Society, enthusiastically endorsed the idea, attributed to Paracelsus, that there was a 'sympathetic powder' that could cure wounds at a distance by, for example, rubbing it on the sword that caused the wound (Gjertsen 1989: 108–9). This hypothesis found contrastive support, since patients treated with sympathetic powder recovered more quickly than patients whose wounds were dressed by doctors and nurses in the normal way. The real reason for this contrast (we now suppose) was that the doctors and nurses inadvertently infected the wounds they were trying to heal, while the patients sympathetically treated were in effect left alone (only the swords were treated), so were less likely to be infected. Additional data, comparing sympathetic treatment and no treatment at all, would have prevented the mistaken inference.

Doubtless there is a way of glossing all of this in terms of the same cause, same effect principle, but I submit that this is not the way we actually tend to think about the impact of complex assemblages of contrastive data. Thus in cases like that of Semmelweis what we find more intuitive is to see inquirers getting into a position where only one of their potential causes could explain the mass of the evidence. In effect, Semmelweis converted the question of which is the best explanation of the original data into the question of which is the *only* explanation of the richer set. We often decide between competing hypotheses by looking for additional data that will discriminate between them. Perhaps in some extreme cases that discrimination works through the refutation of one of the hypotheses; but what seems far more common is that the additional evidence, though logically compatible with both hypotheses, can only be explained by one of them. This eliminative process shows how delicate questions about what makes one explanation lovelier than another can sometimes be finessed by a mechanism that remains within the ambit of Inference to the Best Explanation. Often it is a process of manipulation that makes such finessing possible, as in the case of Semmelweis's control over the fever by means of disinfectant (cf. Clarke 2001, who deploys this argument in the context of entity realism).

Contrastive inference is thus in part an eliminative method, where putative causes are ruled out because they fail to account for evidential contrasts. I use the word 'account' in a pretence of neutrality, but the point is that we find it natural to think of this failure in explanatory terms. So far as the deterministic principle goes, a feature common to fact and foil may nevertheless be a cause of the fact, but it will not explain the contrast and it is in terms of that failure that we often eliminate that causal candidate. The deterministic principle does not entail that factors shared by fact and foil are not causes of the fact, which is just as well, since for example oxygen is a

cause of the dry match lighting when struck, even though it is also present in the case of the match that fails to light because it was damp or because it was not struck. Of course the failure of a factor to explain a contrast, because it fails to mark a difference between fact and foil, does not entail that the factor is not a cause either, but it often means that we have reason to believe it is not a cause and, as I suggested in chapter 5, we often make this judgment in an explanatory idiom, ruling out putative causes because of their failure to explain new contrasts. A shared factor does not for that reason fail to cause, but it does fail to explain.

Another general reason why explanationist thinking may be more natural than thinking directly about causes concerns the subjunctive aspect of inferential thinking. Our thinking about what to infer is shot through with consideration of what causes *would* account for various effects and of what effects the causes *would* have. And it seems to me that this sort of subjunctive thinking is often easier or anyway more natural when we gloss 'account for' as explain rather than as cause. Would general 'epidemic influences' cause childbed fever or the difference in its occurrence between the two maternity wards? Who knows? Maybe such influences are like the oxygen around the match and so would be a cause. But to the question of whether such an influence would explain the contrast the answer is clear.

Similarly, when a putative cause does mark a difference between fact and foil, we can say with confidence that it would explain the contrast, even in cases where we might be quite unsure whether it would cause the effect, because the cause here is far from a sufficient condition. Jones's syphilis (and Smith's lack of syphilis) would explain why Jones rather than Smith contracted paresis, but would it cause Jones's paresis? (Recall that most people with syphilis do not go on to contract paresis.) And in the case where we think subjunctively about possible effects, we also find it convenient to think in explanatory rather than in bluff causal terms, because it is often easier to say what a factor would explain than it is to say what it would cause. Thus, as we saw in the last chapter, asking what a causal hypothesis would explain is often the way we determine relevant evidence.

Let us pause to sum up the line of argument so far in this section. My strategy is to defend the guiding claim here by parlaying the claim that we often think about inductive relations in physical rather than logical terms – causation rather than deduction – into the desired conclusion that we often think about inductive relations in explanatory terms. In causal inference, we are thinking both about what would and would not cause the effects we observe and about what effects the putative causes would and would not have. What I have been arguing is that even in the context of simple contrastive inference we often find ourselves thinking about these questions by asking of the effects, what would and would not explain them, and by asking of the putative causes, what they would and would not explain. I have suggested that this is so because of our tendency to initiate inference by

asking why-questions about our data, because we appeal to a notion of explanatory failure to discriminate between competing hypotheses, and because explanationist thinking supports our tendency to weigh evidence in subjunctive terms.

I have been focusing on a case for saying that even simple contrastive inferences really are often conducted in explanationist terms by appeal only to those simple inferences. But the case of construing these inferences in those terms, and indeed the case for explanationism generally, is substantially strengthened when one takes account of additional factors that contribute to inferential judgment. As I have already noted, many of them are naturally glossed in terms of Inference to the Best Explanation, and this provides an additional reason to suppose that the simple contrastive cases work that way too, since it supports a more unitary and plausible picture of our cognitive activities.

I will here only be able very briefly to sketch this sort of holistic case for Inference to the Best Explanation, but I want to touch on three factors: mechanism, unification and background belief. The pattern of contrastive inference that I have described depends on finding a difference, not on describing a mechanism. And while being able to give a mechanism can be a very substantial epistemic advantage, we clearly do make inferences from differences even when we are not in a position to offer a mechanism (or not much of one) linking putative cause and effect. Indeed inferences not backed by mechanism may nevertheless be quite firm, as in the link between smoking and lung cancer, and a difference in effect leads to the expectation of a causal difference even when no mechanism is in sight. Thus when Semmelweis found that mothers who delivered in the street contracted childbed fever much less frequently than those who made it into the maternity wards, he supposed that this difference was causally relevant even though at this stage of his investigation he had no idea what the underlying mechanism here could be. The case of Semmelweis also illustrates the epistemic value of finding a mechanism, since he went on to describe a mechanism of infection from cadaverous matter, from medical students who conducted autopsies before examining pregnant women in a way that would introduce cadaverous particles into the mother's vascular system. (This also confirmed Semmelweis's prior inference about street births, since rather as in the case of the swords and sympathetic powder, in those cases the women were not examined and so were not infected (Semmelweis 1860: 80–101).) Providing a mechanism is an explanatory virtue, and one that naturally complements contrastive inference, as well as applying in cases where the data are not contrastive. Thus it seems to me that the important (though not invariable) role of providing explanatory mechanisms in our inferential practice strengthens the general case for saying that explanatory considerations guide inference, as well as the specific case for the guiding claim in the context of contrastive inference.

Similar points apply to unification. I have in mind here a very broad concept, incorporating the considerations of scope, simplicity and consilience. Just how this notion should be analyzed is a question well beyond the scope of the present discussion, but it does appear that unification is an explanatory virtue (Friedman 1974; Kitcher 1989). Explanations or patterns of explanation that explain more and more diverse phenomena, explanations that do more to reveal the unity beneath the superficially messy phenomena, are explanations that provide greater understanding. It also seems clear that unity is an inferential virtue, or rather a basket of inferential virtues, and these are naturally described in explanationist terms (Thagard 1978). Like the virtue of mechanism, the virtue of unification naturally complements the basic contrastive mechanism and also applies to non-contrastive data.

Like mechanism and unification, the basic contrastive mechanism does not appeal to background belief, although Mill's discussion of the application of the method does implicitly give it a role. As we have seen, he admits that we will never find cases where there is only a single difference between fact and foil, but allows that we may ignore differences 'such as are already known to be immaterial to the result' (1904: III.VIII.3), and this knowledge will come from our background. In any event, it is clear that the background beliefs we bring to a particular inquiry have an enormous effect on the inferences we draw. These inferential influences are diverse and poorly understood, but it seems clear that a number of them can be given an explanationist gloss.

Here are some examples. First, as we noted in our discussion of prior probabilities in the last chapter, beliefs that govern our next inference may have been formed earlier on the basis of explanatory considerations. Second, the structure of the background will play a major role in determining the unificatory virtues of a new candidate explanation, since the same explanation will add to unity in one background context but detract from it in another. So explanatory loveliness is in part relative to background. A third and rather different inferential-explanatory role for the background exists because the background will incorporate particular explanatory standards. Thus a background might include a ban on explanations that appeal to teleology, to action at a distance or to irreducibly indeterministic processes, and it might privilege certain types of properties (e.g. 'primary properties'), marking them as providers of a particularly lovely explanation (cf. Day and Kincaid 1994). Or what counts as a lovely explanation may be determined in part by previous explanations that serve an exemplary function, as Kuhn describes it (esp. 1970), or by more general 'styles of reasoning' (Hacking 1982, 1992). Variation in explanatory standard should be seen as occurring at diverse levels of generality, from features peculiar to small scientific specialties to those that may apply to almost the entire scientific community at a particular time. The cardinal virtues of unification and mechanism that I have flagged I think span very many scientific

backgrounds, past and present, but their interpretation is bound to vary, and there may even be scientific traditions in which they do not figure.

The background thus should be seen as affecting judgments of loveliness in two different ways: for a given standard, how lovely an explanation is will depend in part on what other explanations are already accepted, and the standard itself will be partially determined by the background. The importance of the background in inference, and the plausible suggestion that what counts as a lovely explanation is thus context sensitive, is entirely compatible with Inference to the Best Explanation, as I construe it. That account maintains that loveliness is a guide to likeliness, but it does not require that standards of loveliness are unchanging or independent of background belief.

That is enough. Suppose that the guiding claim and Inference to the Best Explanation are indeed telling us something important about our inferential practices. Why do we think in this way? Why don't we just go for the likeliest cause, where judgments of likeliness have nothing to do with explanation? This is too big a question for this book properly to address, but I offer one speculation as a potential very partial answer. In this section I have asserted that we often think about inference in physical rather than logical terms, by constructing causal models and simulations rather than by investigating logical relations between statements or propositions. I take it that this thinking in the material mode is an aid to effective inference, but limiting one's attention to causal relations is restrictive, as compared to what is in a sense the universality of logical relations. My speculation is that by moving from thinking just about causation to thinking about explanation, we in effect compensate for this limitation, by being able to incorporate into our inferential thinking the very diverse range of relevant considerations and in a unified framework. Causes are not lovely or ugly, but explanations are, and this enables us to take account of factors that are not part of the causal relation itself but do bear on inference. By deploying Inference to the Best Explanation we gain the advantages of both logical and of causal thinking, the best of both worlds.

Having finished delivering her paper, the philosopher invited discussion by asking, 'Now was that false, or just trivial?'. It is not easy to avoid those horns. In the case of Inference to the Best Explanation, we face triviality if we end up saying no more than that scientists often infer what they judge to be the likeliest of competing explanations. We avoid triviality by insisting the explanatory considerations are a guide to inference, that loveliness is a guide to likeliness. But what about truth? In this chapter I've tried to convince you that the guiding claim is true, by means of a diverse selection of arguments, in the context of a particular challenge, the catch-22. We would like to show a good match between explanatory and independently characterized inferential virtues, as provided by other accounts of inference, but in doing this we invite the rejection of the guiding claim, in favor of the

terms of the independent characterization. I've pursued three strategies for avoiding the catch-22. First, there is improved coverage: the various ways in which an appeal to explanatory considerations provides a fuller account of inductive practices than other accounts on offer. Secondly, there is our obsessive interest in explanation and the psychological research that strongly suggests that we let explanatory considerations guide our inferences, sometimes to a fault. Thirdly, I've argued that the way we think about potential causes and effects, especially in the context of contrastive inference, is naturally structured in explanatory terms. We tend to frame the causal questions as why-questions, we naturally think our way through a method of eliminating competing hypotheses in explanatory terms, and explanatory thinking suits our subjunctive train of inferential thought. In addition, I suggested that we can give a coherent account of the inferential roles of mechanism, unification and background belief if we construe them in a context in which explanatory considerations are an important guide to inference. Not the only guide, but my intent is that the version of Inference to the Best Explanation I here defend is strong enough to avoid triviality, but weak enough to be true.

Chapter 9

Loveliness and truth

Voltaire's objection

The main aim of this book is to articulate and defend Inference to the Best Explanation as a partial answer to the descriptive problem of induction, the problem of giving a principled account of the way we actually go about making non-demonstrative inferences. As I argued in chapter 1, this is different from the justificatory problem, the problem of showing that our inductive practices are reliable or truth-tropic. After all, we might actually be unreliable reasoners. It is perhaps difficult to take this possibility seriously in the case of mundane inferences we have made hitherto about observed medium sized dry goods, but it is of course a possibility that at least epistemologists find it very easy to take seriously as regards the unobserved and the unobservable. And it may well seem that Inference to the Best Explanation makes the justificatory problem particularly recalcitrant, since it may seem particularly implausible that explanatory considerations should be a reliable guide to the truth. Indeed some may find this so implausible that, finding they cannot shake the belief that our actual methods are pretty reliable, they refuse to accept even the descriptive pretensions of Inference to the Best Explanation. That is one way that the descriptive and justificatory problems are related, and it is one reason for me to discuss the bearing of Inference to the Best Explanation on the justificatory problem of induction, to argue that it does not raise any special problem of justification (though the ordinary Humean problem is more than trouble enough). Another is simply that it is such a natural question to ask, even if you are by now irredeemably committed to the idea that explanatory considerations guide our inferences. Suppose this is in fact the way we think: what bearing does this have on the question of whether we in general and scientists in particular are in a successful truth business?

In chapter 4, I flagged two general reasons one might doubt that Inference to the Best Explanation could give an account of a reliable inferential practice, reasons indeed for thinking it very unlikely that using loveliness as a guide to likeliness should be reliable. The first, 'Hungerford's objection'

('Beauty is in the eye of the beholder'), was that explanatory loveliness is too subjective and variable to give a suitably objective account of inference. The other, 'Voltaire's objection', was that Inference to the Best Explanation makes the successes of our inferential practices a miracle. We are to infer that the hypothesis which would, if true, provide the loveliest explanation of our evidence, is therefore the explanation that is likeliest to be true. But why should we believe that we inhabit the loveliest of all possible worlds? If loveliness is subjective, it is no guide to inference; and even if it is objective, why should it line up with truth?

In reply to Hungerford's objection, the first thing to note is that reliable inference is itself audience relative. Warranted inference depends on available evidence, and different people have different evidence: inferential variation comes from evidential variation. Moreover, as we have seen, inference also depends strongly on background beliefs, and these too will vary from person to person. It might, however, be argued that all background variation, and indeed all inductive variation, ultimately reduces to evidence variation, since a difference in beliefs, if rational, must reduce to a difference in prior evidence. But this is unlikely. It is not plausible to suppose that all scientific disagreements are due either to a difference in evidence or to irrationality. This point has been developed with particular power by Kuhn (esp. 1977: ch. 13). For example, such disagreements sometimes stem from different judgments of the fruitfulness of scientific theory, of its ability to solve new puzzles or to resolve old anomalies. What counts is not just what the theory has explained, but what it promises to explain. This is clearly something over which rational investigators may differ; indeed epistemic divergence among its members may perform an important cognitive function for a scientific group, by allowing it to hedge its bets. A community whose members do not always jump the same inferential way may be more reliable than one in lockstep. And the prospects of accounting for these inferential differences entirely in terms of differences in available evidence are dim. Moreover, as I will argue in the next chapter, warranted inference may depend not only on the content of the evidence, but when it was acquired. Evidence that was available to the scientist when she constructed her theory may provide less support than it would have, had it been predicted instead. I will account for this difference by appeal to an inference to the best explanation that brings out the importance of distinguishing between the objective support that a theory enjoys and scientists' fallible assessments of that support. This distinction reveals another source of inferential variation that does not reduce to evidential variation. So the simple claim that explanation is audience relative will not show that Inference to the Best Explanation could not describe a reliable form of inference, since reliable inference would be audience relative as well, in diverse ways.

Hungerford's objection can also be defanged from the other side, by arguing that explanatory considerations do not have the strong form of

relativity that the objection suggests. The explanatory factors I mentioned in the last chapter – unification, mechanism, precision and so on – involve us in no more relativity on the explanatory side than they do on the inferential side, since they are the same factors in both cases. But the compatibility of Inference to the Best Explanation with a reasonable version of the interest relativity of explanation is perhaps clearest in the case of contrastive explanation. As I noted in chapter 3, a contrastive analysis of why-questions illuminates the interest relativity of explanation by analyzing a difference in explanatory interest in terms of a difference in foil choice. Two people differ in what they will accept as a good explanation of the same fact, since one is interested in explaining that fact relative to one foil while the other has a different contrast in mind. Jones's syphilis will explain why he contracted paresis for someone who is interested in understanding why he, rather than Smith, who did not have syphilis, has paresis, but not for someone who wants to know why Jones contracted paresis when other people with syphilis did not. My account of contrastive explanation demystifies the phenomenon of interest relativity in two ways. First, by taking the phenomenon to be explained to be a contrast rather than the fact alone, it reduces the interest relativity of explanation to the truisms that different people are interested in explaining different phenomena and that a good explanation of one phenomenon will not in general be a good explanation of another. Secondly, as the difference condition on contrastive explanation shows, these differences in interest will require explanations that cite different but compatible elements of the causal history of the fact. This is no embarrassment for Inference to the Best Explanation, since that account allows us to infer many explanations of the same fact, so long as they are compatible. Moreover, the account allows that different people are sometimes warranted in inferring different contrastive explanations, since a difference in foil may correspond to a difference in experimental controls, and these differences clearly may be epistemically relevant. A difference in interest may correspond to a difference in evidence.

I conclude that Hungerford's objection has yet to be made in a form that threatens Inference to the Best Explanation. Explanation is audience relative, but so is inference, and we have been given no reason to suppose that the one relativity is more extreme than the other. Let us turn now to Voltaire. Why should the explanation that would provide the most understanding if it were true be the explanation that is most likely to be true? Why should we live in the loveliest of all possible worlds? Voltaire's objection is that, while loveliness may be as objective as you like, the coincidence of loveliness and likeliness is too good to be true. It would be a miracle if using explanatory considerations as a guide to inference were reliably to take us to the truth.

The first point to make in response to Voltaire's objection is that we should not be asked to prove too much. Induction by its nature is not a form of inference that is guaranteed to yield only truths from truths. Moreover,

Inference to the Best Explanation is meant to be descriptively adequate, and the psychological research we discussed in the last two chapters brings out some sources of systematic unreliability in our actual inferential practices. (One might mount a parallel argument from the history of science, as in the so-called 'pessimistic induction' from the falsity of past scientific theories (Laudan 1984).) If Inference to the Best Explanation can account for some of this, that is to its credit. It is no objection to explanationism that it does not make us out to be more reliable than we actually are.

Still, I take it that our inductive practices are reasonably reliable, certainly better than random guessing. And at this level, one should perhaps say that Voltaire is right: even the moderate reliability of Inference to the Best Explanation would require a miracle. For we are talking about induction here and there is a sense in which the success of induction is miraculous or inexplicable on any account of how it is done. This is one way of putting the conclusion of David Hume's great skeptical argument. Hume himself used an over-simple 'More of the Same' description of inductive inference but, as I noted in chapter 1, his skeptical argument does not depend on the particular description he gave. Whatever account one gives of our non-deductive inferences, there is no way to show a priori that they will be successful, because to say that they are non-deductive is just to say that there are possible worlds where they fail. Nor is there any way to show this a posteriori since, given only our evidence to this point and all a priori truths, the claim that our inferences will be successful is a claim that could only be the conclusion of a non-deductive argument and so would beg the question. In short, the impossibility of justifying induction does not depend on a particular account of our practices, but only on the fact that they are inductive. Consequently, in the absence of an answer to Hume, any account of induction leaves inductive success miraculous. That is very bad news, but of course it is no worse news for Inference to the Best Explanation than any other account of induction. Inference to the Best Explanation would not have us infer that the loveliest possible world is actual; at most, it has us infer the loveliest of those worlds where our observations hold. And what Hume showed was that the claim that any system of choosing from among those worlds is reliable is indefensible.

At this level of generality, there are only two positions that could claim an advantage over Inference to the Best Explanation. One is a deductivist account of scientific method, such as Karl Popper's, which takes Hume to have shown that we must abjure induction altogether. This is a brave and direct response to Hume, but I take it that no such account can be true to our actual inferential practices: we clearly do indulge in induction. The other is an account which grants that we use induction, but would have us use it more sparingly than Inference to the Best Explanation allows. A good example of this is Bas van Fraassen's 'constructive empiricism' (1980). In brief, van Fraassen's view is that we restrict inductive inferences to claims about

observable phenomena. When a scientist accepts a claim about unobservable entities and processes, what she believes is only that the claim is 'empirically adequate', that its observable consequences are true. This contrasts with the realist version of Inference to the Best Explanation that we have been considering in this book, since that account sanctions inferences to the truth (or the approximate truth) of the best explanation, whether it appeals to observables or not. So van Fraassen might claim that his account makes inductive success less miraculous than does Inference to the Best Explanation, for the simple reason that it requires fewer miracles.

I do not find constructive empiricism an attractive account. This is not the place for a detailed assessment, but I will voice two complaints. (I will have more to say about van Fraassen in the last section of this chapter and in chapter 11.) First, it is not clear that the account is consistent. Van Fraassen takes the view that what makes a claim observable is not that the claim employs only observation terms but that, however 'theory-laden' the description, what is described is something that our best theories about our sensory capacities tell us we can observe. To take one of his examples, we may describe a table as a swarm of electrons, protons and neutrons without making the claim about the table non-observational (1980: 58). We cannot see a single particle, but we can see a swarm. Since the claim that my computer is now on a table is observable, and I am making the requisite observations, van Fraassen would have me believe that this claim is literally true. He would then also have me believe that 'my computer is on a swarm of particles' is literally true, not just empirically adequate. But I do not see how I can believe this true unless I believe that particles exist, which I take it is just the sort of thing I am not supposed to believe. (Am I supposed to believe that swarms of particles exist but individual particles do not?) My second complaint is that constructive empiricism gives a poor description of our actual practices, since we do actually infer the truth of claims about the unobservables. Most scientists do believe in electrons, protons and neutrons and the claims they make about them, not just that these claims have true observable consequences; each of us believes that other people have had pains, itches and visual impressions, though none of us can observe other people's phenomenal states. Moreover, the inferential path to unobservables is often the same as to unobserved observables. In these two sorts of case, the reasons for belief can be equally strong, so the suggestion that we infer truth in one case but not the other seems perverse. Perhaps there could be creatures whose inductive mechanisms made them constructive empiricists, but they would be different from us.

Whatever one's view about the general merit of van Fraassen's position, however, it cannot claim an advantage over Inference to the Best Explanation with respect to Voltaire's objection, since the objection is not that it is inexplicable that explanatory considerations should lead to so many correct inferences, but that there should be any connection between

explanatory and inferential considerations at all. (Nor do I think that van Fraassen claims such an advantage.) Voltaire's objection is not that Inference to the Best Explanation would make our inferences insufficiently parsimonious, but that it would make the success of the inferences we do make inexplicable. Moreover, as we will see in chapter 11, van Fraassen himself claims that there is such a connection, that the best explanation is one guide, not to truth in general, but to the truth of observable consequences (cf. 1980: 23, 71). If someone claimed that patterns in tea leaves foretell the future, one might object that the connection is inexplicable, to which it would be no reply to say that the leaves are only reliable guides to the observable future.

So, a response to Voltaire's objection is to say that it reduces to Hume's problem. This is hardly a solution, but it would show that Inference to the Best Explanation is in no worse shape than any other account of induction. Leaving Humean skepticism to one side, however, might one claim that Inference to the Best Explanation would make the reliability of induction somehow more surprising than some other account? That would be the case, for example, if there was an incoherence in explanationism that does not arise in other accounts of induction. I can think of two quite different forms this incoherence might be supposed to take, but careful readers will by now know why I hold neither of them to be realized in explanationism properly construed. The first is that Inference to the Best Explanation would not be epistemically effective, since an actual explanation must be true, so one would have to know the truth before one could infer an explanation. According to this objection, explanationism gets things backwards, because we must infer before we explain. As we saw in chapter 4, however, this objection is based on a misunderstanding of Inference to the Best Explanation. Perhaps actual explanations must be true, but the account has us infer to the best potential explanation, a hypothesis that would explain if true. There is no incoherence here. The second claimed source of incoherence is probabilistic: someone who uses Inference to the Best Explanation will violate Bayesian constraints on belief revision and so will be susceptible to a dutch book (van Fraassen 1989: 160–70). As we saw in chapter 7, however, explanationism can be understood in a way that respects Bayesian constraints and indeed in a way that serves a Bayesian approach to inference. The probabilistic incoherence that the dutch book argument displays only occurs if explanatory considerations are supposed to boost the posterior probability after conditionalization has taken place; but this leaves us free to engage in the coherent use of explanatory considerations earlier in the process, for example in the determination of priors and likelihoods (cf. Harman 1999: ch. 4; Okasha 2000).

Indeed far from suffering from an incoherence that other accounts avoid, Inference to the Best Explanation in fact inherits whatever justification various other accounts of inference would provide. As we saw with the

catch-22 of the last chapter, the structural similarity between explanationism and other accounts of inference presents a peculiar challenge to the descriptive task of showing that explanatory considerations are actually driving inference; but in the context of justification and Voltaire's objection these similarities are an advantage. Thus insofar as Bayesian inference avoids Voltaire's objection, so can Inference to the Best Explanation, since we may see the latter as a way of realizing the former. Similarly, if the method of difference strikes you as avoiding the objection, so should Inference to the Best Explanation, in light of the structural similarity between simple contrastive inference and contrastive explanation, a similarity I exploited in chapters 3 to 6. Moreover, as we noted in the last chapter, our explanatory standards and the inferences they support are highly sensitive to our background beliefs, so it seems that Inference to the Best Explanation has the resource to capture the way reliable patterns of inference depend on taking this background into account. Since those background beliefs were themselves generated through explanatory inference, we have a kind of feedback between judgments of likeliness and judgments of loveliness. Successful inferences become part of the background, and influence what counts as a lovely explanation and thus influence future inferences. This is as it should be: as we learn more about the world, we not only know more but we also become better inferential instruments.

Voltaire's objection thus appears ill founded. First, our actual inductive practices are far from perfectly reliable, so it is no criticism of explanationism that it would not make them so. Secondly, although Inference to the Best Explanation does not solve the Humean problem of induction, nothing else does either. Thirdly, there is no reason to believe explanationism is incoherent and, fourthly, indeed it can capture normatively attractive features of Mill's methods, Bayesianism and the essential role of background belief to inference. To these points I would add an inversion of the worry I expressed earlier about the relationship between the normative and the descriptive. Hume notwithstanding, we believe that our inductive methods are pretty reliable. The worry was that a belief that the reliability of Inference to the Best Explanation would be miraculous would thus undermine confidence in its descriptive adequacy as well. 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 Voltaire's objection.

The two-stage process

Nevertheless, I do not wish to put Voltaire to bed quite yet. For van Fraassen has raised a further and interesting objection to Inference to the Best Explanation that I wish to discuss, the objection from what I will call 'underconsideration' (1989: 142–50). Before doing so, however, it will be

useful to say something further about a plausible structure for inferences to the best explanation, a structure that distinguishes hypothesis generation from hypothesis selection. On this view, the mechanism by which we settle on which of the many possible causes to infer has two stages. The first is the process of generation, the result of which is that we only consider a small number of possible causes; the second is the process of selection from among those live candidates. Explanatory considerations play a role in both stages, and this comes out particularly clearly in the context of contrastive inference. The obvious way to select causes from differences left by a particular evidential contrast is to perform more experiments. As we have seen, this is what Semmelweis did. The initial contrast in mortality between the two maternity divisions might have been due to the difference in exposure to the priest, in birth position, or in infection by cadaveric matter. But Semmelweis found that only eliminating the third difference made a difference to the mortality rates, so he inferred that this difference marked a cause. Even if the three differences were equally good explanations of the initial contrast, the third was the best explanation of the total evidence. Semmelweis did not make his inference until he was able to make this discrimination. Thus additional evidence performs an eliminative function. Typically, this process will not leave only one candidate in the running, but then the diverse explanatory considerations mentioned in the last chapter, considerations that include mechanism and unification, come into play.

But an account of this process of selection from the live candidates is only half of the story of the way we handle the problem of multiple causes or hypotheses. The other half is that we only consider a small portion of the actual and possible causes in the first place. We never begin with a full menu of all possible causal differences, because this menu would be too large to generate or handle. Yet the class of differences we do consider is not generated randomly. We must use some sort of short list mechanism, where our background beliefs help us to generate a very limited list of plausible hypotheses, from which we then choose. While the full menu mechanism would have only a single filter, one that selects from all the possible differences, the actual short list mechanism we employ has two stages, one where a limited list of live candidates is generated, the other where a selection is made from this list.

The need to use a short list rather than a full menu raises a question about the scope of Inference to the Best Explanation. We cannot say that the differences that do not make it onto the list are dismissed because they are judged to provide only inferior explanations. They are not dismissed on any grounds, because they are never considered. This raises a challenge for Inference to the Best Explanation. How can it account for the processes by which short lists are generated? The principles of generation that solve much of the problem of multiple differences seem to depend on judgments of plausibility that do not rest on explanatory grounds: judgments of likeliness,

but not of loveliness. A biological analogy may clarify what is at issue. The Darwinian mechanism of variation and selection is also a short list mechanism. Natural selection does not operate on a full menu of possible variations, but only on the short list of actual variations that happen to occur. Here the process of generation is fundamentally different from the process of selection, so an account of the reasons only some types of individuals reproduce successfully would not seem able also to explain why only some types of individuals appear in a population in the first place. Similarly, the challenge to Inference to the Best Explanation is that the processes of generation and selection of hypotheses are fundamentally different, and that only the mechanism of selection depends on explanatory considerations.

I want to resist this restriction on the scope of Inference to the Best Explanation, by suggesting how explanatory considerations can play a role in the generation of potential explanations as well as in the subsequent selection from among them. Let us extend the biological analogy. Darwin's mechanism faces the anomaly of the development of complex organs. The probability of a new complex organ, such as a wing, emerging all at once as a result of random mutation, is vanishingly small. If only a part of the organ is generated, however, it will not perform its function, and so will not be retained. How, then, can a complex organ evolve? The solution is an appeal to 'preadaptation'. Complex organs arose from simpler structures, and these were retained because they performed a useful though perhaps different function. A wing could not have evolved all at once, and a half-wing would not enable the animal to fly, but it might have been retained because it enabled the animal to swim or crawl. It later mutated into a more complex structure with a new function. In one sense, then, mutations are not random. Some complex structures have a much higher probability of occurring than others, depending on whether they would build upon the simpler structures already present in the population. Mutations are not directed in the sense that they are likely to be beneficial but, since the complex organs that occur are determined both by the random process of genetic variation and the preadaptations already in place, only certain types of complex organs are likely to arise.

Preadaptations are themselves the result of natural selection, and they form an essential part of the mechanism by which complex organs are generated. So natural selection plays a role in both the generation and the selection of complex organs. Similarly, the mechanism of explanatory selection plays a role both in the generation of the short list of plausible causal candidates and in the selection from this list. The background beliefs that help to generate the list are themselves the result of explanatory inferences whose function it was to explain different evidence. (This is like the Bayesian point that today's priors are yesterday's posteriors.) We consider only the few potential explanations of what we observe that seem reasonably plausible, and the plausibility judgments may not seem to be

based on explanatory considerations; but they are, if the background beliefs that generate them are so based. Those beliefs now serve as heuristics that guide us to new inferences, by restricting the range of actual candidates, much as preadaptations limit the candidate organisms that are generated (Stein and Lipton 1989). So Inference to the Best Explanation helps to account for the generation of live candidates because it helps to account for the earlier inferences that guide this process. The analogy to the evolution of complex organs also serves in another way. By exploiting preadaptations, the Darwinian mechanism yields the result that old and new variations fit together into a coherent, complex organ. Similarly, the mechanism of generating hypotheses favors those that are extensions of explanations already accepted, and so leads towards a unified general explanatory scheme. This scheme is a complex organ that could not have been generated in one go, but it is built on earlier inferences, themselves selected on explanatory grounds and guiding the generation of additional structures.

In the evolution of complex organs, the process of building on available preadaptations mimics what would be the result of a mechanism where natural selection operated on a full array of ready made options, but only crudely. We find evidence of this difference in the imperfect efficiency of complex organs. We also find much greater retention of old structures, traces of old fins in new wings, than we would if the complex organ had been selected from a comprehensive set of possible organs with a given function. Similarly, we should expect that the mechanism of considering only a short list of candidate explanations will generally yield different inferences than would have been made, had every possibility been considered before selecting the best. For example, we should expect to find more old beliefs retained under the short list mechanism than there would be if we worked from a full menu of explanatory schemes, or from a random selection. Our method of generating candidate hypotheses is skewed so as to favor those that cohere with our background beliefs, and to disfavor those that, if accepted, would require us to reject much of the background. In this way, our background beliefs protect themselves, since they are more likely to be retained than they would be if we considered all the options. We simply tend not to consider hypotheses that would get them into trouble. The short list mechanism thus gives one explanation for our apparent policy of inferential conservatism (cf. Quine and Ullian 1978: 66–8; Harman 1986: 46).

Is the best good enough?

The two-stage picture of inference – first generation of a short list of hypotheses, then selection from that list – is plausible and explanationism has something to say about both stages. But the fact that we only generate a limited list of candidate hypotheses raises an interesting challenge to the claim that such a practice would be a reliable way of discovering the truth,

even if we grant ourselves considerable inductive powers. Van Fraassen has raised this challenge specifically in the context of Inference to the Best Explanation; as we will see, the challenge turns out to be largely independent of that context, but it is interesting for our purposes nevertheless. I will call his the argument from 'underconsideration'.

The argument has two premises. The *ranking* premise states that the testing of theories yields only a comparative warrant. Scientists can rank the competing theories they have generated with respect to likelihood of truth. The premise grants that this process is known to be highly reliable, so that the more probable theory is always ranked ahead of a less probable competitor and the truth, if it is among the theories generated, is likely to be ranked first, but the warrant remains comparative. In short, testing enables scientists to say which of the competing theories they have generated is likeliest to be correct, but does not itself reveal how likely the likeliest theory is. The second premise of the argument, the *no-privilege* premise, states that scientists have no reason to suppose that the process by which they generate theories for testing makes it likely that a true theory will be among those generated. It always remains possible that the truth lies rather among those theories nobody has considered, and there is no way of judging how likely this is. The conclusion of the argument is that, while the best of the generated theories may be true, scientists can never have good reason to believe this. They know which of the competing theories they have tested is likeliest to be true, but they have no way of judging the likelihood that any of those theories is true. On this view, to believe that the best available theory is true would be rather like believing that Jones will win the Olympics when all one knows is that he is the fastest miler in Britain.

The argument from underconsideration is clearly different from the radical Humean problem of induction. The upshot of Hume's argument is that all non-deductive evaluation is unjustifiable. By contrast, the argument from underconsideration concedes very substantial inductive powers, by granting scientists the ability to rank reliably whichever competing theories they generate. Indeed these powers are almost certainly stronger than any sensible scientific realist would wish to claim. This only seems to strengthen the underconsideration argument, however, since it appears to show that even these generous powers can not warrant belief in any scientific theory.

The argument from underconsideration is much more similar to an argument from underdetermination. According to one version of this argument, scientists are never entitled to believe a theory true because, however much supporting evidence that theory enjoys, there must exist competing theories, generated or not, that would be as well supported by the same evidence. This is an argument from inductive ties or stalemates. Like the argument from underconsideration, it is an intermediate form of skepticism, since it grants scientists considerable inductive powers, but the two arguments are not the same. The argument from underconsideration does

not exploit the existence of inductive ties, though it may allow them. On the other side, the argument from underdetermination does not assume any limitations on the scientists' powers of theory generation. Roughly speaking, whereas the underdetermination argument depends on the claim that scientists' inductive powers are excessively coarse-grained, the underconsideration argument focuses instead on the claim that they are only comparative. Moreover, the argument from underdetermination is in one sense more extreme than the argument from underconsideration. The underdetermination problem would remain even if scientists knew all the possible competing hypotheses and all possible data, whereas the underconsideration problem would disappear if they only knew all the competitors. Nevertheless, the similarities between the two arguments are substantial. Towards the end of this section I will suggest that some of the objections to the argument from underconsideration also threaten the argument from underdetermination.

Construing it as he does as an argument against Inference to the Best Explanation, in van Fraassen's hands the argument from underconsideration can be seen as an instance of Voltaire's objection. The argument would show that the two-stage process would make the reliability of Inference to the Best Explanation miraculous, since it would be miraculous if we consistently included a true hypothesis in the generation stage, given the no-privilege premise. But we need to clarify the connections between Inference to the Best Explanation, van Fraassen's constructive empiricism, and the argument from underconsideration, before turning to a critical assessment of the argument itself.

What then is the relationship between constructive empiricism and Inference to the Best Explanation? They are widely supposed to be incompatible. Certainly champions of Inference to the Best Explanation tend to be realists and van Fraassen develops his case against Inference to the Best Explanation as part of his argument for constructive empiricism. But the two views can be made compatible, by allowing a constructive empiricist version of Inference to the Best Explanation. We could do this by construing 'correct' as empirically adequate rather than as true and by allowing that false theories may explain. This is not the version of Inference to the Best Explanation I have been developing in this book, but at least van Fraassen's own accounts of inference and explanation allow it, and it would preserve the core explanationist idea that explanatory considerations are a guide to inference.

Is Inference to the Best Explanation especially vulnerable to the argument from underconsideration, more vulnerable than other accounts of inference? Van Fraassen's discussion gives this impression, since he deploys the argument specifically against explanationism. Moreover, Inference to the Best Explanation does seem particularly vulnerable, since it seems that 'best theory' can only mean 'best of those theories that have been generated'. But

the argument says nothing about using explanatory considerations as a guide to inference, so it will apply to any account of inference that admits two stages and is compatible with the no-privilege and ranking premises. Conversely, not every version of Inference to the Best Explanation will accept these premises. In particular, an explanationist can endorse a two-stage inferential process without maintaining that the explanatory virtues that operate in the selection phase are comparative rather than absolute. As I observed in chapter 4, Inference to the Best Explanation might be more accurately if less memorably called 'Inference to the Best Explanation if the Best is Sufficiently Good'.

Finally, what is the relationship between the argument from underconsideration and constructive empiricism? Once again, van Fraassen's discussion may give a false impression, since one might suppose that the argument forms part of his general case for favoring constructive empiricism over realism. Yet the argument seems clearly to work against the constructive empiricist too, if it works at all. The ranking premise is no less plausible for evaluation with respect to empirical adequacy than it is with respect to truth, and so far as I can tell, van Fraassen himself accepts it. Similarly, we have a constructive empiricist version of the no-privilege premise, to the effect that scientists have no reason to suppose that the means by which they generate theories for testing in itself makes it likely that an empirically adequate theory will be among those generated. Recalling that empirically adequate means adequate to everything observable and not just everything observed, this too will seem plausible to someone who endorses the realist version of the premise and, once again, van Fraassen appears to accept it. Constructive empiricism can itself be seen as being based in part on a kind of intermediate skepticism, to the effect that our inductive powers extend only to the limits of the observable, but this form of skepticism is orthogonal to the one articulated by the argument from underconsideration. The argument from underconsideration is thus especially salient neither as part of an argument for constructive empiricism nor as an argument against Inference to the Best Explanation. Nevertheless, it is of considerable interest, and worth considering here in its own right.

Several quick replies to the argument from underconsideration immediately suggest themselves. We may simply deny either or both of the premises. That is, we may insist either that scientists are capable of absolute and not only comparative evaluation or that their methods of theory generation do sometimes provide them with good reason to believe that the truth lies somewhere among the theories they have generated. These responses may well be correct but, baldly asserted, they lead to an unsatisfying standoff between those who believe in absolute evaluation or privilege and those who do not. Moreover, it seems undeniable that scientists' actual evaluative practises do include a strong comparative element, and one that is reflected in the most popular accounts of

confirmation. Examples of this include the use of 'crucial' experiments and the distribution of prior probabilities between the available hypotheses (Sklar 1985: 151–3).

Another obvious reply would be to concede some force to the skeptical argument but to deny that it undermines the rationality of science. As we have seen, the ranking assumption grants to the scientist considerable inductive powers. In particular, it allows that theory change is a truth-tropic process, so that later theories are always likelier to be correct than those they replace. Thus we might maintain that science is a progressive activity with respect to the aim of truth, even if scientists are never in a position rationally to assert that the best theory of the moment is actually true. (This view would be a kind of inductively boosted Popperianism.) More ambitiously, it might be argued that this truth-tropism even justifies scientific belief, by appealing to the scientist's desire to avoid ignorance as well as error. But the cost of these truth-tropic approaches is high, since there are various aspects of scientific activity that appear to require absolute evaluations. The most obvious of these is the practical application of science. In order to decide whether to administer a drug with known and serious side-effects, one needs to know how likely it is that the drug will effect a cure, not merely that it is likelier to do so than any other drug. Absolute evaluations also seem indispensable to 'pure' research, for example to the decision over whether it is better to develop the best available theory or to search for a better alternative.

The quick replies I have mentioned are not to be disdained, but they concede too much to the argument from underconsideration. The nub of the argument is the claim that there is an unbridgeable gap between comparative and absolute evaluation. This gap is, however, only a plausible illusion.

The most straightforward way to eliminate a gap between comparative and absolute evaluation would be by exhaustion. If the scientist could generate all possible competitors in the relevant domain, and he knew this, then he would know that the truth is among them. Given the reliability that the ranking premise grants, he would also know that the best of them is probably true. This brute force solution, however, seems hopeless, since it takes a wildly exaggerated view of the scientist's abilities. Even granting that we can make sense of the notion of all possible competitors, how could the scientists possibly generate them all?

But collapsing the distinction between relative and absolute evaluation does not require exhaustion. Scientists do not have to know that they have considered all the competitors, only that one of those they have considered must be true, and for this they need only a pair of contradictories, not the full set of contraries. It is enough that the scientists consider a theory and its negation, or the claim that a theory has a probability greater than one-half and the claim that it does not, or the claim that X is a cause of some phenomenon and the claim that it is not, or the claim that an entity or process

with specified properties exists or it does not. Since scientists are plainly capable of considering contradictories and the ranking premise entails that, when they do, they will be able to determine which is true, the argument from underconsideration fails.

The skeptic has two natural replies to this objection from contradictories. The first is to modify and restrict the ranking premise, so it concedes only the ability to rank contraries, not contradictories. But while the original ranking premise is epistemically over-generous, it is not clearly over-generous in this way. Scientists do, for example, compare the likelihood of the existence and non-existence of entities, causes and processes. So the skeptic would owe us some argument for denying that these comparisons yield reliable rankings while accepting the reliability of the comparisons of contraries. Moreover, it is not clear that the skeptic can even produce a coherent version of this restricted doctrine. The problem is that a pair of contraries entails a pair of contradictories. To give a trivial example, (P&Q) and not-P are contraries, but the first entails P, which is the contradictory of not-P. Indeed, all pairs of contraries entail a pair of contradictories, since one member of such a pair always entails the negation of the other. Suppose then that we wish to rank the contradictories T1 and not-T1. If we find a contrary to T1 (say T2) that is ranked ahead of T1, then not-T1 is ranked ahead of T1, since T2 entails not-T1. Alternatively, if we find a contrary to not-T1 (say T3) that is ranked ahead of not-T1, then T1 is ranked ahead of not-T1, since T3 entails T1. So it is not clear how to ban the ranking of contradictories while allowing the ranking of contraries.

The second natural reply the skeptic might make to the objection from contradictories would concede contradictory ranking. For in most cases, only one of a pair of contradictories would mark a significant scientific discovery. Not to put too fine a point on it, usually one member of the pair will be interesting, the other boring. Thus if the pair consists of the claim that all planets move in ellipses and the claim that some do not, only the former claim is interesting. Consequently, the skeptic may concede contradictory ranking but maintain that the result will almost always be that the boring hypothesis is ranked above the interesting one. In short, he will claim that the best theory is almost always boring, so the scientist will almost never be in a position rationally to believe an interesting theory.

This concession substantially changes the character of the argument from underconsideration, however, and it is a change for the worse. Like most important skeptical arguments, what made the original argument from underconsideration interesting was the idea that it might rule out reasons for belief, even in cases where the belief is in fact true. (Compare Hume's general argument against induction: he does not argue that the future will not resemble the past, but that, even if it will, this is unknowable.) With the concession, however, the argument from underconsideration reduces to the claim that scientists are unlikely to think of the truth. The idea that scientists

are only capable of relative evaluation no longer plays any role in the argument, since ranking of contradictory theories has collapsed the distinction between relative and absolute evaluation, and the argument reduces to the observation that scientists are unlikely to think of interesting truths, since they are hidden behind so many interesting falsehoods.

So the revised argument is substantially less interesting than the original. But the situation is worse than this. For scientists do in fact often rank interesting claims ahead of their boring contradictories. The revised argument thus faces a dilemma. If it continues to grant that scientists are reliable rankers, then the fact that interesting claims often come out ahead refutes the claim that scientists do not generate interesting truths. If, on the other hand, reliable ranking is now denied, we have lost all sense of the original strategy of showing how even granting scientists substantial inductive powers would be insufficient for rational belief.

The argument from underconsideration depends on a gap between relative and absolute evaluation. I have suggested that contradictory ranking closes that gap and that the argument cannot be modified to reopen it without substantial loss of interest or force. What I will argue now is that the original argument is fundamentally flawed even if we restrict our attention to the ranking of contraries. Given an uncontroversial feature of the way scientists rank theories, the two premises of the argument from underconsideration are incompatible. That feature, which we have already emphasized in earlier chapters, consists of background beliefs, especially background theories. These are claims already accepted, if only tentatively, at the time when a new theory is tested. They influence the scientists' understanding of the instruments they use in their tests, the way the data themselves are to be characterized, the prior plausibility of the theory under test, and bearing of the data on the theory. (The importance of background theories and their bearing on realism have been emphasized by Richard Boyd in many articles (e.g. 1985).)

Scientists rank new theories with the help of background theories. According to the ranking premise of the argument from underconsideration, this ranking is highly reliable. For this to be the case, however, it is not enough that the scientists have any old background theories on the books with which to make the evaluation: these theories must be *probably true*, or at least probably approximately true. If most of the background theories were not even approximately true, they would skew the ranking, leading in some cases to placing an improbable theory ahead of a probable competitor, and perhaps leading generally to true theories, when generated, being ranked below falsehoods. The ranking premise would be violated. So the ranking premise entails that the background is probably (approximately) true. The problem for the argument from underconsideration then appears on iteration. These background theories are themselves the result of prior generation and ranking, and the best of the theories now being ranked will form part of

tomorrow's background. Hence, if scientists are highly reliable rankers, as the ranking premise asserts, the highest ranked theories have to be absolutely probable, not just more probable than the competition. This is only possible if the truth tends to lie among the candidate theories the scientists generate, which contradicts the no-privilege premise. Hence, if the ranking premise is true, the no-privilege assumption must be false, and the argument from underconsideration self-destructs.

Given the role of background in theory evaluation, the truth of the ranking premise entails the falsity of the no-privilege premise. Moreover, since the ranking premise allows not only that scientists are reliable rankers, but also that they know this, the situation is even worse. If a scientist knows that her method of ranking is reliable, then she is also in a position to know that her background is probably true, which entails that she is capable of absolute evaluation. Thus the knowledge that she is capable of comparative evaluation enables the scientist to know that she is capable of absolute evaluation too, and the claim of the ranking premise that the scientist knows that she is only capable of reliable comparative evaluation must be false.

So the initially plausible idea that scientists might be completely reliable rankers yet arbitrarily far from the truth is an illusion. Might the skeptic salvage his case by weakening the ranking premise, as he was tempted to do in response to the objection from contradictories? I do not think this will help. Of course, if ranking were completely unreliable, the skeptic would have his conclusion, but this just takes us back to Hume. The point of the argument from underconsideration was rather to show that the skeptical conclusion follows even if we grant scientists considerable inductive powers. So the skeptic needs to argue that, if scientists were moderately but not completely reliable rankers, the connection between the best theory and the truth would be severed. Our skeptic has not, however, provided us with such an argument, and there is good reason to believe that no sound argument of this sort exists. For the level of reliability seems to depend, not just on the degree of reliability of the prior ranking of background theories, but on their verisimilitude, their approximation to the truth.

To see this, suppose that reliability did depend only on the reliability of the prior ranking process by which the background theories were selected. Consider now two isolated scientific communities that are equally reliable rankers, but who in the past generated quite different ranges of candidate theories and so come to have quite different backgrounds. One community was lucky enough to generate true theories, while the other was uninspired enough to generate only wildly erroneous ones. If present reliability depended only on prior ranking, we would have to suppose that these two communities are now equally reliable rankers of new theories, which is clearly incorrect. The general point is that the level of reliability a background confers depends on its *content*, not just on the method by which it was generated, and that what matters about the content is, among other

things, how close it is to the truth. Consequently, even though scientists are in fact only moderately reliable rankers, this does not sever the connection between relative and absolute evaluation. Even moderately reliable ranking is not compatible with the claim that scientists' methods may leave them with theories that are arbitrarily far from the truth. In other words, even moderately reliable ranking requires moderate privilege.

The moral of the story is that certain kinds of intermediate skepticism, of which the argument from underconsideration is one example, are incoherent. Because of the role of background beliefs in theory evaluation, what we cannot have are inductive powers without inductive achievements. At the beginning of this discussion, I distinguished the argument from underconsideration from the better known argument from underdetermination. Having seen what is wrong with the former, however, it appears that a similar objection applies to the latter, and I want now briefly to suggest why this may be so.

The central claim of the argument from underdetermination is sometimes expressed by saying that, however much evidence is available, there will always be many theories that are incompatible with each other but compatible with the evidence. This version of underdetermination, however, ought not to bother the realist, since it amounts only to the truism that the connection between data and theory is and always will be inductive. Like the argument from underconsideration, an interesting version of the underdetermination argument is an intermediate skepticism which attempts to show that rational belief is impossible even granting the scientist considerable inductive powers. Such a version of the underdetermination argument is an argument from inductive stalemates. The central claim is that, although some theories are better supported by the evidence than others, for any theory there must exist a competitor (which scientists may not have generated) that is equally well supported, and this situation remains however much evidence the scientist has. The argument thus allows that scientists are reliable rankers, but insists that the ranking will not discriminate between every pair of competing theories. In particular, it is claimed that this 'coarse' ranking is such that, however much evidence a scientist has, there exist competitors to the highest ranked theory which, if they were considered, would do just as well. Consequently, even if one of the theories the scientist has actually generated is ranked ahead of all the others, he has no reason to believe that this one is true, since he has only avoided a stalemate through lack of imagination.

Coarse ranking is not quite the same as moderately reliable ranking; the difference is roughly that between a degree of ignorance and a degree of error. Nevertheless, the objection from the background seems to apply here too. Even coarse ranking requires that most of the background theories be close to the truth. If they were not, we would have more than a failure of discrimination; we would have misranking. In other words, even if the

underdeterminationist is correct in claiming that there will always in principle be ties for the best theory, this does not support the conclusion that the theories we accept may nonetheless be arbitrarily far from the truth. To get that conclusion would require abandoning the concession that coarse ranking is reliable, as far as it goes, and we are back to an undiscriminating Humean skepticism about non-demonstrative inference.

The underdeterminationist might respond to the objection from the background by going global. He could take the unit of evaluation to be the full set of candidate beliefs a scientist might endorse at one time, rather than a particular theory. The point then would be that there are always stalemates for the best total set of beliefs. By effectively moving the background into the foreground, the objection from the background appears to be blocked, since what is evaluated now always includes the background and so cannot be relative to it. At the same time, the argument appears able to grant the scientist considerable inductive powers, since it can allow that not all consistent sets are equally likely or equally ranked, and that the higher ranked sets are more likely to be correct than those ranked below them.

I do not think this response is successful. One difficulty is that the global version of the underdetermination argument does not respect the fact that scientists' actual methods of evaluation are local and relative to a (revisable) background. Consequently, although the argument makes a show of granting scientists some sort of inductive powers, it does not grant reliability to the methods scientists actually employ. The reliability of the actual practice of local ranking relative to background cannot be accommodated within this global version without ruining the argument since, as we have seen, local reliability requires that the background be approximately true, the consequence the underdeterminationist is trying to avoid.

A further and related difficulty with the global argument is that it appears tacitly to rely on an untenable distinction between methodological principles and substantive belief. The argument suggests a picture in which the principles of evaluation float above the varying global sets of candidate beliefs, permitting a common scheme of ranking to be applied to them all. Since beliefs about inductive support (such as what is evidence for what) are themselves part of the scientists' total set of beliefs, however, this picture is untenable. What are we to put in its place? If we could say that all the sets shared the same principles, this would perhaps suffice for the argument, but we cannot say this. The problem is not simply that these principles will in fact vary, but that the very notion of a division of the elements of a global set into those that are methodological principles and those that are substantive beliefs is suspect.

There are two reasons for this suspicion. Notice first that, unlike the principles of deductive inference, reliable principles of induction are contingent. (This is the source of the Humean problem.) A pattern of non-demonstrative inference that generally takes us from truth to truth in this

world would not do so in some other possible worlds. Moreover, although this is perhaps somewhat more controversial, the principles also appear to be a posteriori. Given all this, it is difficult to see why they are not tantamount to substantive claims about our world. A second reason for treating the distinction between principle and belief with suspicion pushes from the other side and appeals to the main theme of this discussion of the argument from underconsideration: the role of the background in evaluation. Given this role, it is unclear on what basis one is to deny that the substantive theories in a global set are themselves also principles of evaluation.

The intermixture of methodological principle and substantive belief, in part a consequence of the essential role of background belief in theory evaluation, makes it unclear how even to formulate the global argument, and in what sense the argument grants the scientist reliable inductive powers. The intermixture of principle and belief is also perhaps the root cause of the failure of the two forms of intermediate skepticism I have been considering: it explains why it proves so difficult to grant the reliability of evaluation without also admitting the correctness of theory.

'Of course! Why didn't I think of that!' The distinction between being able to generate the right answer and seeing that an answer is correct once someone else has proposed it is depressingly familiar. Meno's slave-boy (or the reader of the dialogue) might never have thought of doubling the square by building on its diagonal, but he has no trouble seeing that the answer must be correct, once Socrates suggests it. And it is apparently no great leap from this truism that there is a distinction between generation and evaluation, between the context of discovery and the context of justification, to the thought that powers of evaluation are quite distinct from powers of generation, that we might be good at evaluating the answers we generate, yet bad at generating correct answers. Hence the thought that scientists might be reliable rankers of the conjectures they generate, yet hopeless at generating conjectures that are true, or close to the truth. Yet this thought turns out to be mistaken, falling to the elementary observation that the scientists' methods of evaluation work relative to a set of background beliefs and that these methods can not be even moderately reliable unless that background is close to the truth. Hence the failure of the argument from underconsideration and of at least some versions of the argument from underdetermination. Of course, in particular cases scientists fail to generate answers that are even approximately correct, but the idea that they might always so fail even though their methods of evaluation are reliable is incoherent. Scientists who did not regularly generate approximately true theories could not be reliable rankers.

What is the bearing of these considerations on scientific realism, on the idea that science is in the truth business? Both the arguments from underconsideration and from underdetermination threaten the view that scientists may have rational grounds for believing that a theory is at least

approximately true; insofar as these arguments have been turned, the realist who believes in the existence of such grounds will be comforted. It is important, however, to emphasize what has not been shown. I have argued against certain intermediate skepticisms, but have suggested no answer here to wholesale inductive skepticism. Moreover, I have not tried to show that all intermediate arguments are untenable. In particular, I have not in this discussion of the argument from underconsideration criticized van Fraassen's own intermediate position, which depends in part on the claim that scientists' inductive powers extend only to statements about the observable. As we have seen, on this view what the scientist is entitled to believe is not that theories are true, but only that they are empirically adequate, that their observable consequences are true. The objection from the background would gain purchase here if it could be shown that, in order for scientists reliably to judge the empirical adequacy of their theories, their background theories must themselves be true, not just empirically adequate. I suspect that this is the case, but I have not tried to show that here.

The role of the background in theory evaluation is something of a two-edged sword. It defeats some skeptical arguments, but it also shows both that the realist must take care not to exaggerate scientists' inductive powers and how much even a modest realism entails. Even the most fervent realist cannot afford to claim that scientists are completely reliable rankers since this would require that all their background beliefs be true, a hopelessly optimistic view and one that is incompatible with the way the scientific background changes over time. The objection from the background drives home the point that realists must also be thoroughgoing fallibilists, allowing the possibility of error not just about theory and the data which support it, but also about the assessment of support itself. The argument of this section also shows that the realist cannot maintain that scientists are good at evaluation while remaining agnostic about their ability to generate true theories. Reliable evaluation entails privilege, so the realist must say that scientists do have the knack of thinking of the truth. This ability is, from a certain point of view, somewhat surprising, but it remains in my view far more plausible than the extreme ignorance, substantive and methodological, that a coherent critic must embrace.

But perhaps the idea of privilege should not appear so surprising after all. From the point of view of a tradition in the philosophy of science that treats the generation of hypotheses as a mysterious process prior to the rational mechanisms of testing, a talent for generating true theories would be mysterious, but that picture of the context of discovery is untenable, and one that I have already firmly rejected. Theory generation is highly constrained by background, and insofar as the background approximates the truth, we should not be so surprised that our powers of generating true theories are substantially better than guesswork. Moreover, it might be argued that the striking successes of our best scientific theories actually provide empirical

support for privilege; after all, success is much more likely with privilege than without it. This line of thought is closely related to the so-called miracle argument for realism, an application of Inference to the Best Explanation that we will investigate in chapter 11. It is also related to the issue of the putative special probative force of successful predictions, as opposed to data that a theory is generated in order to accommodate, our subject in the next chapter.

Voltaire's objection was that Inference to the Best Explanation would make the reliability of induction peculiarly problematic. The argument from underconsideration can be seen as a version of that objection. And my dominant strategy in answering the objection in its various forms in this chapter has been to argue that Inference to the Best Explanation is in no worse shape from this normative point of view than other accounts of induction. That strategy applies also to the argument from underconsideration since, as we have seen, the argument would apply not just to Inference to the Best Explanation but to pretty much any account of induction that appeals to a plausible two-stage generation/selection process. But here I have also argued for an instability in an intermediate conception of our inductive powers, an instability which if we reject wholesale inductive skepticism will push us to account for the process of theory generation as an essential and rationally constrained component of our inductive practices. It is thus to the credit of Inference to the Best Explanation that it is applicable to the process by which hypotheses are generated, as well as to the process by which they are selected.

Chapter 10

Prediction and prejudice

The puzzle

The main purpose of the model of Inference to the Best Explanation is to provide a partial solution to the problem of description by giving an illuminating account of the black box mechanism that governs our inductive practices. As we have seen, however, this account also bears on problems of justification. On the one hand, there are questions about whether an explanationist inferential practice is justifiable as a way of discovering truths; these are questions we considered in the last chapter. On the other hand, there are questions about the use of inferences to the best explanation to provide justifications, not just in science and in ordinary life, but even in philosophy. In this chapter and the next, we will consider two such philosophical inferences to the best explanation. One is perhaps the best known argument of this form, the argument for scientific realism on the grounds that the approximate truth of predictively successful scientific theories is the best explanation of that predictive success. The other, the subject of this chapter, is the related but narrower argument that a theory deserves more inductive credit when data are predicted than when they are accommodated, on the grounds that only in the predictive case is the correctness of the theory the best explanation of the fit between theory and evidence.

In a case of *accommodation*, the scientist constructs a theory to fit the available evidence. If you are by now inclined to accept Inference to the Best Explanation as an account of scientific inference, read 'fit' as 'explain'. My discussion in this chapter, however, will not depend on this. If you are so recalcitrant as to continue to prefer some other account of the basic relation of inductive support, you may substitute that notion. For example, if you are a hypothetico-deductivist, a case of accommodation is one where a theory is constructed so as to ensure that it, along with the normal complement of auxiliary statements, entails the evidence. In particular, my discussion about the putative advantage of prediction over accommodation will not rest on the claim that there is any deductive difference between the two cases, say that only predictions require deduction. To simplify the exposition, then, let us

suppose that the explanations in question are deductive ones; the difference between explanatory and non-explanatory deductions will not affect the course of the argument. In a case of successful *prediction*, the theory is constructed and, with the help of auxiliaries, an observable claim is deduced but, unlike a case of accommodation, this takes place before there is any independent reason to believe the claim is true. (Because of this last clause, my notion of prediction is narrower than the ordinary one, closer perhaps to the idea of novel prediction.) The claim is then independently verified. Successful theories typically both accommodate and predict. Most people, however, are more impressed by predictions than by accommodations. When Mendeleyev produced a theory of the periodic table that accounted for all sixty known elements, the scientific community was only mildly impressed. When he went on to use his theory to predict the existence of two unknown elements that were then independently detected, the Royal Society awarded him its Davy Medal (Maher 1988: 274–5). Sixty accommodations paled next to two predictions.

Not all predictions provide as much inductive support or confirmation as Mendeleyev's, and some accommodations give stronger support than some predictions. Is there, however, any general if defeasible advantage of prediction over accommodation? We may usefully distinguish two versions of the claim that there is such an advantage. According to the 'weak advantage' thesis, predictions tend to provide more support than accommodations, because either the theory or the data tend to be different in cases of prediction than they are in cases of accommodation. For example, it might be that the predictions a scientist chooses to make are those that would, if true, provide particularly strong support for her theory. According to the 'strong advantage' thesis, by contrast, a successful prediction tends to provide more reason to believe a theory than the *same* datum would have provided for the *same* theory, if that datum had been accommodated instead. This is the idea that Mendeleyev's theory would have deserved less confidence had he instead accommodated all sixty-two elements. It seems to be widely believed, by scientists and ordinary people, if not by philosophers, that even the strong advantage thesis is correct. The purpose of this chapter is to determine whether this belief is rational.

However pronounced one's intuition that prediction has a special value over accommodation, it is surprisingly difficult to show how this is possible; so difficult that a number of philosophers have claimed that at least the strong thesis is false (e.g. Horwich 1982: 108–17; Schlesinger 1987). The content of theory, auxiliary statements, background beliefs and evidence, and the logical and explanatory relations among them, are all unaffected by the question of whether the evidence was accommodated or predicted, and these seem to be the only factors that can affect the degree to which a theory is supported by evidence. The difference between accommodation and prediction appears to be merely temporal or historical in a way that cannot

affect inductive support. Moreover, the view that prediction is better than accommodation has the strange consequence that ignorance may be an advantage: the person who happens not yet to know about a certain datum would come to have more reason to believe her theory than someone who knew the datum all along. The problem is not merely that it seems impossible to analyze a preference for prediction in terms of more fundamental and uncontroversial principles of inductive support, but that the preference seems to conflict with those principles.

We can make the case against the strong thesis more vivid by considering a fictitious case of twin scientists. These twins independently and coincidentally construct the same theory. The only difference between them is that one twin accommodates a datum that the other predicts. If there really were an epistemic advantage to prediction, we ought to say that the predictor has more reason to believe the theory than the accommodator, though they share theory, data and background beliefs. This is counterintuitive, but things get worse. Suppose that the twins meet and discover their situation. It seems clear that they should leave the meeting with a common degree of confidence in the theory they share. If they came to the meeting with different degrees of rational confidence in their theory, at least one of them ought to leave with a revised view. But what level should they settle on: the higher one of the predictor, the lower one of the accommodator, or somewhere in between? There seems no way to answer the question. Moreover, if there is a difference between prediction and accommodation, then the twin who should revise her view when she actually meets her sibling must not revise simply because she knows that someone like her twin might have existed. If revision were in order merely because of this possibility, then the difference between accommodation and prediction would vanish. Whenever data are accommodated, we know that there might have been someone who produced the theory earlier and predicted the data instead. But how can the question of whether there actually is such a person make any difference to our justified confidence in the theory? Any adequate defense of the putative difference between prediction and accommodation will have to explain how an actual meeting could be epistemically different from a hypothetical meeting. Those who reject the distinction seem on firm ground when they maintain that no such explanation is possible.

The case against the strong thesis seems compelling. Nevertheless, I will argue that the strong thesis is correct, that evidence usually provides more reason to believe a theory when it is predicted than it would if the evidence were accommodated. My argument will hinge on a distinction between the objective but imperfectly known credibility of a theory and the actual epistemic situation of the working scientist. Before I give my own argument, however, I will canvass some other plausible defenses of the strong thesis. We will find that none of them are acceptable as they stand but, in light of my own solution, we will be in a position to see the germs of truth they contain.

One of the attractive features of the puzzle of prediction and accommodation is that it requires no special philosophical training to appreciate. People are quick to catch on and to propose solutions. Some bite the bullet and deny that there is any difference. In my experience, however, most believe that prediction is better than accommodation, and give a number of different reasons. Three closely related ones are particularly popular. First, a theory built around the data is ad hoc and therefore only poorly supported by the evidence it accommodates. Second, only evidence that tests a theory can strongly support it, and accommodated data cannot test a theory, since a test is something that could be failed. Third, the truth of the theory may be the best explanation of its predictive success, but the best explanation of accommodation is instead that the theory was built to accommodate; since we accept the principle of Inference to the Best Explanation, we then ought only to be impressed by prediction.

None of these reasons for preferring prediction to accommodation are good reasons as they stand. Accommodating theories are obviously ad hoc in one sense, since 'ad hoc' can just mean purpose-built, and that is just what accommodating theories are. To claim, however, that a theory that is ad hoc in this sense is therefore poorly supported begs the question. Alternatively, 'ad hoc' can mean poorly supported, but this is no help, since the question is precisely why we should believe that accommodating theories are in this sense ad hoc. To assume that accommodating theories are ad hoc in the sense of poorly supported is to commit what might be called the '*post hoc ergo ad hoc*' fallacy. The simple appeal to the notion of an ad hoc theory names the problem but does not solve it.

The second common reason given for preferring prediction is that only predictions can test a theory, since only predictions can fail. We are impressed by an archer who hits the bullseye, not by the archer who hits the side of a barn and then draws a bullseye around his arrow (Nozick 1983: 109). But this argument confuses the theory and the theorist. A theory will not be refuted by evidence it accommodates, but that theory would have been refuted if the evidence had been different. Similarly, a theory will not be refuted by later evidence that it correctly predicts, though it would have been refuted if that evidence had been different. The real difference is that, although the predicting scientist might have produced her theory even if its prediction were to be false, the accommodating scientist would not have proposed *that* theory if the accommodated evidence had been different. It is only in the case of prediction that the scientist runs the risk of looking foolish. This, however, is a commentary on the scientist, not on the theory. In the case of archery, we want the bullseye drawn before the volley, but this is because we are testing the archer's skill. When we give students a test, we do not first distribute the answers, because this would not show what the students had learned. In science, however, we are supposed to be evaluating a theory, not the scientist who proposed it.

Finally we come to the appeal to an overarching inference to the best explanation. In the case of accommodation, there are two explanations for the fit between theory and data. One is that the theory is (approximately) true; another is that the theory was designed to fit the data. We know that the second, the accommodation explanation, is correct and this seems to pre-empt the inference to the truth explanation. In the case of prediction, by contrast, we know that the accommodation explanation is false, which leaves the truth explanation in the running. In one case, the truth explanation can not be the best explanation, while in the other it might be. This is why prediction is better than accommodation. This account has considerable intuitive force, but also a number of weaknesses. The most important of these is that it is unclear whether the accommodation explanation actually does pre-empt the truth explanation (Horwich 1982: 111–16). They certainly could both be correct, and to assume that accepting the accommodation explanation makes it less likely that the theory is true is once again to beg the question against accommodation. The issue is precisely whether the fact that a theory was designed to fit the data in any way weakens the inference from the fit to the correctness of the theory.

The fudging explanation

I turn now to my own argument that prediction is better than accommodation. Notice first that, because of the 'short list' mechanism we employ, which I discussed in the last chapter, there is some tendency for a theory that makes a successful prediction to be better supported overall than a theory generated to accommodate all the same data the first theory did, plus the datum the first theory predicted. The predicting theory must have had enough going for it to make it to the short list without the predicted datum, while the accommodating theory might not have made it without that datum. If this is so, then the predicting theory is better supported overall, even if we assume that the specific epistemic contribution of the datum is the same in both cases. But for just this reason, the short list consideration does not itself argue for even the weak advantage thesis, since it does not show any difference in the support provided by prediction or accommodation itself, but at most only that the predicting theory will tend to have stronger support from other sources.

Let us consider now an argument for the weak advantage thesis. It is uncontroversial that some data support a theory more strongly than others. For example, heterogeneous evidence provides more support than the same amount of very similar evidence: variety is an epistemic virtue. Again, evidence that discriminates between competing theories is more valuable than evidence that is compatible with all of them. Similarly, the same set of data may support one theory compatible with the set more strongly than another since, for example, one theory may be simpler or in some other way

provide a better explanation than the other. And there are many other factors that affect the amount of support evidence provides. Because of this, it is not difficult to find certain advantages in the process of prediction.

A scientist can choose which predictions to make and test in a way that she cannot choose the data she must accommodate. She has the freedom to choose predictions that will, if correct, provide particularly strong support for her theory, say because they will increase the variety of supporting evidence for the theory, or will both support the theory and disconfirm a competitor. This freedom is an advantage of prediction. Of course just because scientists can choose severe tests does not mean that they will, but since scientists want to convince their peers, they have reason to prefer predictions that will, if correct, provide particularly strong support. Another advantage of prediction concerns experimental design. Much experimental work consists of using what are in effect Mill's eliminative methods of induction to show that, when an effect occurs as a theory says it should, what the theory claims to be the cause of the effect is indeed the cause. The sorts of controls this requires will depend on what cause the theory postulates. As a consequence, data that were gathered before the theory was proposed are less likely to have proper controls than data gathered in light of the theory's predictions, and so less likely to give strong support to the theory. Our discussion of Semmelweis in chapter 5 is a good example of this. Semmelweis was lucky enough to start with suggestive contrastive data, in the form of the difference in mortality due to childbed fever in the two maternity divisions of the hospital in which he worked. But he was able to get much better data after he proposed his various hypotheses, by designing experiments with careful controls that helped to discriminate between them. Theories thus provide guides to the sort of data that would best support them, and this accounts for a general advantage that predicted data tend to have over accommodated data.

So the weak advantage thesis is acceptable. Even if the actual support that a particular datum gives to a theory is unaffected by the time at which the datum is discovered, the data a scientist predicts tend to provide more support for her theory than the data she accommodates, because she can choose her predictions with an eye to strong support, and because she can subject her predictions to the sorts of experimental controls that yield strong support. My main quarry in this chapter, however, is the strong advantage thesis, and here considerations of choice and control do not help much. The freedom the scientist has to choose predictions that, if true, will provide strong support, does nothing to show that these predictions provide more support than the very same data would have, had they been accommodated. Similarly, the extra controls that the scientist can impose in cases of prediction should not be taken to underwrite the strong thesis, since an experiment without these controls is a different experiment, and so we ought to count the resulting data different as well. To defend the strong thesis, we need a new argument.

When data need to be accommodated, there is a motive to force a theory and auxiliaries to make the accommodation. The scientist knows the answer she must get, and she does whatever it takes to get it. The result may be an unnatural choice or modification of the theory and auxiliaries that results in a relatively poor explanation and so weak support, a choice she might not have made if she did not already know the answer she ought to get. In the case of prediction, by contrast, there is no motive for fudging, since the scientist does not know the right answer in advance. She will instead make her prediction on the basis of the most natural and most explanatory theory and auxiliaries she can produce. As a result, if the prediction turns out to have been correct, it provides stronger reason to believe the theory that generated it. So there is reason to suspect accommodations that do not apply to predictions, and this makes predictions better.

What I propose, then, is a 'fudging explanation' for the advantage of prediction over accommodation. It depends not so much on a special virtue of prediction as on a special liability of accommodation. Scientists are aware of the dangers of fudging and the weak support that results when it occurs. Consequently, when they have some reason to believe that fudging has occurred, they have some reason to believe that the support is weak. My claim is that they have such a reason when they know that the evidence was accommodated, a reason that does not apply in cases of prediction. It must be emphasized that this does not show that all accommodations have less probative force than any prediction. Similarly, it does not show that all accommodations are fudged. What it does show, or so I shall argue, is that the fact that a datum was accommodated counts against it in a way that would not apply if that same datum were predicted. The fudging explanation thus supports the strong advantage thesis.

An analogy may clarify the structure of the fudging explanation. Consider a crossword puzzle. Suppose that you are trying to find a word in a position where some of the intersecting words are already in place. There are two ways you can proceed. Having read the clue, you can look at the letters already in place and use them as a guide to the correct answer. Alternatively, you can think up an answer to the clue with the requisite number of letters, and only then check whether it is consistent with the intersecting letters. The first strategy corresponds to accommodation, the second to prediction. The first strategy is, I think, the more common one, especially with difficult puzzles or novice players. It is often only by looking at the letters that are already in place that you can generate any plausible answer at all. If, however, you are fortunate enough to come up with a word of the right length without using the intersections, and those letters are then found to match, one might hold that this matching provides more reason to believe that the word is correct than there would be if you had adopted the accommodating strategy for the word. If that is the case, why is it so? The only plausible explanation is that, when you accommodate, you have some

reason to believe that the constraints that the intersections supply may pull you away from the best answer to the clue. Of course, assuming that the letters in place are correct, they do provide valuable information about the correct answer. But you have this information under both strategies, by the time you write in your word. It is only in the case of accommodation that the intersecting letters could possibly pull you away from the best answer to the clue, since it is only in this case that these letters are used in the process of generating the answer.

One might, of course, reject my suggestion that the two crossword strategies differ in the amount of support they provide. One might claim that there could be only a single word of the right length that both matches the intersections and fits the clue, or that one can simply see with either strategy how well a word fits the clue, so one need not take the difference in strategies into account. I do not think this is the case for all crossword puzzles, but nothing I will have to say about fudging in science depends upon this. The point of the crossword analogy is not to convince you that the fudging explanation is correct, but simply to illuminate its structure. As I hope will become clear, the case for saying that the motives for fudging support the strong thesis is much stronger in science than it is in crossword puzzles, because science offers so much more scope for fudging and because it is harder to detect. So let us return to science.

We can begin to flesh out the fudging explanation by distinguishing theory fudging and auxiliary fudging. Just as a theory may be more strongly supported by some data than by others, different theories may, as we have already noted, receive different degrees of support from the same data. This follows from the fact that there are always, in principle, many theories that fit the same data. If our inductive principles did not distinguish between these theories, the principles would never yield a determinate inference. So some theories must be more strongly confirmable than others, by a common pool of data that provides some support to all of them. When a theory is fudged, the result may be a theory that is only weakly confirmable. With special clauses and epicycles to handle particular accommodations, the theory becomes more like an arbitrary conjunction, less like a unified explanation. The result is that data that may support parts of the theory do not transfer support to other parts and the theory as a whole is only weakly supported. Another factor that affects the credibility of a theory is its relation to other non-observational claims the scientist already accepts. For example, as we discussed in chapter 8, a theory that is compatible with most of these background beliefs or, even better, coheres with them to produce a unified explanatory scheme is more credible than a theory that contradicts many of them or that depends on idiosyncratic causal mechanisms. The need for accommodation may force the scientist to construct a theory that fits poorly into the background and is thus harder to support. Prediction does not supply this motive to abjure the theory that fits best into the background.

When data that need to be accommodated threaten to force the scientist to construct a fudged theory, he can avoid this by tampering instead with the auxiliaries, but this too may result in weakened support. The theory itself may be elegant, highly confirmable in principle, and fit well into the background, but the accommodated data may only be loosely connected to it. This occurs because the scientist is forced to rely on auxiliaries that have little independent support, or approximations and idealizations that are chosen, not because they only ignore effects that are known to be relatively unimportant, but because they allow the scientist to get what he knows to be the right answer.

In short, in theory fudging the accommodated evidence is purchased at the cost of theoretical virtues; in auxiliary fudging it is at the cost of epistemic relevance. In both cases, the result is an inferior explanation. The fudging explanation suggests that the advantage of prediction over accommodation ought to be greatest where there is the greatest scope for fudging, of either sort, so we can provide some defense for the account by noting that this corresponds to our actual judgments. For example, we should expect the difference between prediction and accommodation to be greatest for complex and high level theories that require an extensive set of auxiliaries, and to decrease or disappear for simple empirical generalizations since, the more auxiliaries, the more room for auxiliary fudging. I think that this is just what we do find. We are more impressed by the fact that the special theory of relativity was used to predict the shift in the perihelion of Mercury than we would have been if we knew that the theory was constructed in order to account for that effect. But when it comes to the low level generalization that all sparrows employ a distinctive courtship song, we are largely indifferent to whether the observation that the song is employed by sparrows in some remote location was made before or after the generalization was proposed. Similarly, among high level theories, we ought to judge there to be a greater difference between prediction and accommodation for theories that are vague or loosely formulated than we do for theories with a tight and simple mathematical structure, since vaguer theories provide more scope for theory fudging. Here again, this seems to be what we do judge. The fudging explanation also correctly suggests that the difference between accommodation and prediction ought to be roughly proportional to the number of possible explanations of the data we think there are. In a case where we convince ourselves that there is really only one possible explanation for the data that is, given our background beliefs, even remotely plausible (but not where we have simply only come up with one such explanation), fudging is not an issue and accommodation is no disadvantage. (I owe this point to Colin Howson.) And if the scientist is ever justifiably certain that she would have produced the same theoretical system even if she did not know about the evidence she accommodated, that evidence provides as much reason for belief as it would have, had it been predicted.

The fudging explanation gains some additional support from the way it handles two borderline cases. Consider first a case where a scientist has good independent reason to believe that some future event will occur, and ensures that the theory he constructs will entail it. It seems fair to say that those who admit a difference between prediction and accommodation would place this on a par with accommodation, even though it is a deduction of a future event. This is also entailed by the fudging explanation. The second case is the converse of this, an old datum, but one not known to the scientist when she constructs her theory and deduces that datum. This should be tantamount to a prediction, as the fudging explanation suggests.

The fudging explanation is related to Karl Popper's requirement that scientific theories must be incompatible with some possible observation (Popper 1959). The falsifiability requirement does not directly discriminate between accommodation and prediction since, from a logical point of view, a theory that has only accommodated may be as falsifiable as one that has made successful predictions. This is why it was wrong to say simply that predictions are better than accommodations because only predictions are tests. Moreover, the difference in inductive support I am seeking between prediction and accommodation is precluded by Popper's deductivist philosophy, since he abjures the notion of inductive support altogether. Nevertheless, some sort of falsifiability is an important theoretical virtue, and there is sometimes the suspicion that a theory that only accommodates does not have it. The theoretical system may be so vague and elastic that it can be fudged to accommodate any observation. By contrast, a system that can be used to make a prediction will often be tight enough to be falsifiable in the broad sense that the theory, along with independently plausible auxiliaries, is incompatible with possible observations. An astrologer who explains many past events does not impress us, but one who consistently made accurate predictions would give us pause. At the same time, the fudging explanation has considerably broader application than Popper's requirement. It also accounts for the difference between accommodation and prediction in the more interesting cases where the theories meet a reasonable version of the falsifiability requirement. Even if we are convinced that the accommodating theory is falsifiable, the suspicion remains that there was some fudging that weakened the data's inductive support, a suspicion that does not arise in the case of prediction. The use of unfalsifiable theories to provide accommodations is only the limiting case of a general phenomenon.

As I hope this discussion has made clear, what makes the difference between accommodation and prediction is not time, but knowledge. When the scientist doesn't know the right answer, she knows that she is not fudging her theoretical system to get it. The fact that her prediction is of a future event is only relevant insofar as this accounts for her ignorance.

We can make essentially the same point in terms of the distinction between generating a theory and testing it. It follows from the definitions of

accommodation and prediction that only accommodated data can influence the process of generation, and this is the difference that the fudging explanation exploits. What is perhaps somewhat less clear is that the fudging explanation is compatible with the claim that the objective degree of support that a theory enjoys is entirely independent of the time at which the datum is observed. We may suppose that this objective support the theory enjoys is precisely the same, whether the datum was accommodated or predicted. Nevertheless, the fudging explanation and the strong advantage thesis it underwrites may still be correct, because a scientist's actual epistemic situation gives him only fallible access to this objective support. As a consequence, information about whether a datum was predicted or accommodated is relevant to his judgment. But perhaps the best way to develop this important point and more generally to extend the case for the fudging explanation is to consider some objections.

First of all, it might be claimed that there is no real motive for fudging in the case of accommodation, since the scientist is free to choose any theoretical system that accommodates the available evidence. She can choose a natural and plausible combination of theory and auxiliary statements that will therefore be highly confirmed by the accommodated data. The main weakness of this objection to the fudging explanation is that it severely exaggerates the ease with which a scientist can produce an accommodating system. Influenced by Quine, philosophers of science are eager to emphasize the underdetermination of theory by data, the fact that there are always in principle many theoretical systems that will entail any given set of data (Quine 1951). This is an important point, but it may blind the philosopher to the actual situation of the working scientist, which is almost the opposite of what underdetermination suggests. Often, the scientist's problem is not to choose between many equally attractive theoretical systems, but to find even one. Where sensible accounts are scarce, there may be a great temptation to fudge what may be the only otherwise attractive account the scientist has been able to invent. The freedom to choose a completely different system is cold comfort if you can't think of one. And, in less dire straits, where there is more than one live option, all of them may need to be fudged to fit. This brings out an important point that the pejorative connotations of the word 'fudge' may obscure: a certain amount of fudging is not bad scientific practice. At a particular stage of research, it may be better to work with a slightly fudged theory than to waste one's time trying unsuccessfully to invent a pure alternative. Fudging need not be pernicious, and this is consistent with my argument that we ought to be more impressed by prediction than by accommodation.

Another line of thought that seems to undermine the fudging explanation begins with the idea that having extra evidence available when a theory is being constructed ought to make it easier to construct a good theory. But then it seems misguided to say that the fact that this evidence was available for

accommodation somehow weakens the amount of support it provides. I accept the premise, with some reservations, but I reject the inference. Extra data can be a help, because they may rule out some wrong answers and suggest the right one. We are not, however, directly concerned with how easy it is to generate a theory, only with the reasons for believing a theory that has been generated, and we are not comparing the support a theory enjoys with this extra initial data to the support it enjoys without them. When a theory correctly predicts these data, it is not remotely plausible to claim that they would have provided more support if only the scientist had known about them when she constructed her theory. When the data are accommodated, however, this fact is reason to discount them (to some degree), since they provided a motive for fudging. This would not be plausible if evidence never misled us, leading us to an incorrect inference, but it often does.

The objections we have just considered were that neither prediction nor accommodation provide a motive for fudging. A third objection is that they both do. This has most force when focused on the auxiliary statements. For high level theories, the route from theory to observation is long and often obscure. The only way to deductively tie the theory to the observational consequences will be with an elaborate and purpose-built set of auxiliary statements, including the battery of approximations, idealizations and *ceteris paribus* clauses needed to make observational contact. This looks like fudging with a vengeance, and it seems to apply to prediction and accommodation alike. The natural response to this objection is to insist on distinguishing the manipulation that is necessary to get what one knows to be the right answer from the manipulation needed to generate any observational consequences at all. It is often difficult to get from a general and abstract theory to any empirical consequences, and so predictions may involve implausible auxiliaries that weaken the support the evidence eventually provides. In his ignorance, however, the predictor will try to use the best auxiliaries he can find, even if it turns out that this yields a false prediction. It is only the accommodator who will be tempted to use less plausible auxiliaries just to get the right result. So the difference between prediction and accommodation remains.

A fourth objection is that my explanation proves too much. As I observed at the beginning of this chapter, almost all theory construction involves some accommodation: a scientist rarely if ever constructs a theory without some data to which it is supposed to apply. This does not eliminate the question of the difference between prediction and accommodation, which is whether the fact that a particular datum was accommodated means there is any less reason to believe the theory than there would have been, had the datum been predicted. It might, however, seem to ruin the fudging explanation. If virtually every theory accommodates, then the fudging explanation seems to have the consequence that almost all predictive theories have been fudged, in which case there seems to be no general asymmetry between accommodation

and prediction. My reply is twofold. First, while theory fudging may lead to a kind of global weakness in the theory that would make it implausible, poorly confirmable, and so only weakly supported by the evidence, this need not be the case with auxiliary fudging, since different auxiliaries may be used for different deductions. The fact that bad auxiliaries have been used to make an accommodation does not entail that they will be used to make a prediction. Indeed, the scientist will have a motive to use better auxiliaries when she turns to prediction. So the weak support provided by the accommodations is compatible with strong support from the successful predictions generated by the same theory.

My second and complementary reply brings out a new aspect of the fudging explanation. Accommodation is indirect evidence of fudging and, while this evidential link is quite general, it is also defeasible. I have already suggested that suspicions may be allayed in cases where the theory is low level, requiring few auxiliaries, where it has a tight mathematical structure, or where it seems that there is no other possible explanation. Another way of allaying the suspicion of fudging is precisely by means of successful prediction. A fudged theory is one where the accommodation does not make it very likely that predictions will be successful; conversely, then, the success of predictions makes it less likely that the theory was fudged. Thus it turns out that the difference between accommodation and prediction according to the fudging explanation is not entirely a question of the potential liabilities of accommodation: prediction also has the special virtue that it may minimize them. This explains and defends what seems to be a common intuition that accommodations are most suspect when a theory has yet to make any tested predictions, but that the accommodations gain retrospective epistemic stature after predictive success. At the same time, we should not expect successful predictions to eliminate the relative weakness of accommodations entirely, in part because there may have been a switch in auxiliaries.

A fifth objection to the fudging explanation concedes the force of theory or auxiliary fudging in the case of accommodation, but maintains that there is a corresponding liability that applies only to prediction, namely 'observational fudging'. When a scientist makes a prediction, she has a motive to fudge her report of her future observation to fit it. There is no such motive in the case of accommodation, since the data are already to hand, and since in this case she can fiddle with the theoretical side to establish the fit between theory and evidence. According to this objection, accommodation and prediction both run risks of fudging, albeit of different sorts, so there is no clear advantage to prediction over accommodation. I agree that the risk of observational fudging is real. Moreover, it may be harder to detect than theory or auxiliary fudging, since most members of the scientific community can not check the investigator's raw data. (I owe this point to Morton Schapiro.) Observational fudging, however, applies to accommodation as well as to prediction. In accommodation there is sometimes, and perhaps

usually, a mutual adjustment of theory and data to get a fit; in a successful prediction, only the data can be fudged, so the asymmetry remains. Of course, in cases of prediction where there is reason to believe that the observation has been badly fudged to score a success, we do not give the prediction much credit.

Scientists have ways of minimizing the risk of observational fudging. The most familiar of these is the technique of double-blind experiments. If a scientist wants to test the efficacy of a drug, he may ensure that he does not himself know which subjects are receiving the drug and which a placebo, so that he does not illicitly fulfill his own prophecy when he diagnoses the subjects' reactions. Inference to the Best Explanation gives a natural account of the value of double-blinds. When an experiment is done without one, the inference that would otherwise be judged the best explanation may be blocked, by the competing higher level explanation that the fit between theory and reported data is due to unreliable reporting, skewed to make the theory look like the best explanation. Performing the experiment with a double-blind eliminates this competing explanation. The technique of double-blind experiments provides a useful analogy to the virtues of prediction as revealed by the fudging explanation. Just as the traditional double-blind is a technique for avoiding fudged observation, I am suggesting that the general technique of prediction provides an analogous technique for avoiding fudged theory. In one case, the double-blind improves the reliability of the observational reports; in the other it improves the reliability of the theoretical system. Of course data gathered without the benefit of a double-blind may be unbiased, just as an accommodating system may not in fact include any fudging. In both cases, however, there is a risk of fudging, and the techniques of double-blinds and of prediction both reduce the risk that the process of inquiry will be polluted by wish-fulfillment.

Actual and assessed support

There is a final and particularly important objection I will consider. On my account, accommodation is suspect because it provides indirect evidence of fudging. It might be claimed that this evidence is relevant only if we know something about the theory's track record, but not much about the detailed content of the theory, auxiliaries and data. The objection is that the information that the data was accommodated is irrelevant to the investigator himself. He has no need for indirect evidence of this sort, since he has the theory, data and auxiliaries before him. He can simply *see* whether the theoretical system has been fudged and so how well his theory is confirmed by the evidence, whether that evidence is accommodated or predicted (Horwich 1982: 117). In this case, to put weight on the indirect evidence provided by knowing the data were accommodated is like scrutinizing the tracks in the dirt when the pig itself is right in front of you (cf. Austin 1962:

115). This objection does allow the fudging explanation some force. It admits that information about whether evidence was predicted or accommodated is relevant for someone who is unfamiliar with the details of the theory. It is also consistent with the claim that accommodating theories tend to be fudgier than predicting theories, hence that predictions are usually worth more than accommodations. It does not, however, allow that this information is relevant to the scientist who does the accommodating, or to any other scientist familiar with the details of the theoretical system and the evidence that is relevant to it.

According to this 'transparency objection', information about whether a datum was predicted or accommodated is irrelevant to a direct assessment of the credibility of a theory because, while credibility or support does depend on whether the theoretical system has been fudged, this is something that can be determined simply by inspecting the theory, the data and their relation. We cannot, alas, simply observe whether a theory is true, but we can observe to what extent it is supported by the evidence. My reply is that even the investigator himself cannot merely observe the precise degree of support his theory enjoys. Inductive support is translucent, not transparent, so the indirect evidence that the fudging explanation isolates is relevant to both scientist and spectator. What I am suggesting is that we need to distinguish between actual and assessed inductive support, between the extent to which the data actually render the theory probable and the scientist's judgment of this. As I formulated it at the beginning of this chapter, the strong advantage thesis claims that a predicted datum tends to give more reason to believe a theory than the same datum would have provided for the same theory, if that datum had been accommodated. It does not claim that the actual support for the theory is different, only that our assessment of this support ought sometimes to be different. Let us see how this can be so.

It ought to be uncontroversial that there is a distinction between actual and assessed support, since it is a familiar fact of ordinary as well as scientific life that people misjudge how well their views are supported by their evidence. Moreover, without this distinction, it is difficult to see how we could account for the extent of scientific disagreements. We should expect support to be only translucent, even with respect to those factors that depend only on the content of the theory, auxiliaries and evidence. A scientist may misjudge the variety of the evidence or the simplicity of his theory. He may also be wrong about the plausibility of the auxiliaries, since these are rarely all made explicit. But the case for saying that scientists are only fallible judges of the actual support their theories enjoy is even stronger than this, since actual support also depends on additional factors that go beyond the content of the theoretical system and the evidence. First, it depends on the relationship between that system and the scientist's complex web of background beliefs. Second, it depends on the existence of plausible competing theories, theories the scientist may not know. Lastly, as Kuhn has suggested,

it depends on the fruitfulness of the theory, on its promise to solve new puzzles or old anomalies. This may be an objective matter, and it is a consideration in judging the credibility of a theory, but it is certainly not 'given'. For all these reasons, it seems clear that support is translucent, not transparent. Scientific claims are all fallible, whether they concern theory, observation or the bearing of one on the other.

Since support is translucent, the fudging explanation applies to the judgments of the scientist herself. The actual support that evidence gives to theory does not depend on the information that the evidence was accommodated or predicted but, since we can only have imperfect knowledge of what this support is, the information is epistemically relevant. It is worth scrutinizing the tracks in the dirt if you cannot be certain that what you see is a pig. Fudging need not be a conscious process, so the scientist should not assume that she is not doing it just because she is not aware that she is. The mechanism by which scientists judge credibility is not conscious either, as I emphasized in chapter 1, or else the problem of describing their inductive practices would not be as extraordinarily difficult as we have found it to be. So there is plenty of room for unconscious and undetected fudging. This is why the indirect evidence that information about whether data were accommodated or predicted provides remains relevant to scientists' assessment of support.

The failure to acknowledge the gap between actual and assessed credibility has a number of probable sources. One is the legacy of a too simple hypothetico-deductive model of induction, since deductive connections are meant to be the model of clarity and distinctness. But once we appreciate how complex judgments of credibility are, we should be happy to acknowledge that these are fallible. If the fudging explanation is along the right lines, we should also conclude that judgments of credibility have an empirical component, since the question of whether evidence has been accommodated is an empirical one. Another reason the distinction between actual and assessed support has been elided, or its importance missed, is the tradition of stipulating an artificial division between the context of discovery and the context of justification, between the ways theories are generated and the ways they are evaluated. The difference between accommodation and prediction is a difference in generation, but what the fudging explanation shows is that this is relevant to the question of evaluation.

The assumption that support is transparent to the investigator, that no distinction need be drawn between actual and assessed support, is an idealization, very similar to the idealization epistemologists sometimes make that people are deductively omniscient, so that they know all the deductive consequences of their beliefs. Such an idealization may be useful for certain issues, but it obscures others. It would, for example, yield an absurd account of the nature of philosophy or of mathematics. Again, to give a scientific analogy, the assumption of transparency is like a biological idealization that

all members of a species are identical. This is a simplification that might serve various biological purposes, but it would obscure the mechanism of natural selection, by ignoring what makes it possible. Similarly, the idealization of transparency, though perhaps sometimes useful, obscures the mechanism that explains why prediction is better than accommodation, and the reason this information is relevant to scientists' assessments of the merits of their theories.

Only accommodations influence the generation of theory, and theory influences only the generation of predictions. These two observations have provided the basis for my argument for the epistemic advantages of prediction over accommodation. Predictions are better, because the theory under investigation may be used to select those that would give particularly strong support, especially because they would discriminate between this theory and its rivals. Accommodations are worse, because they may lead to a theory that is only weakly supportable. In accommodations, the scientist knows the correct answers in advance, and this creates the risk, not present in predictions, that he will fudge his theory or his auxiliaries. Moreover, since a scientist's judgment of the degree of support evidence provides is only a fallible assessment of its actual value, the indirect evidence of fudging that the fact of accommodation provides is relevant to the scientist's own judgment of support. It leaves the scientist with less reason to believe his theory than he would have had, if he had predicted the datum instead.

One clear limitation of my account is that it does not include anything like a precise characterization of the features that make one theoretical system fudgier than another. I have not provided such a characterization because I do not know how to do so, but my argument does not depend on it. It is enough that we have good reason to believe that different theoretical systems enjoy different degrees of support from data they all fit, and that the requirements of accommodation may be in tension with the desire to produce a highly confirmable theoretical system. Another limitation is that I have not shown that the fudging explanation accounts for the true extent of the contrast between prediction and accommodation. Perhaps most of us feel a greater contrast than my account appears to justify. There are then various possibilities. One might simply say that this excess is psychologically real, but rationally unjustifiable. There is some advantage to the fact of prediction, but it is weaker than many suppose. Secondly, one might argue that my account justifies a greater contrast than it initially seems to do. Two of the ways this might be done are by stressing the way the case for the weak advantage thesis combines with the case for the strong thesis, or by emphasizing the retrospective credit that successful predictions grant to earlier accommodations. Finally, one might argue that there is some independent and additional reason why predictions are better than accommodations that I have not taken into account. My story does not rule out such a thing but, given the difficulty in coming up with any

justification at all for the strong advantage thesis, I am content to have given one.

In the first section of this chapter, I criticized several plausible accounts of the difference between accommodation and prediction. From the perspective of the fudging explanation, we are now in a position to see the germs of truth they contain. Recall first the claim that accommodations are worse than predictions because accommodations are ad hoc. My objection was that this leaves the question unanswered or begged, because either calling a theoretical system ad hoc simply means it is designed to accommodate, or because it means it is only weakly confirmable. The fudging explanation provides an independent reason why accommodating systems tend to be ad hoc in the second sense, only weakly supported by the evidence they accommodate.

Another account I initially rejected was that predictions are better because only they test the theory, since a test is something that can be failed. My objection was that, while it may be that only predictions test the scientist, an accommodating theory can be as falsifiable as a predicting one. In such a case, if the accommodated evidence had been different, the theory would have been disconfirmed. As we have seen, however, in extreme cases there will be a suspicion that an accommodating theory is in fact unfalsifiable, a suspicion that does not arise in the case of prediction. More generally, the fudging explanation is related to the notion that we should test the scientist, and that she should test herself. It is not simply that she ought to run the risk of being wrong, though that is a consequence of the real point. She should place herself in a situation where, as in the case of a standard double-blind experiment, she does not know the right answer in advance.

At long last, we return to the claim that an overarching inference to the best explanation underwrites the strong advantage thesis. In its standard form, that account claims that prediction is better because the best explanation for predictive success is truth, while the best explanation for accommodation is instead that the theory was designed for that purpose. My objection was that it was not clear that the accommodating explanation pre-empts the truth explanation, making it less likely. It is true that only in cases of accommodation can we explain the fit between theory and data by pointing out that the theory was designed for the purpose, but whether this makes it less likely that the theory is correct is just what is at issue. By contrast, it is clear that the fudging explanation competes with the truth explanation. Insofar as we may reasonably infer the explanation that the fit between theory and data in the case of accommodation is due to fudging, this undermines our confidence in the inference from fit to truth. So the best explanation account of the difference between accommodation and prediction can be salvaged by replacing the accommodation explanation with the fudging explanation. In cases of accommodation, the inference from fit to truth may be blocked by the inference from fit to the conclusion that the

theoretical system is ad hoc, in the epistemically pejorative sense. In addition, we may now say that we have vindicated the original claim that the accommodation explanation and the truth explanation are competitors. We have shown this in two steps, by arguing that accommodation is evidence of fudging, and that fudging results in weakened support. For reasons that will become clear in the next chapter, however, it is perhaps better to say simply that the fudging explanation competes with the ordinary inferences to the best explanation that scientists are otherwise inclined to make, rather than speaking of a special, philosophical 'truth explanation'. Seen in this way, the fudging explanation should sometimes act as a brake on inference, preventing scientists from accepting what they otherwise would judge to be an explanation lovely enough to infer.

This view of the way an overarching inference to the best explanation accounts for the epistemic distinction between prediction and accommodation links up with several features of Inference to the Best Explanation that I have discussed earlier in this book, of which I will mention only two, both of which relate to material discussed in the last chapter. The first is that it allows us to give a more nuanced response to Hungerford's objection. The objection is that explanatory loveliness is audience relative in a way that makes it unsuitable for an objective notion of inductive warrant. My main response was that, as the structure of contrastive explanation shows, loveliness is relative, but in an innocent way, since different interests will lead to different but compatible inferences. Armed now, however, with the distinction between actual and assessed support, we may add that while actual loveliness is not strongly relative, assessed support may be. Different scientists may judge incompatible theories loveliest, if at least one of them is wrong. But scientists also have ways of minimizing the dangers of this sort of misassessment, and the special value they place on novel prediction over accommodation is one of them. The fudging explanation also brings out one consequence of the short list method that we must employ in inference. If Inference to the Best Explanation worked from a full menu of all possible explanations of the evidence, the fudging explanation would not apply. The scientist would have no motive to fudge her explanation to make the accommodation, since she would have before her every possible explanation, from which she could choose the best. Since, however, this can never be her position, and since she often must scramble to generate even one plausible candidate, the fudging explanation applies. Thus the epistemic distinction between accommodation and prediction is one symptom of the way our actual inferential techniques only imperfectly mimic a full menu mechanism.

I end this chapter by observing that the fudging explanation also gives us an answer to the puzzle of the twin scientists, which seemed to show that there could be no difference between prediction and accommodation. We had two scientists who happen to come up with the same theory, where one accommodates evidence the other predicts. After they compare notes, they

must have a common level of confidence in the theory they share. The difficulty was that it seemed impossible to say what level this should be, and how meeting a predictor could be any different for the accommodator than knowing what all accommodators know, namely that if someone had produced the same theory on less evidence, her prediction of the balance of the evidence would have been successful. The answer to this puzzle is now clear. The accommodator ought to worry that she had to do some fudging, but his suspicion is defeasible. One of the things that can defeat it, though not common in the history of science, is meeting a corresponding predictor. If the accommodator meets someone who predicted data she only accommodated, with the same theoretical system, this shows that she almost certainly did not fudge to make those accommodations. The predictor, ignorant as he was of the data he predicted, had no motive to fudge his theoretical system to get those results; consequently, the fact that he came up with just the same system provides strong independent evidence that the accommodator did not fudge either. At the same time, the fact that any accommodator knows that the same theory could have been constructed earlier is not enough, since such a construction might have then required arbitrary and unmotivated fudging. A predictor might have come up instead with a different and better theoretical system. If an actual meeting takes place, however, the twins should leave it sharing the higher confidence of the predictor. The accommodator, like all scientists, has only fallible knowledge of the actual credibility her theory enjoys, and the meeting gives her some additional evidence that ought generally to lead her to increase her assessment.

Chapter 11

Truth and explanation

Circularity

We have just seen one way Inference to the Best Explanation can be used to provide a justification for a particular aspect of our inductive practices. We tend to give more credit to successful predictions than to accommodations, and the fudging explanation shows how this preference can be justified, by showing how it follows from a more general preference for the best explanation. Accommodations are often worth less than predictions, because only they have to face the possibility that the best explanation of the fit between theory and data is that the theoretical system was fudged. In this chapter, we will consider a second overarching application of Inference to the Best Explanation to a problem of justification, but now what is to be justified is not just one rather special aspect of our inductive practices, but the general claim that they take us towards the truth.

When a scientist makes an inference to the best explanation of the sort we have discussed in the previous chapters, I have taken it that she infers that the claim in question is (at least approximately) true, whether the claim concerns something observed, something unobserved but observable, or something unobservable. Even if this descriptive claim is accepted, however, it leaves open the question of justification. Do we have any reason to believe that the inferences to the best explanation that scientists make really are truth-tropic, that they reliably take scientists towards the truth? To have a familiar label, let us call someone who accepts both that Inference to the Best Explanation provides a good description of scientists' inductive practices, and that these practices are truth-tropic, a 'scientific realist'. There is a well-known argument for scientific realism that itself has the form of an inference to the best explanation. In its simplest version, the argument is that we ought to infer that scientific theories that are predictively successful are (approximately) true, since their truth would be the best explanation of their success. As Hilary Putnam has famously put it, an account that denies that our best scientific theories are at least approximately true would make their success a miracle (1975: 73; 1978: 18–22). Moreover, since these theories are them-

selves the results of inferences to the best explanation, that form of inference is truth-tropic. This argument from predictive success to truth and to truth-tropism is often known as the miracle argument for realism.

A realist who endorses the miracle argument will take it that Inference to the Best Explanation is a generally legitimate form of inductive inference, and so will endorse the inferences of that form that scientists make. The miracle argument itself, however, is not supposed to be a scientific inference; instead it is a philosopher's argument, an additional inference of the same form, whose conclusion is that the scientific inferences are truth-tropic. What is the intuition behind this argument? Suppose that you find yourself in the middle of the woods with a detailed topographical map that may or may not be accurate. You then proceed to navigate by means of the map, and find that everything the map says you ought to find – rivers, lakes, roads, peaks and so on – is just what you do find, though of course you only see a small portion of the terrain the map purportedly depicts. What is the best explanation of the success you have achieved with this map? One possibility is that your success is a fluke: the map happens to be correct in those few details you have checked, but it is generally wrong. In this case, however, your success is inexplicable: it is simply good luck. Another possibility, however, is that the map is generally accurate, both in the parts you have checked and in those you have not. Its general accuracy would explain why you have had success using it, and this is why you would infer that the map is accurate. Similarly, the predictive success of a scientific theory does not entail that the theory is correct, since every false statement has many true consequences. So one 'explanation' for predictive success is that the predictions that have been checked just happen to be some of the true consequences of a false theory. This, however, is really no explanation at all, since it is just to say that the predictive success is a fluke. By contrast, if the theory is true, this would explain its predictive success. So, since Inference to the Best Explanation is a warranted form of inference, we have reason to infer that the theory is true. This is an inductive argument, so it does not prove that successful theories are true, but it does, according to its proponents, provide a good reason for believing that they are true, and so that the form of inference that led to them, namely Inference to the Best Explanation, is a reliable guide to the truth.

I want to consider whether the miracle argument is cogent, taking seriously its pretension to be both distinct from the particular inferences scientists make and a real inference to the best explanation, as that form of inference has been developed and defended in this book. The most obvious objection to the miracle argument is that it begs the question (Laudan 1984: 242–3; Fine 1984: 85–6). Who is the argument supposed to convince? It is an inference to the best explanation, so it has no force for someone who rejects inferences to the best explanation, either because, like Popper, she rejects inductive inferences altogether or because, while she accepts some

form of induction, she does not accept inferences to the best explanation. More surprisingly, it does not even have force for everyone who accepts some version of Inference to the Best Explanation. Recall that, as we saw in chapter 4, Inference to the Best Explanation requires that we distinguish between potential and actual explanations. We cannot say that we are to infer that the best actual explanation is true, since to say that something is an actual explanation is already to say that it is true, so this method would not be effective. Instead, we must say that we are to infer that the best potential explanation is true, or that it is an actual explanation. Because of this feature of Inference to the Best Explanation the model may be co-opted by someone who denies that inferences generally are or ought to be inferences to the truth of the inferred claim. As we saw already in chapter 9, one might claim that the best potential explanation is a guide to inference, yet that what we ought to infer is not that the explanation is true but only, for example, that its observable consequences are true. Such a person would not be moved by the miracle argument, at least not in the direction its proponents intend. Perhaps the truth of a theory is the best explanation of its predictive success but, if the theory traffics in unobservables, the claim that it is true is not an observable claim, so the most that an instrumentalist will accept is that all the observational consequences of the theory are true. The realist is trying to argue from a theory's past observational successes to its truth, not from past observational successes to future ones. So there is at most one sort of person who ought to be impressed by the miracle argument, and this is someone who both accepts Inference to the Best Explanation and accepts that this form of inference is truth-tropic. In other words, the miracle argument can only have force for a scientific realist. This, however, is precisely the view that the miracle argument was supposed to justify. In short, the miracle argument is an attempt to show that Inference to the Best Explanation is truth-tropic by presupposing that Inference to the Best Explanation is truth-tropic, so it begs the question.

The circularity objection does I think show that the miracle argument has no force for the non-realist. The interesting remaining question is whether it may yet be a legitimate argument for someone who is already a scientific realist, giving her some additional reason for her position. To answer this, it is useful to compare the miracle argument to the general problem of justifying induction. The miracle argument is an attempt to justify a form of inductive inference, and so we should have expected it to run up against the wall that Hume's skeptical argument erects against any such attempt. More particularly, the circularity problem for the miracle argument is strikingly similar to the problem of trying to give an inductive justification of induction. Most of us have the gut feeling that the past success of induction gives some reason for believing that it will continue to work well in the future. This argument, however, is itself inductive, so it cannot provide a reason for trusting induction. The counter-inductive justification of counter-

induction, which claims that counter-induction will work in the future because it has failed in the past, strikes us as obviously worthless, but the only relevant difference between this and the inductive justification of induction seems to be that the latter confirms our prejudices (Skyrms 1986: sec. II.3). The miracle argument has just the same liability as the inductive justification of induction: it employs a form of inference whose reliability it is supposed to underwrite, and this seems to show not only that the argument is of no value against the skeptic, but that it has no value for anyone.

Objections from circularity are among the most elegant and effective tools in the philosopher's kit, but I am not sure whether the circularity objection is conclusive against either the inductive justification of induction or the miracle argument. One reason I hesitate is that, for all its intuitive appeal, circularity is very difficult to characterize. We cannot appeal simply to the psychological force of an argument, since people often reject perfectly good arguments and are often persuaded by circular arguments. Many are convinced by Descartes's argument that every event must have a cause, since a cause must have at least as much 'reality' as its effect, and this would not be so if something came from nothing (1641: Meditation Three). Yet, as Hume observed, the argument is circular, since it implicitly treats 'nothing' as if it would be a cause, which would only be legitimate if it were already assumed that everything must have a cause (1739: Sec. 1.3.3). Nor will it do to say that an argument is circular just in case its conclusion is somehow already present among its premises. This runs the risk of counting all deductive arguments as circular. In fact, a deductive argument may be legitimate even though its conclusion is logically equivalent to the conjunction of the premises. The classic free will dilemma is an example of this. Either determinism is true or it is not; if it is, there is no free will; if it is not, there is no free will; therefore there is no free will. The premises clearly entail the conclusion and, on a truth-functional interpretation of the conditionals, which seems not to diminish the force of the argument, the conclusion entails each of the premises, so the conclusion is logically equivalent to the conjunction of the premises. Yet this is, I think, one of the stronger arguments in the history of philosophy, even though its conclusion is incredible. A better account of circularity is possible if we help ourselves to the notion of a good reason. We cannot say that an argument is circular if there is less reason to accept the premises and rules of inference than there is to accept the conclusion, since this condition is met by the non-circular argument that the next raven I see will be black since all ravens are black. But we might say that an argument is circular just in case the conclusion is an essential part of any good reason we might have to accept one of the premises or rules of inference. I suspect, however, that any account of circularity that relies on the notion of good reason is itself circular, since we cannot understand what it is to have a good reason unless we already know how to distinguish circular from non-circular arguments.

Another reason I hesitate to dismiss the inductive justification of induction is that there are clear cases where we can use inductive arguments to defend a form of inductive inference. I will give two examples from personal experience. The first is the charting method. Many years ago, I spent a summer working as a clerk on the London Metal Exchange. Although my responsibilities were menial, I took an interest in the way dealers formed their views on the future movements of the markets in the various metals they traded. The dealers I spoke to cited two different methods. The first, the physical method, is straightforward. It consists of monitoring those conditions which obviously affect the supply and demand of the metal in question. For example, a strike at the large Rio Tinto copper mine was taken as evidence that copper prices would rise. The second, the charting method, is more exotic. Chartists construct a graph of the prices of the metal over the previous few months and use various strange rules for projecting the curve into the future. For example, two large hills (whose peaks represent high prices) separated by a narrow valley would be taken as an indication that prices would rise, while a small hill followed by a larger hill suggested that prices would fall. The dealers provided no explanation for the success of this method, but many of them put stock in it. My second example is persistence forecasting. During the summer after the one on the Metal Exchange, I spent several weeks hiking and climbing in the Wind River mountain range of Wyoming. In these travels I met another hiker who turned out to be a meteorologist. After some talk about the weather, our conversation turned to predicting it. She told me that there had been studies comparing the reliability of various predictive methods and that the technique of persistence forecasting still compared favorably with other methods. Persistence forecasting, it turns out, is the technique of predicting that the weather will be the same tomorrow as it was today.

Perhaps the charting method should be of more interest to the anthropologist than to the epistemologist, but it does provide a clear case of an inductive method whose reliability could be assessed with an inductive argument based on past performance. The method is implausible, but an impressive track record would lead us to give it some credit. Similarly, though I do not know whether the meteorologist was being entirely honest about its performance, persistence forecasting is a technique whose reliability can be sensibly assessed by considering how well it would have worked in the past. This example is particularly attractive to the defender of the inductive justification of induction, since, unlike charting, persistence forecasting is itself a kind of primitive enumerative induction. The general point of both examples, however, is to show that we can legitimately assess inductive methods inductively. One reason for this may be that the premises of such arguments can 'track' their conclusions (Nozick 1981: ch. 3, esp. 222–3). Thus the past successes of persistence forecasting or charting may provide a reason to believe they will work in the future as well, because if

they were not going to work in the future, they probably wouldn't have worked in the past. On this attractive conception of induction, a good inductive argument turns us into a reliable instrument for detecting the truth value of its conclusion, and it seems that these inductive justifications of inductive policies could do this.

If charting and persistence forecasting had strong track records, this would not impress an inductive skeptic, but it would impress us, since we already accept induction. Similarly, the fudging explanation as an argument that predictions are better than accommodations might not move someone who rejected Inference to the Best Explanation altogether, but it still has some force for the rest of us. What this suggests is that the notion of circularity is audience relative. The same argument may beg for one audience, yet be cogent for another. So while the inductive justification of induction has no force for an inductive skeptic, it may yet have some value for us (cf. Black 1970). There is nothing illegitimate about giving arguments for beliefs one already holds. Yet we are pulled by conflicting intuitions in this case. On the one hand, the Humean argument for the circularity of the inductive justification does not remove the strong feeling that the past successes of induction bode well for its future performance. On the other hand, we also feel that the pathological ineffectiveness of the inductive justification of induction against someone who does not already accept induction shows that it also provides no good reason for those of us who do accept it. This is why being circular may be more debilitating than merely having premises your opponent does not accept. Someone who is inclined to doubt my veracity will not be convinced of something on my say so, and he should not be moved by my additional statement that my testimony is true. Someone else may have reason to trust me, but my additional statement provides him with no additional reason. In short, the debilitating consequences of circularity do not always disappear when an argument is used to preach to the converted.

The fact that arguments from the track records of charting and persistence forecasting would beg against an inductive skeptic does not show that these arguments are circular for us, yet the fact that the track record of induction would beg against the skeptic seems to show that it also begs for us. What is the difference? In the former cases, it is easy to see how there could be people who initially reject charting and persistence forecasting, yet accept induction. So there are people who start by having no reason to trust these specific methods, yet ought, by their own principles, to accept the arguments from track record. In the case of the inductive justification of induction, by contrast, it is hard to see how the argument could have any force for anyone who did not already accept its conclusion, since such a person could only be an inductive skeptic. The underlying idea here is that, unless an argument could be used to settle some dispute, it can have no force even for people who already accept its conclusion.

But the track record of induction could be used to settle a dispute, not over whether induction is better than guesswork, but over the *degree* of reliability of induction among disputants who endorse the same principles of induction. These people would give different weight to the inductive justification, but they would all give it some weight. Consider someone who is much too optimistic about his inductive powers, supposing that they will almost always lead him to true predictions. Sober reflection on his past performance ought to convince him to revise his views. If he admitted that he had not done well in the past, yet claimed for no special reason that he will be virtually infallible in the future, he would be inductively incoherent. Similarly, someone who is excessively modest about his inductive powers, though he gives them some credit, ought to improve his assessment when he is shown how successful he has been in the past. And the fact that the inductive justification can help to settle disputes over reliability is enough, I think, to show that even people who already believe that induction has the degree of reliability that the argument from its track record would show, can take that argument to provide an additional reason for holding that their belief is correct. So the inductive justification of induction, while impotent against the skeptic, is legitimate for those who already rely on induction. If the problem of induction is to show why the skeptic is wrong, the inductive justification of induction is no solution, but it does not follow that the justification has no application. Someone who has no confidence in my testimony will not be moved if I add that I am trustworthy, but someone who already has some reason to trust me will have an additional reason to accept my claim if I go on to say that I am certain about it.

There is an additional reason for saying that the inductive justification is legitimate for those who already accept induction. Consider again charting and persistence forecasting, but imagine now that we have two machines that generate predictions, one about metal prices, the other about the weather. In fact, one machine runs a charting program, the other a persistence program, but we do not know this. All we notice is that these machines generate predictions, and we then might use our normal inductive techniques to evaluate their reliability, much as we use those techniques to determine how reliable a drop in barometric pressure is as a predictor of tomorrow's weather. Well, some organisms are predicting machines too. Suppose we find alien creatures who predict like mad, though we do not yet know what mechanism they use. Still, we may look at how well these creatures have done in the past to assess their future reliability. Our procedure will not be undermined when we later discover their predictive mechanism is the same as ours, though this will show that we have unwittingly given an inductive justification of induction.

Our problem is to determine when an argument that is circular for some may nevertheless provide a reason for belief among those who already accept its conclusion. My suggestion is that this form of preaching to the

converted is legitimate when there might also be someone who does not already accept the conclusion of the argument but who would accept the premises and the rules of inference. In short, an argument provides an additional reason for those who already accept its conclusion, when there is a possible audience who does not already accept it, yet for whom the argument would not be circular. This condition does not give us an analysis of circularity, but it seems to solve our problem. The condition is not satisfied by those who are taken in by Descartes's argument that every event has a cause, since they do not realize that the argument includes a tacit premise they do not accept. It is, however, clearly satisfied for many deductive arguments. The condition is also met by the cases of charting and persistence forecasting, and I have argued that it is also met by the inductive justification of induction, where that argument is taken as one to the degree of reliability of our practices. Since we already accept a method of induction, we may use an inductive argument to assess its reliability, since there could be (indeed there are) optimistic and pessimistic people who would accept both its premises and its rule of inference, yet do not already accept its conclusion. This result accounts for our ambivalence about Hume's argument against the inductive justification, our feeling that while he has shown that the argument is circular, it nevertheless has some probative value. Circularity is relative to audience, and the inductive justification of induction is circular for an audience of skeptics, yet not among those who already accept that induction is better than guessing.

We may now return to the miracle argument, which says we ought to infer first that successful theories are true or approximately true, since this is the best explanation of their success, and then that Inference to the Best Explanation is truth-tropic, since this is the method of inference that guided us to these theories. The objection to this argument that we have been considering is that it is circular, just like the inductive justification of induction. Indeed it might be claimed that the miracle argument is just the inductive justification dressed up with a fancy theory of induction. But we have now seen that, while the inductive justification begs against the inductive skeptic, a case can be made for saying that it is not circular in the narrower context of a dispute among those who already accept induction. Similarly, we must concede that the miracle argument has no force against those who do not already accept that inferences to the best explanation are truth-tropic. The question before us is whether we may extend the analogy with the inductive justification, and so salvage something from the miracle argument by claiming that at least it is not circular for those who already accept a realist version of Inference to the Best Explanation.

One obvious difference between the simple inductive justification of induction and the miracle argument is that only the former is enumerative. The inductive justification says that past success provides an argument for future success, but the miracle argument says that past success provides an

argument for truth, not just for future success. This, however, is just the difference between enumerative induction and Inference to the Best Explanation: what we have are parallel applications of different methods of induction. So the miracle argument can be defended against the charge of circularity in the narrower context in the same way as the inductive justification was defended, by showing that the argument can be used to settle disputes. For the inductive justification, the dispute was over degrees of reliability. The miracle argument, in a more sophisticated form, could also be used for this dispute. Even successful theories make some false predictions, so the literal truth of the theory cannot be quite the best explanation of its actual track record. More perspicuously, perhaps, the dispute can be seen as one over degree of approximation to the truth. Scientific realists may disagree over how effectively or how quickly scientists get at the truth or over the relative verisimilitude of competing theories, and something like the miracle argument could be used to address this sort of dispute.

There is also another dispute that the miracle argument might help to settle. In the last chapter, I argued that the distinction between prediction and accommodation is epistemically relevant, because scientists are only imperfect judges of the actual inductive support their theories enjoy. When a theory is constructed to accommodate the data, there is a motive to fudge the theory or the auxiliary statements to make the accommodation. This fudging, if it occurs, reduces the support that the data provide for the theory by the scientist's own standards but, since fudging is not always obvious, she may miss it and so overestimate the credibility of her theory. When a theory makes correct predictions, however, this provides some evidence that the theory was not fudged after all, and so that her assessment was correct. The issue here is not over the reliability of scientists' inductive principles, considered abstractly, but over the reliability of their application. The question is how good scientists are at applying their own principles to concrete cases. This is clearly something over which scientists with the same principles could disagree, and it is a dispute that the miracle argument could help to settle, since the predictive success of accepted theories is, among other things, a measure of how well scientists are applying principles that they agree are reliable. So I conclude that, while the miracle argument is no argument against the inductive skeptic or the instrumentalist, the circularity objection does show that realists are not entitled to use it. The argument is circular against non-realists, but not for realists themselves.

A bad explanation

To say that the truth explanation does not fall to the circularity objection is not, however, to say that it does not fall. It faces another serious problem, the bad explanation objection. As I have argued in this book, an inference to the best explanation is not simply an inference to what seems the likeliest

explanation, but rather the inference that what would be the loveliest explanation is likeliest. And how lovely is truth as an explanation of predictive success? According to the bad explanation objection, not lovely at all. Even if the miracle argument is not hopelessly circular, it is still a weak argument, and weak on its own terms. This is so because the argument is supposed to be an inference to the best explanation, but the truth of a theory is not the loveliest available explanation of its predictive success; indeed it may not be an explanation at all. In a way, this is the reverse of the circularity objection. According to that objection, the miracle argument is illegitimate precisely because it would be an inference to the best explanation, and so would beg the question, since only a realist could accept it. According to the bad explanation objection, the miracle argument is not warranted by Inference to the Best Explanation, so even a realist ought not to accept it. It is simply not good enough.

Van Fraassen has given what he claims is a better explanation of predictive success (1980: 39–40). His explanation is neo-Darwinian: scientific theories tend to have observed consequences that are true because they were selected for precisely that reason. Scientific method provides a selection mechanism that eliminates theories whose observed consequences are false, and this is why the ones that remain tend to have observed consequences that are true. This mechanism makes no appeal to the truth of theories, yet explains the truth of their observed consequences. (This selection explanation is similar to the accommodation explanation we considered in the last chapter: no appeal to the truth of theories is needed to explain the fit with accommodated data.) The realist, however, seems to have a simple response. According to Inference to the Best Explanation, we are to infer the loveliest of *competing* explanations, but the truth explanation and the selection explanation are compatible, so we may infer both. The scientific environment may select for theories with observed consequences that are true and the theories thus selected may be true. Van Fraassen's explanation does not deny that successful theories are true, it just does not affirm this. So it appears that the realist is free to accept van Fraassen's account yet also to make the miracle argument.

Van Fraassen will surely say that this is too quick. The selection explanation is logically compatible with the truth explanation but, once we infer the selection explanation, we ought to see that it deprives us of any reason we may have thought we had for inferring the truth explanation. The one explanation pre-empts the other. But why should this be? When my computer did not work, I did not infer that the fuse was blown, since I noticed that the computer was unplugged. These two explanations are logically compatible, since they both could be true, but the plug explanation is known on independent grounds to be correct, and it takes away any reason I would have had to infer that the fuse has blown. Once I accept the plug explanation, there is nothing left for the fuse to explain. But I want to argue

that this is not the relevant analogy. The realist will accept the selection mechanism, but this does not explain everything that inference to truth would explain.

To see this, notice that a selection mechanism may explain why all the selected objects have a certain feature, without explaining why each of them does (Nozick 1974: 22). If a club only admits members with red hair, that explains why all the members of the club have red hair, but it does not explain why Arthur, who is a member of the club, has red hair. That would perhaps require some genetic account. Similarly, van Fraassen's selection account may explain why all the theories we now accept have been observationally successful, but it does not explain why each of them has been. It does not explain why a particular theory, which was selected for its observational success, has this feature. The miracle argument, by contrast does explain this, by appealing to an intrinsic feature of the theory, rather than just to the principle by which it was selected.

There is a second noteworthy difference between the blown fuse explanation and the truth explanation, namely that only the truth explanation can 'track'. If the computer is unplugged, then it would not work whether or not the fuse was blown. By contrast, even if the theory was selected because of its observational successes, it might not have been successful if it were not true. So the miracle argument both explains something not accounted for by the selection explanation and may meet a tracking requirement on inference. This, along with the compatibility of the two accounts, suggests that the selection explanation does not pre-empt the truth explanation.

Yet perhaps even this explanatory difference and the possibility of tracking do not remove the feeling of pre-emption. For the truth explanation was motivated by the thought that the success of science would be miraculous if its theories were not largely true, but the selection explanation seems to remove the miracle. It is no miracle that all the members of the club have red hair, if this was a criterion of selection. But there is a natural reply to this thought. The real miracle is that theories we judge to be well supported go on to make successful predictions. The selection mechanism does not explain this, since it does not explain why our best supported theories are not refuted in their next application. Constructive empiricism assumes that scientific canons of induction yield theories that will continue to be empirically successful in new applications, but it does not explain why this should happen. The truth explanation, by contrast, does provide some sort of explanation of a theory's continuing predictive success. If our inductive criteria are truth-tropic, then well supported theories tend to be true, and so they will tend to generate true predictions. This assumes that our criteria are generally truth-tropic, so that is not explained, but the miracle argument does explain our continuing observational success and, as we noted in chapter 2, an explanation may be sound and provide understanding even though what does the explaining is not itself explained.

The selection explanation and the truth explanation account for different things. The selection explanation accounts for the fact that, at any given time, we only accept (in either the realist's or the constructive empiricist's sense) theories that have not yet been refuted. It assumes nothing about our inductive powers; indeed it is an explanation that Popper might give. The truth explanation, by contrast, accounts for two other facts. First, it explains why a particular theory that was selected is one that has true consequences. Secondly, it explains why theories that were selected on empirical grounds then went on to more predictive successes. The selection explanation accounts for neither of these facts. I conclude that it does not pre-empt the truth explanation. The miracle argument cannot be defeated on the grounds that the selection explanation is a better explanation of observational success and so blocks the inference to truth as the best explanation.

This may dispose of the argument that the truth explanation is a bad explanation because the selection explanation is better, but it does not dispose of the bad explanation objection altogether. Inference to the Best Explanation, correctly construed, does not warrant inferring the best explanation at all costs: as I have noted before, the best must be good enough to be preferable to no inference at all. Perhaps the likeliest thing is that continuing observational success is just inexplicable, a brute fact. Perhaps it is really no less likely that a false theory should be observationally successful than that a true one should be. Or perhaps there are competing explanations quite different from the selection explanation, that are at least as good as the truth explanation and so would block an inference to it.

How lovely, then, is the truth explanation? Alas, there is a good reason for saying that it is not lovely at all. The problem is that it is too easy. For any set of observational successes, there are many incompatible theories that would have had them. This is our old friend, underdetermination. The trouble now is that the truth explanation would apply equally well to any of these theories. In each case, the theory's truth would explain its observational success, and all the explanations seem equally lovely. A very complex and ad hoc theory provides less lovely explanations than does a simple and unified theory of the same phenomena, but the *truth* of the complex theory is as lovely or ugly an explanation of the *truth* of its predictions as is the explanation that the truth of the simple theory provides. In either case, the explanatory value of the account lies simply in the fact that valid arguments with true premises have only true conclusions. So the truth explanation does not show why we should infer one theory rather than another with the same observed consequences. To appreciate this point, it is important to remember that the proponent of the miracle argument holds that his explanation is distinct from the first-order explanatory inferences scientists make. Those inferences, construed as inferences to the best explanation, do distinguish between theories with the same observed consequences, since not every such theory gives an equally lovely explanation of the evidence. This is one of the

main strengths of Inference to the Best Explanation as an account of those inferences. But the proponent of the miracle argument, as I have construed her position, insists that the truth explanation, applied to a particular theory, is distinct from scientific explanations that the theory provides. She is entitled to this, if she wants it. After all, we may suppose that most of those scientific explanations are causal, but the truth explanation is not. The truth of a premise in a valid argument does not cause the conclusion to be true. But the price she pays for this separation is an exceptionally weak explanation, that does not itself show why one theory is more likely than another with the same observed consequences.

The same point can be made in terms of explanations that compete with the truth explanation. In this case, the competitor is not the selection explanation; rather there are many 'underdetermination competitors', namely all those theories incompatible with the original theory whose truth the truth explanation is supposed to defend but which would have enjoyed the same empirical success to date as that theory has done (which does not require that they be empirically equivalent). Unlike the selection explanation, these are logically incompatible with the truth explanation and so clearly competitors. The truth of any one of those theories would also explain the success of the original theory, rather as the truth of special relativity would explain the success of classical dynamics. And insofar as the truth explanations are taken to be distinct from first-order scientific explanations, it would seem that all these explanations would be equally lovely. This leaves us with a stalemate, blocking inference to the truth explanation.

The intuition behind the miracle argument is that it would be a miracle if a highly successful theory were false; but once we take these underdetermined competitors seriously, a miracle no longer seems required for our successful theory to be false. Quite the opposite: it would rather be a miracle if the truth did not lie instead somewhere among the innumerable underdetermined competitors. And here it is difficult not to suspect that the original plausibility of the miracle argument is just an instance of philosophers falling for the ubiquitous fallacy of ignoring base rates, which we discussed in chapters 7 and 8 (Kahneman *et al.* 1982: 154). The classic case, as we saw, was as follows. A test for a disease has a false negative rate of nil: nobody who has the disease will test negative. The false positive rate is 5 percent: five out of a hundred of the people who do not have the disease will nevertheless test positive. And one in a thousand people have the disease. You give your patient the test and the result is positive. What is the probability that your patient has the disease? When staff and students at Harvard Medical School were asked, most said it is 0.95. The correct answer is just under 0.02. If you give the test to a thousand people, about fifty of them will test positive even though they don't have the disease, as against the one poor person who has the disease. The probability that your patient has the disease is much higher after the test – from one in a thousand to about

one in fifty – but that is still pretty unlikely. The medics who thought that the probability is 0.95 were wildly out because they ignored the base rate, which should have told them that results for healthy people are swamping results for the ill. Beware false positives, even if unlikely, for rare conditions.

As Colin Howson has suggested (2000: 54), the miracle argument for realism seems to work the same way, with realists playing the role of the probabilistically challenged medics. On this analogy, being true is the disease and making lots of true and precise predictions is testing positive for it. The false negative rate is nil: no true theory makes false predictions. The false positive rate is low, since relatively very few false theories are so empirically successful. So we are inclined to infer that highly successful theories are likely to be true. What we ignore is the base rate that, to put it crudely, the vast majority of theories are false, so even a very small probability that a false theory should make such successful predictions leaves it the case that the great majority of successful theories are false. Most false theories are unsuccessful, but alas what counts is that most successful theories are false. Beware false positives, even if unlikely, because the truth is rare.

Why are so many of us prone to commit a base rate fallacy here? We apparently focus on the comparison between the relatively few false theories that are successful and the many false theories that are unsuccessful, while ignoring the comparison between the very few true theories and the many more false theories that are successful. Thus when Putnam memorably says that it would take a miracle for a false theory to make all those true predictions, this way of putting it directs our attention to the very low proportion of successful theories among false theories, but this has the effect of diverting our attention away from the low proportion of true theories among successful theories that is more to the point.

That philosophers of science should be so susceptible to this blindspot is particularly surprising, given how familiar we are with the problems of underdetermination of theory by data. Howson's explanation for our tendency to be impressed by the miracle argument is that we give the theory in question a high probability prior to its empirical successes (2000: 57). But this may not be the full story of our blindspot, or anyway the most perspicuous way of telling it. In the case of the test for the disease, the source of the medics' fallacy does not seem to be that they exaggerate the probability that their patient has the disease before running the test, but rather that they ignore that probability after the test. Medics act as if the test makes the prior probability irrelevant. Something similar may explain our tendency to be impressed by the miracle argument. It is not just that the successful theories to which the miracle argument is applied happen to be ones to which we gave a high prior. (We are, after all, impressed by the *form* of the miracle argument even when the theory in question is not specified.) What seems to be going on is rather that the miracle argument encourages us

to assess the reliability of empirical success as a test for truth by estimating its false positive rate (the chance that a false theory is successful), which we rightly judge to be low. We here ignore how incredibly unlikely it would be that, prior to testing, a given theory should be true. And this has the effect of hiding from our view all those other theories that would be just as successful even though they are false. Our tendency to think in explanationist terms may even contribute to this mistake since, as I suggested in chapter 8, the base rate does not affect the loveliness of the explanation, in this case of the truth explanation. How well the truth of our theory would explain its empirical successes is unaffected by the number of other theories whose truth would do the same. Our perception, however, is easily altered. Underdetermination arguments force us to face up to those nasty false but successful theories, and our intuitions flip. Here explanationist thinking may be operating on the other side, against the miracle argument, since the depressing base rate information is encoded by all those competing underdetermined explanations that block inference by denying the title of 'best' to the truth explanation.

Underdetermination, that adverse base rate, is the great bane of realists. In the case of first-order scientific inferences it is not immediately paralyzing, because we may, as Howson suggests, have given our successful theory a high prior probability, even if the miracle argument provides no justification for this practice. Indeed it is unclear whether the miracle argument really adds anything to the first-order evidence for the truth of our theories. That evidence is tantamount to the predictive success the truth explanation accounts for and, as we have seen, insofar as the truth explanation is distinguished from the first-order causal explanations it seems if anything a weaker basis for inference. But perhaps the miracle argument has one advantage, an advantage of conglomeration (cf. Psillos 1999: 79). For if truth were the best explanation of success, then the best explanation in turn of that would presumably be what I called 'privilege' in chapter 9, a general knack for discovering scientific truths. Thus our successes across the scientific board would support the hypothesis that our inductive practices are truth-tropic, information which would apply to each application of those methods. And might mean that our grounds for saying that a given theory is true extend beyond the proprietary evidence in its favor, being boosted by our other scientific successes. This, however, seems to me at best lukewarm comfort, given the other difficulties that we have found the miracle argument to face.

The scientific evidence

The miracle argument is an inference to the best explanation but one that is supposed to be distinct from the multifarious inferences to the best explanation that scientists make. Can explanationists defend realism instead by appeal to the structure of those first-order inferences? In particular, is there

any reason, based on the scientific evidence and the structure of inference, to prefer a realist version of Inference to the Best Explanation over some instrumentalist surrogate, such as van Fraassen's constructive empiricism? This is a large question I cannot adequately answer here, but I would like briefly to identify some of the realist's resources.

Let us focus on the causal inferences that I have emphasized in this book. Van Fraassen construes these inferences realistically when the inferred causes are observable, but not otherwise. Is there any reason to say instead that all causal inferences are inferences to actual existence of the causes, observable or not? Nancy Cartwright has argued that we must be realists about causal explanation, because these explanations 'have truth built into them' (1983: 91). To accept a causal explanation is to accept that the cause really exists so, insofar as we infer explanations that traffic in unobservable causes, we must also infer the existence of unobservable causes. The idea here is that a causal account only actually explains what it purports to explain if the causal story it tells is true. I agree, but then I am a realist and I find it very implausible to say that any false (and not even approximately true) explanation is an actual explanation, whether the explanation is causal or not. I do not think that you understand why something happens if the story you tell is fiction. Cartwright herself, however, holds that actual explanations need not be true, so long as they are not causal explanations (1983: essay 8), and she does not succeed in showing why someone like van Fraassen should feel compelled to agree that even causal explanations have truth built into them. Why not say instead that scientists tell causal stories that enable them to make accurate predictions, and that these explanations are actual explanations, so long as they are empirically adequate? Cartwright has not, so far as I can see, shown why the instrumentalist cannot have a non-realist model of causal explanation.

Cartwright's argument also has a peculiar structure. She takes van Fraassen to be challenging the realist to say what it is about the explanatory relation that makes the fact that something would explain a truth 'tend to guarantee' that the explanation is true as well. In the case of causal explanation, she claims, this challenge can be met, because we can only accept the explanation if we accept that the entities and processes it describes actually occur (1983: 4–5). But this seems a non sequitur. Even if only real causes can explain, this does not show why the fact that a putative cause would explain a truth tends to guarantee that it actually does.

There are, however, some other arguments for saying that someone who, like van Fraassen, construes inferences to observable causes realistically ought to do the same for unobservable causes. I will briefly consider three. The first is the 'same path, no divide' argument. As I have tried to show in earlier chapters, the structure of causal inference is the same, whether the cause is observable or not. Mill's description of the Method of Difference obscures this, since it suggests that we must have independent access to the

cause we infer, but Inference to the Best Explanation brings it out, by allowing inference to the existence of causes, as well as to their causal role. So there is a prima facie case for saying that all these inferences should be construed in the same way: granting the truth-tropism of inferences to observable causes, we ought also all to be realists about inferences to unobservable causes, since the inferences have the same form in both cases.

To resist this, the instrumentalist must claim that there is some principled epistemic distinction between inferences to the observable and to the unobservable, even if the paths are the same. On first hearing, this may sound plausible: the existence of the unobservable seems by its nature more speculative than the existence of the observable. This intuition might be strengthened by appeal to underdetermination. Only theories that traffic in unobservables have truth values underdetermined by all possible evidence. Nevertheless, the claim is misguided. The relevant distinction, if there is one, is between the observed and the unobserved, not between the observable and the unobservable. What counts for our actual epistemic situation is not ideal underdetermination by all possible evidence, but the much greater actual underdetermination by the evidence we now have. But neither the realist nor the instrumentalist is willing to abjure inferences to the truth of the unobserved, since this would make the predictive application of science impossible. To show that scientists are not entitled to infer unobservables, it would at least have to be shown why these inferences are all more precarious than inferences to the observable but unobserved, but no good reason has been given for this.

Inferences to the unobserved are risky but, if they are to the observable, there is at least in principle a way to determine whether they were successful. In cases of inferences to the unobservable, however, it might be claimed that we never can know if we were right, and it is this that makes such inferences intrinsically more speculative than inferences to the observable. Does this line of thought underwrite the epistemic relevance of ideal underdetermination? I think not. It is not clear why an inference we cannot check is therefore illegitimate, or even less likely to be correct than one we can check. Moreover, the realist need not concede that these inferences cannot be checked. To suppose that we cannot go on to test them begs the question. These tests will never prove that a claim about the unobservable is true, but neither will any observation prove that things really are as they appear to be.

Van Fraassen does not, however, generally argue that inferences to the truth of unobservable claims are always unwarranted, or even less well warranted than inferences to unobserved observables. Instead, he claims that they are unnecessary, on grounds of parsimony (1980: 72ff.). The realist and the constructive empiricist infer exactly the same claims about observables, but the realist also makes additional inferences about unobservables; these additional inferences are unnecessary to the scientist, so he should not make them. The more he infers, the greater risk he runs of being wrong, so he

should only make the inferences he needs. To this, I think the realist ought to give two replies, one easy, one more difficult. The easy one is that he is satisfied if the additional claims he wishes to infer are warranted, whether or not they are strictly required for scientific practice. He may add that, if they can be warranted, such inferences ought to be made, since science is, among other things, the attempt to discover the way the world works, and it is agreed on all sides that the mechanisms of the world are not entirely observable. The more difficult reply is that while the realist does stick his neck out further than the constructive empiricist, simply because he believes more, one of the things he gains is greater warrant than the constructive empiricist for the observable claims they share. If this is so, it is a telling argument for realism; but how could it be?

A scientist who is a constructive empiricist engages in two different types of inference. In the case of observable general claims, he infers the truth of the best explanation, from which he then goes on to deduce predictions. In the case of unobservable theories, however, he only infers the empirical adequacy of the best explanation, and he then also infers the truth of the predictions that follow from them. The realist, by contrast, always employs the first schema, whether the claims are observable or not. Of course the constructive empiricist can cover both his schemata under one description, by saying that he always infers the empirical adequacy of the best explanation, since observable claims are empirically adequate just in case they are true. This, however, leaves the difference, which is simply that, in the first case the lemmas on the road to prediction are believed to be true, while in the second case they are not. Similarly, the constructive empiricist really has two different models of actual explanation. For explanations that only appeal to observables, they must be true to be actual explanations, but for explanations that appeal to unobservables, they must only be empirically adequate. Thus, for observable explanation, all actual explanations must be logically compatible, while for unobservable explanation, actual explanations may be incompatible, though they must have compatible observable consequences. (What should the constructive empiricist say about the many scientific explanations that appeal to both observable and unobservable causes?)

I suggest that the constructive empiricist's bifurcation of both inference and explanation leave him with less support for his predictive claims from unobservable explanations than the realist has. Call this the 'transfer of support' argument. Consider first the case that the constructive empiricist and the realist share. What reason do they have for believing that the predictions they derive from observable causal explanations that are well supported by the evidence will be true? One reason, perhaps, is that this form of inference has been predictively successful in the past. But there is also another reason. We have found that our method of inferring causes has been a reliable way of discovering actual causes, and predictions deduced from information about actual causes tend to be correct. If the argument of this

book is correct, this method prominently includes inferences to the best contrastive explanation and, in the case of inferences to observable causes, we have good reasons to believe that the method is effective. In the cases of causes that are observed but not initially known to be causes, Inference to the Best Explanation inherits the plausibility of the Method of Difference, since in these cases, as we have seen, the two forms of inference have the same structure. In the cases of inferences to observable but unobserved causes, one signal reason we have to trust our method is that we often subsequently observe the causes it led us to infer. If this were not so, we would have substantially less confidence in this extension of the Method of Difference to unobserved (but observable) causes. I infer that the fuse is blown, because none of the lights or electrical appliances in the kitchen is working, and then I go into the basement and see the blown fuse. If we never had this sort of vindication of our causal inferences, we would have much less confidence in them. Fortunately, we enjoy this vindication all the time, and this is something to which both the realist and the constructive empiricist can appeal. But consider now what reason the constructive empiricist has for believing that the predictions he derives from an unobservable causal explanation that is well supported (in his sense) will be true. If he has any reason for this, it can only be that this form of inference to unobservable causes has been predictively successful in the past, a reason the realist has as well. The realist, however, has an additional reason to trust the predictions he generates from the unobservable causes he infers. His reasons for trusting his method in the case of observables also supply him with reasons for trusting his method in the case of unobservables, since it is the same method. In particular, just as his success in sometimes observing the causes he initially inferred supports his confidence in his method when he infers unobserved but observable causes, so it supports his confidence when he infers unobservable causes, because it gives him reason to believe that his method of inference is taking him to actual causes, from whose description the deductions of predictions will tend to be true.

The realist's justified confidence in his predictions comes, in part, from his justified confidence in the existence of the causes his theory postulates, and this confidence comes from his success in subsequently observing some of the causes Inference to the Best Explanation led him to infer. For him, the observed success of Inference to the Best Explanation in locating causes supports the application of that method to unobserved causes and the predictions we generate from their description, whether the causes are unobserved because we were not in the right place at the right time, or because we are constitutionally incapable of observing them. The constructive empiricist cannot transfer this support from the observable to the unobservable case, since he uses a different method of inference in the two cases, and since his method of prediction in the unobservable case does not travel through an inference to the existence of causes. The realist does

run a greater risk of believing a falsehood, since he believes so much more, but the benefit of his ambition is that he has better reason than the constructive empiricist to trust his predictions. This argument, if sound, hits the constructive empiricist where it matters most to him, in the grounds for believing claims about the observable but as yet unobserved.

What can the constructive empiricist say in reply? He might try to deny the transfer of support for the realist's method from observable to unobservable inference, but this would require an argument for an epistemic divide between inferences to unobserved observables and inferences to unobservables, an argument that I have suggested has not been provided. Alternatively, he might try to appropriate this transfer of support for his own method of generating predictions by means of unobservable theories. Crudely, he might pretend to be a realist until he makes his prediction, with all the confidence in it that the realist deserves, but then at the last minute cancel his belief in the theory itself. Perhaps this has all the advantages of theft over honest toil, but the accomplished thief does sometimes end up with his booty and out of jail. And this sort of strategy is often perfectly reasonable. I may believe something because it follows from a much larger claim I also believe, but if I can arrange a wager on the consequence alone, I would be silly to bet instead on the larger claim, on the same terms. But the realist need not deny this. It is enough for him to have shown that his confidence in the consequence depends in part on his having good reasons for believing the truth of the larger claim, even if he need not make practical use of that more ambitious inference. For he has shown that he would not have the degree of justified confidence he does have in his predictions, unless he also had good reason to believe his theory, which is what he set out to show.

I want to try out one more argument for realism and against instrumentalism or constructive empiricism, an argument for the initially surprising claim that we sometimes have more reason to believe a theory than we had to believe the evidence that supports it. This claim is not essential to scientific realism but, if it is acceptable, it makes such realism extremely attractive. We may call the argument the 'synergistic' argument. Recall that the 'same path, no divide' argument for realism was in part that the epistemic divide, if there is one, falls between the observed and the unobserved, not between the observed and the unobservable, which is where the instrumentalist needs it. But now I want to question whether all beliefs about unobserved observables are more precarious than beliefs about what is actually observed. One reason for thinking this false is the theory-ladenness of observation. We see with our theories and our dubitable ordinary beliefs, not just with our eyes, so our observational judgments presuppose the truth of claims that go beyond what we have actually observed. In some respects, these observational judgments are like inferences, which it may be possible to analyze as inferences to the best explanation, where the explanations we infer are explanations of our

experiences. As Mill remarked, 'in almost every act of our perceiving faculties, observation and inference are intimately blended. What we are said to observe is usually a compound result, of which one-tenth may be observation, and the remaining nine-tenths inference.' (1904: IV.I.2)

Observational judgments themselves have an inferential component, so the need for inference to form judgments about the unobserved but observable does not show there to be a principled epistemic divide between the observed and the unobserved. Mill goes on to make a further interesting claim:

> And hence, among other consequences, follows the seeming paradox that a general proposition collected from particulars is often more certainly true than any one of the particular propositions from which, by an act of induction, it was inferred. For each of those particular (or rather singular) propositions involved an inference from the impression on the senses to the fact which caused that impression; and this inference may have been erroneous in any one of the instances, but cannot well have been erroneous in all of them, provided their number was sufficient to eliminate chance. The conclusion, therefore, that is, the general proposition, may deserve more complete reliance than it would be safe to repose in any one of the inductive premises. (1904: IV.I.2)

This seems to me right: we may have more confidence in an inferred generalization than we initially had in the evidence upon which it was inferred. The evidence is never certain, and our justified confidence in it may change. After we infer the generalization, our confidence in each of our data will improve, since it will inherit additional support from the inferred generalization. Moreover, the same point may apply to theories more ambitious than simple generalizations, which appeal to unobservables. When a theory provides a unified explanation of many and diverse observational judgments, and there is no remotely plausible alternative explanation, we may have more confidence in the theory than we had in the conjunction of the evidence from which it was inferred. But if we may have more reason to believe a theory involving unobservables than we initially had for the observations upon which it is based, the claim that we have insufficient reason to infer the truth of a theory would be perverse. The instrumentalist must admit we have sufficient reason to accept our observational judgments, since he wants to say we are entitled to infer the predictions they support, so if we may have even more reason to infer a theory, that is reason enough.

It would be desirable to have detailed historical examples to help to make out the claim of the synergistic power of evidence, but I do not attempt this here. There are, however, simple and interesting cases where I think the claim is plausible. I am thinking of situations where we have reason to believe that the data were fudged, though not fabricated. One example may

be Gregor Mendel's pea experiments on the laws of inheritance; another may be Robert Millikan's oil-drop experiments on the discreteness and value of the charge of electrons. In both cases we have good reason to believe there was fudging, in Mendel's case because his data are statistically too good to be true, in Millikan's because we have his laboratory notebooks and so can see how selective he was in using his raw data. One might take the position that what this shows is that Mendel and Millikan actually had no good reason to believe their theories, but I think this is wrong. While each of their observational claims was highly dubitable, taken together they gave more support for their theories than the claims themselves enjoyed before the theories were inferred. In Millikan's case, in particular, his preference for the results on certain days would be unwarranted in the absence of his theory but, given that theory, he may have been reasonable in supposing that the 'bad' data were the result of irrelevant interference with an extremely delicate experiment.

This, then, is a sketch of three promising arguments for realism. Unlike the miracle argument, they appeal to the structure of scientific inferences to the best explanation, rather than to an overarching inference of the same form. They also differ from the miracle argument in having some force against forms of instrumentalism that grant that we have reason enough to accept both the truth of our observational judgments and the truth of the predictions of well supported theories, but deny that we have sufficient reason to infer the approximate truth of the theories themselves. The first argument, the 'same structure, no divide' argument, was that we can have the same sort of warrant for claims about unobserved causes as we have for observed causes, since the path of causal inference to the best explanation is the same in both cases and the instrumentalist has not made out a principled epistemic distinction between claims about observables and claims about unobservables. The second, the 'transfer of support' argument, was that only the realist can account for the actual degree of support observable predictions from theories enjoy, since only she can account for the support that the observed successes of inferences to the best explanation in locating observable causes provides for the application of the same form of inference to unobservable causes and so to the predictions we generate from their description. Finally, the 'synergistic' argument was that, because of the inferential structure implicit in our observational judgments and the unifying power of the best explanation, we may have more reason to believe a theory than we initially had for believing the data that supports it. I do not expect that the instrumentalists will now all collapse: instrumentalism, like other forms of skepticism, is a philosophical perennial, and most instrumentalists have already seen variants of the arguments I have just sketched, yet still stand by their positions. These arguments, however, may be enough to show that the failure of the miracle argument for realism, at least as an argument against anyone, does not show that all arguments for realism are bound to

beg. Moreover, they suggest that, even though the overarching inference to the best explanation upon which the truth explanation relies is too undiscriminating to make a strong case for realism, and even though there are non-realist versions of Inference to the Best Explanation, the structure of actual causal inferences that Inference to the Best Explanation illuminates shows this account of inference to be a friend to realism after all.

I treat my philosophical intuitions with considerable and perhaps excessive respect. If the intuition refuses to go away, even in the face of an apparently good argument against it, I either look for further arguments that will make it go away, or I find a way to defend it. Endorsing a philosophy I cannot believe does not interest me. This book is a case in point. Some years ago, I wrote a dissertation attacking Inference to the Best Explanation, but my belief that the account is fundamentally right would not go away, so I wrote a book defending it, and now a second edition in which I have attempted to develop and strengthen that defense. Or witness my heroic defense in the last chapter of the view that predictions are better than accommodations, in spite of the apparently strong arguments on the other side. I could not get rid of the belief that predictions are better, so I found an argument. But what about the miracle argument? In the first two sections of this chapter, we found that it is almost entirely without probative force. Yet I am left with the intuition that underlies it. If I were a scientist, and my theory explained extensive and varied evidence, and there was no alternative explanation that was nearly as lovely, I would find it irresistible to infer that my theory is approximately true. It would seem miraculous that the theory should have these explanatory successes, yet not have something importantly true about it. As Darwin said about his theory, 'It can hardly be supposed that a false theory would explain, in so satisfactory a manner as does the theory of natural selection, the several large classes of facts above specified' (Darwin 1859: 476). But this intuition does not depend on the miracle argument. It is not that the truth of the theory is the best explanation of its explanatory or predictive success; it is simply that the theory provides the best explanations of the phenomena that the evidence describes. We find inference compelling in such a case because we are creatures that judge likeliness on explanatory grounds. This is why Inference to the Best Explanation gives a good account of our actual inferential practices. This does not show that all arguments for scientific realism must beg the question, but it does suggest that, in the end, the best evidence for scientific realism is the scientific evidence, and the structure of the methods scientists use to draw their inferences from it.

Conclusion

This inquiry into Inference to the Best Explanation has been an attempt at both articulation and defense. I began by distinguishing the problem of description from the problem of justification and by making a case for the importance and autonomy of the descriptive project. The history of epistemology has been driven by skeptical arguments, and this has resulted in what is perhaps an excessive focus on justification and a relative neglect of the apparently more mundane project of principled description. I also urged a certain conception of the descriptive project, one that focuses on the deductive underdetermination of our inferences by the evidence available to us and attempts to discover the black box mechanisms that we actually employ to generate determinate inferences. The model of Inference to the Best Explanation is to be seen as a putative description of one of these mechanisms, perhaps a central one.

In my attempt to articulate and defend this conception of inference, I urged the importance of the distinction between potential and actual explanation. Inference to the Best Explanation must be Inference to the Best Potential Explanation, instances of which are inferences that the best potential explanation is an actual explanation. Without a conception of potential explanation, the account would not be epistemically effective and could not explain how explanatory considerations could be a guide to the truth; with it, we can see more clearly just what the model can and cannot provide by way of solutions to problems of justification. I also made the distinction between the explanation most warranted by the evidence – the likeliest explanation – and the explanation which would, if true, provide the most understanding – the loveliest explanation. The model tends to triviality if we understand 'best' as likeliest, since the sources of our judgments of likeliness are precisely what the model is supposed to illuminate. Inference to the Loveliest Explanation, by contrast, captures the central idea that the explanatory virtues are guides to inference, so I urged that we construe the model in this ambitious and interesting form. I do not maintain that Inference to the Best Explanation describes the only form of induction or that all other forms reduce to it, but I have tried to defend the claim that we do often use

how well a hypothesis would explain as a barometer of how likely it is that the hypothesis is correct.

To make out the central claim that explanatory considerations are a guide to inference requires identifying explanatory virtues and showing them to match inferential virtues. I attempted the identification and matching tasks primarily through my analysis of contrastive explanation, but also by appeal to the notions of causal mechanism, unification and the roles of background belief. The structural similarity between the Difference Condition on contrastive explanation and Mill's Method of Difference also enabled me to argue that Inference to the Best Explanation can be more than Inference to Likeliest Cause since, by looking for explanations that would exclude appropriate foils, we are led by explanatory considerations to what are in fact likeliest causes. This idea was developed with a detailed examination of the structure of Ignaz Semmelweis's research into the cause of childbed fever. That case strongly suggests not just that there is a good match between explanatory and inferential virtues, but that the former is a guide to the latter. The guiding claim was further defended by appeal to the way it gives a natural and unitary account of diverse inferential practices, the way it chimes with research in cognitive psychology that suggests that we rely extensively – sometimes too extensively – on explanatory considerations as a guide to inference, the way it makes sense of our disposition to think through inferential problems in causal rather than logical terms, and by the way it lends itself to an independently plausible two-stage mechanism involving the generation of candidate hypotheses and then a selection from among them.

Much of my work of defending Inference to the Best Explanation has itself been contrastive, consisting in developing favorable comparisons between that account and other familiar models of induction. Thus I argued that explanationism improves on the hypothetico-deductive model by avoiding various counterexamples (brought out for example by the raven paradox) and by bringing out aspects of inference about which hypothetico-deductivism is silent. I argued that it improves on Mill's method on grounds both of applicability and of scope. I have also in this edition compared Inference to the Best Explanation and Bayesianism, arguing that explanationism takes account of certain features of inference about which Bayesianism has little to say, such as the context of theory generation, and that explanatory considerations can be seen not as an alternative to Bayesianism but as a psychologically realistic way of realizing Bayesian calculation.

The guiding claim that lies at the heart of this book is a descriptive claim, a claim about how we actually make inferences; but I have also considered problems of justification. Hungerford's objection is that since explanatory beauty is in the eye of the beholder, loveliness cannot provide a suitably objective guide to inference. My reply to this challenge is that, on the one hand, there is substantial inferential variability and that, on the other, my contrastive analysis of explanation shows how what counts as a good

explanation can be genuinely interest relative without thereby being subjective in a sense that would make explanatory considerations unsuitable guides to inference. Voltaire's objection is that the model is too good to be true since, even if loveliness is objective, we have no reason to believe that we inhabit the loveliest of all possible worlds, no reason to believe that the explanation that would provide the most understanding is also the explanation that is likeliest to be true. I replied that in fact explanationism inherits the presumptive reliability of other approaches, especially Mill's methods and Bayesianism. At a Humean level of skepticism, of course, the success of any inductive policy is miraculous, but from this point of view Voltaire's objection does not succeed in showing that the explanationist model of inference is less plausible than any other. I also suggested that the match between explanatory and inferential considerations need not be fortuitous, since our explanatory standards are malleable and may have evolved so as to track the truth. Another justificatory challenge to Inference to the Best Explanation is that even if the loveliest of the generated hypotheses were also the likeliest of that group, we would have no warrant to infer that hypothesis, since we have no warrant to claim it likely that any of the hypotheses we have generated is correct. In reply I argue that such a gap between our powers of comparative and of absolute inductive evaluation turns out on examination not to be possible.

The case for Inference to the Best Explanation must of course also rest on its power to give a natural description of many and various sorts of inference we actually make. I hope that the examples scattered through the book help to show this. One of the main attractions of the model is that it accounts in a natural and unified way both for the inferences to unobservable entities and processes that characterize much scientific research and for many of the mundane inferences about middle sized dry goods that we make every day. It is also to its credit that the model gives a natural account of its own discovery, that the model may itself be the best available explanation of our inductive behavior since, as we have seen, that inference must itself be inductive and moreover an inference to a largely unobservable mechanism. Furthermore, the model offers a satisfying explanatory unification of our inductive and explanatory practices, and one that casts light on the point of explanation and the reason this activity occupies so large a part of our cognitive lives.

In the last two chapters of this book, I turned to the question of whether the explanationist model helps to solve some of the problems of the justification of induction. The results were mixed. It can be used to justify particular aspects of our inferential practices, such as our preference for predictions over accommodations; and the structure of scientific inference that it reveals can be used to provide some arguments for adopting a realist rather than an instrumentalist stance towards scientific theories. We found, however, that the miracle argument for realism – the argument that

predictively successful theories are likely to be approximately true since the truth of a theory is the best explanation of its predictive success – has serious liabilities, since it begs against the opponents of scientific realism and does not appear to provide a good enough explanation even to supply those who are already realists with much additional support for their position. Moreover, I do not see how Inference to the Best Explanation helps us to solve the big one, the Humean problem of induction. The model does offer a description of our inductive behavior that is importantly different from the simple extrapolation model that Hume assumed, and so perhaps undermines the theory of habit formation that Hume himself adopted as part of his 'skeptical solution', but his skeptical problem seems discouragingly insensitive to the particular description we favor. I hope to have more to say about this Humean predicament in future work.

I am acutely aware that this book has only scratched the surface of our inductive practices or even of one particular and partial model of those practices, both because of the hardness of the surface and the softness of my nails. It has included many arguments, undoubtedly of unequal value. Certainly much more needs to be said about what makes one explanation better than another and about which aspects of our inductive behavior the explanationist model can cover and which it cannot. Also, while I have defended the model, I also believe that it may have enjoyed rather more support than it really deserved, because it has remained for so long little more than an attractive slogan, in part because of the general neglect to distinguish clearly between the plausible but relatively banal view that we make inductions by means of inferences to the likeliest explanation and the exciting but by no means obviously correct view that we employ inferences to the loveliest explanation. Nevertheless, I hope that I have said enough to convince some readers that this exciting view merits further development. I also take some comfort in the otherwise discouraging fact that an account of our inductive practices does not have to be very good to be the best we now have.

Bibliography

Achinstein, P. (1992) 'Inference to the Best Explanation: Or, Who Won the Mill–Whewell Debate?', *Studies in the History and Philosophy of Science*, 23, 349–64.

Ajzen, I. (1977) 'Intuitive Theories of Events and the Effects of Base-Rate on Prediction', *Journal of Personality and Social Psychology*, 35, 303–14.

Austin, J. L. (1962) *Sense and Sensibilia*, Oxford: Oxford University Press.

Ayer, A. J. (1956) *The Problem of Knowledge*, Harmondsworth: Penguin.

Barnes, E. (1994) 'Why P rather than Q? The Curiosities of Fact and Foil', *Philosophical Studies*, 73, 35–55.

Barnes, E. (1995) 'Inference to the Loveliest Explanation', *Synthese*, 103, 251–77.

Ben-Menahem, Y. (1990) 'The Inference to the Best Explanation', *Erkenntnis*, 33, 319–44.

Bird, A. (1998) *Philosophy of Science*, London: UCL Press.

Black, M. (1970) 'Induction and Experience', in L. Foster and J. W. Swanson (eds) *Experience and Theory*, London: Duckworth, 135–60.

Boyd, R. (1985) '*Lex Orandi est Lex Credendi*', in P. Churchland and C. Hooker (eds) *Images of Science*, Chicago: University of Chicago Press, 3–34.

Braithwaite, R. B. (1953) *Scientific Explanation*, Cambridge: Cambridge University Press.

Brody, B. (1970) 'Confirmation and Explanation', in B. Brody (ed.) *Readings in the Philosophy of Science*, Englewood Cliffs: Prentice-Hall, 410–26.

Bromberger, S. (1966) 'Why-Questions', in R. G. Colodny (ed.) *Mind and Cosmos*, Pittsburgh: University of Pittsburgh Press, 86–111.

Campbell, D. (1974) 'Evolutionary Epistemology', in P. A. Schilpp (ed.) *The Philosophy of Karl Popper*, LaSalle: Open Court, 413–63.

Carroll, J. (1997) 'Lipton on Compatible Contrasts', *Analysis*, 57, 3, 170–8.

Carroll, J. (1999) 'The Two Dams and That Damned Paresis', *British Journal for the Philosophy of Science*, 50, 65–81.

Carter, K. C. and Carter, B. R. (1994) *Childbed Fever: A Scientific Biography of Ignaz Semmelweis*, Westport: Greenwood Press.

Cartwright, N. (1983) *How the Laws of Physics Lie*, Oxford: Oxford University Press.

Casscells, W., Schoenberger, A. and Grayboys, T. (1978) 'Interpretation by Physicians of Clinical Laboratory Results', *New England Journal of Medicine*, 299, 999–1000.

Chihara, C. S. (1987) 'Some Problems for Bayesian Confirmation Theory', *British Journal for the Philosophy of Science*, 38, 551–60.

Chomsky, N. (1965) *Aspects of the Theory of Syntax*, Cambridge, MA: MIT Press.
Chomsky, N. (1986) *Knowledge of Language*, New York: Praeger.
Churchland, P. M. and Hooker, C. A. (eds) (1985) *Images of Science*, Chicago: University of Chicago Press.
Clarke, S. (2001) 'Defensible Territory for Entity Realism', *British Journal for the Philosophy of Science*, 54, 701–22.
Cummins, R. (1989) *Meaning and Mental Representation*, Cambridge, MA: MIT Press.
Curd, M. and Cover, J.A. (1998) 'Lipton on the Problem of Induction', in M. Curd and J. A. Cover (eds) *Philosophy of Science: The Central Issues*, New York: Norton, 495–505.
Darwin, C. (1859) *On the Origin of Species*, reprinted 1961, New York: Collier.
Day, T. and Kincaid, H. (1994) 'Putting Inference to the Best Explanation in its Place', *Synthese*, 98, 271–95.
Descartes, R. (1641) *Meditations on First Philosophy*, Donald Cress (trans.), Indianapolis: Hackett (1979).
Dreske, F. (1972) 'Contrastive Statements', *Philosophical Review*, 82, 411–37.
Earman, J. (1992) *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*, Cambridge, MA: MIT Press.
Fine, A. (1984) 'The Natural Ontological Attitude', in J. Leplin (ed.) *Scientific Realism*, Berkeley: University of California Press, 83–107.
Friedman, M. (1974) 'Explanation and Scientific Understanding', *Journal of Philosophy*, LXXI, 1–19.
Garfinkel, A. (1981) *Forms of Explanation*, New Haven: Yale University Press.
Gettier, E. (1963) 'Is Justified True Belief Knowledge?', *Analysis*, 23, 121–3.
Giere, R. (1988) *Explaining Science: A Cognitive Approach*, Chicago: University of Chicago Press.
Gigerenzer, G., Todd, P. and the ABC Research Group (2000) *Simple Heuristics that Make Us Smart*, New York: Oxford University Press.
Gilovich, T., Griffin, D. and Kahneman, D. (eds) (2002) *Heuristics and Biases: The Psychology of Intuitive Judgment*, Cambridge: Cambridge University Press.
Gjertsen, D. (1989) *Science and Philosophy: Past and Present*, Harmondsworth: Penguin.
Glymour, C. (1980a) 'Explanations, Tests, Unity and Necessity', *Nous*, 14, 31–50.
Glymour, C. (1980b) *Theory and Evidence*, Princeton: Princeton University Press.
Goodman, N. (1983) *Fact, Fiction and Forecast*, 4th ed., Indianapolis: Bobbs-Merrill.
Grossner, M. (1970) 'Adams, John Couch', in C. G. Gillespie (ed.) *Dictionary of Scientific Biography*, New York: Charles Scribner's Sons, 53–4.
Hacking, I. (1982) 'Language, Truth and Reason', in M. Hollis and S. Lukes (eds) *Rationality and Relativism*, Oxford: Blackwell, 48–66.
Hacking, I. (1992) ' "Style" for Historians and Philosophers', *Studies in the History and Philosophy of Science*, 23, 1–20.
Hanson, N. R. (1972) *Patterns of Discovery*, Cambridge: Cambridge University Press.
Harman, G. (1965) 'The Inference to the Best Explanation', *Philosophical Review*, 74, 88–95.

Harman, G. (1973) *Thought*, Princeton: Princeton University Press.
Harman, G. (1986) *Change in View*, Cambridge, MA: MIT Press.
Harman, G. (1999) *Reasoning, Meaning, and Mind*, Oxford: Oxford University Press.
Hempel, C. (1965) *Aspects of Scientific Explanation*, New York: Free Press.
Hempel, C. (1966) *The Philosophy of Natural Science*, Englewood Cliffs: Prentice-Hall.
Hitchcock, C. (1999) 'Contrastive Explanation and the Demons of Determinism', *British Journal for the Philosophy of Science*, 50, 585–612.
Horwich, P. (1982) *Probability and Evidence*, Cambridge: Cambridge University Press.
Howson, C. (2000) *Hume's Problem: Induction and the Justification of Belief*, Oxford: Oxford University Press.
Howson, C. and Urbach, P. (1989) *Scientific Reasoning: The Bayesian Approach*, LaSalle: Open Court.
Hume, D. (1739) *A Treatise of Human Nature*, D. F. and M. J. Norton (eds), Oxford: Oxford University Press (2000).
Hume, D. (1748) *An Enquiry Concerning Human Understanding*, T. Beauchamp (ed.), Oxford: Oxford University Press (1999).
James, W. (1897) 'The Will to Believe', in his *The Will to Believe, and Other Essays in Popular Philosophy*, New York and London: Longmans, Green & Co., 1–31.
Jevons, W. S. (1877) *Elementary Lessons in Logic*, 6th ed., London: Macmillan.
Jones, M. (1994) 'Comments on "Contrastive Why Questions" by Eric Barnes', American Philosophical Association Central Division Meeting, Kansas City.
Kahneman, D., Slovic, P. and Tversky, A. (eds) (1982) *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press.
Kitcher, P. (1989) 'Explanatory Unification and the Causal Structure of the World', in P. Kitcher and W. Salmon (eds) *Scientific Explanation*, Vol. 13, *Minnesota Studies in the Philosophy of Science*, Minneapolis: University of Minnesota Press, 410–505.
Kitcher, P. and Salmon, W. (eds) (1989) *Scientific Explanation*, Vol. 13, *Minnesota Studies in the Philosophy of Science*, Minneapolis: University of Minnesota Press.
Kuhn, T. (1970) *The Structure of Scientific Revolutions*, 2nd ed., Chicago: University of Chicago Press.
Kuhn, T. (1977) *The Essential Tension*, Chicago: University of Chicago Press.
Langer, E. J. (1975) 'The Illusion of Control', *Journal of Personality and Social Psychology*, 32, 311–28.
Laudan, L. (1984) 'A Confutation of Convergent Realism', in J. Leplin (ed.) *Scientific Realism*, Berkeley: University of California Press, 218–49.
Lepin, J. (ed.) (1984) *Scientific Realism*, Berkeley: University of California Press.
Lewis, D. (1986) 'Causal Explanation', in *Philosophical Papers*, Vol. II, New York: Oxford University Press, 214–40.
Lipton, P. (1987) 'A Real Contrast', *Analysis*, 47, 207–8.
Lipton, P. (1990) 'Prediction and Prejudice', *International Studies in the Philosophy of Science*, 4, 1, 51–65.
Lipton, P. (1991) 'Contrastive Explanation', in D. Knowles (ed.) *Explanation and its Limits*, Cambridge: Cambridge University Press, 247–66.
Lipton, P. (1993a) 'Making a Difference', *Philosophica*, 51, 39–54.

Lipton, P. (1993b) 'Is the Best Good Enough?', *Proceedings of the Aristotelian Society*, XCIII, 89–104.
Lipton, P. (2000) 'Tracking Track Records', *The Aristotelian Society*, Supplementary Volume LXXIV, 179–206.
Lipton, P. (2001) 'Is Explanation a Guide to Inference?', in G. Hon and S. S. Rakover (eds) *Explanation: Theoretical Approaches and Applications*, Dordrecht: Kluwer, 93–120.
Lipton, P. (forthcoming) *The Humean Predicament*, Cambridge: Cambridge University Press.
Mackie, J. L. (1974) *The Cement of the Universe*, Oxford: Oxford University Press.
Maher, P. (1988) 'Prediction, Accommodation, and the Logic of Discovery', in A. Fine and J. Lepin (eds) *Philosophy of Science Association 1988*, Vol. 1, East Lansing: Philosophy of Science Association, 273–85.
Mill, J. S. (1904) *A System of Logic*, 8th ed., London: Longmans, Green & Co.
Newton-Smith, W. H. (1981) *The Rationality of Science*, London: Routledge.
Niiniluoto, I. (1999) 'Defending Abduction', *Philosophy of Science*, 66 (Proceedings), S436–51.
Nisbet, R. E. and Ross, L. (1980) *Human Inference: Strategies and Shortcomings of Social Judgment*, Englewood Cliffs, NJ: Prentice-Hall.
Nozick, R. (1974) *Anarchy, State, and Utopia*, New York: Basic Books.
Nozick, R. (1981) *Philosophical Explanations*, Cambridge: Harvard University Press.
Nozick, R. (1983) 'Simplicity as Fall-Out', in L. Cauman *et al.* (eds) *How Many Questions?*, Indianapolis: Hackett, 105–19.
Okasha, S. (2000) 'Van Fraassen's Critique of Inference to the Best Explanation', *Studies in the History and Philosophy of Science*, 31, 691–710.
Peirce, C. S. (1931) *Collected Papers*, C. Hartshorn and P. Weiss (eds), Cambridge, MA: Harvard University Press.
Popper, K. (1959) *The Logic of Scientific Discovery*, London: Hutchinson.
Popper, K. (1962) *Conjectures and Refutations*, London: Routledge & Kegan Paul.
Popper, K. (1972) *Objective Knowledge*, Oxford: Oxford University Press.
Psillos, S. (1999) *Scientific Realism: How Science Tracks Truth*, London: Routledge.
Psillos, S. (2002) 'Simply the Best: A Case for Abduction', in A. C. Kakas and F. Sadri (eds) *Computational Logic: Logic Programming and Beyond 2002*, Berlin: Springer-Verlag, 605–26.
Psillos, S. (2003) 'Inference to the Best Explanation and Bayesianism', in F. Stadler (ed.) *Institute of Vienna Circle Yearbook* 10, Dordrecht: Kluwer.
Putnam, H. (1975) *Mathematics, Matter, and Method*, Cambridge: Cambridge University Press.
Putnam, H. (1978) *Meaning and the Moral Sciences*, London: Hutchinson.
Quine, W. v. O. (1951) 'Two Dogmas of Empiricism', *Philosophical Review*, 60, 20–43.
Quine, W. v. O. (1969) 'Natural Kinds', in his *Ontological Relativity*, New York: Columbia University Press, 114–38.
Quine, W. v. O. and Ullian, J. (1978) *The Web of Belief*, 2nd ed., New York: Random House.
Rappaport, S. (1996) 'Inference to the Best Explanation: Is it Really Different from Mill's Methods?', *Philosophy of Science*, 63, 65–80.

Ruben, D. (1987) 'Explaining Contrastive Facts', *Analysis*, 47, 1, 35–7.
Ruben, D. (1990) *Explaining Explanation*, London: Routledge.
Salmon, W. (2001a) 'Explanation and Confirmation: A Bayesian Critique of *Inference to the Best Explanation*', in G. Hon and S. S. Rakover (eds) *Explanation: Theoretical Approaches and Applications*, Dordrecht: Kluwer, 61–91.
Salmon, W. (2001b) 'Reflections of a Bashful Bayesian: A Reply to Peter Lipton', in G. Hon and S. S. Rakover (eds) *Explanation: Theoretical Approaches and Applications*, Dordrecht: Kluwer, 121–36.
Schlesinger, G. (1987) 'Accommodation and Prediction', *Australasian Journal of Philosophy*, 65, 33–42.
Scriven, M. (1959) 'Explanation and Prediction in Evolutionary Theory', *Science*, 130, 477–82.
Semmelweis, I. P. (1860) *The Etiology, Concept, and Prophylaxis of Childbed Fever*, K.C. Carter (trans.), Madison: University of Wisconsin Press (1983).
Sklar, L. (1985) *Philosophy and Spacetime Physics*, Berkeley: University of California Press.
Skyrms, B. (1986) *Choice and Chance*, 3rd ed., Belmont: Wadsworth.
Sober, E. (1986) 'Explanatory Presupposition', *Australasian Journal of Philosophy*, 64, 2, 143–9.
Sosa, E. and Tooley, M. (eds) (1993) *Causation*, Oxford: Oxford University Press.
Stein, E. and Lipton, P. (1989) 'Evolutionary Epistemology and the Anomaly of Guided Variation', *Biology and Philosophy*, 4, 33–56.
Temple, D. (1988) 'The Contrast Theory of Why-Questions', *Philosophy of Science*, 55, 1, 141–51.
Thagard, P. (1978) 'The Best Explanation: Criteria for Theory Choice', *Journal of Philosophy*, 75, 76–92.
Tversky, A. and Kahneman, D. (1984) 'Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgement', *Psychological Review*, 91, 293–315; reprinted in T. Gilovich, D. Griffin and D. Kahneman 2002: 19–48.
Urbach, P. (1985) 'Randomization and the Design of Experiments', *Philosophy of Science*, 52, 256–73.
van Fraassen, B. C. (1980) *The Scientific Image*, Oxford: Oxford University Press.
van Fraassen, B. C. (1989) *Laws and Symmetry*, Oxford: Oxford University Press.
Vogel, J. (1990) 'Cartesian Skepticism and Inference to the Best Explanation', *The Journal of Philosophy*, 87, 658–66.
Williams, B. (1978) *Descartes*, Harmondsworth: Penguin.
Woodward, J. (1984) 'A Theory of Singular Causal Explanation', *Erkenntnis*, 21, 231–62.
Worrall, J. (1984) 'An Unreal Image', *British Journal for the Philosophy of Science*, 35, 1, 65–80.

Index