Inference to the Best Explanation in Science

David Walker

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Doctor of Philosophy in the Faculty of Arts, Department of Philosophy, September 2008.

Word count: 78114

I declare that the work in this dissertation was carried out in accordance with the regulations of the University of Bristol. The work is original, except where indicated by special reference in the text, and no part of the dissertation has been submitted for any other academic award. Any views expressed in the dissertation are those of the author.

SIGNED: David Walker DATE: 30/3/09

Abstract

This thesis defends inference to the best explanation (IBE) by giving an account of explanatory 'loveliness' in science. I begin by presenting IBE in generic form and showing how it out-performs rival accounts of induction. I then trace a path through the early literature which emphasises the role of background belief in determining loveliness. I then introduce crucial features of Lipton's account of IBE. I argue that Lipton's remarks on loveliness, though minimal, support the background-dependent view and that, appropriately construed, the view does not trivialise IBE.

I then consider 'Hungerford's objection', that loveliness is too subjective to guide inference, and 'Voltaire's objection', that loveliness is not a guide to truth. Finding Lipton's responses inadequate, I introduce Kuhn's account of science as a way to do better. I argue that the background theory of Kuhnian science, exemplars specifically, provides scientists with a standard of loveliness. Exemplars are endorsed by the entire scientific community; thus loveliness is not subjective. The reliabilist interpretation of Kuhn shows that exemplars approach the truth; thus loveliness is truth-tropic. Hungerford's objection reappears, arguing that loveliness is now relative to paradigms. I argue that in fact it's relative to puzzle-solving context, which is unproblematic. I then present an independent case for Kuhnian IBE drawing on work by McMullin on the Copernican revolution and McAllister on aesthetics in Kuhnian science.

I next consider Psillos' presentation of the 'no miracles' argument (NMA) for scientific realism. Against Psillos, I argue that Maxwell's version of the NMA is an IBE. I then outline the definitive Boyd–Psillos version, its naturalist background, and the place within it of Kuhnian IBE. I endorse Psillos' defence of the NMA against two central objections, but come to a slightly different conclusion about the argument's status. I close with some speculative remarks about naturalism and loveliness.

Acknowledgements

Thanks are due first and foremost to my supervisor, Alexander Bird. If there are any good ideas in this thesis, they are greatly inspired by his advice and criticism, if not borrowed wholesale. I am grateful to have had such a helpful mentor and ruthless editor.

I would also like to thank in advance my examiners, Samir Okasha and Stathis Psillos. I apologise to them both for my inevitable waffling and for my use of endnotes (a last-minute technical glitch made footnotes unavailable).

I acknowledge the assistance of a University of Bristol Postgraduate Arts Faculty Scholarship which allowed me to undertake the study of which this thesis is the result.

The content herein has benefited from discussion with many people, but none have been more generous than Joe Morrison and Huginn Thorsteinsson. I thank them both, and I also thank Zoe Drayson, Simone Duca, Ellen Clarke, Guillaume Rochefort-Maranda, Chris Gifford and Robin Brown, fellow PhD students with whom I have talked about and forgotten about philosophy on many occasions. All of them have made my time at Bristol considerably more enjoyable.

Parts of chapter 3 were presented at a departmental work-in-progress seminar in January 2007 and at a postgraduate seminar the same month; I thank all staff and postgraduates who contributed to the subsequent discussions.

Over the last three years, I have accompanied my study with undergraduate seminar teaching. I extend my thanks to those students who regularly made it such good fun. Teaching also made me realise what I knew already, that an academic department cannot run without its secretaries. Duly, I thank Susan Frost and Debra Hughes, whose assistance has been invaluable on so many occasions.

For making inference to the best explanation such an engaging and accessible topic to study, I thank the late Peter Lipton, whom I am disappointed never to have met.

For diverting me from my work, I thank Dan Gilbert, Lucy Davies, Sam Marsh, Rich Hayton, Andy Gee, Anna Hughes, Tom Abdy, Bella Smith, Anna Jones, James Scott, Hannah Taggart, Jon Webber, Suzi Wells, Anthony Everett, Seiriol Morgan, David Harris, Sean Cordell, the Paramount Comedy channel, my Fender DG-3 acoustic and more CDs than I care to mention.

For unfailing love and support I thank my mother Sue Walker, my grandparents Irene and Peter Williams, and Hannah Neate.

This thesis is dedicated to my mother and my father, Bob Walker (1946-2004).

Table of contents

Introduction

On 20th June 2008, the Guardian newspaper reported that the Office for National Statistics had recorded the biggest monthly rise in retail sales since comparable records began in 1986. Retailers confirmed that May 2008 had seen shoppers spending generously on luxury items such as garden furniture and barbecues, summer fashions and a variety of electronic goods. The surprise was even greater since every other indicator suggested the economy was in dire straits. The paper's economics editor Larry Elliott identified four possible explanations:

1. **The economic woes are overstated**: retailers in particular had been reporting relatively healthy figures despite the overall downturn.
2. **The figures were misinterpreted**: officials had analysed and re-analysed the data, but some thought that seasonal adjustment or rising inflation had not been accounted for.
3. **The warm weather**: a mini-heatwave during May brought shoppers out, and encouraged spending on summer clothes and entertainment.
4. **The last hurrah**: consumers were determined to have some fun before the shrinking economy forced them to tighten their belts.

Explanation 1 was discounted, since there was too much other evidence suggesting the economy really was in trouble. Explanation 2 was also eliminated: there's always the chance that complex data will mislead, but the ONS employs experts, uses sophisticated methods, and has a track-record of reliability. Explanation 3 was plausible, but the May figures showed an even bigger increase in spending than is normal for a late-spring heatwave. Explanation 4 was the best explanation. It accounted for the rise in a way compatible with other contemporary evidence. It employed a simple and familiar psychological phenomenon that made the data understandable, and integrated them into a wider pattern of socio-economic activity. It also fitted with other things economists knew about similar events in previous years. Thus they inferred that explanation 4 really does explain the rise in retail sales.

According to one account of how we reason, this example is typical. We encounter some suggestive phenomenon, consider a number of potential

explanations, select the best, and infer it. That account is called *inference to the best explanation.*

This thesis defends inference to the best explanation (IBE). I endorse IBE as a general account of inductive inference, but my specific focus is on IBE in science. The central claim of IBE is that explanatory virtues – features that make for good explanations, and make one explanation better than another – guide inference. The main motivation behind my defence is the need to account for the way in which this happens. There have been various attempts to do this before, but most have dealt with isolated virtues, abstracted from their scientific context, on the one hand, and poorly-articulated or generic accounts of science on the other. Such efforts may be illuminating, but they are incomplete to the extent that they fail to show how, when returned to their natural habitat, the explanatory virtues studied are actually put to work. In a nutshell, some accounts succeed in showing *that* certain explanatory virtues guide scientific inference, but few, if any, have succeeded in showing *how* they do so. The centrepiece of my defence of IBE tries to remedy this defect. It takes the most fully-developed account of IBE, due to Peter Lipton, and the most detailed account of the structure of science, due to Thomas Kuhn, and explains how the former works in the context of the latter. I do not focus on any particular explanatory virtues; rather I try to show that IBE offers the best account of scientific inference whatever the explanatory standards of science may be.

Our discussion begins in a traditionally philosophical way: with a distinction. Inference can be deductive or inductive, but what's the difference? Chapter 1 begins by suggesting it's far from clear-cut. Happily, this doesn't hold us back, and I proceed to survey three attempts to account for (what we may legitimately call) induction that compete with IBE: enumerative induction, hypothetico-deductivism and Bayesianism. All are found wanting since they fail to account for basic features of the way we respond to evidence and the way evidence supports our responses. The following section introduces IBE in schematic form and shows how it succeeds where its competitors failed. In particular, it shows that IBE can account for when we make an inference from certain evidence and when we refrain from inferring. This is crucial for an account of induction; we are highly sensitive to those features of our situation that tell us when an inference is justified. Simply having gathered some evidence is not enough; we are very selective and do not just infer willy-nilly. Further, IBE can account for the kind of inference we make: explanatory considerations reveal why

inference is sometimes cautious, sometimes bold, and justifiably so. Chapter 1 then considers the role of Hume's problem, before taking a whirlwind tour of some early landmarks in the development of IBE. These give the account further motivation but their main purpose is to bring out those parts of the IBE debate that dominate the subsequent discussion. Chief among these is the role of background belief, or in scientific terms, background theory, in determining the nature and function of explanatory virtues.

Chapter 2 concentrates on Lipton's account of IBE. Lipton's contribution to the IBE discussion cannot be overstated. His ideas on IBE are known throughout epistemology and philosophy of science, almost always endorsed as definitive of IBE, and among supporters of IBE, endorsed full-stop. But Lipton's account is seldom subjected to any kind of holistic scrutiny, resulting in a widespread lack of understanding as to what IBE really amounts to. This ignorance generates both misguided criticism from its detractors and a degree of blind faith from its proponents. In order that this thesis doesn't follow suit, chapter 2 isolates Lipton's key positive claims about the inferential procedure of IBE: the two-stage process of generating hypotheses and selecting from among them, the distinction between likeliness and loveliness, and the crucial claim that loveliness is a guide to likeliness. A hypothesis' loveliness is its degree of explanatory virtue, or the amount of understanding it would create, if true. Lipton stresses that loveliness is IBE's central notion; without the claim that loveliness is a guide to likeliness IBE becomes trivial and loses its identity. Thus chapter 2 looks closely at the essential role of loveliness at both the generation and selection stages.

Lipton's remarks on loveliness return us to the issue of the inferential background. As chapter 1 suggested, fit with background knowledge isn't simply one criterion of loveliness; it also operates on a meta-level: a hypothesis must 'fit with background' in the sense that background knowledge determines the meaning of other loveliness criteria and influences their application. It's this feature of loveliness that allows IBE to have the kind of sensitivity to context that enables it to account so effectively for our actual inductive behaviour. In fleshing out Lipton's comments to emphasise the role of background knowledge, I don't propose a reductive analysis of loveliness; criteria of loveliness may still be fruitfully individuated and investigated on their own terms, although as noted above, it's not a task I undertake. A further worry, that by revealing the intimate connection between loveliness and background we

reduce IBE to triviality, is dealt with by considering two interesting criticisms of Lipton's position, due to Eric Barnes.

Lipton's claim is not only that we judge likeliness on grounds of loveliness, but also that this method is reliable. Thus loveliness has some connection with truth. Lipton leaves this connection largely unanalysed, meaning IBE lacks crucial normative force. This is no big criticism: Lipton's descriptive bias is remarkably productive, and he is entirely justified in his reluctance to identify explanatory virtues 'one by one' and show how they match up with truth. But it is a deficit worth narrowing, and I attempt to do so in chapter 3. My focus is on what Lipton calls Hungerford's objection, that loveliness is too subjective to guide inference, and Voltaire's objection, that loveliness is not a guide to truth. I note that although Lipton's responses remove some force from both objections, his descriptive approach neglects their normative clout. I answer both objections by considering Kuhn's account of science, and in doing so tackle the normative issue head-on.

Kuhn fully realises the fundamental role played by well-established theories in the practice of science, scientific inference in particular. These theories Kuhn calls exemplars, and he describes them (and their various explanatory applications) as forming the kind of theoretical background apt to determine scientific standards of loveliness: they generate understanding and provide exemplary solutions to scientific problems, against which others should be judged. The place of exemplars within the Kuhnian structure gives a natural solution to Hungerford's objection and, once we've separated that structure from its anti-realist associations, Voltaire's objection too. Appealing to the Kuhnian structure generates a new version of Hungerford's objection: loveliness is relative to certain scientific communities. Far from fatal, this objection actually illuminates the way loveliness works in science. The latter sections of chapter 3 give arguments for thinking IBE particularly plausible as an account of Kuhnian inference. Especially supportive here is work by Ernan McMullin and James McAllister, based on analysis of historical episodes of theory choice in science.

Chapters 4 and 5 change tack slightly and look at the philosophical use of IBE as an argument for scientific realism, the view that science aims at truth; but the issues are not unrelated to what's gone before. This is clearest when chapter 4 looks at the influential realist views of Richard Boyd and Stathis Psillos. Their arguments for realism depend on a general philosophical approach – naturalism – that my development of IBE supports. Further, the account of science generated by their

approach bolsters my claims about IBE; in turn, my claims add detail to their view and demonstrate its compatibility with Lipton's work. Boyd and Psillos use IBE to propose the 'no miracles' argument for scientific realism, the essence of which is that science's manifest empirical success is only explicable if we assume that its theories are, in some moderate sense, true. Grover Maxwell is credited with being the first to argue this; I devote a section of chapter 4 to arguing not only that Maxwell provides a more fully-fledged precedent to the modern view than many realise, but also that in terms of IBE, his argument has virtues that contemporary realists neglect. However, the chapter begins by arguing that framing the realism/anti-realism dispute in terms of explanation, and using IBE to justify the acceptance of realism, realists ignite a debate that only they can win.

This theme recurs in chapter 5, which considers the objection that realism doesn't offer the best explanation of the success of science after all, alongside the objection that the no miracles argument is circular, using the very form of inference it sets out to justify. Having endorsed Psillos' account of how the naturalist epistemological stance makes circularity unproblematic, I introduce Lipton's critique of the no miracles argument to bring out its real status as part of the defence of realism. The argument, it turns out, isn't much of an argument at all, in that it cannot convince non-realists of realism. But this doesn't mean it's useless. On the contrary, the no miracles argument demonstrates the many benefits of the philosophical stance that generates it. The thesis closes with a section drawing out some consequences of the combination of naturalism and the account of scientific IBE given in chapter 3. There are some speculative conclusions about IBE in philosophy, but the core message is that despite the failure of the no miracles argument, the naturalist, realist supporter of Kuhnian IBE is in a position of enviable strength.

Chapter 1

Inference to the best explanation

1. Introduction: deduction and induction

Inference is the practice of justifying and accepting claims on the basis of other claims. Sometimes the claims are explicit: seeing the notice that the library is closed on weekends, I infer that I won't be able to borrow a book this Saturday. Sometimes the claims are more implicit: seeing your frown and fixed gaze I infer you are concentrating. Sometimes it's not even clear that a *claim* prompts the inference at all: seeing the darkening sky, I infer it's about to rain. Sticking with our initial characterisation, there are two types of inference: deductive and inductive (here I use 'inductive' and related terms in the broad sense to describe all non-deductive inference). It is common for philosophers to describe deductive inference in the following ways: the premises entail the conclusion or make it necessary; it is impossible for the premises to be true and the conclusion false; deductive conclusions are absolutely certain, or proven. Correspondingly, it is common to see (good) inductive inference described in the following terms: the premises support the conclusion, or give reason to believe it; it is possible for the premises to be true and the conclusion false; inductive conclusions are more probable in the light of their premises.

This makes deduction and induction look straightforwardly distinct, but in fact it's not so clear. There are at least four examples that show the line between deduction and induction to be far more blurred than the orthodox view suggests. I consider them here not only to allege that philosophical tradition is often misguided, but also to emphasise that it's wrong to judge accounts of induction against deductive standards (a temptation that remains strong even though it is often warned against by modern writers). Quite apart from being an inappropriate measure – induction and deduction are complementary, used in different circumstances, to do different jobs – we had better not start out expecting induction to 'measure up' to deduction if we aren't clear about what deduction actually is. The view that we are more certain of

what deduction is, and that it is in some way preferable to induction because it delivers more reliable conclusions, should thus be jettisoned straight away.

The first example that suggests deduction and induction aren't clearly distinct is this. It's a well-known point of logic that every necessary proposition entails every other. In an argument constructed out of random necessary propositions, it is impossible for the premises to be true and the conclusion false. We might call such an argument 'deductive', but it would be far from making its conclusion certain. We might be certain of the conclusion (it is a necessary proposition, after all), but the argument would not have brought this about. It would not be persuasive; in particular, we would not feel compelled to accept the conclusion in the way we would if exposed to a well-constructed inductive argument, even though such arguments are supposed to confer less support on their conclusions than deductive ones.

A second example shows that deductive structure or rules are not sufficient either. We might feed premises from the more speculative end of science into a deductive argument to see what they entail, and feel justifiably confident that if the premises are true then the conclusion is true. But as a matter of fact the premises are uncertain, so the conclusion is at least as uncertain. Further, subsequent scientific research might confirm the conclusion and reveal a falsehood amongst the premises. What's deductive about the argument now? Further doubt is cast on the orthodox distinction by our third example: complex deductive arguments. Complex mathematical proofs proceed by steps, each of which entails the next, but we are seldom certain of the conclusions because of the intricacy of the arguments, and with some justification. There are cases in which such proofs have been accepted for some years before a fault has been discovered and the argument revised. Once again the argument is supposedly deductive, but it is now even clearer that awarding it the name does nothing to distinguish it from an inductive argument in terms of persuasiveness. We are much more certain of a simple inductive argument with basic observations for premises (the sun has risen every day in the past) and a weak generalisation for a conclusion (the sun rises every day). Indeed we might say that such a conclusion is closer to being necessitated by its premises than in the mathematical case, or that it's harder to conceive of such premises being true and the conclusion false.

Looking at the distinction from the inductive side, we find the fourth example of inductive/deductive blurring. Consider inductive arguments with conclusions that

are necessary because they result in reference-fixing. Imagine an inductive argument expressing the scientific research that led to the conclusion 'water = H_2O'. If we fix the reference of 'water' as H_2O, only using the term correctly whenever H_2O is present, then it turns out that wherever 'water' appears in the premises it simply means 'H_2O'. So the premises cannot be true without the conclusion being true and (trivially) the premises necessitate the conclusion. It's an inductive argument with what are typically considered to be key deductive features. Similarly, if part of our philosophy of science is the view that the laws of nature are metaphysically necessary, then any inductive argument from science will have a necessary conclusion, providing it has true premises. Once again, such an argument would confer absolute certainty, exhibit a strong relation of entailment or necessitation between premises and conclusion, and be perfectly truth-transmitting. Yet it's apparently inductive.

So what *is* the difference between deduction and induction? We've been using the terms to mark a distinction of some kind, so there must be something to it. It could be that deductive rules, whatever they may be, are accepted a priori while inductive rules, whatever they may be, are accepted a posteriori. This would mean that deductive arguments are justified a priori while inductive arguments are not. This fits with the fundamental intuition that we are somehow compelled to accept the conclusions of deductive arguments, while the acceptance of inductive conclusions is a matter of judgement, based on the evidence in their favour (which is not to say that, if this is a good basis for the deductive/inductive distinction, some arguments traditionally thought to be deductive won't turn out to be inductive and vice versa). This is only part of the story, but we must leave the question behind. Assuming that we may call 'inductive' any argument that involves a judgement about whether or not to accept its conclusion, we must focus our attention on induction, and some notable attempts to describe it.

2. Accounts of induction

The question of how we judge whether to accept an inferential conclusion, and under what conditions we do so correctly, is at the heart of the philosophical discussion of induction. A successful account of induction must provide an answer to both descriptive and normative aspects of the question; it must both accommodate our actual inductive practices and justify them or show they are rational.

One account of induction is inference to the best explanation, according to which explanatory considerations are the basis for our judgements about whether to accept inductive conclusions. This is the account that will be discussed and developed in this thesis. First though, we will look at some rival accounts: enumerative induction, hypothetico-deductivism and Bayesianism. This will help us get a clearer idea of what induction actually is and help us, when the time comes, to judge whether inference to the best explanation offers the best account of it. Before we start, it's worth noting that in this competition, each of the following is better described as naming a family of accounts rather than a single, unified position (cf. Norton 2003).[1] Few, if any, would support the basic inductive principle exploited by each family as an adequate account of induction by itself. Nevertheless, I will continue to write as if 'enumerative induction', 'hypothetico-deductivism' and 'Bayesianism' nominated real and plausible analyses of our inductive practices. Certainly, the benefits I discuss under each heading should accrue to any variant; likewise, any criticism should apply to all members of the relevant family.

2.1. Enumerative induction

The most basic account of induction is enumerative induction, according to which we infer from observed instances of the form 'A is B' (and none of the form 'A is not-B'), to a generalisation of the form 'all A's are B'. For example, we observe instances of military conflict in which innocent civilians are harmed, and infer that innocent civilians are harmed in all military conflicts (crudely, A = military conflict, B = innocent civilian-harming). This model seems to capture important features of our inductive practice, almost to the point of triviality. It expresses the important idea that induction is a matter of projecting the patterns of the past into the future, and that generalisations or laws are supported by their positive instances. It also makes room for the idea that strength of belief in a conclusion is proportional to the amount and variety of evidence in its favour, although it is silent on exactly how many (kinds of) supporting instances are sufficient for a conclusion to be inferred. Nevertheless, enumerative induction seems like a promising basis for an account of induction: if we observe enough A's that are also B, and no A's that are not B, we infer that all A's are B (or, more weakly, that the next A we observe will be B), and this seems justified.

However, there are notorious problems with enumerative induction. Most notable amongst these are the following pair, both of which show that enumerative induction under-describes our actual inferential practices. The first problem was formulated by Nelson Goodman ([1983] 2000), who argued that under enumerative induction there is no restriction on the number of hypotheses supported by a given range of observations. In Goodman's original example, he introduces the artificial predicate 'grue', which applies to all things observed before time *t* that are green and all things not observed before *t* that are blue.[2] Then he notes that, assuming *t* has not yet elapsed, all our observations of emeralds support both the hypothesis 'all emeralds are green' and the hypothesis 'all emeralds are grue'. If we set *t* at, say, midnight tonight, then for every piece of evidence we currently have stating that a given emerald is green we also have a piece stating that that emerald is grue; the two hypotheses are equally well supported. In Goodman's own words, "the prediction that all emeralds subsequently examined will be green and the prediction that all will be grue are alike confirmed by evidence statements describing the same observations. But if an emerald subsequently examined is grue, it is blue and hence not green" (Goodman [1983] 2000: 556-557). It seems our emerald-observations serve to confirm two hypotheses with contradictory consequences. "Moreover, it is clear that if we simply choose an appropriate predicate, then on the basis of these same observations we shall have equal confirmation, by [enumerative induction], for any prediction whatever about other emeralds – or indeed about anything else" (ibid.: 557).

Goodman shows that enumerative induction cannot be a good account of our inductive practices. In reality, we have little trouble distinguishing the hypotheses that our evidence genuinely confirms from those it does not. To put it another way, we are not prepared to project all observed regularities into the future. For example, if I observe that all students in the philosophy common room are wearing jeans, I do not infer that the next person to walk in will be wearing jeans, let alone that all philosophy students wear jeans (when they use the common room), even though under enumerative induction I am licensed to do so. There is obviously some restriction on the hypotheses confirmed by their positive instances that enumerative induction does not capture, resulting in its being massively over-permissive. It would have us infer any old conclusion on the basis of any old evidence, so enumerative induction cannot be right.

The second problem with enumerative induction is the raven paradox, due to Carl Hempel (1965). We observe numerous black ravens and make the generalisation that all ravens are black. So far so good; but the generalisation that all ravens are black is logically equivalent to the generalisation that all non-black things are non-ravens (intuitively, if it's true that all ravens are black, then we will never find, say, a red thing that is also a raven; if it's red it must not be a raven, since all ravens are black). According to enumerative induction, 'all non-black things are non-ravens' is confirmed by its positive instances, so observations of, say, red poppies confirm it, just as observations of black ravens confirmed 'all ravens are black'. Again, all seems well, until we consider that observations that confirm a proposition also confirm any logically equivalent proposition. Now the problem is clear: under enumerative induction, observations of red poppies, or any other non-black non-ravens, confirm that all ravens are black. This is clearly unacceptable, as red poppies obviously have nothing to do with the truth-value of any claim about ravens. Once again, enumerative induction is shown to be massively over-permissive, giving our hypotheses inductive support where actually there is none.

2.2. Hypothetico-deductivism

Despite capturing those aforementioned features of our inductive behaviour (projection of past patterns into the future, inference of generalisations supported by positive instances, proportioning of belief to evidence), enumerative induction falls well short of an adequate account of induction. Time to consider an alternative: hypothetico-deductivism, most closely associated with Hempel (though not even he supported the characterisation I use here). According to hypothetico-deductivism, evidence confirms a hypothesis if a statement of that evidence is deducible from the hypothesis. For example, our hypothesis might be that all badgers feed nocturnally, from which we may deduce that if x is a badger, then x feeds nocturnally. Our hypothesis receives confirmation if the next badger we see feeds nocturnally, and disconfirmed if we find one that feeds during the day. Although it has the obvious flaw of being silent on how our hypotheses are generated in the first place, hypothetico-deductivism has several appealing features. It captures the important idea that hypotheses are confirmed if their predictions are successful. Further, it makes

this process of confirmation depend on deduction, an advantage since (foregoing discussion notwithstanding) it de-mystifies the notion of confirmation.

But perhaps the most significant advantage of hypothetico-deductivism over enumerative induction is that it allows for 'vertical' inference. Vertical inferences involve claims that refer to entities, occurrences, processes and so on not contained in statements of the evidence that prompted the inference; they 'go beyond' the evidence. As we've seen, enumerative induction would only have us infer statements generalising about the evidence gathered, so the proposition inferred is only ever as sophisticated as the evidence in its favour. This severely limits induction in a way not reflected in real life. The content of our inferences is not constrained by the content of the immediate evidence; we are frequently much more creative, making reference to unobserved and unobservable stuff in order to gain more insight into the phenomenon at work. The priority given to the *hypothesis* in hypothetico-deductivism – we frame a hypothesis *and then* check its consequences – reflects this. Our hypothesis may be about anything we like: as long as it doesn't generate false consequences, it may be inferred. The absence of restrictions on the content of our inductive inferences (other than the one just mentioned – outlandish hypotheses are more likely to have false consequences) is a strong reason to favour hypothetico-deductivism. When we consider the massive scope of induction, we see that an account of vertical inference is essential.

Perhaps the most typical vertical inferences are those referring to unobservable entities and processes. These may be everyday claims about the mental activity underlying others' behaviour or sophisticated scientific claims about electrons. In fact, another reason to favour hypothetico-deductivism is that many believe it accurately represents inductive practice in science. Frequently in modern science, the instances of a theory are unobservable (if we can make sense of the idea of 'instances' at all), but hypothetico-deductivism allows theories to be confirmed by their observable logical consequences. An example is Eddington's confirmation of Einstein's theory of relativity: his observations of the positions of certain stars during a solar eclipse matched those he had deduced from the theory, which is at such a level of generality as to make its direct confirmation by 'instances' impossible.

But for all its benefits, hypothetico-deductivism shares the problems of enumerative induction; it would have us infer almost anything on the basis of almost anything. Goodman's problem affects hypothetico-deductivism because it places no

restriction on the hypotheses from which we may try to derive confirming consequences. We may hypothesise that all emeralds are green or that all emeralds are grue and have both claims confirmed by the same observations, despite the fact that the two hypotheses entail contradictory claims about unobserved emeralds, so cannot both be right. Similarly, Hempel's raven paradox affects hypothetico-deductivism, its impact even clearer than in the case of enumerative induction. Recall that our hypothesis 'all badgers feed nocturnally' entailed that if *x* is a badger then *x* feeds nocturnally, and so is confirmed by our observation of a badger that also feeds nocturnally. Our hypothesis also entails that if *y* doesn't feed nocturnally then *y* is not a badger (true if the hypothesis is true), which is confirmed by our observation of a robin skilfully catching flies in the midday sun. But if confirmation occurs when observation matches entailment, then our observation of a robin confirms our hypothesis about badgers. Of course the observation is completely irrelevant, meaning hypothetico-deductivism cannot be the correct account of how we go about induction (indeed, since it sanctions vertical inferences, hypothetico-deductivism is even more permissive than enumerative induction, which ignored such inferences altogether).

Hypothetico-deductivism's attempt to identify confirmation with entailment has failed; hypotheses are not confirmed every time we observe something they entail, otherwise confirmation is both ubiquitous and meaningless. A good account of induction must make sense of confirmation without entailment; it would help if it could also accommodate the related idea that degree of belief is proportional to the evidence in a hypothesis' favour. Our next candidate, Bayesianism, represents an attempt to provide such an account.

2.3. Bayesianism

Bayesianism uses the mathematical axioms of probability to calculate the correct degree of belief in a hypothesis after evidence has been obtained. It works by assigning probabilities to the chosen hypothesis, the evidence, and the evidence given the hypothesis, *before* the evidence is obtained, and feeding these figures into Bayes' theorem, a consequence of the axioms of probability:

$$p(h/e) = \frac{p(h) \times p(e/h)}{p(e)}$$

Here, *h* is the hypothesis and *e* the evidence, so *p(h)* is the probability of the hypothesis, *p(e)* is the probability of the evidence, and *p(e/h)* is the probability of the evidence given the hypothesis; these figures determine *p(h/e)*, the probability of the hypothesis given the evidence. The probabilities are all subjective: they represent, in figures between 0 for certainly false and 1 for certainly true, the degree of belief we attach to the propositions chosen to occupy the brackets. Bayesianism states that if we assign values to *p(h)*, *p(e)* and *p(e/h)* then rationally speaking we may assign only one value to *p(h/e)*, namely that determined by applying the theorem. To illustrate, let's say we decide on the probabilities of some *h*, *e* and the value of *p(e/h)*, and use the theorem to determine *p(h/e)*. We then conduct an experiment to test *h*, which results in the occurrence of *e*. According to Bayesianism we must now change our degree of belief in *h* from the *p(h)* we fed into the theorem to the *p(h/e)* we got out. In other words, given our rational commitment to calculating the probability of *h* on the assumption of *e*, now that *e* has actually occurred we have no choice but to adjust our degree of belief in *h* in the Bayesian way.

In Bayesian terminology, *p(h)* and *p(e)* are known as the 'prior probabilities' (or 'priors' for short), for the obvious reason that they are assigned prior to the occurrence of *e*. Correspondingly, *p(h/e)* is known as the 'posterior probability' of *h* as it represents the probability of *h* after *e* has occurred. The remaining value, *p(e/h)*, is commonly known as the 'likelihood' of *e* given *h*, because it represents how likely *h* would make *e* if it were true. Changing one's degree of belief in *h* from its prior to its posterior probability is known as 'conditionalisation'. Conditionalisation makes sense as part of an account of induction. If the posterior probability of *h* is greater than its prior probability – if *e* supports *h*, makes its truth more likely – then increasing our degree of belief in *h* is the natural, rational response. Bayesianism has degree of belief vary proportionally with degree of confirmation, which gives it great intuitive plausibility.

Alexander Bird (1998: 204-205) notes that Bayesianism also explains some more subtle features of inductive inference. Firstly, it accommodates our preference for hypotheses that seem (on background knowledge) likely before being tested. Secondly, it allows hypotheses that strongly suggest some evidence to receive greater confirmation if that evidence is obtained. Lastly, Bayesianism expresses our tendency to award greater confirmation to a hypothesis if we think the supporting evidence is

by itself highly unlikely. Bayesianism can do all this thanks to the way Bayes' theorem relates the priors and likelihood. It says that *p(h)* and *p(e/h)* are to be multiplied together and divided by *p(e)*, so high values for *p(h)* and *p(e/h)* and a low value for *p(e)* translate into a high value for *p(h/e)*.

Bayesians sometimes argue that they can solve the raven paradox (cf. Lipton 2007a: 81-82) and have attempted a response to Goodman's grue problem. Success here would be a major point in Bayesianism's favour, but alas the view suffers from other problems, on which I concentrate here. Firstly, it has been argued that Bayesianism is once again over-permissive. It allows a hypothesis to be confirmed by any of its logical consequences, as long as all the probabilities involved are neither 0 nor 1 but somewhere in between. The problems associated with confirmation by logical consequence have been discussed in relation to hypothetico-deductivism. Secondly, it has also been argued that Bayesianism is too strict. There is debate over the degree to which a hypothesis is confirmed by evidence obtained before it was formulated, but most parties agree that such evidence offers *some* confirmation. Yet Bayesianism cannot allow for this, as such 'old' evidence will have a prior probability of 1, so cannot affect posterior probability.

Further objections concern the guiding principle of Bayesianism – that we attach probabilities to propositions and use them to adjust our degrees of belief. The first, and perhaps most obvious, of these objections is that Bayesianism provides no constraints on the values we may attach to the priors. As mentioned above, these are subjective probabilities, so in principle there is nothing stopping you and I from attaching wildly different values to the same propositions, and coming out with wildly different values for *p(h/e)*. This is a damaging result for a model of inference that claims to express the rational rules of belief-updating. If two people allocate sufficiently different priors, Bayes' theorem recommends two different degrees of belief in *(h/e)*, even though both parties have identical evidence.

To this criticism, Bayesians have a couple of responses lined up. The first is to argue that in reality, we don't attach vastly different priors to *h* and *e*, and this may be because our priors are derived from the results of earlier applications of Bayes' theorem. Generally, this may be true, but there is an obvious problem: not every allocation of priors can depend on a previous Bayesian calculation, otherwise we have an infinite regress of Bayesian calculations. The Bayesian may then turn to evidence – from the sciences, for example – that even without previous Bayesian calculations, the

allocation of priors is a matter of considerable agreement. This seems right: in concrete applications of Bayes' theorem there's no reason to expect priors to diverge such that the theorem becomes useless, and we would regard priors at the extremes of the probabilistic range to be absurd without good supporting reasons. If such reasons were supplied, they might prompt a discussion resulting either in agreement over one extreme prior, or agreement on some more moderate figure. But this response opens the door to a new criticism: that there's more to (scientific) reasoning than is expressed by Bayes' theorem. It's starting to look like we need an account of how we judge the probability of certain statements (in the absence of Bayesian influence). Such an account threatens to become the real story behind rational inductive inference, making Bayes' theorem look insignificant.

Bird illustrates this with the example of the neutrino, an unobserved particle that was postulated to account for unexpected patterns of energy and momentum displayed after certain subatomic reactions. Deciding on a prior probability for the neutrino hypothesis would have been a difficult matter, dependent on the judgement of several factors. Bird gives some idea of the complexity of the situation:

> "How likely is the existence of an unobserved particle? How successful have postulations of similar particles been hitherto? How well established are any theories with which this hypothesis might conflict? How much would such theories need to be altered to accommodate this hypothesis? How good are alternative explanations – such as dropping the law of conservation? How likely is it that we have thought of all the possible explanations? Do we know this field well enough to estimate the chances of some as yet not thought of alternative hypothesis successfully explaining the same data?" (Bird 1998: 211).

The Bayesian scientist would have to ask similar questions when estimating the prior probability of the unusual energy distributions and the likelihood of that evidence given the neutrino hypothesis. All of which goes to show that the allocation of priors is a complex business about which Bayesianism is completely silent. The problem that we started with – that if I attach a high prior to one hypothesis, I might maintain a high degree of belief in it even though the evidence suggests it's false – is brought into sharper focus. But the more worrying problem for the Bayesian is that their view misses the point behind inductive inference and the real story is elsewhere, in some account of how we formulate hypotheses and carry out the kind of judgements mentioned in the quotation above. This problem is due to Bayesianism's being non-ampliative: it gives us a rule telling us how to update our degrees of belief in the light of new evidence, but is silent as to how inferential content might exceed a mere

restatement of the evidence. Moreover, a good account of induction must tell us when we are entitled to accept a hypothesis. Given priors and likelihoods, Bayesianism tells us how much support a hypothesis should have on the basis of evidence, but not when there is sufficient support for belief (cf. Psillos 2007: 445-447). We still have every reason to favour Bayes' theorem as a part of rational inference, but it looks more like a consistency constraint on that inference rather than a full account of it (cf. Lipton 2007a: 83). Fulfilling the Bayesian constraint depends on some other, very important reasoning that Bayesianism neither describes nor justifies. Consequently it cannot be the correct account of induction.

I conclude that there are serious problems with enumerative inductive, hypothetico-deductive and Bayesian accounts of induction. They have been considered here in crude form, but I suggest that modifications will either prove unsuccessful or will cause such accounts to lose their identity. My critique of Bayesianism has set the stage for what I believe is the correct account of induction: inference to the best explanation (IBE). In one way or another, IBE will be the main concern of the remainder of this thesis; the next section introduces it.

3. Inference to the best explanation

On a superficial level, IBE defines itself. It is an account of inference that puts the *explanation* of evidence centre-stage. According to IBE, we infer what would, if true, be the best explanation of the evidence. In order to qualify as best, an explanation must meet three basic conditions. Firstly, it must be sufficiently good to merit being inferred, that is, it must be sufficiently plausible and interesting to make inference tempting, otherwise we may suspend judgement. Secondly, an explanation must be sufficiently better than its competitors. If two or more explanations have equal or similar claim on being the best, then again judgement may be suspended. Thirdly, an explanation must explain a sufficiently large range of data. If it explains only limited evidence, or is inconsistent with previous data, then we may choose not to infer.

In chapter 2, I describe in some detail how Peter Lipton (2004) takes these basic ingredients and develops IBE into a fully-fledged account of induction. In the course of that discussion, much is said about why we might favour IBE; in the following three sections I note the respects in which it improves on the accounts just discussed.

3.1. IBE and enumerative induction

One key advantage of IBE over enumerative induction is that it accounts for when we choose to make an inference and when we do, which inference we make. We don't infer whenever we find ourselves with any old collection of evidence, and inference isn't always prompted by collections of a certain size and type. Consequently, enumerative induction has nothing to say about the conditions under which we infer. This is a big problem: we are highly sensitive to what we might call the 'epistemic circumstances'. Sometimes we infer confidently on the basis of a mere handful of instances, sometimes we refrain from inferring until numerous instances are in and other considerations have been satisfied. IBE can account for this sensitivity, since some phenomenon or pattern of phenomena is not always best explained by extrapolating to a generalisation (cf. Harman 1965: 90-91; Lipton 2004: 66). It may rain every summer's night in Bristol for a month and I wouldn't infer that it rains every summer's night, not even in Bristol. But I might infer that there's an area of low atmospheric pressure above south-west England. Or, aware of the complexity of UK weather systems, I may make no inference at all. A palaeontologist on a dig may find several dinosaur vertebrae of the same size and weight. She wouldn't infer that all dinosaur bones are that size and weight, but she may infer that several like beasts perished in that location. Or, again, she may make no inference at all, perhaps because she feels further evidence is needed. In both cases, enumerative induction is silent about why we don't extrapolate from the evidence to a simple generalisation. IBE accounts for it: we refrain from inference when, under the circumstances, there's no good explanation available, and when we make an inference, we choose the best explanation, which isn't necessarily just an extension of the observed phenomena.

However, as was the case with hypothetico-deductivism, the signal advantage IBE offers over enumerative induction is that it accounts for vertical inferences. Enumerative induction relies on a principle of 'more of the same'; it only licenses the inference that whatever we've observed previously will continue in the same manner. But in reality few of our inferences are like this, as the previous paragraph suggests. We freely, and, it seems, justifiably infer from observations to conclusions referring to unobserved observables or unobservables. If the best explanation of some observable

phenomenon would make reference to such things, then under IBE, it may legitimately be inferred. Upon seeing your red nose, watering eyes and frequent sneezing on a winter's day, I infer that you have a cold. I haven't observed your cold virus, and neither have I observed anyone else's when they've displayed the same symptoms (although I could've done so, given the right equipment). Yet all my inferences from observed symptoms to colds, including yours, are justified. This is because a cold is the best explanation of the symptoms. On an afternoon walk through my local park, I smell freshly-cut grass and see that the grass is shorter than yesterday. I infer that it was cut this morning. My inference is justified because even though I can't now observe it, the cutting is the best explanation of the evidence.

Now for a case involving unobservables. Given some problems with the mathematical description of the behaviour of energy, the scientist postulates a new sub-atomic particle of a certain size, weight, spin etc, which enables the description to fit the observed phenomena. After further experiments, this fit leads the scientist to infer that the particle exists. Under IBE this inference is justified because (given the mathematical assumptions) the particle is the best explanation of the phenomena. IBE's ability to account for vertical inferences to the unobservable is particularly important if it's to provide a good account of inference in science. It's a fact of scientific practice that scientists infer from observed phenomena to unobservable entities that would account for them. Those inferences are not fanciful; they are justified on the evidence available, and scientists report as much. Further, inferences to the unobservable play an important role in science's considerable empirical success. Thus any theory of induction must account for them. That IBE can do so, and with a notion – best explanation – that scientists suggest is actually at work in their inferences, is a major point in its favour.

3.2. IBE and hypothetico-deductivism

But hypothetico-deductivism can also account for vertical inferences; can IBE do better? Yes. IBE's first advantage over hypothetico-deductivism connects to its first advantage over enumerative induction. IBE accounts for our sensitivity to the epistemic circumstances surrounding inference, so it accounts for when we make a vertical inference and when we stay horizontal, i.e. on the observational level. Hypothetico-deductivism lacks the resources to explain this. Consider another

example. Seeing frost on my window one morning, I may infer that it's very cold outside, or that it's 0°C or below. The inferences are compatible, but something may prevent me from inferring an estimate of the temperature. I may feel cautious about adding the extra detail, and this will probably be because I know that frost isn't always best explained by sub-zero temperatures, since it can persist when temperatures are slightly above zero. So instead of inferring a figure in degrees centigrade that I can't directly observe, I may just combine my observation of frost with the feeling of cold by the window, the draught coming in under the door etc, and infer only that there's frost on the window because it's cold outside. This inference is explanatory but horizontal – nothing unobserved or unobservable is mentioned. But if I'd recently heard a weather forecast which predicted sub-zero temperatures, I may feel more confident about inferring that it's 0°C or below. In this case, the extra unobservable detail is worth the risk, and I end up with a better explanation of the frost. The extra detail gives more understanding of the evidence, which now includes my weather forecast observations.

This example and those above give an idea of how explanatory considerations account for the kinds of inference we make and what evidence we make inferences from, about which hypothetico-deductivism is silent. Note especially that the evidence that prompts our inferential mechanisms to kick in isn't the only evidence relevant to inference. Also relevant are various other assessments of our epistemic position, particularly what else we know about the situation and other similar situations. These factors are naturally accounted for by explanatory considerations, roughly because by accounting for our sensitivity to circumstances, they show how previous experiences, or more specifically, background knowledge, influences our inferential behaviour.

Lipton (2004: 67) diagnoses hypothetico-deductivism's problems, which are twofold. First, hypothetico-deductivism is strictly an account of confirmation, not of inference. It tells a story about when evidence confirms a hypothesis; it is silent about what that hypothesis may legitimately say and when it may be inferred. This is precisely what IBE aims to explicate, as shown by the above examples. The second, related, issue is that hypothetico-deductivism has nothing to say about the 'context of discovery'. For many years, the study of scientific inference was divided into two main areas: the 'context of discovery' and the 'context of justification'. The former was concerned with how scientists arrive at hypotheses, the latter with when

hypotheses are confirmed and may be rationally accepted. The context of justification was thought to be the proper focus of philosophy: the discovery of a logic of inductive support or principles dictating when inference is legitimate. The context of discovery was thought to be of psychological interest only. However, investigation of science revealed the distinction between contexts to be vague or even non-existent, so modern philosophy of science takes it less seriously. Indeed, given that discovery and justification are in fact closely linked, an account of scientific inference that fails to explicate one or the other is flawed, and one that gives a unified account of both is promising. Hypothetico-deductivism stays firmly in the context of justification; it makes no attempt to tell us about how scientists arrive at the hypotheses they test. Under IBE meanwhile, discovery and justification are both guided by explanatory considerations: we generate various candidate explanations of some evidence, constrained by other things we know, judge which is the best, and infer it on those grounds.[3] Sometimes a candidate that would be well-supported is not even generated, and sometimes we don't infer an explanation even when we know it's well-supported. Thus IBE accounts naturally for features of inference that hypothetico-deductivism can't handle.

The supporter of hypothetico-deductivism might think I'm being unfair here; she might argue that her favoured model does have something to say about hypothesis generation after all. Take Hempel's position for example. His support of hypothetico-deductivism came as part of an explanationist package, meaning he thought that hypotheses can only be confirmed by data they would explain: evidence *e* confirms hypothesis *H* if and only if *H* would explain *e*. So for Hempel, hypothesis generation is a matter of explanatory considerations. This is right, but it's a dead-end for the hypothetico-deductivist who follows Hempel. Hempel's account of explanation, the deductive-nomological model (cf. Hempel 1965), adds nothing to his account of confirmation. Under hypothetico-deductivism, confirmation comes when a deduction is satisfied by observation; under the deductive-nomological model, explanation comes when a (lawlike) hypothesis (plus specific factual statements) entails that observation. More intuitively: under hypothetico-deductive confirmation, the hypothesis comes first and is confirmed (or disconfirmed) by appropriate observations; for deductive-nomological explanation, the observation comes first and is explained by an appropriate hypothesis.[4] Thus on Hempel's picture, explanatory considerations may determine hypothesis generation, but there's nothing to say about

explanatory considerations, beyond what we already know about confirmation. That which gets confirmed is that which would explain, and that which explains is that which would get confirmed. This is an interesting attempt to break down the discovery/justification distinction, but it doesn't help the hypothetico-deductivist fight back against the supporter of IBE in the battle to account for the inferences we actually make.

It was easily dismissed, but the hypothetico-deductivist's attempted response suggests something interesting. Plausibly, Hempel's hypothetico-deductivism fails to account for inference not because it's really an account of confirmation, nor because it's silent about the context of discovery, but rather because it employs a bad account of explanation. Hempelian hypothetico-deductivism, like IBE, only sanctions the inference of explanations; thus it tells us something about hypothesis generation: we must generate explanations. The trouble is, what it tells us about explanation is largely wrong. Many of the problems with hypothetico-deductivism could be solved by a better notion of explanatory relevance than that provided by the deductive-nomological model.[5] Non-black non-ravens wouldn't confirm 'all ravens are black' if the hypothesis 'all ravens are black' didn't explain observations of non-black non-ravens. What's needed is some way of distinguishing black ravens from everything else such that 'all ravens are black', along with certain statements about the observation-situation, explains only black ravens. On Hempel's explanationist assumption, confirmation would then fall into place. The explanation relation promises to clarify the confirmation relation in a way that confirmation doesn't promise for explanation. Unfortunately for the hypothetico-deductivist, replacing the deductive-nomological model with a better account of explanation would make hypothetico-deductivism a footnote to IBE. Hypothetico-deductivism would merely formalise the confirmation relation assumed by IBE, all the interesting work being done elsewhere by an independent account of explanation.

This point is driven home by the fact that IBE can avoid problems such as the raven paradox only on the condition that it *doesn't* employ the deductive-nomological model of explanation. Lipton's defence of IBE against the raven paradox (2004: chapter 6) relies heavily on his causal-contrastive model of explanation; if 'all ravens are black' is read as making an oblique causal claim – all ravens have some feature causing their plumage to appear black – then it does not explain my observations of red poppies, since such raven-features do not cause poppies to appear red.[6] Thinking

of explanation in terms of causes offers a similarly natural defence of IBE against the grue problem (cf. Lipton 2004: 91-94). Briefly, assuming our causal picture of the world is approximately correct (there's no reason to assume it isn't correct in the relevant respects), the property of grueness cannot be an effect, apt to be explained by noting its causes, while greenness can (we may explain it by citing features of an emerald's microstructure, which cause it to reflect light of a certain frequency, which stimulates certain retinal receptors, etc). The deductive-nomological model does not make such responses available. In fact, were that model to be plugged into IBE, IBE would collapse into a version of hypothetico-deductivism (cf. Lipton 2000: 186). So IBE must avoid the deductive-nomological model, not just to score points over its competitors, but to maintain its very identity.

Thus, seemingly, the deductive-nomological model of explanation is the root cause of hypothetico-deductivism's problems and IBE's successes (I've mentioned the raven paradox and the grue problem, but other problems, such as the tacking paradox, in either its conjunctive or disjunctive form, may be solved by a decent account of explanatory relevance too).[7] Fortunately for IBE, other models of explanation are available. To summarise, hypothetico-deductivism, despite Hempelian efforts, fails where IBE succeeds; consequently, it cannot rival IBE in the competition to account for our inductive practices.

3.3. IBE and Bayesianism

As a descriptive account of induction, IBE faces stiffer competition from Bayesianism. Once again though, IBE can claim several advantages. Firstly, recall that Bayesianism did not allow old evidence to confirm a hypothesis. Positive evidence obtained before a hypothesis is formulated offers some (though perhaps not much) confirmation, but Bayes' theorem dictates that that evidence has a prior of 1, and so will not raise a hypothesis' posterior in the appropriate way. This problem arises from the structure of Bayes' theorem. IBE features no such strict mechanism for attaching probabilities to hypotheses and their evidence, which means it is not restricted to confirming hypotheses only with new evidence. In short, confirmation may come from anywhere, so long as the appropriate explanatory relations obtain. Thus, crucially, IBE can account for the way in which hypotheses gain support from evidence enshrined in our background knowledge. That a hypothesis explains such

evidence counts in its favour. Such evidence isn't new; it's old evidence seen by new lights, but it offers confirmation in a way cancelled out by Bayesian demands.

Where IBE has a real advantage over Bayesianism, however, is in the fixing of the priors. By bringing subjective probability into the issue of inductive inference, Bayesianism definitely gets something right, but it pays a heavy price by failing even to attempt a story about how we fix the priors. IBE is perfectly placed to give an account of such matters. Put simply, our assessment of hypotheses' probability in relation to one another and to the evidence, and arguably our assessment of the probability of the evidence alone, is guided by explanatory factors (cf. Okasha 2000, Lipton 2004: chapter 7). We award a high value to $p(e/h)$ if h would provide a good explanation of e – a good potential explanation makes the evidence it would explain likely. Likewise a high $p(h)$ is given if h coheres with our background beliefs – we assess it as independently likely, given what else we know. If these values are divided by a low $p(e)$ – awarded because we would find e surprising, perhaps even inexplicable given current beliefs – then the posterior $p(h/e)$ will be correspondingly high.[8] Thus we end up with higher degrees of belief in hypotheses that explain better; this is done not out of a misplaced affection for good explanations, but out of respect for the claim that explanation, plausibility and probability are closely linked epistemic notions. We know something is plausible if it would explain well, and positive assessment of plausibility means positive assessment of probability.

Of course these links are defeasible – we realise that many poor explanations are highly probable in virtue of giving little new information. But even Bayesians must acknowledge that we are not in the habit of framing uninteresting hypotheses, much less inferring them. Not only do explanatory factors promise to reveal how we determine the priors, thus plugging the hole in Bayesianism, they do so in a way that minimises the possibility of radical variance in probability judgements, since judgements of explanatory merit are based on shared background knowledge. (This last claim is commonly made but seldom explicated. Large parts of this project are devoted to articulating the relationship between explanatory virtue and background knowledge.)

Sometimes philosophers get nervous about explanatory considerations influencing matters Bayesian. This is usually because, following Bas van Fraassen (1989), they think such considerations distort the Bayesian calculation. Since Bayes' theorem represents a tenet of rationality, anything that interferes with it must lead us

towards irrationality. Indeed we *are* rationally compelled to update our degrees of belief in accordance with Bayes' theorem, but IBE is compatible with this. Van Fraassen's 'Dutch book' argument showed that, if we give a boost to the posterior probabilities of hypotheses that are also good explanations, we dispose ourselves to accept a set of bets that is certain to lose us money. But explanatory considerations don't violate Bayesianism in the way van Fraassen thinks. As suggested above, Lipton's conciliation of Bayesianism and IBE (2004: chapter 7) argues that IBE provides a heuristic for 'realising' the Bayesian calculation.[9] He argues that we find thinking in terms of abstract, isolated probabilities difficult, whereas we find explanatory thinking natural. Determining the Bayesian priors is easier if we let our grasp of explanatory connections guide our evaluation. So far from fixing the priors, running the theorem, conditionalising appropriately, *and then* increasing the posterior probability if the hypothesis is a good explanation, we factor in explanatory value *before* making the calculation. To the supporter of IBE, as to everyone else, Bayes' theorem represents the way in which we should, and sometimes do, think about evidential support. One way of being a Bayesian is to use IBE. Thus Bayesianism and IBE are not in direct competition.[10]

However, Lipton's view of 'Bayesian abduction' ('abduction' is a synonym for IBE: see section 5.1) does raise an issue of priority (noted briefly above): are explanatory considerations merely at the service of Bayesian conditionalising, or does Bayes' theorem merely formalise a certain feature of explanatory inference?[11] There's little doubt that Bayes' theorem represents a rule of rational belief updating, but it's a rule that IBE can accommodate. Is the theorem central to a descriptive account of induction, or is it just one of several constraints that in practice we respect via IBE?[12] Perhaps this question is the modern equivalent of the debate over whether enumerative induction is a special case of IBE or vice versa (see section 5.2). Lipton himself tries to avoid such issues, saying that IBE can only be part of the full story about inductive inference and there may be no fundamental account from which others derive their force. But by arguing that Bayesianism and IBE are complementary he reignites the debate. Opposition to IBE from supporters of (some kind of) enumerative induction is minimal and diffuse. Bayesianism, however, enjoys considerable support (though there is debate between Bayesians as to what, Bayes' theorem aside, Bayesianism actually is). Whatever Lipton's views, proponents of IBE and Bayesianism – and both are numerous – tend to think that their own account gets

to the bottom of inductive inference. Lipton argues for compatibility, but the question may still be asked: which one's *really* doing the work? To put it another way, Lipton's proposal of Bayesian abduction invites a Bayesian retort: abductive Bayesianism. Lipton argues that IBE provides a heuristic for realising the Bayesian calculation, but the competitive Bayesian might argue that IBE is *merely* a heuristic, one among many others (cf. Psillos 2007: 447).[13]

I do not mean to criticise Lipton's irenic approach, or his stance on issues of priority between accounts of induction, and I certainly don't mean to galvanise a Bayesian attack on IBE or, worse still, provoke idle debate. Rather I mean to make the following point: perhaps there is no competition between IBE and Bayesianism, but if there is competition, then the victor is far from obvious. I'm inclined to think that Lipton's argument is right but that IBE provides a central heuristic, perhaps even the only heuristic, for establishing Bayesian priors and running the Bayesian calculation. There is considerable evidence in favour of our thinking in terms of explanation when we infer; some of which we've already seen. This suggests we are explicitly explanationist and only implicitly Bayesian. In other words, Bayes' theorem describes one aspect of the inferential rationality that we aim towards, but it is only by using IBE that we come to obey it, albeit imperfectly. I don't say that our respect for Bayes' theorem is coincidental; rather, I suggest we have developed, perhaps under evolutionary pressures, a sophisticated and effective method of thinking about hypotheses, evidence and probability such that we may approximate Bayesian requirements. That method is IBE.[14] As an account of induction, Bayesianism competes more closely with IBE than enumerative induction and hypothetico-deductivism, but I have suggested that IBE has the edge, and hope that my case will be strengthened by the discussion that follows.

The relationship between IBE and Bayesianism, and Lipton's conciliatory story in particular, have generated much discussion since the latter was included in Lipton (2004). But such issues will not concern us further. We've seen that IBE is in a strong position with respect to rival accounts of induction, but we still don't know much about it. This is partly because attempts to say more about IBE beyond the superficial are usually controversial. If any consensus has been reached, it is seldom over a solution to the following obvious and important problem: inference aims at truth, but why should we think that when we infer best explanations, we thereby infer the truth? This question is often aimed at IBE as if it were a decisive blow: it might be nice to

believe good explanations, but when it comes to matters of truth, quality of explanation is surely irrelevant. However, this accusation is simply a variant of a much more famous and less tractable problem that affects all accounts of induction: Hume's problem.

4. Hume's problem

David Hume ([1748] 1999) exposed this key philosophical problem, otherwise known as the problem of induction. He noted that if we want induction to deliver knowledge, it had better be a justified form of inference. In other words, we had better have some reason to think that induction reliably delivers true conclusions, or that our belief in its conclusions is justified. He then argued that we have no such reason. This is because arguments for induction must themselves be inductive, meaning we have to assume what we seek to establish, namely that induction is a reliable form of inference. Therefore all attempts to justify induction will, sooner or later, involve a circular argument; induction can never be given an acceptable formal justification. This is the case no matter how we account for induction; the circularity holds simply in virtue of the fact that it is an account of induction (i.e. not deduction) that we seek to justify.

An example will make Hume's problem clearer. Suppose I have a glass of red wine. I believe that the wine will stain the carpet if I spill it. What's my justification for this belief? Quite obviously, it's the fact that I've experienced instances of carpets being stained by spilt red wine, heard reports of such instances, and so on. But I'm only justified in thinking that these past instances support my belief in the red wine staining on this occasion if I have some reason to think that induction, the method of reasoning from past instances to present or future ones, delivers reliable conclusions. Why might I think that? Well, the method doesn't justify itself; I know this because sometimes it has given me false beliefs from true evidence. So I appeal to experience: induction has almost always been a reliable method of reasoning in the past. In fact, when I come to think of it, my life is peppered with applications of induction, both implicit and explicit – they seem essential to my continued existence, and the successful instances vastly outweigh the unsuccessful. So I've every reason to think that induction will deliver the truth on this occasion. All would be well, except this is itself an inductive argument. Just as I tried to argue from past instances of red wine

staining to the next instance of red wine staining, I'm now trying to argue from past instances of inductive success to the next instance of inductive success. And just as I didn't have a justification for induction in the first case, I still don't have one now. It seems I must assume that induction is justified in order to make an argument that justifies induction. My argument will therefore be circular.

Hume's problem can be seen another way. Above I justified my belief that the red wine will stain the carpet by mentioning previous instances of red wine staining carpets, but I may add to that justification a statement such as 'nature is uniform' or 'the future will resemble the past'. This seems to improve my argument. In fact, it makes it deductive; from the statements of previous spillages/stainings and a statement of uniformity I can deduce that all future instances of spilt red wine will result in stained carpets. Perhaps I can justify my belief that the red wine will stain after all. But I only have reason to believe this conclusion if I have reason to believe the statement of uniformity. Happily, I do: I have lots of evidence that the future will resemble the past. In the past, what were once future cases of x always turned out to resemble past cases of x. That is to say, in the past, the future resembled the past. But what reason have I got for thinking that in the future, the future will resemble the past? Well, in the past, the future resem… oh dear! It seems that my experience can only support a statement of uniformity if I already have reason to believe that past cases are a good guide to future ones. That is, I must assume that nature is uniform in order to establish that nature is uniform. Once again my attempt to justify induction is circular.

Thus we end up in a peculiar situation. Our inductive practices are massively successful yet we remain to some extent mystified as to why they yield success. Indeed, induction seems to ensure its own mysteriousness; any account we try to give of it will be unable to explain its success simply because it's an account of induction. Here the relevance of Hume's problem to IBE becomes clear. The problem of finding a link between best explanations and truth is Hume's problem reformulated for IBE. It's perfectly natural to want to connect the process of inferring the best explanation with the success we achieve thereby, just as it's perfectly natural to want to explain the success of, say, the process of generalising from past instances to the next instance or all instances. But to ask, as many have, that IBE demonstrate a cast-iron link between best explanations and truth is to ask it to make inductive success explicable. As Hume showed, this is impossible. So the claim that best explanation is

a poor guide to truth is no knockout blow to IBE; given Hume's problem, IBE is no worse off than any other account of induction.

Nevertheless, we might still want some reason to think that in inferring the best explanation, we at least increase the likelihood of finding the truth. The challenge of giving such a reason will reappear several times in this project, and is a central motivation for my new defence of scientific IBE in chapter 3. Now though, we return to the task at hand: the articulation of IBE.

5. IBE before Lipton

In this section I survey four key contributions to the development of IBE made before Lipton first published his ground-breaking book *Inference to the Best Explanation* (1991, 2nd edn. 2004). These give the account much-needed motivation. They also give some sense of the dimensions of the IBE discussion, and emphasise those that will be of special importance to the present project.

5.1. Peirce and abduction

The history of IBE can be traced back to the work of Charles Sanders Peirce. Before Peirce, it was commonly thought that inferences came in two forms, necessary (deductive) and probabilistic, with the probabilistic class exhausted by various forms of enumerative induction. Reference was made to explanatory inference, but only as part of a more general 'method of hypothesis', or as a footnote to more sophisticated forms of enumerative induction. Peirce was the first to subdivide probabilistic inferences into two kinds. Alongside induction, he thought, was what he called 'abduction', a form of inference which allowed the explanatory value of a hypothesis to influence its acceptance.[15] He defined it thus:

> "The surprising fact, C, is observed;
> But if A were true, C would be a matter of course,
> Hence, there is reason to suspect that A is true." (Peirce 1998: 5.189)

Whither explanation? The key feature is the removal of surprise about fact C. Surprise about a phenomenon is removed by revealing some feature or features of its prior circumstances that made it likelier, or perhaps inevitable. Were we to have known about those features, the phenomenon would have been expected. Thus there

is a close link between removal of surprise and provision of understanding; we understand why a phenomenon occurred when we know about the conditions that brought it about. Hypothesis A does this for C in Peirce's definition: if the conditions stated by A were in fact in place, C would naturally follow. Thus we see that for Peirce, hypothesis A is an *explanatory* hypothesis; A renders C understood and unsurprising, given the circumstances. More crucially though, it's the explanatory virtue of A that suggests we should accept it. Peirce's definition of abduction is not a Bayesian statement of evidential support: it does not state merely that A suggests C will occur, so if C occurs, A is supported in some appropriate degree. Nor is Peirce's definition merely hypothetico-deductive in Hempel's sense: he agrees that C confirms A if A would explain C, but satisfaction of a prediction is not his concern. Rather, Peirce talks about how A would make C "a matter of course": it's the fact that A would make C a matter of course, i.e. would explain C, which gives us reason to accept it. The "hence" in the definition is crucial: it's A's ability to explain C that leads us to suspect that it's true. This is why Peircean abduction is a true forerunner of IBE.

Thus Peircean abduction shares with IBE the claim that if a hypothesis (best) explains a phenomenon then *on those grounds* we should accept it. But Peirce anticipated the modern IBE debate in other ways. He thought that in science, one arrived at hypotheses by abduction, derived predictions from them by deduction, and tested them by (enumerative) induction. This suggests an awareness of the role of explanatory considerations in the generation of hypotheses, an important feature of IBE. Perhaps more interestingly, in Peirce's scientific methodology, abduction held the trump card. If testing predictions fails to distinguish between competing hypotheses, and one explains the evidence better than the others, then there is reason to infer that hypothesis ahead of its competitors.[16] This is one reason why today, IBE claims to be the correct account of scientific inference. Scientists sometimes have to decide between competing theories that empirical testing cannot separate and which even mention unobservables. IBE is ideally placed to account for these choices, since it allows non-empirical features to license inference in a way that doesn't make scientists' decisions arbitrary or merely pragmatic, which scientists deny they are and the empirical success of science tells against. Thus Peirce's description of abduction highlighted a key role for IBE, one that's nowadays often used in its defence.

Influential as Peirce's work was, any movement to popularise explanationist philosophy of science was quickly smothered by the rise of positivism in the 1920s. Inspired mainly by the work of Ernst Mach in the late 19th and early 20th century, and the 18th century empiricism of Hume, positivists believed that "the possibility of observational and/or experimental verification [is] the defining characteristic of all scientific statements" (Ray 2000: 245). Therefore "our knowledge of the physical world is derived entirely from sense experience, and the content of science is entirely characterized by the relationships among the data of our experience" (ibid.). Consequently, positivists rejected metaphysics; there was to be no role in science for theoretical entities or causal connections, simply because statements referring to such things could not be checked against observation.

A further belief of positivists was that science, and in particular physics, offered the model of correct procedure for all intellectual activity. Thus the abandonment of metaphysics, and a focus on empirical (phenomenal) investigation, was the hallmark not only of science but also philosophy. As a result, explanation was out of the methodological picture on all fronts. The purpose of scientific laws was not to explain the evidence, but rather to make generalisations about it, or even just to summarise it. But this didn't mean a statement's explanatory value was merely irrelevant to its acceptance. Since explanatory value was non-verifiable, any explanatory statement carried an implicit metaphysical commitment; further, explicit metaphysical commitments (such as unobservable entities) were usually licensed for their explanatory value. Thus for positivists, explanations were to be actively eliminated from science and philosophy.

Given that positivism dominated philosophy of science from the 1920s until the early 1960s, it's hardly surprising that Peirce's abductive ideas weren't followed through. However, when positivism came to grief (neither observation nor language proved capable of playing the roles positivism required), the stage was set for explanation to be reappraised. Ground-breaking in this respect was Hempel and Oppenheim (1948), and Hempel's wider programme that culminated in his (1965). However, Hempel was working with largely positivist assumptions, a key goal of his project being to make explanation empirically acceptable. The same could be said of another important text on explanation, Braithwaite (1953). As explanation re-entered the picture, so did the idea of explanatory inference, likewise encouraged by the failure of positivist attempts to formalise inductive logic. Into this newly-liberated

arena came Gilbert Harman (1965). Taking inspiration from Peirce, he coined the phrase 'inference to the best explanation', developing an account that he argued was the true basis of inductive inference, more fundamental than any kind of enumerative induction. This provocative claim helped make IBE a popular focus of attention in epistemology and philosophy of science.

5.2. Harman and enumerative induction

Harman chose the name '[the] inference to the best explanation' because he believed other names, including abduction, had misleading suggestions that his preferred terminology could avoid. But in any case, under his basic definition (which is still current), IBE is more than Peircean abduction. The latter could accommodate the idea of competition between hypotheses, and the corresponding idea of comparing hypotheses to see which best explains, but Harman incorporates them into his definition of IBE: "one infers, from the premise that a given hypothesis would provide a 'better' explanation for the evidence than would any other hypothesis, to the conclusion that the given hypothesis is true" (Harman 1965: 89). (A note on the inverted commas: Harman set another trend by acknowledging the need to define what makes for better explanation, and then saying nothing more about it.) Plausibly, his recognition of the role of competition in abduction is one reason why he chose to re-christen it; 'inference to the best explanation' makes it clear that only the best will do.

But Harman's really distinctive contribution is to invert the traditional view and argue that warranted enumerative induction is a special case of IBE. His argument is that enumerative induction under-describes inference on three counts: it lacks the resources to describe even straightforward examples, it cannot provide clear conditions of warrant, and it disguises the importance of true inferential lemmas in gaining knowledge. By contrast, IBE succeeds with respect to all three, grounding Harman's claim that any instance of warranted enumerative induction may be re-described as an instance of IBE but not vice versa. IBE is the more basic account of knowledge and enumerative induction is no longer interesting in its own right (where an enumerative induction is warranted, it is only because it is a special case of IBE).

The examples Harman uses are simple enough. The detective infers that the butler committed the crime because it's the best way to explain the various fragments

of evidence. No past regularity is involved here and the inference is only defeated if a better explanation becomes available. Harman concedes that in this case, a complicated enumerative induction might account for the detective's decision, but "it is difficult to see how one would go about filling in the details of such an inference" (ibid.: 90). The implication is that such a redescription would be either implausibly complex or an IBE in all but name. But Harman is bolder still about his other examples. The scientist infers the existence of subatomic particles from experimental data; we infer that those around us have certain mental states when they exhibit certain behaviour. Harman claims that such inferences – notable as inferences to the unobservable – are not describable in terms of enumerative induction. Thus "even if one permits himself the use of enumerative induction, he will still need to avail himself of at least one other form of nondeductive inference" (ibid.).[17]

Harman then argues that enumerative induction is superfluous. His argument has already been discussed above (section 3.1). Unlike enumerative induction, IBE provides clear conditions of warrant; it explains and justifies our decision sometimes not to make an inference having observed a regularity, and accounts for our choice of inference when we do. If all our observed A's have been B's, we are warranted in inferring the hypothesis that all A's are B's only if the hypothesis is the best explanation of the observations. If another explanation is better (e.g. our sample is biased) we should infer that; if no hypothesis can establish superiority, the observed regularity warrants no inference. Thus enumerative induction cannot account for all inductive inferences, and those it can account for are better accounted for by IBE. Harman concludes that IBE is the fundamental account of inference and enumerative induction, where it's appropriate, is a special case of IBE.

In arguing this, Harman makes a key observation: "in practice we always know more about a situation than that all observed A's are B's, and before we make the inference, it is good inductive practice for us to consider the total evidence" (ibid.). The observed regularity by itself prompts no inferential activity; the cogs start to whirr only when we consider the regularity in the light of other evidence. The plausibility of the various explanatory hypotheses is determined by the total evidence, not just the evidence – the regularity – at the centre of the inference. This observation will prove to be of central importance in the development of IBE offered here.

Another of Harman's insights comes in his final argument for IBE, based on the importance of true lemmas (intermediate inferential conclusions) in gaining inferential

knowledge. This argument is Harman's response to Gettier (1963), designed to show what else is needed for knowledge beyond the requirement that a belief be true and justified. Harman's claim is that in order for a belief to count as knowledge, it must be true, justified, and arrived at via only true lemmas. Harman describes a Gettier case in which he sees a notice saying that Stuart Hampshire will read a paper at Princeton tonight. He warrantedly infers that Hampshire will read a paper at Princeton tonight, and from that, that Hampshire will read a paper (somewhere) tonight. But the former belief is false: Hampshire has cancelled. The latter belief is not false, since Hampshire happens to be reading a paper at N.Y.U. tonight. But even though Harman's belief that Hampshire will read a paper (somewhere) tonight is true and warranted, he does not know this fact about Hampshire. Harman claims this is because he inferred the belief from a false lemma, viz. that Hampshire will read a paper at Princeton tonight.

Harman explains that lemmas are at work in less unusual cases, and in order for us to gain knowledge, none of them can be false. Whenever we obtain knowledge from an authority, in person or via the printed word, we do so on the condition that "the utterance is there because it is believed and not because of a slip of the tongue or typewriter" (ibid.: 93). Obviously the testimony must be true, but this lemma must also be true; "even if the slip of the tongue or the misprint has changed a falsehood into truth, by accident, we still cannot get knowledge from it" (ibid.). For Harman, knowledge depends on a lemma ruling out mistakes, even fortuitous ones. Another example returns to the inference of mental states from behaviour. My knowledge that your hand hurts when you pull it away from a hot stove depends on the lemma that the pain is responsible for the pulling. If this is false, I do not know your hand hurts, even if, by coincidence, it does.

The point of rehearsing Harman's Gettier cases is not to revive a tired epistemological tradition, but to motivate his final claim about IBE: describing these cases as IBEs exposes the essential role of lemmas and thus accounts for the presence or absence of inferential knowledge. We infer that the expert testimony is true because it's best explained as an expression of sincere belief without mistakes. I infer that your hand hurts because your actions are best explained as the result of sudden pain. If truth-telling and pain weren't the best explanations of the testimony and the behaviour, i.e. if the lemmas weren't true, we wouldn't have knowledge. This leads Harman to criticise the attempt to describe the inferences as cases of enumerative induction: "when the inferences are described as basically inductive, we are led to

think that the lemmas are, in principle, eliminable. They are not so eliminable" (ibid.). Enumerative induction thus makes the acquisition of inferential knowledge mysterious, while IBE, to its credit, clarifies the process.

This, plus the remarks on total evidence affecting what we infer and when, reveals that Harman was on the tail of something important about IBE. His concern for the Gettier problem and his desire to establish the priority of IBE over enumerative induction may have obscured it, but his message is this: the main advantage of IBE is that it accounts for the role of the *background* in inference. Harman describes this background in terms of the 'total evidence' of a situation, and in terms of the 'intermediate lemmas' involved in inference. These are two different ways of saying that our inductive practices depend not just on what our inferences directly concern, but also on what else we know. IBE has us infer explanations, and for a hypothesis to explain, i.e. provide understanding, it must stand in the right kind of relation to all kinds of beliefs, not just ones about the evidence at hand. Choosing the *best* explanation amplifies the importance of such background considerations. What's more, this is entirely apt. We often think of an inference being appropriate 'under the circumstances'; given the same (immediate) evidence and a different situation, a different inference may be warranted. A recurring theme of the present project is that IBE exposes the role of background knowledge in sanctioning our inferences; no rival theory has quite the same capacity to account for the sensitivity of inductive inference to context- or situation-specific features. IBE pushes the background into the foreground, and this has been in its favour ever since Harman's seminal paper.[18]

5.3. Thagard and criteria of best explanation

Harman explicitly avoids the issue of what makes one explanation better than another. Paul Thagard (1978) bravely tackles this least tractable part of IBE head-on, and in doing so greatly illuminates the account. He uses real examples of theory choice from different branches of science to show that three explanatory criteria – consilience, simplicity and analogy – are used to evaluate scientific theories. Thus IBE offers the best account of scientific reasoning, since it "accounts for many different aspects of scientific reasoning and applies to examples from different sciences...,

represents the importance of competition among theories..., [and captures] the multi-dimensional character of scientific-theory evaluation" (Thagard 1978: 91-92).[19]

Thagard says consilience "is intended to serve as a measure of *how much* a theory explains" and that "a theory is said to be consilient if it explains at least two classes of facts" (ibid.: 79, where he also defines consilience more precisely, alongside its comparative counterpart). He explains that facts are organised into classes by scientists themselves, in a broadly intuitive way: "we, like Newton and Huygens, have no difficulty in deciding that reflection and refraction constitute more than one application of the theory of light... On the other hand, we would probably say that the distribution of species of finches and the distribution of tortoises on the Galapagos islands are not facts of different classes and, hence, amount to only one application of the theory of evolution" (ibid.: 80). Thagard says "the inductive logician must take this organisation [of facts into classes] as given, just as do the scientists whose arguments are studied" (ibid.).

Thagard's first example of consilience comes in Huygens' argument for the wave theory of light, and the development of it by Young and Fresnel. Huygens could explain "classes of facts concerning the propagation, reflection, refraction, and double refraction of light", but Young added "facts concerning color" and Fresnel "improved the argument still further by explaining various phenomena of diffraction and polarization" (ibid.: 81). Secondly, Thagard cites Lavoisier's argument for the oxygen theory of combustion, which showed that it explained facts concerning the increase in weight of burning bodies, which made it more consilient than the competing phlogiston theory. His third example of consilience is Darwin's theory of natural selection. He notes that Darwin "cites a large array of facts... concerning the geographical distribution of species, the existence of atrophied organs in animals, and many other phenomena" which are "inexplicable on the then-accepted view that species were independently created by God" (ibid.: 77). Thagard also points out that Darwin uses the terminology of consilience, quoting the famous passage from *The Origin of Species* in which he argues that, if it were false, the theory of natural selection would not explain "the several large classes of facts above specified" (Darwin, quoted in ibid.).

Huygens, Lavoisier and Darwin all argue for their theories on the grounds that they explain more facts than their competitors. Thus there's good grounds for calling them instances of IBE. Something like consilience must feature in any good account

of IBE; note that consilience captures the third basic condition best explanations must meet in order to be inferred, namely that they explain a sufficiently large range of data (see section 3).

Thagard's second criterion of best explanation is simplicity. Many scientists have believed simplicity to be crucial to theory choice. As such it has received a lot of philosophical attention, little of which has resulted in agreement as to what simplicity is or how it is applied. Thagard's definition of simplicity concerns the auxiliary hypotheses needed by a theory to explain the evidence. He begins, "the explanation of facts *F* by a theory *T* requires a set of given conditions *C* and also a set of auxiliary hypotheses *A*. *C* is unproblematic, since it is assumed that all members of *C* are accepted independently of *T* or *F*' (ibid.: 86). Thus auxiliary hypotheses *A* are Thagard's focus. He goes on, "an auxiliary hypothesis is a statement, not part of the original theory, which is assumed in order to help explain one element of *F* or a small fraction of the elements of *F*' (ibid., italics removed). He gives the example of Huygens' assumption that some light waves are spheroidal in order to explain the unusual refraction of light in Iceland crystal; otherwise, Huygens saw light waves as spherical. In order to explain Snell's law of refraction, Huygens also assumed that the speed of light is slower in denser media. These assumptions "were not independently acceptable at the time of Huygens, so they do not belong in *C*; and they were not used to explain any phenomena besides those mentioned, so they must be placed in *A* rather than *T*" (ibid.). Thagard then claims "simplicity is a function of the size and nature of the set *A* needed by a theory *T* to explain facts *F*' (ibid.). Roughly, the smaller the *A* the simpler (and better) the explanation, but Thagard admits this notion of simplicity is not "neatly quantitative,... a qualitative comparison, application by application, must be made" (ibid.: 87). Simplicity depends not only on the number but also on the kind of auxiliary assumptions made. If two sets of assumptions have no members in common, adjudication between their respective theories on grounds of simplicity may be very difficult.

Thagard claims that this account fits with the notion of simplicity used in the arguments of Fresnel and Lavoisier. Referring again to the inference of new theories in optics, he says, "Fresnel accused the Newtonian theory of needing a new hypothesis, such as the doctrine of fits of easy transmission and easy reflection, for each phenomenon that it explained, whereas the wave theory uses the same principles to explain the phenomena" (ibid.: 86-87). During his argument for the oxygen theory

of combustion, Lavoisier "criticizes the phlogiston theory for needing a number of inconsistent assumptions to explain facts easily explained by his theory" (ibid.: 87). Thagard shows that for Fresnel and Lavoisier, simplicity was an explanatory virtue. For them, a theory "not only must explain a range of facts; it must explain those facts without making a host of assumptions with narrow application" (ibid.).

Again, Thagard's claim is that scientific inference is IBE and that his criteria of best explanation are being employed. He also illustrates how simplicity puts a constraint on consilience. The two criteria should not be sought independently. A theory may explain many classes of facts but need numerous auxiliary assumptions to do so. Likewise, a theory may require only a few auxiliary assumptions but explain few classes of facts. Both theories would be undesirable. The best explanations are consilient *and* simple; weighing these two criteria against each other is a complex matter that will be resolved differently in each case. As mentioned above, Thagard thinks this accurately reflects scientific practice, and that "capturing the multi-dimensional character of scientific-theory evaluation is yet another virtue of the view that scientific inference is inference to the best explanation" (ibid.: 92).

Thagard emphasises this by noting that his final criterion of best explanation, analogy, "may be at odds with both consilience and simplicity" (ibid.). He characterises analogical inference as follows:

> "Suppose A and B are similar in respect to P, Q, and R, and suppose we know that A's having S *explains* why it has P, Q, and R. Then we may conclude that B *has* S is a promising explanation of why B has P, Q, and R. We are not actually able to conclude that B has S; the evidence is not sufficient and the disanalogies are too threatening. But, the analogies between A and B increase the value of the explanation of P, Q, and R in A by S" (ibid.: 90).

Thagard's formulation captures the way in which the scientists in his examples used the criterion of analogy. He says "Darwin used the analogy between artificial and natural selection... as one of the grounds for belief in his theory. Huygens, Young, and Fresnel each used the analogies between the phenomena of sound and those of light to support the wave theory of light" (ibid.: 89). But those scientists were also at pains to point out disanalogies; putting analogy in terms of explanation accounts for when they felt analogical inferences were legitimate.

Thagard notes that analogy not only suggests candidate explanations, it also improves them. Explanations that employ familiar models provide greater understanding, and he sees this as crucial to the arguments of Darwin and Huygens: "the explanatory value of the wave hypothesis is enhanced by the model taken over

from the explanation of certain phenomena of sound. Similarly, the explanatory value of the hypothesis of evolution by means of natural selection is enhanced by the familiarity of the process of artificial selection" (ibid.: 91). Thagard admits that explanation is not "reduction to the familiar", but that, other things being equal, the explanations given by a theory are better if it "introduces mechanisms, entities, or concepts that are used in established explanations" (ibid.).[20]

Thagard shows that the theory of evolution by natural selection, the oxygen theory of combustion, and the wave theory of light, were all developed and inferred using IBE, with consilience, simplicity and analogy as criteria of best explanation. Thagard's argument is successful enough for Lipton to endorse his account almost wholesale during his subsequent development of IBE. He frames his discussion in terms of mechanism, unification and background belief, but the claims are almost identical (cf. Lipton 2004: 122-123, 138-140). What I wish to stress is that, as we shall see when we discuss those passages of Lipton, all three of Thagard's criteria are informed by theoretical background. Analogy is obviously so influenced, since explanations are generally better if they cohere with explanations already accepted. Consilience is also theory-dependent in obvious ways: judgements about which facts belong in the same class and which are separate will focus on what our pre-existing theories tell us about how the world carves up. Simplicity is dependent on background in that what qualifies as a disputed auxiliary hypothesis A or an accepted condition C is fixed by what else scientists believe.

5.4. Ben-Menahem and the evolution of explanatory standards

The theme of IBE's dependence on background knowledge is picked up by Yemima Ben-Menahem (1990), in our final contribution to the development of IBE before Lipton. Ben-Menahem's point is simply that "our evaluation of explanatory power... is informed by *empirical considerations*" (Ben-Menahem 1990: 322). Thus, "there is a legitimate inference to the best explanation; that is, it is rational to regard the theory which best explains as the most credible" (ibid.), where 'credible' means roughly 'likely to be true or approximately true'. As will become clear, this is an idea this thesis takes very seriously.

Ben-Menahem sets up the problem for IBE as follows. She starts with a simplifying assumption: the minimal requirement on explanation is that the explanans

(that which does the explaining) entail the explanandum (that which is explained). She admits "this is a gross idealization" and "the deductive model is assumed only in order to emphasize that even if formal criteria [of explanation] exist, they cannot, on their own, confer plausibility on explanations complying with them... there is still room to distinguish better from worse deductive explanations" (ibid.: 321). She then notes two points over which supporters and opponents of IBE agree, firstly that meeting the minimal requirement does not guarantee that an explanation is true, and secondly that a theory compatible with the evidence is preferable to one refuted by it (given specified standards of compatibility and incompatibility). Ben-Menahem calls theories compatible with the evidence "minimally supported", and says that, in view of the above, "there seems to be no dispute over the status of a minimally supported explanatory theory vis à vis a refuted rival or vis à vis the ideal truth" (ibid.: 322). Where there is dispute, however, is over whether, given a range of minimally supported explanations, the best explanation qua explanation is also the most likely to be true. She concludes that if this is where the controversy over IBE lies, then "IBE typically involves explanatory merits which go beyond minimal explanatory power and minimal support" (ibid., italics removed). This reconstruction of the problem was taken up by Lipton and is now characteristic of the debate over IBE.

Ben-Menahem notes three typical examples of IBE: a court deciding a case on circumstantial evidence, a scientist preferring an action-by-contact theory to one involving action-at-a-distance, and three historians – "the historian of ideas, the Marxist, and the sociology-or-psychology-oriented historian" (ibid.: 323) – offering competing explanations of the same historical event, e.g. the development of Newton's thought. In all three cases, she claims, the choice of best explanation can be understood as determined by empirical factors. The judge decides to convict the accused because although the evidence does not prove his guilt, it describes a scenario which resembles others in his experience where a suspect behaved similarly and committed the crime. The judge also brings other empirical findings to bear when judging the plausibility of any competing explanation the accused offers in his defence. Likewise, Ben-Menahem claims, the scientist prefers the action-by-contact theory over the action-at-a-distance theory because of his evaluation of the successes of previous action-by-contact theories. "And the historian, too, bases his preference for a certain type of explanation on a more general belief in the impact of socio-economic, ideological, metaphysical, religious or psychological factors" (ibid.).

Thus IBE (note that these are *not* examples of enumerative induction – the judgements are about competing explanations) is governed by explanatory standards that develop along empirical lines, "how crimes are usually committed, what seems to be the nature of physical interaction, what factors dominate intellectual development, etc" (ibid.). These standards also justify the inference: B en-Menahem makes the crucial observation that they will improve over time; "explanatory merits are justly considered relevant to an assessment of plausibility precisely because our standards for explanatory merit *evolve with the rest of our empirical knowledge*" (ibid.: 324). She underlines her point by noting that in the examples, "if the judge or the scientist or the historian were to deem one explanation superior but another more credible, they would seem unreasonable" (ibid.). Our judgements of credibility (likely truth) and our judgements of quality of explanation are inseparable.[21]

Ben-Menahem also spots the kind of circularity to which critics of IBE often object and which was discussed in relation to Hume's problem. In the process of judging explanations, "we move both from assessment of plausibility to judgement of explanatory power and, vice versa, from assessment of explanatory power to judgements of plausibility" (ibid.: 323). That is, we use the same background knowledge as arbiter in both cases; it determines what we consider a good explanation *and* what we consider plausible. That background knowledge is itself the result of IBEs which required judgements of plausibility and explanatory power. So, say the critics, our inference is unjustified because it's circular. Not so, says Ben-Menahem; these are merely "the pitfalls of inductive reasoning in general: there is no guarantee of truth and there is always the possibility that the case under consideration, rather than being an instance of the rule, will turn out to be a counter-example that suggests a change in our generalizations" (ibid.: 323-324). That is, on any account of induction, justifying an inference will involve recourse to previous instances of successful inductive reasoning. Sometimes we may go wrong, but that is no likelier on IBE than on any rival account.

Returning to the point about empirical evolution of explanatory standards, Ben-Menahem warns against the traditional philosophical focus, which ignores this aspect of explanation in favour of 'timeless' structural features and their supposed relationship with truth. She claims that "in real cases of disagreement over the explanatory power of a theory, the dispute is hardly ever over the *structure* of an adequate explanation" (ibid.: 325). This is because there are "detailed, non-structural

standards which each particular discipline develops and which, as a matter of course, vary with time and context" (ibid.). Such shared standards remove the need for scientists to debate the link between explanatory power and truth so that they can focus instead on the debate over which explanation is the best. Their explanatory standards "have to be justifiable if considered to be rational, but given that justification is in principle possible, the question concerning the connection between explanatory force and credibility no longer arises" (ibid.: 326). Thus in order adequately to analyse IBE, the philosophical attention must turn from "structural features which are general enough to characterize all adequate explanations" (ibid.: 325) to field-specific criteria which "cannot be formulated once and for all by the philosopher" (ibid.). I endorse this idea, and Ben-Menahem's approach finds clear expression in my defence of IBE in chapter 3.[22] Ben-Menahem notes the historical support for the view:

> "There were very few controversies in the history of science which centered on the issue of the credibility of the best explanations or on the structural characteristics of adequate explanations, but there were numerous controversies about the non-structural standards. Moreover, even the seemingly formal standards change over the years. The shady reputation of teleological explanation is not simply a result of philosophical disfavor. Changes in our description of the world led us to reconsider our desiderata for an adequate explanation. Similarly, no philosophical argument could have legitimated probabilistic explanation as effectively as the lesson taught us by quantum mechanics. The history of science reveals that the methods of science are evolving with the rest of science" (ibid.: 325-326).

Ben-Menahem's insights are central to this thesis. My defence of IBE argues that scientists define and refine their explanatory standards in such a way that IBE becomes the best available route to truth. Those standards are relative to certain disciplines during certain periods and are determined by nothing except the current most successful way of doing science. (This does not rule out the possibility that specific criteria such as Thagard's have a place in IBE, but it does suggest that such criteria are unlikely to have a single, static interpretation across the whole of science at all times.)

Ben-Menahem's work marks a great leap forward. She illuminates the way explanation must be interpreted if we are to make sense of the claim that in inferring the best explanation, we thereby act rationally. But now we have some questions. How do explanatory standards change? How is the link with empirical developments maintained? How do explanatory criteria regulate inference across communities? As just mentioned, chapter 3 will answer these questions, but those answers rely heavily

on Lipton's account of IBE. Chapter 2 will therefore be devoted to a discussion of his work.

6. Summary

The purpose of this chapter has been to present IBE and give it some prima facie motivation. Section 1 introduced the distinction between deductive and inductive inference, suggesting that it's harder to characterise than many realise. Assuming a rough-and-ready definition of induction (inductive rules are accepted a posteriori), section 2 discussed three attempts to describe it – enumerative induction, hypothetico-deductivism and Bayesianism – concluding that all had serious faults, none of them able to capture central features of the way we consider evidence and make inferences. This set the stage for IBE, which was sketched in section 3 and shown to succeed where its competitors failed. One problem it couldn't overcome was Hume's problem, briefly discussed in section 4, but here IBE is in good company, since its rivals are just as unable to explain their success without circularity. Section 5 added detail to the sketch of IBE by discussing four notable developments towards a good account, due to Peirce, Harman, Thagard and Ben-Menahem. The key theme that came out of that discussion was that standards of good explanation depend on other things we know, and thus adapt to changing empirical circumstances.

Endnotes

[1] Norton (2003: 652-662) thinks that "virtually all" inductive systems in the literature are covered by his three families of 'inductive generalization', 'hypothetical induction' and 'probabilistic accounts', which correspond to the three kinds considered here. He includes IBE under the heading of 'hypothetical induction'. This is confusing. Psillos thinks that "IBE [itself] should be considered as an inferential *genus*" (2007: 444).

[2] The time-index in Goodman's example is dispensable and often thought to mislead. 'Grue' may be defined as 'green and observed or blue and not observed'; the same result follows. It's worth noting here that some doubt that Goodman's 'new riddle' is really a problem.

[3] This is emphasised in the discussion of Lipton's 'two-stage' process, described in chapter 2.

[4] This is not to suggest that for Hempel only predictions can confirm and only new hypotheses can explain. New hypotheses may be confirmed by old data, and old hypotheses can explain new observations.

[5] Ruben (1993) notes that, among others, Brody has suggested adding a causal requirement to the D-N model that may remedy this problem.

[6] Lipton (2007a) offers a different solution to the raven paradox. It relies on no account of induction, but rather exploits a general reliabilism about justification, in conjunction with a truth-tracking condition, inspired by Nozick. Interestingly, it may not be altogether distinct from the solution in Lipton (2004), in that it could be cashed out in terms of a causal-contrastive model of explanation, with causation analysed counterfactually.

[7] For a summary of the problems with the deductive-nomological model, see Newton-Smith (2000).
[8] Lipton (2004: 115-116) makes the plausible claim that the prior of the evidence can be fixed by explanatory considerations in this way. Okasha (2000) argues that the priors of the hypothesis and the evidence given the hypothesis can be fixed by explanatory considerations, but does not mention the prior of the evidence.
[9] Lipton acknowledges the influence of Okasha (2000) on this aspect of his book. Ben-Menahem (1990: 330) anticipates this kind of defence of IBE against the Dutch book argument, while Day and Kincaid (1994: 286) make precisely Lipton's claim, albeit in less detail.
[10] Hitchcock (2007) makes positive and negative comments about the prospects for Bayesian abduction. Psillos (2007) is negative about Bayesianism in general, and thus negative about Lipton's quest for conciliation.
[11] Although Peircean abduction is only a part of a full account of IBE, this thesis will follow tradition and, where necessary, use 'abduction' interchangeably with 'IBE'.
[12] Okasha says "a fundamental, and unresolved question is whether the Bayesians are *explaining*, or just *representing* these [scientific methodological] strategies" (2000: 706).
[13] Douven (2005) is one such character, agreeing that explanatory considerations can help the Bayesian, but arguing that Lipton has failed to establish that they're generally useful, much less that they ought to be used. Psillos (2007) argues from such considerations to a dilemma facing Lipton's project: "*either* accommodate (relatively easily) IBE within Bayesianism, but lose the excitement and most of the putative force of IBE *or* endorse an interesting version of IBE but radically modify Bayesianism" (2007: 448).
[14] Okasha (2000: 706-709) gives further reasons to favour IBE in the way described. In doing so, he diagnoses another source of error in van Fraassen's critique of IBE: in addition to assuming that explanatory value can only boost the posterior, he considers it only in the context of justification, ignoring its role in the context of discovery.
[15] Hacking (1983) notes that in his later writings Peirce became less enthusiastic about abduction.
[16] Peirce's pragmatism involved the view that truth is largely a matter of what we find worthy of belief (after proper inquiry). Since we generally find good explanations worthy of belief, for Peirce there was no problem of connecting such explanations with truth.
[17] Ennis (1968) offers purported counterexamples to Harman's claim that all cases of enumerative induction are cases of IBE. Harman (1968) shows how IBE can accommodate all three.
[18] Harman developed his own thoughts about IBE in his (1973). He emphasised that the inference be to the best of competing explanations, and developed the link to background knowledge via the idea of 'inference to the best total explanatory account' (1973: 158-161). He stresses that inductive inference seeks a coherent total set of beliefs, our inferential conservatism leading us sometimes to reject background beliefs rather than a foreground inferential conclusion. IBE found another early champion in N. R. Hanson, whose (1972) had already included a defence of abduction when first published in 1958 (see especially chapter IV). Elsewhere, and for reasons that will become clearer as we progress, the debate tended to split along realist/anti-realist lines, the former group generally positive about IBE, the latter generally negative. This is especially unsurprising given the zeal with which IBE was used to develop a defence of scientific realism based on the 'no miracles' argument, notably by Putnam (e.g. 1975) and Boyd (e.g. 1984). Laudan (e.g. 1981) is representative of the anti-realist stance. A flurry of notable criticism came IBE's way in the early 1980s. I discuss van Fraassen's (1980, 1989) arguments briefly elsewhere (see chapter 3, section 1.2.1, and 3.3); Hacking (1983) joins van Fraassen in rejecting IBE on the grounds that explanation is not a guide to truth. Their reasons are broadly positivist: explanations may be satisfying, even heuristically useful, but reasons for belief are given by empirical, especially predictive, success alone. Cartwright (1983, essay 5) agrees, though her realism about theoretical entities (which she shares with Hacking), and her recognition that causal explanation carries an existential commitment, mean she thinks 'inference to the likeliest cause' is legitimate. Inspirational to the pro-IBE counter-attack was Newton-Smith (1981). His realist account of science held that explanations were the goal, and his 'good-making features' of theories (1981: 226-232) are all explanatory virtues. His book contains no sustained discussion of IBE, but he does note his support for it, both as a scientific and a philosophical method.
[19] As a follower of van Fraassen, Thagard thinks explanation is a pragmatic notion, and thus doesn't think that his three criteria provide reasons to believe a theory is true.
[20] Fumerton (1980) argues against Thagard. He rightly points out that criteria such as Thagard's do not have an a priori justification, but wrongly thinks this means they are justified by enumerative induction: successful past cases of consilient, simple theories that use analogies justify the inference of further ones. He argues that all purported instances of IBE can thus be reduced to instances of enumerative induction. But the supporter of IBE can argue that Thagard's criteria are justified by IBE; the best

explanation of past theories' success is that they were consilient, simple and used analogies. The circularity involved here is merely Hume's problem.

[21] Fumerton (1980) may object that Ben-Menahem has described three examples in which explanatory virtues are justified by enumerative induction. But as footnoted above, there is a straightforward response available to the supporter of IBE.

[22] I do not fully endorse Ben-Menahem's position. I am agnostic about the prospects for the traditional philosophical project of finding some link between structural features of explanation and truth. It depends on a proper analysis of the relevant scientific inferences and theories.

Chapter 2

Lipton's account of IBE

1. Introduction: Lipton's project

In 1991, in the Introduction to the first edition of *Inference to the Best Explanation*, Lipton noted that IBE was widely-endorsed but under-articulated; his self-imposed brief was to "flesh out the slogan and give the model the detailed assessment it deserves" (Lipton 2004: 2). That supporters of IBE multiplied still more rapidly after Lipton's book was published is testament to the plausibility of his proposals and the skill with which he argues for them. The second edition of 2004 not only improves upon the original, but also ensures that Lipton's definitive ideas will be a focal point of epistemological debate for some time to come.

This chapter extracts the main structural features of Lipton's account of IBE, in particular the 'two-stage' process of generation and selection, and the guiding role of 'loveliness' (Lipton's word for explanatory virtues) at both stages. It also considers the nature of loveliness, trying to expand on Lipton's comments to get a clearer idea of what IBE's core notion amounts to. The main outcome of that discussion is that loveliness is heavily dependent on background belief, as was suggested in chapter 1.

Lipton's basic definition of IBE deviates little from Harman's: "beginning with the evidence available to us, we infer what would, if true, provide the best explanation of that evidence" (ibid.: 1). However, Lipton believes that "IBE cannot be the whole story about inference: at most, it can be an illuminating chapter" (ibid.: 4). Later, he says "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" (ibid.: 62). This is a clear departure from Harman, and Lipton's modesty allows his account to avoid criticisms to which a similarly sophisticated Harman would be vulnerable. One obvious advantage is that, with IBE one species of inductive inference among many, it is not as susceptible to counterexamples. However, at various points in his book, Lipton tends to forget his admission of limited scope and write as if he were defending IBE as *the* account of

inductive inference. This benign vacillation finds its way into the present discussion (Lipton's attitude is perhaps best attributed to an optimism about the scope of IBE, which I share).

One ambiguity that isn't reflected here is between Lipton's defence of IBE as an account of everyday inference and as an account of scientific inference. Throughout his book, Lipton mixes talk of belief-forming with talk of theory choice, sprinkling his discussion with examples of both. His final chapters are devoted to issues in the philosophy of science – novel prediction and scientific realism – but the bulk of the early discussion concerns IBE's ability to compete in the epistemological marketplace of general inductive theories. Lipton realises he's multi-tasking, even suggesting a third dimension to the discussion: "in addition to accounting for scientific and everyday inference, Inference to the Best Explanation has a number of distinctively philosophical applications" (ibid.: 67). His ambition is admirable, and indeed there's no reason to suppose that IBE can't offer a unified account of scientific and everyday inference – it's certainly not necessary that one should be dealt with separate from the other. But even though Lipton's approach is harmless, a division of labour yields greater productivity. Thus my defence of IBE in chapters 3, 4 and 5 focuses on the scientific case, if only in the hope that separating the two strands will lead to a better grip on whether IBE genuinely can account for inference across the board, and if so, what the scientific can tell us about the everyday.[1]

Lipton aims to spell out the IBE slogan by developing the following ideas:

> "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)" (ibid.: 56).

Lipton notes that in pursuing these themes, he assumes "inferential and explanatory realism", that is, "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" (ibid.: 57). These assumptions hardly sound controversial, until we consider scientific inference. The question of whether science aims at truth is hotly contested; the variety of anti-realist positions in the philosophy of science show that it's quite possible to believe all kinds of things about science except that it aims at, and sometimes attains, approximate truth. Lipton states his realist assumptions in the interests of full

disclosure, but he need not have done: realism is built into IBE no matter what Lipton says about it. As I argue in chapter 4, anti-realist versions of IBE are incoherent since they allow that what is inferred may be in some respects false. In this case what's inferred isn't an explanation, but rather a hypothesis that, *if* true, *would* explain.

That Lipton often misses this is surprising given his early emphasis on the distinction between potential and actual explanations (ibid.: 57-59). Potential explanations would explain the evidence if they were true, and actual explanations really do explain it. Lipton notes that Inference to the Best Actual Explanation is both implausible, since it would make us infallible, and trivial, since it would give no account of how explanatory factors are a guide to truth. IBE must be described as Inference to the Best Potential Explanation, or more accurately as the inference "that the best of the available potential explanations is an actual explanation" (ibid.: 58). Here Lipton seems fully aware that IBE commits us to the truth (or approximate truth) of the best explanation: actual explanations are (approximately) true. From the fact that a certain potential explanation would be the best explanation, we infer that it is indeed an explanation. Any account of inference that has us infer anything less is not a version of IBE.

The "an" in the last quotation is also worth noting at the outset. Evidence can usually be explained in several different, compatible ways. IBE is the inference that the best available potential explanation is one of these actual explanations. Users of IBE are not thereby committed to the falsity of every other explanation, only the falsity of those incompatible with the one they've inferred. To put it another way, IBE "does not require that we infer only one explanation of the data, but that we infer only one of *competing* explanations" (ibid.: 62). IBE is the inference that the best of a pool of competing potential explanations is an actual explanation (more is said below on the idea of a 'pool' of explanations). IBE doesn't force us to make this inference though; as Lipton notes, we may remain agnostic if none of our potential explanations are good enough to be inferred.

2. Likeliness, loveliness and the two-stage process

With the preliminaries out of the way, we may now turn to Lipton's crucial distinction between likely and lovely explanations. Lipton defines it thus: "the

explanation that is most warranted [is] the 'likeliest' or most probable explanation... the one which would, if correct, be the most explanatory or provide the most understanding [is] the 'loveliest' explanation" (ibid.: 59). The best explanations will be both likely and lovely, but the two sets of criteria can favour different explanations. Lipton attempts to show this with two quick examples: the 'dormitive powers' explanation of opium's tendency to induce sleep (likely but unlovely) and conspiracy theories (lovely but unlikely). I agree with Lipton's distinction, but I'm doubtful that these examples support his point. The dormitive powers explanation does display minimal loveliness, and faced with unlovelier competitors, e.g. mere paraphrases of the evidence or hypotheses with ad hoc clauses, we may infer it. Conspiracy theories are often lovely only at first glance; they are notorious for being unable to explain relevant evidence that the theorist has chosen to ignore, at least not without ad hoc clauses.

Likeliness and loveliness can be shown to mark distinct standards of explanatory quality, recommending different explanations for inference, but it's an open question whether any explanation is likely but completely unlovely or lovely but completely unlikely. Plausibly, the difference is only ever one of degree; if so, this might support IBE: if only explanations with some degree of likeliness can be lovely and only partially lovely explanations can have any degree of likeliness, this speaks in favour of the idea that explanatory considerations guide inference. It's intriguing to wonder whether our explanation-forming practices are naturally truth-linked, but the question cannot be pursued here (related issues inform my defence of IBE and the subsequent discussion of realism in chapters 3 to 5). It's also worth bearing in mind that whether or not we infer a minimally likely or lovely explanation depends in large part on the comparative likeliness and loveliness of other potential explanations we consider. If the competition is weak, a potential explanation does not have to be very good to be good enough, at least for the time being.

Lipton recognises the attractiveness of Inference to the Likeliest Potential Explanation as a definition of IBE: "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" (ibid.: 60). But this version of IBE must be rejected. A good account of inductive inference must tell us on what basis we judge one potential conclusion likelier (to be true) than another; baldly stating that we infer the likeliest explanation will not do: "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" (ibid.). In order to give an illuminating account of inductive inference that takes explanatory factors seriously, IBE must be defined as Inference to the Loveliest Potential Explanation. I endorse Lipton's definition; hereafter, 'IBE' will be taken to be synonymous with 'Inference to the Loveliest Potential Explanation'.[2]

(Note the following. IBE's main descriptive claim, that explanatory virtue is a guide to inference, and its main normative claim, that explanatory virtue is a guide to truth, can both be translated as 'loveliness is a guide to likeliness'. Lipton uses this phrase often, and it can be read in two ways: 'as a matter of fact we use loveliness to guide us towards judgements of likeliness', and 'as a matter of fact loveliness is a good indicator of truth'. The problem is that 'likeliness' may mean 'objective chance' or 'subjective probability'. In general, it's clear which meaning Lipton intends, but the ambiguity means he sometimes reads as though he's claiming that if loveliness is a guide to judgements of likeliness, then it's a guide to truth. This doesn't follow. Establishing that we do in fact judge likeliness on grounds of loveliness does not establish that we do so reliably. Indeed, another part of the problem is Lipton's reliabilism (see chapter 3, section 4), which allows him to move from the descriptive case that IBE is good account of our inductive practices to the normative case that loveliness is a guide to truth (we use IBE and live successful lives, so loveliness must put us in touch with the world in the appropriate way). I am sympathetic to this move, but have tried to eliminate Lipton's ambiguity. If it creeps back in at any point, I ask the reader to be charitable.)

We now know that IBE is the *inference that the loveliest of a pool of competing potential explanations is an actual explanation (where the loveliest is lovely enough to be inferred).*[3] To this description of the structure of IBE Lipton adds an account of the 'two-stage process', which tells us more about what happens once we've encountered some evidence that stands in need of explanation.[4] At the first stage of the process we generate a pool of competing potential explanations of that evidence. At the second stage we select the loveliest of that pool and infer it. Lipton's key point is that explanatory considerations – criteria of loveliness – play a role at *both* the generation and selection stages. That they play a role at the selection stage is obvious; IBE tells us to infer the best explanation, so we need some account of how we detect it, and we've already seen why loveliness must play this role. That it also plays a role at the generation stage is

less obvious, and more controversial. There is a strong descriptive reason to agree: we don't formulate all possible explanations of the evidence before we start the selection stage, yet those we do consider are not generated randomly. Even if we were capable of generating a full list of possibilities, to do so would be not only a huge waste of cognitive resources but also self-defeating: we'd spend so long at the generation stage that we'd never make it to selection. Instead generation is rigidly constrained; we select only from a shortlist of 'live' candidates. As Lipton puts it, "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" (ibid.: 149). We only ever allow a comparatively small number of potential explanations into our pool, namely those we have some prior reason to think might be actual (or at least no prior reason to exclude). Thus loveliness guides generation as well as selection.

Lipton's language in the above quotation reminds us of our key theme. He describes the shortlist mechanism as guided not by explanatory considerations or loveliness but by *background beliefs*. This is natural: even more obviously than at the selection stage (because there's no comparative assessment of competitors), at the generation stage loveliness is a standard formed prior to the inference in question. Lipton frequently acknowledges the role of background beliefs and prior inferences in the determination of loveliness, and rightly so. Experience teaches us what makes for lovely explanations: it informs us what criteria of loveliness mean and how to apply them in different situations. Experience also furnishes us with an increasing stock of beliefs with which new explanations ought to fit. Lipton clarifies this idea via an analogy with Darwinian variation and selection. In nature, complex organs don't arrive fully-formed, but neither do they evolve from complex-organ-parts; such parts would not perform their function properly and so would not be retained. Instead, complex organs evolve from simpler ones that played a somewhat different role. For example, "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" (ibid.: 150). The predecessor of the wing then mutated into the more complex structure that enables flight. This phenomenon, known as 'preadaptation', means that even though genetic variation is random, some types of complex organ – those that derive from earlier, simpler organs with different functions – are more likely to arise than others.

Just as natural selection plays a role in the generation and selection of complex organs, loveliness plays a role in the generation and selection stages of IBE. It might seem that, in generating the shortlist of candidate explanations, we are constrained not by loveliness but by simple plausibility judgements. But as Lipton explains (and as Ben-Menahem noted), such judgements are based on loveliness, since "the background beliefs that help to generate the list are themselves the result of explanatory inferences whose function it was to explain different evidence" (ibid.; see also 139-140). Background beliefs are like preadaptations: they limit the range of outcomes to those more likely to be useful, and they do so by defining a standard of loveliness. Lipton extends the analogy: by exploiting preadaptations, evolution populates nature with complex yet coherent organisms in which old and new variations fit comfortably together. With the generation stage of IBE constrained by previously inferred explanations that provide a standard of loveliness, our inferential practices tend towards a unified explanatory scheme that retains and extends what we already accept. This is sometimes called 'inferential conservatism', and there are independent grounds for thinking it's a policy we follow. If the mechanism by which we generate potential explanations is constrained by loveliness, then IBE can account for inferential conservatism. We simply seek to uphold the standard of loveliness generated by the background: "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… [M]ore old beliefs [are] 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" (ibid.: 151).[5]

Lipton is right that accommodating and explaining inferential conservatism counts in IBE's favour. It seems that we do indeed seek to preserve our background beliefs by only considering for inference those explanations that cohere with them. But there is a downside: if the generation stage of IBE is so restricted, how can it account for those occasions when new kinds of explanation are inferred? Such occasions are unusual, but they are real; it's particularly urgent that IBE should account for them if it's to fulfil its aim of describing scientific inference. Sometimes, science infers theories radically at odds with prevailing standards for theory assessment, and such theories score dramatic successes. Perhaps such instances receive excessive attention, but IBE had better have some story to tell about them,

and the restrictions on the generation stage seem to deny it the necessary resources. The challenge has a normative dimension: ought we to endorse a form of inference that threatens to limit the inferential freedom that science sometimes successfully exploits?

This is an interesting worry. We never generate a full list of possible explanations prior to inference, or generate candidates at random, but there are occasions when we loosen the constraints on the generation stage and consider potentially disruptive explanations. Although Lipton realises the importance of the background in determining loveliness (see especially ibid.: 138-140; his discussion of the generation stage uses 'explanatory considerations' and 'background beliefs' interchangeably), he stops short of fully articulating how this manifests itself in science. My account of IBE in chapter 3 explains this in a way that also shows how science's inferential conservatism is compatible with what I call 'revolutionary inference'.

Lipton's account of the two-stage process brings out the importance of background beliefs in relation to loveliness. Background beliefs, themselves inferred via IBE, form a coherent explanatory scheme. By doing so, they enforce a certain standard: they tell us that only some potential explanations are lovely (i.e. would, if true, yield understanding); these are the ones we generate. From that pool of live candidates, we select the one that fits best with the aforementioned standard of loveliness, and infer that it's an actual explanation. Explanatory loveliness guides us throughout the two-stage process in virtue of being determined by background beliefs. The following example illustrates this idea, bringing out other attractive features of IBE along the way, and giving Lipton's account some descriptive support (the example is inspired my own experience but has been embellished in places).

2.1. An example: police activity in Bristol

Not long after I moved to Bristol to start my PhD, I began to notice a large amount of police activity in the city. I frequently saw patrol cars, heard their sirens and sometimes encountered police officers in the street. Almost without realising, I'd started to consider potential explanations early on. Perhaps I was imagining it, or maybe it was just a coincidence. As I made further observations of cars, sirens and officers, in several locations over a period of weeks, I rejected these hypotheses; I

wasn't making it up, and the instances were too many and varied for it to be mere coincidence. My inclination then was to infer that I'd moved to a city with a high crime rate, until a friend suggested the hypothesis that the local police force was very good at responding to call-outs. I then generated a third candidate: hoax reporting of crime was popular in Bristol. I acknowledged that these could be compatible (perhaps Bristol's crime rate was high *and* the police were very responsive, or perhaps hoax calls were popular *but* so was real crime), but I construed them as competing, so for example I construed my friend's suggestion as 'Bristol's crime rate and the frequency of hoax calls are average for a big city but the police are very responsive' (this was how he intended it).

I did some research in an attempt to distinguish the three potential explanations; it turned out Bristol's crime rate wasn't especially high but neither was there any indication that the city's police force was unusually responsive. Statistics also showed that hoax calls were no more frequent in Bristol than any other comparable city. What else could explain the data? Perhaps I was simply more aware of the police. A friend had just joined the force, and I had recently been the victim of a minor mugging, so perhaps my frame of mind caused the observations to stick in my memory. This might also have tricked me into applying a kind of confirmation bias: once I'd started to see police cars and officers and hear sirens, I was primed to see and hear more. This was plausible, but could not fully account for the observations. They were too numerous and diverse to be the result of clouded judgement, and I wasn't the kind of person who regularly allowed circumstance to compromise his objectivity! A few years spent studying philosophy had also sensitised me to confirmation bias, so these explanations did not seem very lovely.

Another potential explanation of my police-observations was that I had moved to live near a police station; yet another was that my regular routes around the city took me near to police stations. In these cases, frequent observations of police activity wouldn't be surprising, but a glance at a local map showed that I wasn't regularly passing such locations. Later, when I had developed some local knowledge, I inferred an explanation I'd first formulated near the start of my investigation. My new house was located near one of Bristol's high-crime neighbourhoods; further, my regular walks took me through the city centre and along arterial routes that the police used to respond to most crimes. The explanation of why I was hearing police sirens

and seeing officers and their cars was that I was spending a lot of time in areas with a high volume of police traffic.

This example displays several features that support Lipton's account of IBE, but I want to isolate three that are relevant to our discussion: loveliness at the generation stage, the subjunctive aspect of IBE, and loveliness at the selection stage.

2.1.1. Loveliness at the generation stage

All the hypotheses I considered to explain my police observations had some degree of loveliness; each potential explanation could have created understanding by accommodating my evidence within my explanatory scheme. Looking back, I could have considered the hypotheses that I was undergoing some kind of extended hallucination, that Bristol police were out to impress me with their response to crime, or that their patrol cars were unusually luminous, their sirens unusually loud, and their uniforms unusually snazzy. But to consider these, and countless other possibilities, would have been a waste of time. Not only is there no independent evidence in their favour, but also they fail to cohere with my background beliefs (I'd gamble on their not cohering with yours either – they're not just a waste of time *for me*). Take the hallucination hypothesis. To the best of my knowledge, it is impossible to sustain a hallucination over such a long period, but even if it were possible, I would've had no reason to think one had been induced, and every reason to think that if one were induced, I would experience symptoms other than a heightened awareness of crimefighting.

This may sound like a judgement of likeliness, but likeliness is a matter of warrant given the evidence, and the evidence I sought to explain told me nothing about hallucinations. Rather, it's a judgement about the relationship between the hypothesis and other things I already believe – a judgement of plausibility. As Lipton noted, judgements of plausibility are judgements of loveliness, since they are based on background beliefs, i.e. the results of previous applications of IBE. My concern was that my explanation fit with the overall explanatory scheme I had discovered in the world. Recall also that loveliness is a matter of potential understanding. I didn't even consider the hallucination hypothesis as an explanation of my observations because my background beliefs wouldn't let me. I had the resources to formulate it – I knew about hallucinations and couldn't rule out the possibility that I had unwittingly

experienced one – but didn't, because even to entertain the hypothesis would've disrupted my background beliefs and thus been inimical to understanding. In fact, far from creating understanding, the hypothesis would've caused confusion. My avoidance of outlandish possibilities such as the hallucination hypothesis was due to the influence of loveliness on the generation stage of my inference.

But suppose my search for explanation was going badly, I had exhausted my supply of potential explanations, and there was no evidence to favour that which I in fact inferred. If my desire for explanation remained strong, I might have tentatively formulated the hallucination hypothesis at the generation stage. But I would still not have inferred it because it wouldn't have got through the selection stage. The reason is the same: despite being driven to scrape the barrel of potential explanations, I would rather make no inference at all than infer such an unlovely hypothesis. Given a pool of similarly desperate potential explanations, it might be the loveliest available, but it is not lovely enough to be inferred. Even though I've had to formulate it, the hallucination hypothesis remains highly implausible, and what little understanding it gives of my police-observations would come at the cost of great confusion elsewhere. Why did I only hallucinate police officers, sirens and cars? What on Earth caused it? How come it lasted so long? Why were there no side-effects? Suppose my research was thorough: these questions and others would then be even more baffling given that no other like hallucination was ever known to have occurred. In this situation, I would gamble on there being a lovelier explanation out there that I hadn't thought of yet, the costs of staying in inferential limbo being less than those of inferring a hypothesis that conflicts with so many other things I believe. Loveliness shows the hallucination hypothesis to be too far-fetched.

2.1.2. The subjunctive aspect

Returning to the original example, the second thing to notice is that potential explanations suggest a 'research programme' for inference by telling us what evidence is relevant to their acceptance or rejection. We suppose they are true, see what would follow and check to see if that's the case. If my friend's hypothesis, that Bristol police are especially responsive to crime, were true, then it would be evident in their statistics: if the force was so effective, crime rates would be low. It would also be advertised on their website and emblazoned over printed literature. In fact, none of

these things were the case. I rejected the hypothesis quickly because it directed me straight to the sources of evidence that would confirm or disconfirm it. On the other hand, I accepted the 'high volume of police traffic' explanation because its consequences were true: I did live near a neighbourhood with frequent crime, I did pass through areas with a naturally high police presence, and my journeys did take me along major roads. That inference proceeds via this kind of subjunctive assessment – what *would* be the case if such-and-such were true? how good an explanation *would* this be if true? – speaks in favour of IBE.[6] Lipton acknowledges this: "there seems no reason why an inferential engine has to work in this way… 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" (ibid.: 65).[7]

By starting with an explanatory hypothesis we are encouraged to ask appropriate subjunctive questions, and this makes inference more efficient by directing our attention towards the right kind of evidence. When we finally take the plunge, our inference is better supported than it would have been otherwise: it accounts for the original evidence *and* it's consistent with the other things we've learnt. I was persuaded to infer the traffic explanation because it not only explained all my original observations, but also was consistent with those collected in the process of eliminating other hypotheses: Bristol crime statistics, the location of local police stations, everything I'd seen on the police website and in related literature, and my reliability as an observer. Cases such as mine show that using explanation as a guide to inference makes matters of entailment and confirmation easy to understand and put into practice. IBE thus offers a partial explanation of why our inductive practices are so efficient.[8]

2.1.3. Loveliness at the selection stage

It might now seem that under IBE we generate a pool of potential explanations, consider what would be the case if each were true, and find evidence against them until there's only one left which both explains the original evidence and is consistent with whatever else we've found. The subjunctive research programme aspect of IBE thus threatens to make IBE purely eliminative. In the eliminative picture loveliness all but vanishes: IBE is a matter of discriminating between potential explanations with

empirical evidence alone. Loveliness is still at work at the generation stage, but in order for the core idea of explanatory considerations being a guide to truth to be fully realised, loveliness needs to constrain the selection stage too. And indeed it does.

The question of what would be the case if a potential explanation were actual is not the only subjunctive question we ask at the selection stage. We also ask how good an explanation would be if it were true. Even if a hypothesis explains all original data and is consistent with subsequent research, it still needs to count as lovely on other grounds. This is obvious if more than one candidate passes those tests, but loveliness isn't only applied as a tie-breaker. If only one hypothesis remains, loveliness is still required to check that it's good enough to be inferred. This enables us to avoid inferring bad hypotheses, but it also insures against the possibility of there being other, better explanations that we haven't thought of. If an explanation scores particularly well on criteria of loveliness at the selection stage, our general knowledge of explanatory standards should persuade us that it's as good as we're likely to get. If an explanation is found to be unlovely, we'll go back to the drawing board. Thus the fully eliminative picture is not accurate.[9]

This is the third important feature of my example: loveliness was essential at the selection stage; the fact that the traffic hypothesis offered to explain the data was not enough. Just as at the generation stage, loveliness at the selection stage largely comes down to fit with background beliefs. The traffic hypothesis does especially well here. There are various reasons for this, but it's largely due to the fact that, unlike, say, the hallucination hypothesis, it postulates no unusual causes, and thus sees my observations as instantiating the kind of worldly regularity expressed in the coherence of my background beliefs. According to the traffic explanation, my observations are part of the causal patterns found in the world. My background beliefs reflect those patterns, and I accept the explanation largely because the patterns aren't disrupted. Extraordinary and long-term hallucinations aren't the kind of thing that causes observations of police activity; regularly being near city centres, major roads and dodgy neighbourhoods is just the kind of thing that causes such observations. What's more, I know this. Thus the traffic hypothesis is both plausible and lovely: it extends my understanding of the world by fitting my observations into its causal structure.[10]

Here we touch on issues that resurface later. Eric Barnes (1995) criticises Lipton's account of IBE on the grounds that it reduces loveliness to fit with causal structure of the world (see section 4.2). This makes loveliness a guide to likeliness, but

at the cost of trivialising loveliness. Barnes' claim is that the links between loveliness, plausibility and coherence with background beliefs expressing causal structure are just too close. If our background beliefs accurately reflect the world, especially in its causal aspects, then judgements of plausibility based on those beliefs are going to track likeliness pretty closely. But what about loveliness? The danger is that we explain it away, and Lipton's ambition to show how explanatory considerations are a guide to truth goes unfulfilled, simply because there's nothing substantive to say about loveliness. The danger is even more acute given that Lipton's attempt to say something substantive, considered shortly, reinforces the idea that, ultimately, loveliness really is a matter of coherence with the causal patterns reflected in our background beliefs.

To hint at one way in which loveliness might be something other than agreement with a causally coherent background, consider the following. One potential explanation of my police observations that's also consistent with my later findings is that I simply spend lots of time near major thoroughfares. This explanation is a less detailed version of the traffic explanation that I actually inferred; in fact, it's entailed by that explanation. Even though it's more mundane, it would still have explained all my observations. But I wouldn't have inferred it when there were more interesting candidates around, because even though it's correct, it's comparatively uninformative, i.e. less lovely. In the event, I didn't even generate it, and this is typical of such possibilities: they fail to make the list of live candidates because even though they meet the constraints on the generation stage, they're just too dull. We might generate and even infer them somewhere down the line, but only once we've exhausted the more interesting possibilities we can think of. While there are potentially informative explanations out there, their more humdrum cousins remain unformulated. We might be inferentially conservative, but not that inferentially conservative.[11]

Putting this in terms of a criterion of loveliness is hard; detail can't be such a criterion, for the simple reason that explanations can be ruined by gratuitous detail. To borrow from accounts of theoretical virtue, it might have something to do with precision, i.e. that the traffic explanation is tailor-made for my observations, whereas the thoroughfares explanation could just as well be offered for many other phenomena. Or it could be to do with unification, i.e. that the traffic explanation unifies my observations with other disparate data – patterns of crime in Bristol, the behaviour of the police, popular routes through the city, etc – while the

thoroughfares explanation, being less detailed, does not have this scope. What we can say for sure is both the traffic explanation and the thoroughfares explanation explain all my observations, are consistent with the evidence that eliminated other candidates, and fit with my background beliefs. Yet the former is inferred ahead of the latter because it is lovelier. Thus there is something to loveliness besides fit with evidence and background belief.

2.2. Lipton on loveliness

We've now seen how Lipton defines IBE and, with the help of an example, how the two-stage process works. We've had to talk about loveliness quite a lot, but so far we've heard little from Lipton about what loveliness amounts to. We know that loveliness is identified with the creation of understanding and thus that it's a matter of 'explanatory considerations' or 'explanatory virtues'. We've also begun the discussion that will form the backdrop to most of this thesis, about the connection that Lipton urges between loveliness and fit with background belief. But there are reasons to think that loveliness isn't exhausted by fit with background belief. Lipton agrees. He acknowledges that if IBE is descriptively correct then loveliness is a guide to likeliness, and this might lead one to conflate the two, especially since, "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 (ibid.: 61). But IBE depends on "an account of explanatory loveliness that is conceptually independent of likeliness" (ibid.). An account purely in terms of fit with background belief is not conceptually independent; we've already seen how the connection between loveliness and plausibility/likeliness, beneficial for the claim the loveliness is a guide to truth, runs the risk of removing explanatory virtues from the picture altogether.

Lipton's desire to say something more distinctive about loveliness is also expressed in his claim that loveliness "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" (ibid.: 66). However, his desire is tempered by the observation that "the weakness of our grasp on what makes one explanation lovelier than another is discouraging" (ibid.: 61). Lipton is right, and perhaps it's outside the scope of his book to do the necessary spade-work, but given how central loveliness is to his account of IBE, there is a certain tension here. He

repeatedly acknowledges the need to say something substantive about loveliness and then excuses himself from doing so, either because he lacks the space, because it's too hard, or because it's tangential to the prevailing discussion and must be postponed (see e.g. ibid.: 122, 139). Lipton also thinks some of the job has been done by his account of causal-contrastive explanation (ibid.: chapters 3 and 5); in particular, he seems to regard meeting his famous Difference Condition (ibid.: 42) as a criterion of loveliness. His views are confused here; for instance, within a single paragraph he claims both that "on the side of explanatory virtues I have been able to [develop]… an account of contrastive explanation" and that there is such a thing as "lovely contrastive explanation" (ibid.: 122). If his account of contrastive explanation is part of his account of loveliness, then there's no such thing as an *un*lovely contrastive explanation. The loveliness claim is the one that has to go: the causal-contrastive model is an account of what constitutes an explanation, not of what makes one explanation better than another. So while Lipton says a lot about explanatory contrasts and quite a lot about background belief, he fails to get to the bottom of loveliness. This doesn't diminish his considerable achievements, but it does mean his ambition to defend a version of IBE in which explanatory virtue is the central notion is not fully met.

I pause here to note that this thesis does not discuss Lipton's thoughts on explanation. IBE is an account of inference that states that we infer explanations, irrespective of how the latter term is analysed. As long as the correct account of explanation isn't radically revisionary, i.e. doesn't depart far from our pre-theoretical understanding of explanation, IBE remains plausible. We have no reason to think such an account awaits us, so to defend IBE we do not need to defend a particular model of explanation. Of course, Lipton does defend an account of explanation, and his development of the causal-contrastive model is one of the most engaging passages of *Inference to the Best Explanation*, but it comes dangerously close to dominating his views on IBE. In particular, the contrastive aspects of the much-discussed Semmelweis case (ibid.: 75-90) lead to loveliness being just about forgotten, even though the example is supposed to be IBE in excelsis.[12] The root cause of such oversight is the above-mentioned equivocation between the Difference Condition as constitutive of explanation and as constitutive of loveliness. There's also Lipton's claim that his chapter on contrastive inference marks the beginning of "a specific argument on [IBE's] behalf" (ibid.: 64).[13]

Thus we see that as Lipton moves beyond his basic description of the structure of IBE, the causal-contrastive model starts to play a fundamental role in his account. His more detailed picture of how IBE proceeds depends on his account of contrastive inference, which via the Difference Condition, assumes causal-contrastive explanation. There's nothing wrong in this, but it serves to hide the fact that the reliance on causal-contrastive considerations is an artefact of Lipton's construal of IBE, rather than an essential part of the account per se (consider what would happen if we plugged the unification model of explanation into IBE).[14] Thus I leave explanation to one side; this focuses our attention on loveliness, the driving force behind IBE, and what we can say about it without endorsing a particular view on what it means for A to explain B (which is not to say that models of explanation don't rightly influence what loveliness amounts to).[15]

So what *does* Lipton say about loveliness? Well, the need to suggest a match between explanatory and inferential virtues leads Lipton to endorse the following criteria: "mechanism, precision, scope, simplicity, fertility or fruitfulness, and fit with background belief" (ibid.: 122).[16] He observes that there's widespread agreement that these virtues are both inferential and explanatory (even though accounts of them in either guise are controversial).[17] Thus his argument for the matching claim is minimal, consisting only of the list of virtues and admittedly uncontroversial claims about their dual identity (ibid.: 122-123). Sadly, Lipton's account of loveliness doesn't go much further. The following comments on simplicity, mechanism, and fit with background belief are the extent of his discussion.

Lipton says the following about simplicity: "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" (ibid.). This is usually attributed to unification, a virtue absent from Lipton's initial list, but introduced later as a criterion of loveliness (ibid.: 138). Indeed, Michael Friedman's influential 1974 paper on unification is referenced right after the above quotation; so what does Lipton think about *simplicity*? The later discussion reveals all: unification is an umbrella virtue, covering simplicity alongside scope (another from the initial list) and consilience. According to Lipton, unification contributes to loveliness because "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" (ibid.: 139). He endorses Friedman

([1974] 1988) and Kitcher ([1989] 2002) on unification, and Thagard's work showing that unification is an explanatory virtue (see chapter 1, section 5.3) but this is where his discussion ends.[18]

Lipton says little more about the virtue of mechanism. "We understand a phenomenon better when we know not just what caused it, but how the cause operated" (ibid.: 122). He illustrates this with the Semmelweis case: the cause of childbed fever in the maternity ward was infection by cadaverous particles, introduced via the mechanism of examination by medical students who had just performed autopsies. Lipton rightly claims that this supports the general case for explanation as a guide to inference, and the more specific case for contrastive inference that he champions. But as far as loveliness goes, it's not very illuminating. Lipton devotes part of his brief discussion to arguing that finding a mechanism isn't necessary for good explanatory inference, and his account of contrastive inference doesn't depend on it. This is obvious. An inference may be contrastive without being virtuous, and such an inference may be virtuous (i.e. be a good explanatory inference) without having the virtue of mechanism. Lipton's analysis of mechanism as a criterion of loveliness is disappointing.

Lipton's remarks on the virtue of fit with background belief are considerably more interesting (we were looking for help with the other components of loveliness, but some elucidation is useful here too). He argues that this virtue may count as inferential and explanatory because background beliefs may include beliefs about acceptable explanations. Science has, at one time or another, rejected explanations involving appeals to teleology, action at a distance and irreducibly indeterministic processes. Sometimes these bans have been revoked. This suggests that loveliness is contextual, "since the same hypothesis may provide a lovely explanation in one theoretical milieu but not be explanatory in another" (ibid.: 123).[19] Whether an explanation is a good one, and whether or not it is inferred, depend on the prevailing standards of loveliness determined by background belief. This point anticipates closely the account of loveliness I provide in chapter 3. That account exploits the contextuality of loveliness in science to show how it's both objective and truth-tropic enough to guide inference. Notably though, my account does not depend on beliefs about explanation and loveliness being part of background belief. Loveliness may guide inference without scientists having any beliefs about the kinds of explanation they accept; in fact, science might benefit from the absence of such beliefs.

Before clarifying the role background belief in IBE, let's remark on Lipton's comments on loveliness. Lipton realises there must be more to loveliness than fit with background belief: "*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... the core idea of Inference to the Best Explanation is that explanatory considerations are a guide to inference" (ibid.: 56, my italics). He has defined IBE as Inference to the Loveliest Potential Explanation in order to make good on this core idea, so for him to say so little about loveliness is disappointing to say the least. To be sure, there is already a good deal of agreement about explanatory virtues, so there's no harm in simply naming them and giving intuitive definitions. There's a similar amount of sympathy for the idea that explanatory virtues are inferential virtues. But simple endorsement of these popular views is not enough, for two reasons. First, IBE relies too heavily on both claims for Lipton not to say more about them. And second, in any case authors usually accept them only in principle. This typically means they look appealing in the absence of work spelling out what they mean, and if further work makes them unappealing then support will be withdrawn.

Consider simplicity: doubt creeps in about its status as an explanatory virtue whenever a serious analysis gets going. The effort to make simplicity fit the world in the way we want often reveals both simplicity and the world to be far more complex than we thought. Worse still, authors often endorse the claim that explanatory virtues are inferential virtues only on the condition that IBE isn't shown to be false. So general intuitive agreement about the constituents of loveliness and their role in inference can't be used to support IBE when the correctness of loveliness and IBE is itself at stake. We *first* need to be told about loveliness in a non-superficial way; *then* we can use that account to tell whether or not the identity of explanatory and inferential virtues is plausible. This is central to a defence of IBE, and it can be done without engaging in the thankless task (warned against in my Introduction) of singling out abstract explanatory virtues and identifying their connection with truth. The claim that explanatory virtues are inferential virtues *is* the claim that loveliness guides inference, i.e. *the* claim of IBE, so it deserves substantial defence. Lipton remarks, "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" (ibid.: 62). With his account of loveliness, Lipton leaves that exciting promise far from fulfilled.

With the exception of fit with background, I am not going to look any more closely at individual explanatory virtues. However, chapter 3 does do a little better than Lipton with respect to loveliness. Without giving too much away, my defence of IBE offers no close analysis of any particular explanatory virtue. Instead, it claims that standards of loveliness are to be identified with the loveliness-making features exhibited by certain fundamental scientific theories. An analysis of the way inference in science depends on those theories then establishes the claim that the explanatory virtues are inferential virtues.

My account makes progress by dispensing with Lipton's three-stage programme for a defence of IBE (ibid.: 121). Lipton's first stage is identification of the inferential and explanatory virtues. The second stage involves showing that these virtues match, i.e. that the loveliest explanation is the likeliest and vice versa (a perfect match is not required – the claim is not that explanation is our only guide to inference). At the third stage, we show that in practice, loveliness is a guide to our judgements of likeliness. Lipton notes that the first stage is hard, which makes the second stage hard too. But, at least as far as science is concerned, Lipton's three-stage programme may be inverted. This is the strategy adopted in my account of IBE. First, an analysis of scientific inference shows that loveliness, in whatever form the fundamental theories stipulate, guides judgements of likeliness. Second, an analysis of the structure of science and scientific progress yields the result that the loveliest explanation is indeed the likeliest. Third, an analysis of particular scientific inferences enables us to identify the virtues that constitute a particular standard of loveliness. My account makes no attempt on (what is now) the third stage, but makes some progress with the first and second, thus defending IBE via an account of loveliness in a way Lipton does not achieve.

3. The importance of the background

We know that loveliness has a lot to do with background belief, but the issue needs clarification; that is the aim of this section. I propose we make sense of the discussion by claiming that fit with background plays a dual role in determining loveliness. Firstly it's a *criterion* of loveliness: explanations are better to the extent that they fit with our background beliefs. Secondly it's what we might call a *meta-criterion* of loveliness: background beliefs determine what the other explanatory virtues are, and

fix their values. The distinction isn't too controversial; it's often remarked that the prevailing theoretical tradition determines how a science defines the theoretical virtues it seeks. Explanations must have those virtues, and it's likely that *one* of them will be a high level of coherence with the tradition. Thus there are two ways in which a hypothesis can be said to 'fit with background': it may cohere with other things we believe (criterion), and it may meet the other criteria of loveliness that that background defines (meta-criterion).

Consider an example: my guitar is out of tune because it has a faulty tuning peg. This explanation fits quite naturally with my background beliefs but doesn't unify my guitar's being out of tune with any other phenomena. It meets the criterion of fit with background, but not the criterion of unification. Yet if I propose a unifying explanation – my guitar is out of tune because it's an Acme 300, notorious for having faulty tuning pegs – it would meet the criterion because my background beliefs had determined what kind of things ought to be unified (instances of faulty Acme 300s) and how that unification ought to come about (suggesting a chain of events originating in the guitars' manufacture). In order to meet the unification criterion, the explanation must meet the meta-criterion of fit with background – the background that tells me what unification means.

Lipton seems to agree with this two-level view, saying that "background 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" (ibid.: 140). But this is hardly clear: Lipton could be describing the two aspects of the meta-criterion role: background beliefs tell us what the standards are, and determine the loveliness of an explanation in relation to that standard. Separate from this is fit with background as a simple criterion of loveliness. There's not much to say about it on this level beyond what we've already considered, but it's worth taking a paragraph to deal with a potential worry.

The way the background behaves at the selection stage, and especially the constraints it places on the generation stage, suggest that fit with background belief is less an explanatory virtue, more a pre-requisite for any degree of loveliness, and for that matter likeliness, to be present. Further, if an explanation is lovelier to the extent that it fits with background belief, it looks like we could ensure lovely explanations every time just by generating pools of the most mundane possibilities. This puts fit

with background in direct tension with the inference of explanations that create understanding, making it look like a kind of anti-virtue. Similar considerations motivate the claim, noted earlier, that fit with background makes it hard for IBE to account for successful scientific inferences that went against prevailing standards. These objections are not serious. My response, part of my defence of IBE in the next chapter, argues that the sense of tension derives from misunderstandings about what fit with background is and the part it plays in a standard of loveliness. Fit with background is one criterion among many; others insure against favouring mundane explanations. Also, a proper account of background belief will show that it guides inference in much the same way as virtues such as unification, simplicity, scope or consilience, virtues that (controversy over their analysis notwithstanding) promote interesting inferences.

Over the last few pages I've begun to sketch chapter 3's account of IBE. That account exploits the distinction between fit with background as a criterion and meta-criterion to show how loveliness guides scientists towards judgements of likeliness, and accurate ones at that. I now want to add a little colour to that sketch by doffing my cap to some perceptive remarks made by Lipton which anticipate my defence of his views:

> "what counts as a lovely explanation may be determined in part by previous explanations that serve an exemplary function, as Kuhn describes it... 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... 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" (ibid.: 139-140).

Lipton acknowledges that dependence on background makes loveliness *context-sensitive*, and foresees that Kuhn's account of science supports this idea with its notion of scientific traditions guided by theoretical exemplars. Further, he notices that loveliness might thus be relative to different scientific disciplines and sub-disciplines at different times. These are precisely the ideas that are developed in my account of IBE.[20] Fit with the theoretical background of a science, in particular its exemplars, acts as a meta-criterion of loveliness, fixing the interpretation of the first-order criteria of loveliness and the extent to which they're valued in that science. This idea of a plurality of standards of loveliness, each evolving within its specialism, each favouring different criteria of loveliness to different extents at different times, is compatible

with Lipton's development of IBE hitherto. "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" (ibid.: 140).

This is important. The background-dependent, context-sensitive view of loveliness has a fight on its hands, against the intuitive idea that loveliness is somehow timeless, the same features contributing more-or-less continuously to good inference throughout scientific history. This view is reinforced by many discussions of theoretical or explanatory virtues such as simplicity. Two things are worth saying here. First, loveliness is not to be identified with such broadly aesthetic considerations. The word 'loveliness' has aesthetic connotations, and there may be considerable overlap between aesthetic virtue and explanatory virtue, but loveliness is a matter of the latter. Loveliness-making features are features that make for understanding, not aesthetic value. Defenders of IBE do not need to argue for a match between aesthetic virtues and likeliness or truth, though they may do so if those virtues also contribute to better explanation. Secondly, the account of loveliness that Lipton suggests and I develop does not claim that loveliness is *necessarily* pluralistic and diverse. It claims that loveliness is relative to the theoretical background of the various sciences, and that because of this, different sciences may develop different standards of loveliness, and the standard in each science may change over time. But it allows for the possibility that the various standards between the sciences have much in common, and for the possibility that a single science cleaves to a standard that changes little despite changes in its theoretical background. There may be much that is constant about loveliness in science; in making loveliness context-sensitive, IBE merely allows for there not to be.

Psillos (2007) realises the value of a context-sensitive version of IBE, with background beliefs controlling all criteria of generation and selection. He even suggests that background determines the required explanatory relation; this is because of his sense that there's no single correct model of explanation and that we may avail ourselves of different kinds of explanation in different contexts. He says that such contextual considerations reveal "the fine structure of IBE", and that they remove the pressing need to link likeliness and loveliness in the abstract: "rather the connection stands or falls together with the richness and specificity of the relevant information available" (2007: 443).[21] Lipton (2007) agrees that IBE's compatibility with localised warrant diverts attention fruitfully away from Hume's problem (which for IBE is the

problem of linking likeliness and loveliness) and towards more tractable problems of justification. But he disagrees that the general problem of linking loveliness and likeliness thereby becomes less pressing. Given that the background will include general hypotheses, themselves reached by IBE, "there is an apparently coherent global question about the whole edifice and general inferential practice, and for an explanationist this is just the question about the general link between loveliness and likeliness" (2007: 461). We've already learned something about Hume's problem and its relation to the task of connecting loveliness and likeliness. More is to come, but the extent to which the relativity of loveliness to background alleviates the Humean burden will remain moot.

Lipton and Psillos both notice the value of context-sensitive loveliness, which we've been discussing as the idea that fit with background belief is a meta-criterion of loveliness, fixing the values of simplicity, unification, mechanism, and so forth. Now for two points on which I disagree with Lipton about loveliness and background belief.

3.1. Two departures from Lipton

My first disagreement with Lipton is this. Having rightly noted that "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 unity in one background context but detract from it in another", Lipton claims that "explanatory loveliness is *in part* relative to background" (2004: 139, my italics). I think that loveliness is *wholly* relative to background. Lipton's thoughts on unification are fine: recall that he thinks there are several unificatory virtues, and the contribution all of them make to an explanation's loveliness will depend on what else the background claims there is to unify.[22] But this generalises: fit with background as a meta-criterion fixes the interpretations of, and weights attached to, all explanatory virtues. Lipton implicitly recognises this fact with his claim about unification and mechanism potentially being absent from loveliness in some scientific traditions. His explicit claim about fit with background as a meta-criterion is, as quoted above, that "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" (my italics). But once we draw out the consequences of Lipton's

other comments, e.g. those about loveliness in Kuhnian science, we see that the general conclusion cannot be resisted. Given that fit with background plays the meta-criterion role Lipton specifies, we should conclude that background wholly determines loveliness.

The second point on which Lipton and I disagree is the role of background beliefs *about* loveliness. In short, Lipton's preferred view seems to be that on the meta-level, fit with background involves fit with beliefs about loveliness; my preferred view is that it doesn't. Lipton argues twice (ibid.: 122-123, 139) that one way in which fit with background acts as a meta-criterion of loveliness is by requiring explanations to fit with that part of the background that contains beliefs about explanatory standards. This suggests a picture in which scientists articulate their standards of loveliness and are fully aware of their implications, which I think is implausible. On the meta-level, fit with background would come down to fit only with those beliefs that express the standard that scientists' other background beliefs collectively determine.

This is confusing, so let's consider an example. Say I have a certain belief about unification; its content is a specification of the kind of unification I want and the degree to which I want it in relation to other virtues. Now all my candidate explanation has to do is meet that standard, i.e. fit with *that* background belief, in order to have the virtue of unification as I construe it. The same goes for the other explanatory virtues that make up my standard of loveliness.

My belief about unification is formed after I've considered all my other background beliefs; it expresses, in a single belief, the relationship I want my explanation to have with the remainder of my background. But if my candidate explanation would provide the right kind of unification in the right degree, it would do so in virtue of having the right kind of relationship with that remainder. So the belief about unification is dispensable. I might have it, in which case the meta-criterion of fit with background belief is really a matter of fit with the subset of background beliefs that express the requisite standard of loveliness. But equally I might not have it, and explanations would count as lovely in virtue of meeting the standard that my background beliefs collectively instantiate, but about which I have no beliefs at all. This is my preferred view. I hold it for two reasons. First, I advocate epistemological externalism (I realise Lipton does too). Under this view, any awareness we may have of the justification for our beliefs does not contribute to that

justification. Thus whatever beliefs scientists may have about loveliness are irrelevant to whether or not loveliness justifies their inferences. What matters is whether or not those inferences in fact conform to the standards expressed by their background beliefs, i.e. the background theory of their science. The way in which background theories determine a standard of loveliness means there's no need for scientists to have any beliefs about that standard in order for IBE to be effective. All they need is beliefs about the exemplary theories. My second reason for thinking that beliefs about loveliness should not be part of an account of fit with background is that science might benefit from scientists *not* holding beliefs about loveliness. The main reason for this is that in a discipline where standards of loveliness may change, any attachment to one standard will hinder scientists' ability to progress and make inferences under any other. Attachments to a particular standard of loveliness are going to be fewer and weaker in the absence of beliefs about it. Under the right circumstances, a fully articulated standard of loveliness that scientists explicitly endorse may prevent them from carrying out good science.

Hopefully the influence of background belief on loveliness is now clear. My example of police activity, Lipton's minimal account of loveliness, and some suggestive remarks about loveliness' context-sensitivity have brought out the idea that in order to be inferred, potential explanations should both fit with background belief, and fit with other criteria of loveliness, all of which are determined by background belief. Now I turn to two criticisms of IBE due to Barnes (1995), which concern the determination of loveliness by background and the way in which this may enable judgements of loveliness to track judgements of likeliness, and truth. Having replied to Barnes, we may also reply to the critic who claims that loveliness' dependence on background trivialises IBE.

4. Barnes' criticisms

Barnes' criticisms are these:

1. loveliness is not a guide to likeliness
2. loveliness is not an adequate measure of understanding.

The two are related. Regarding 1, he argues that our judgements of loveliness only coincide with our judgements of likeliness because both derive from a single source: a general picture of the world's causal structure. Barnes characterises this picture with what he calls the 'same cause/same effect' principle: "if an event (or conjunction of events) c1 is a cause of e1 in the sense that c1 induces the occurrence of e1, then an occurrence of c2 (an event or conjunction of events intuitively of the 'same type' as c1) will induce e2 (an event intuitively of the 'same type' as e1)" (Barnes 1995: 255). Regarding criticism 2, Barnes argues that loveliness is only an adequate measure of understanding on the condition that we accept this picture; thus it is the causal picture, not the criteria of loveliness, that accounts for potential understanding. The same cause/same effect principle is uncontroversial; what I want to tackle is Barnes' claim that this principle, and others that follow from it, are what give Lipton's criteria of loveliness plausibility, both as creators of understanding and as guides to likeliness. I will argue, with Lipton, that to the extent that Barnes is right, his conclusion is not worrying, but there's also an important sense in which he's wrong. The dependence of loveliness on background belief will help in both cases.

4.1. Criticism 1: mechanism and evidence

Barnes uses the mechanism criterion of loveliness to support his claim that loveliness is not a guide to likeliness (he also discusses the unification criterion, which is mentioned below, and the criteria of precision and elegance/simplicity, which are omitted here as I'm not convinced Lipton is committed to them as constituents of loveliness). Barnes invites us to consider the phenomenon of Harry (a man) hitting Fido (a dog) rather than Bozo (another dog).[23] We are offered two potential explanations: the 'Bite' hypothesis, which claims that Harry hit Fido rather than Bozo because Fido bit him and Bozo didn't; and the 'Childhood' hypothesis, which claims that Harry hit Fido rather than Bozo because Fido reminded him of a childhood pet and Bozo didn't. We judge the Bite hypothesis lovelier, by the lights of the mechanism criterion, since there is a familiar mechanism linking being bitten by a dog with hitting it, namely that we often lash out at those who injure us. The Bite hypothesis thus offers more understanding than the Childhood hypothesis, since in the latter case there is no familiar mechanism linking being reminded of one's childhood dog with hitting the dog that evoked such memories. We also judge the

Bite hypothesis likelier, and again the reason for this is the existence of the familiar mechanism. Thus that which creates greater understanding also tells us what's likelier to be true; loveliness is a guide to likeliness.

So far so good. But Barnes then argues that we only judge the Bite hypothesis lovelier because we implicitly conjoin it to the familiar 'lashing out' mechanism (call it M) that it invokes; by itself it is only as lovely as its competitor. OK, says the defender of IBE, make the conjunction explicit: the conjunction 'Bite hypothesis & M' is lovelier than the Childhood hypothesis. Barnes agrees, but argues that this move reveals that loveliness is not our guide to likeliness. He claims that we judge the conjunction 'Bite hypothesis & M' likelier not because it's lovelier (i.e. has the virtue of mechanism) but because we have independent evidence in favour of M. We know that events like dog bites tend to cause events like hits, so when we conjoin the mechanism to the hypothesis, the conjunction comes out as likely (to be true): "our knowledge of how common this type of mechanism is simply amounts to independent evidence that the Bite hypothesis is true – that bites cause hits" (ibid.: 259). Thus on Barnes' analysis, there's no need to appeal to the greater loveliness of 'Bite hypothesis & M' in order to explain its greater likeliness.

So Barnes' claim is that the Bite hypothesis, taken by itself, is judged likely and lovely because of an independent factor – its invocation of the familiar and plausible (i.e. well-evidenced) mechanism M. He brings this out by conjecturing a mechanism for its rival, the Childhood hypothesis, which is familiar enough but has no independent evidence in its favour. Suppose that Harry's childhood dog once befriended Harry's worst enemy. Through the years, he suppressed his hatred for the dog, until he saw Fido, when the memories came flooding back, causing him to lose control and hit Fido. Call this mechanism M'. Now Barnes claims that the conjunction 'Childhood hypothesis & M'' is just as lovely as 'Bite hypothesis & M'. It provides a moderately detailed causal story linking Harry's seeing Fido to Harry's hitting Fido rather than Bozo; that is, it would, if true, provide as much understanding as the Bite hypothesis. But, Barnes continues, 'Childhood hypothesis & M'' will not be judged as likely because M' does not enjoy the independent support that M does. There's nothing unusual about M' – under the right circumstances, suppressed anger could cause dog-hitting – it's just that "there is no common, independently well known sequence of events typified by those we imagine to constitute mechanism M'" (ibid.: 260). This is the influence of independently accepted causal picture mentioned

above. Again, Barnes concludes that it is the independent evidence for M, and the absence of such evidence for M', that leads us to judge the Bite hypothesis likelier, not the fact that we judge it to be lovelier.

4.1.1. Reply to criticism 1

Straight away, a response should be obvious. Contrary to what Barnes says, we *don't* judge 'Childhood & M" as lovely as 'Bite & M', because the mechanism criterion is relative to background belief. 'Childhood & M" doesn't provide as much understanding as 'Bite & M' because, given our background, the mechanism is less familiar. To be sure, M' is familiar enough in isolation; there's nothing out-of-the-ordinary about unpleasant childhood incidents leading to suppressed anger and subsequent violence. Indeed, M' is familiar enough given the example; we recognise that Harry may well have so hated his childhood pet that he lashed out when reminded of it, and the mechanism introduces no new mysteries into the situation. But M' is *not as familiar as M*, given the circumstances of the case. The familiarity of the mechanism invoked by a hypothesis is a matter of fit with background belief, and our background beliefs tell us that under the circumstances, although both M and M' have some degree of familiarity, M is more familiar than M'. It is more understandable that Harry should hit Fido because Fido bit him than that Harry should hit Fido because Fido reminded him of a hated childhood pet.

Of course, M is more familiar than M' is simply because we've encountered more cases involving M than M', a fact represented in our background beliefs about dogs, bites, bad childhood memories, natural reactions to such memories, etc. In general, we may have no reason to believe that bites cause hits any more frequently than childhood resentments; but in dog-type situations, it *is* more likely that bites cause hits. (Were Barnes to disagree here, he would undermine his own argument. If in dog-type situations, childhood resentments are just as likely to cause hits as bites, then we would no longer judge Bite likelier than Childhood.) Now the situation is back as the supporter of IBE wants it: 'Bite & M' is lovelier than 'Childhood & M" because it meets the mechanism criterion more closely by positing what is, under the circumstances, a more familiar mechanism. Furthermore, the fact that 'Bite & M' employs such a mechanism leads us to judge it likelier than 'Childhood & M". Loveliness is once again a guide to likeliness.

One possible response to this is to argue that M' is no less familiar than M with respect to *dog-type cases*. Suppose the case is as Childhood suggests. The situation is then one in which a chance event brings to mind a memory about which a person feels vengeful. It's then hardly surprising that that person behaves violently, since suppressed anger causes violence. Call this mechanism (suppressed anger causes violence) M". The example now represents a familiar chain of events; in particular, the specified mechanism fits the circumstances. Now 'Childhood & M"' seems to do as well by the lights of the mechanism criterion as 'Bite & M', since an informative and relatively thorough causal story has been told about why Harry acted in the way he did. It's entirely understandable that Harry hit Fido rather than Bozo, because Fido and not Bozo triggered the unhappy memories, which given Harry's suppressed anger, led him to act violently. With Childhood and Bite on an equal footing with respect to loveliness, the above response to Barnes doesn't go through.

But consider the following. First, under the new description of the Childhood scenario, Barnes' criticism doesn't go through either, since Childhood and Bite are equally likely as well as lovely. Hence there's no extra support, independent or otherwise, for M over M" to which Barnes can appeal to explain our likeliness judgement without referring to loveliness. This suggests the second point: the example still supports my account in that loveliness guides our assessment of likeliness. Any evidence for or against a mechanism is incorporated into background belief, which in turn determines our preferred interpretation of the mechanism criterion. 'Childhood & M"' is as lovely as 'Bite & M' since they both utilise familiar mechanisms, and this feature leads us to judge them both likely. The third thing to say to our opponent is that we needn't even agree that 'Childhood & M"' and 'Bite & M' are equally lovely. We have been unduly lenient both with respect to the scope of dog-type cases and the appropriate description of mechanisms. Although we've manufactured an account of the case in which Childhood may compete, the case is now no longer of the dog-type. It's just a case in which a chance event brings to mind a painful memory. And instead of the mechanism being M, we lash out when injured, or M', a childhood betrayal, issuing in suppressed hatred which resurfaced leading to a loss of control, it's now M", suppressed anger causes violence. Thus the mechanism isn't at all linked to the specifics of the case. Now, while redescribing events and the mechanisms behind them at different levels of abstraction can undoubtedly yield understanding, accounts at a higher level of abstraction tend to yield less than those

that contain details appropriate to the case. Thus even though it makes sense of Harry's hitting Fido, 'Childhood & M"' isn't really as lovely as 'Bite & M' or even 'Childhood and M", by the lights of the mechanism criterion. It isn't as lovely as 'Bite & M' because M is a mechanism appropriate to dog-type cases, and it isn't as lovely as 'Childhood and M" because M' is a more informative mechanism since it provides a more detailed explanation of why Harry hit Fido. Our opponent's attempt to make the Childhood hypothesis as lovely as the Bite hypothesis has turned out to be counter-productive.[24]

But now loveliness has fallen out of synch with likeliness. 'Bite & M' is arguably still the loveliest and likeliest hypothesis, but 'Childhood & M" is less likely than 'Childhood & M"', even though it's lovelier because more detailed. Bringing considerations of description and detail into the argument is problematic precisely because detail favours understanding but disfavours probability. By redescribing in more general terms the mechanism Childhood says is behind Harry's hitting Fido, our opponent has decreased our potential understanding of the case, but increased the chances of Childhood being true. Barnes uses just this point to discredit the precision criterion of loveliness (ibid.: 260-261). Lipton hardly discusses this criterion, but it is part of his initial list of explanatory/inferential virtues (see section 2.2). He seems to be referring to it when he says that "we understand more when we can explain the quantitative features of a phenomenon, and not just its qualitative ones" (Lipton 2004: 122). The debate about the benefits of detail is already part of the literature, and there isn't space to discuss it further here. The supporter of IBE needn't be too downhearted though. Sometimes we want detailed, precise explanations, sometimes not. If we do want detail, it's detail appropriate to the case, as I have insisted; unnecessary detail often spoils an explanation, and might stop it from explaining altogether.

The point is twofold. The demand for precision and detail by itself *is* antagonistic to loveliness as a guide to likeliness; but the precision criterion is never by itself – it is just one of many criteria of loveliness. So loveliness may yet be a guide to likeliness when viewed holistically, as a collection of different criteria, even though individual criteria may, under the right circumstances, be in tension with likeliness (and indeed with each other). Further, although Barnes' criticism of the precision criterion is decisive on his construal, there may yet be a more plausible version of the criterion (perhaps Lipton's quantitative/qualitative distinction will help here). The

hope is that this version will make sense of the contribution that precision and detail make to understanding without sacrificing the idea that loveliness is a guide to likeliness.

Summing up, Barnes claimed that we judge 'Bite & M' to be likelier and lovelier than 'Childhood & M" because of independent evidence in favour of M. We responded by saying that 'Bite & M' is lovelier than 'Childhood & M" because the mechanism criterion is determined by background belief. This enables IBE to accommodate Barnes' insight that in situations such as the one he describes, hypotheses' loveliness and likeliness are both due to the plausibility of the mechanism they mention. With loveliness determined by background, that plausibility is not independent of loveliness and likeliness; on the contrary, assessments of the relative plausibility of mechanisms given the details of the situation play a central role in both kinds of judgement. This means loveliness is a guide to likeliness.

Similar thoughts give a response to Barnes' criticism of the unification criterion, which is analogous to his criticism of the mechanism criterion. Barnes uses the example of the 'sympathetic powder' hypothesis about recovery rates among 17th century soldiers (cf. Lipton 2004: 136). It was proposed that victims of sword wounds could be effectively treated by rubbing a certain powder on the offending sword. Indeed, soldiers treated with the powder tended to recover more quickly than soldiers treated by established methods of the time. However, it is now thought that the contrast in recovery rates is due to the fact that those methods were counter-productive; soldiers given normal medical treatment were often thereby harmed, while those treated with the sympathetic powder were simply left alone. Lipton would claim that IBE accounts for this situation: the sympathetic powder hypothesis is rejected because it fails to meet the unification criterion and is thus unlovely.

Barnes articulates Lipton's position by borrowing from Friedman ([1974] 1988), which Lipton endorses. Barnes explains that the sympathetic powder hypothesis fails to unify because it would replace one brute fact (soldiers treated with the powder heal quicker) with another (sympathetic powder causes quicker healing). The latter does not explain the former because it doesn't reduce to other phenomena we already accept; in fact, it decreases understanding because the explanans is utterly mysterious. We also judge the sympathetic powder hypothesis unlikely, and Lipton's claim is that the unloveliness accounts for the unlikeliness. But Barnes argues that our judgement of unlikeliness is accounted for by the fact that the sympathetic powder hypothesis "is

strongly disconfirmed by the totality of existing data, which provides extensive evidence that causes of the sort posited by the powder explanation (action at a distance causing events like wound healings) do not exist" (Barnes 1995: 263-264). So once again, although loveliness and likeliness coincide, Barnes claims the former is not guiding us with respect to the latter, and a general picture of the world's causal patterns determines both types of judgement.

But with the background-dependence of unification revealed, the response is again obvious. The new brute facts that fail to explain old ones are only brute because they fail to fit within the systems of phenomena we already understand. In Barnes' words, the sympathetic powder hypothesis would "increase the number of fundamental mysteries in the world, for it would posit a phenomenon that could not be reduced to those phenomena already accepted as real" (ibid.: 263). The hypothesis fails to provide understanding – it's unlovely by the lights of the unification criterion – in virtue of the fact that our background beliefs enshrine a certain conception of the world, one in which phenomena like that of the sympathetic powder do not figure. But now our judgement of the hypothesis' likeliness cannot be accounted for without reference to its loveliness. Our judgement that it's unlikely is due to the fact that according to our background beliefs, phenomena like the one it posits are brute and mysterious – they cannot be incorporated into existing hierarchies, systems, patterns etc. That is, our judgement that the hypothesis is unlikely is due to the fact that we find it unlovely. Contrary to Barnes' argument, loveliness is our guide to likeliness in the sympathetic powder case and all similar cases.

Because criteria of loveliness are determined by background belief, the evidence for or against the phenomenon mentioned in a potential explanation is not independent of loveliness, as Barnes tries to show. Our judgements of how much understanding a hypothesis would provide take into account the plausibility of the phenomenon alleged to do the explaining. One might think this was obvious, but it's only when we look properly at fit with background as a meta-criterion of loveliness that we can incorporate it into IBE. Evidence relevant to our potential explanations is not independent of judgements of loveliness because it's reflected in background belief, which in turn determines the various criteria of loveliness. Barnes' attempts to account for judgements of likeliness in terms of evidence are successful, but he does not thereby show that loveliness is not a guide to likeliness.

4.2. Criticism 2: causation and understanding

Barnes' second criticism of IBE concerns loveliness as a measure of potential understanding. He returns to his overarching claim that a certain causal picture of the world, characterised by the same cause/same effect principle, is doing the real work behind loveliness. Barnes argues that the mechanism criterion is only an adequate measure of potential understanding on the condition that we accept that causes are in fact transmitted mechanistically. He claims that, were there to be an effect of non-mechanistic causes, then we would understand that effect completely by apprehending those causes. In such a case, it would not be true that a potential explanation meeting the mechanism criterion "would offer more potential understanding than the explanation citing the (actual) non-mechanistic causes, since the understanding offered by the latter is absolute" (ibid.: 271).

Again, Barnes makes an analogous claim on behalf of the unification criterion, arguing that it "depends for its plausibility on the prior assumption that the world does indeed have a unified structure: should the ultimate causes of the world be highly fragmented, it would be a fragmented theory, and not a unified one, that would maximize scientific understanding" (ibid.: 272). Thus he concludes that loveliness, to the extent that it comprises criteria like mechanism and unification, is not a good characterisation of potential understanding. Understanding is a matter of revelation of causal history (Barnes and Lipton both argue), and were the world to have a causal make-up different from the one we assume, lovely explanations would not provide understanding. Accordingly, Barnes suggests that IBE should dispense with loveliness, and account for potential understanding in some way that makes direct reference to causal history, however causation may in fact be.

4.2.1. Reply to criticism 2

Lipton (2004: 123-124) responds to this criticism himself. He denies the claim that revelation of causal history is all there is to provision of understanding, arguing that the description of the causes we cite is also relevant. This is especially obvious in the interest-relativity of explanation: in order to maximise understanding, cited causes must be described in a way appropriate to the interests of the audience. But no matter how we describe causes, our demand that explanations reveal them in a way that

expresses unity or mechanism will not lead to understanding if it turns out that causes work differently from how we assume. Lipton realises this, and says bluntly, "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" (ibid.: 124). Lipton's view is that understanding is tied to unified, mechanistic causes, and if the world turns out to be unlike that in some respects, then our understanding will be correspondingly limited.

This sounds like a concession, but Lipton's reference to possible worlds has an interesting implication. Plausibly, he doesn't just mean that there are various ways things could be, some of which are more amenable to understanding than others; he also means that, as a matter of fact, such possibilities are not actual, and indeed those causally chaotic, disunified worlds are distant from the actual world. Interpretation of possible worlds is always controversial, but if there is to be a measure of closeness to actuality, such fundamental features as causal structure are plausible candidates. While we must acknowledge that causally non-mechanistic, disunified worlds might be closer than we realise, there is no evidence that this is actually the case (except perhaps at the limit), and an overwhelming amount of evidence that it isn't, so it's fair to assume that causation is set up in just the way we think it is, in this world and in all close possible worlds. Thus Lipton's point is that loveliness *is* based on certain causal assumptions, but there's no reason to think that those assumptions aren't correct. Barnes is right that, if our assumed causal picture were false, understanding would be provided by explanations unlovely by current lights. But our causal picture isn't false, so as a matter of fact, lovely explanations provide understanding.[25]

So Barnes is right that the mechanism and unification criteria of loveliness strike us as plausible because they express certain parts of our causal picture. Loveliness as a whole can be seen in this way. Given certain assumptions about a causal model of explanation, a standard of loveliness can be viewed as an articulation of the structures and patterns we perceive in the world. If our perceptions are reliably formed and those structures and patterns really are there, then the worry of Barnes' criticism 2 simply disappears. Loveliness does presuppose a certain causal picture, but that doesn't mean that the causal picture's doing the work with respect to explanation and loveliness should be dropped. Loveliness *just is* that causal picture; it's the causal picture expressed in such a way that it forms a standard of good explanation. Moreover, this fails to be worrying on account of the causal picture being accurate.

Given our view that loveliness is determined by background belief, we might ask the following question about Barnes' criticisms: if loveliness' provision of understanding and the coincidence of likeliness and loveliness are due to background beliefs accurately reflecting the causal structure of the world, then what's the problem?

I agree with Lipton about the likelihood of alternative causal pictures; criteria of loveliness are in fact conducive to understanding because they express what is in fact the causal structure of the world. But with loveliness determined by background beliefs in the context-sensitive way described here, we may admit the possibility that our causal beliefs are mistaken *without* sacrificing loveliness. Let's assume, with Barnes, that there is a real possibility that the world is in fact causally non-mechanistic and disunified. This is compatible with my account of scientific IBE sketched above, which sees loveliness as defined by the theoretical background of a particular science. On that account, loveliness doesn't depend on our present causal assumptions; it depends on the picture of the world given by prevailing theory, whatever that may be. As it happens, theory defines a conception of the world in which criteria such as mechanism and unification are conducive to understanding, but were they to recommend a different picture, loveliness could change accordingly, perhaps dispensing with mechanism and unification altogether. Whatever criteria did express this newly-discovered causally non-mechanistic, disunified world, they would probably seem unlovely in the aesthetic sense of the word, but that's not important (and anyway they would only seem aesthetically unappealing *by our lights*). What's important is that in such a world, they would contribute to understanding. We should agree with Barnes that loveliness is dependent on our causal picture because understanding is identification of causes, but if that causal picture were to change, loveliness would simply change along with it.

It's difficult for us to conceive of a causally chaotic world being *explicable* at all. It seems such a world would be *necessarily* confusing, no matter how accustomed we grew to it. This is because along with explanation, prediction and all kinds of inductive inference would be impossible too. This is the heart of Lipton's point about possible worlds, and I agree. But my claim against Barnes is that the changeability of causal assumptions doesn't show that loveliness is eliminable. Criteria of loveliness *as traditionally conceived*, including those discussed by Lipton, would not promote understanding in a causally chaotic world, but loveliness, as defined by background, can adapt to such circumstances. Loveliness may yet be both universal and constant;

my point is that loveliness is just whatever promotes understanding, in this causally deterministic world or any other.

5. Why IBE is not trivial

Having dealt with Barnes' criticisms, we are now in a position to see why the dependence of loveliness on background belief does not trivialise IBE. The worry would be that the gloss I've put on loveliness threatens the conceptual distinctness of likeliness and loveliness; judgements of loveliness and likeliness both look like judgements of plausibility, in that they're both based on how well a hypothesis fits with what we already know. This was apparent in our reply to Barnes' criticism 1, which talked about how allegedly independent evidence for a certain mechanism or causal pattern may be incorporated into an assessment of loveliness. If 'loveliness' doesn't name a set of specifically explanatory criteria, then IBE's claim that explanatory factors are a guide to truth is reduced to the trivial claim that we judge likely truth on the basis of other things we know.

The worry is answered by our reply to criticism 2. Loveliness is only background-dependent in the sense that the very idea of inductive practice is background-dependent. Likeliness and loveliness are both background-dependent in this sense, but how could it be otherwise? We learn from experience and increase our stock of background knowledge; this inevitably influences every piece of reasoning we subsequently undertake. The idea that this might be for the worse, that loveliness is an independent standard that fickle experience can only prevent us from perceiving, is an example of discredited rationalist thinking. It's our body of background knowledge that tells us what causation is, what inference is, what's likely, what's plausible, what explains, what's a lovely explanation. Against this unavoidable background, the conceptual distinctness of likeliness and loveliness is maintained if we may talk about them in distinct terms, or at least without too many terms overlapping, as indeed we can. We can talk about what a familiar mechanism is, what kind of unification we want, what simplicity means, and so forth. As long as any epistemic notion is conceptually distinct from any other, loveliness *is* conceptually distinct from likeliness (and the whole dialogue about explanatory virtues has not been about nothing).

If loveliness were merely fit with background as a *criterion*, to the elimination of all other criteria, then IBE *would* be trivial. Any hypothesis that didn't violate other

things we accept would count as lovely. But we've been talking about fit with background as a *meta-criterion*, background beliefs determining the content of the various criteria of loveliness. This does not trivialise loveliness, it simply shows that whether any criterion of loveliness is a (good) guide to likeliness can only be discovered a posteriori. There is much evidence, particularly from science (some of which is considered in the next chapter), that there is no a priori link between loveliness and truth and that scientists have used different standards to guide them towards explanations at different times (see chapter 1, section 5.4). This is the real moral of this chapter: the dependence of loveliness on background belief shows only that it's context-sensitive, not that it's at the heart of a trivial model of inference.

6. Summary

This chapter has discussed Lipton's account of IBE and drawn out the theme of loveliness' dependence on background belief. Having set up the core structure of IBE – a two-stage process of *generation* of potential explanations and *selection* from among them, with both stages guided by loveliness (explanatory virtue) – we saw how it worked in a specific case, involving my observations of police activity in Bristol. We then looked at what Lipton says about loveliness, noting that his description of what it was and how it might work betrayed its intimate links to background belief. Section 3 aimed to clarify the discussion by distinguishing two senses of fit with background, one in which it acts as one of a range of criteria of loveliness (ceteris paribus, hypotheses ought to cohere with background beliefs), and another in which it acts as an overarching meta-criterion (hypotheses ought to meet the criteria of loveliness, all of which are determined by background beliefs). The section then sketched out the kind of context-sensitive view of loveliness used to defend IBE in the next chapter, before section 4 turned to Barnes' two criticisms of IBE. Loveliness' ability to accommodate, through its background-dependence, judgements of plausibility based on independent evidence, enabled us to reply to both. Those replies in turn enabled us to respond to the charge that background-dependence trivialises loveliness and therefore IBE. In particular, we were able to note that background-dependence merely shows that loveliness is context-sensitive and only contingently connected to truth.

[1] It's part of Lipton's rhetoric that if IBE can be shown to apply to everyday and scientific inference, the undisputed realism of the former will count against anti-realism about the latter, though it won't banish it altogether.

[2] Lipton (ibid.: 61-62) realises that a more defensible version of IBE may give likeliness a bigger role.

[3] It's worth remembering that this is what IBE really means. Even after Lipton, many authors fail to acknowledge this definition when they press 'IBE' into service.

[4] Grimm (2008) uses Lipton's causal-contrastive model of explanation to account for when evidence stands in need of explanation.

[5] The idea of a background of beliefs generated by loveliness, which in turn define a standard of loveliness, is crucial in Lipton's reply to van Fraassen's 'best of a bad lot' argument (see chapter 3, section 1.2.1) and to a certain extent in his reply to van Fraassen's 'Dutch book' argument (see chapter 1, section 3.3).

[6] Although hypothetico-deductivism recommends the asking of subjunctive questions via explanatory hypotheses, IBE arguably has the upper hand. Alongside such hypotheses, we normally need to rely on other things we independently accept in order to determine evidential relevance. With IBE, these considerations are 'built-in' to the potential explanation thanks to its having been generated by a mechanism constrained by those background beliefs (cf. ibid.: 65).

[7] Lipton (ibid.: 137) argues that this subjunctive process favours IBE because it's easier and more natural to conduct it in explanatory terms rather than brute causal terms, i.e. it's another point in favour of our tendency to think about inference in terms of explanation.

[8] Lipton (2004: 65-66) notes other ways in which we benefit from an inferential interest in explanation, even when it's not our main goal. For example, discovering causes of phenomena enables us to predict and control them. IBE is supported by the fact that it can account for inferences that make an 'explanatory detour'. IBE also justifies our concern with explanation: "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" (ibid.: 66).

[9] Despite his emphasis on loveliness, Lipton occasionally strays into talking about IBE as if it were purely eliminative. This is particularly so when he discusses the famous Semmelweis case. Elsewhere, I argue that the case can be straightforwardly redescribed to bring out the role of loveliness at Semmelweis' generation and selection stages. Bird (2007) argues that cases such as Semmelweis' support an eliminative model he calls 'inference to the only explanation'. Lipton (2007) responds, remarking that Semmelweis' is an ideal case.

[10] Like all inductive inferences, IBEs are defeasible; in particular, they may be defeated upon presentation of a lovelier hypothesis. Elimination is defeasible too: hypotheses are rarely refuted, but shown to be implausible/unlovely, as Lipton notes: "the additional evidence, though logically compatible with both hypotheses, can only be explained by one of them" (2004: 136). In the absence of a reason to believe them, they are abandoned, but reasons may appear later to change our minds (for instance, in my example I may come to doubt my reliability, which would make me doubt that there was a phenomenon to be explained). Note that defeasibility doesn't depend on the appearance of new evidence.

[11] The fact that we go for interesting possibilities rather than boring hypotheses with higher probabilities explains why we do so badly at those psychological tests like Jill the bank teller. It also favours IBE: see Lipton 128-132 on 'explanatory obsessions' and psychological evidence to this effect.

[12] Again, see Bird (2007) and Lipton (2007).

[13] Because I ignore contrastive inference, I also ignore the matter of the similarity between IBE and Mill's methods (see e.g. Lipton 124-128, 132-135). For discussion see e.g. Achinstein (1992) and Rappaport (1996).

[14] It would be interesting to develop an account of IBE based on the unification model. Day and Kincaid (1994: 275-279) consider such an account, even saying that "one *common* rendering of IBE equates explanation with unification" (276, my italics). Though interesting, their discussion is hampered by the all-too-common confusion of unification as model with unification as virtue.

[15] By sidestepping Lipton's views on explanation we also avoid getting bogged down in a sticky debate. He tells us why causes explain in his (2004a), but Achinstein ([1981] 1993) argues that the causal model cannot be illuminating. Salmon ([1990] 2002) argues that unification and causal models are the only kinds viable, and advocates a causal model, but concedes unification might also be needed to account for all scientific explanations. Ruben (1990) argues that unification can only be an explanatory virtue, not the basis for a constitutive account of explanation. This gives a flavour of what awaits the explanation theorist!

[16] Lipton notes the influence of Kuhn's 'five values' (Kuhn 1977): see chapter 3, section 6 here.

[17] See chapter 1, section 5.3.

[18] Friedman and Kitcher develop unification as an account of explanation not as an explanatory virtue, though at several points the language of Kitcher ([1989] 2002) suggests otherwise. Lipton's 'umbrella' view is supported by the fact that all three of Thagard's criteria are easily construed as kinds of unification. Kitcher ([1981] 1988) endorses this by arguing for unification via the Darwinian example Thagard used to illustrate consilience.

[19] In making these comments, Lipton acknowledges his debt to Day and Kincaid (1994). Those authors use the contextuality of IBE to argue that its force qua rule of inference depends on substantive empirical assumptions. Interestingly, Norton (2003) doesn't refer to Day and Kincaid, but expresses the same view: "inferences to the best explanation are licensed by facts pertinent to the local domain that supply us explanatory resources" (658). He argues for the same result for all accounts of induction: there is no general rule of induction; whether an inductive inference is justified depends on facts specific to the case. I'm not sure I agree with Day and Kincaid's and Norton's conclusion. Suffice to say, at present, I'm inclined to think that there *is* a general rule of induction, viz. IBE, and explanatory loveliness, as described here, allows context-specific features to have their proper influence. Justification is context-specific because loveliness is context-specific, but this doesn't mean we have to sacrifice the idea of inductive rules. Even if those rules are just a generalisation from the features common to context-specific inferences, IBE still makes substantive claims, as per the arguments developed here. Psillos (2007: 443) supports this view.

[20] It's worth noting again the influence of Day and Kincaid (1994), whose comments on the contextuality of IBE include the following: "while there may be some constraints on explanation that hold across all scientific domains and epistemic contexts, it is unlikely that those constraints will be the whole story about explanation. In brief, the requirements for good explanation are likely to invoke divergent, domain-specific principles" (1994: 282).

[21] Norton (2003: 666-669) argues that his material theory of induction localises justification in such a way that it avoids Hume's problem, except in a relatively minor form. This is suggestive: could context-sensitive IBE claim the same?

[22] This isn't to claim that unification is a matter of unifying explanations or finding a unifying vocabulary. The confusion of this with the unification of phenomena is a common problem in the discussion of unification.

[23] The example involves contrastive explanation because of Barnes' other interests. It is not important here.

[24] Further, the redescription of phenomena and mechanisms in order to test how much explanation the latter would provide might be evidence in favour of my emphasis on loveliness as determined by background. It's arguably only possible to redescribe in this way because we exploit background belief as a meta-criterion of loveliness which then guides us towards likeliness. We know what kinds of thing would be appropriate under what circumstances and come up with descriptions of the evidence and the mechanism accordingly. Thus our opponent may be undone by her own philosophical manoeuvring.

[25] Ladyman (2005: 334) endorses these points. Interestingly, he goes on to note that if we may legitimately reject the idea of worlds in which there is utter causal disunity and hence total failure of induction, an a priori justification of IBE is closer than Lipton realises.

Chapter 3

A Kuhnian defence of IBE

1. Introduction: two crucial objections

The dependence of explanatory loveliness on background belief should make plausible the claim that judgements of loveliness guide us towards judgements of likeliness. We judge explanations that fit with other things we know to be likely, and they also provide great understanding. But there's a problem: we may *judge* lovely explanations to be likely, but are they really? That is: is loveliness indicative of truth? Further, if loveliness is relative to background beliefs, doesn't that make it dangerously subjective? After all, what's lovely for you might not be lovely for me, if our backgrounds are sufficiently different. If loveliness is subjective, then it can hardly be a guide to inference.

Lipton (2004: 70; 142-163) expresses these worries as Hungerford's objection and Voltaire's objection:

Hungerford's objection: *beauty is in the eye of the beholder*; loveliness is too subjective to be a guide to likeliness. Truth is objective, loveliness is not. Conceivably, different groups of scientists may fail converge on a single explanation when asked to select the loveliest from any given range of competitors.

Voltaire's objection: IBE assumes that we live in *the loveliest of all possible worlds*; this assumption is unwarranted. There's no reason to think that in choosing the loveliest explanation, scientists thereby choose a true explanation. Even if loveliness were a source of agreement (answering Hungerford's objection), it may yet fail to correlate with truth.

I take these to be the two most serious challenges facing IBE. Lipton himself does not respond adequately. In both cases, his response consists of a descriptive argument to the effect that since IBE accounts for our inferential practice, and this practice is reliable, the objections do not raise genuine problems. Although Lipton's

efforts help to diffuse their worries, he fails to acknowledge the objections' real challenge: to provide a normative justification of IBE that would give us reason to think our reliance on loveliness is legitimate. We now look at Lipton's responses.

1.1. Hungerford's objection

Hungerford's objection says that when faced with a range of explanations, we will be unable to agree on which is the loveliest, and inference will be impossible (or at least wildly variant). Lipton (2004: 143-144) responds by making two claims:

(i) inference is audience-relative, so for loveliness to show some subjectivity is, descriptively speaking, an advantage
(ii) loveliness isn't subjective in the way suggested by the objection.

Claim (i) points out that there are reasons, independent of Hungerford's objection, for thinking that inductive inference shows certain kinds of audience-relativity. So if we do find some subjectivity in loveliness, then that would in fact support IBE's claim to be a correct *description* of our inductive practices. Of course, even if they are accurately described as IBE, it remains a further question whether those practices are harmed by their subjectivity (audience-relativity). Thus Lipton makes claim (ii), aiming to show that this sort of subjectivity is not epistemically worrying.

Defending claim (i), Lipton argues that inference is audience-relative in three respects: (a) available evidence, (b) cognitive background, and (c) evaluation of evidence. His argument for (a) is simple: warrant depends on evidence. Different people have different evidence available to them when they infer, so we can expect their conclusions to differ accordingly. Considering (b), Lipton argues that since new inferential conclusions should be optimally consistent with previously established well-warranted beliefs, an audience's cognitive background constrains what they infer. Any variation in background may be reflected in different, but equally warranted, inferences. Such variation may reduce to (a), but need not. For example, background beliefs may be influenced by scientific training, intuition, pragmatic or aesthetic preference. None of these factors need vary in accordance with evidence, yet they hold some sway over inference, particularly when evaluating proposed conclusions. Arguing lastly for (c), Lipton simply notes that our judgements about evidential

support are fallible. Different people may perform different inferences because they award different weights to certain pieces of evidence.

Putting loveliness in the driving seat allows IBE to accommodate (a), (b) and (c), three natural features of inference. IBE would be descriptively inadequate if it tried to eliminate such subjectivity from inductive inference. Lipton concludes that to this extent, Hungerford's objection worries about nothing.

Lipton's defence of claim (ii) is weaker than his defence of claim (i), requiring as it does a pre-existing sympathy with his account of IBE. He begins with the argument, discussed in chapter 2, that at least some explanatory virtues are also inferential virtues, and that any account of those virtues in either guise will present them as equally audience-relative. On this basis, there is no reason to expect the variation in judgements of loveliness to exceed the variation in good inference. Lipton then uses his account of contrastive explanation in support of claim (ii). That account is subjective to the extent that it makes explanation a matter of contrasting the fact to be explained with some 'foil', a relevant alternative state of affairs that happened not to obtain. Differences in foil choice express differences in explanatory interests; our preferred explanations may cite different parts of the causal history of the fact, but as long as they're compatible (as they should be if they cite genuine causes), the subjectivity is not a problem. If we agree with Lipton that loveliness is partly a matter of explaining desired contrasts, then again we see that loveliness is not too subjective to be a guide to inference. Hungerford's objection tries to argue that it is; thus Lipton has defended claim (ii).

Lipton's response to Hungerford's objection consists in arguing for two descriptive claims, first that loveliness helps IBE accommodate the subjectivity of inductive inference, and second that since IBE is descriptively adequate (a claim he supports elsewhere), loveliness can't be a barrier to inference. These moves do much to diffuse the worry of Hungerford's objection. Lipton's guiding thought is this: as we practise it, inductive inference is reliable, so clearly it isn't harmed by any subjectivity inherent in our methods. If IBE offers an accurate account of those methods, it too is unharmed by that subjectivity. This is surely right, but Hungerford's objection retains some force, for two reasons. First, Lipton's response talks about audience-relativity, and audience-relativity does not exhaust subjectivity. Even with all audience-relative considerations made explicit, it may yet be a subjective matter which potential explanation is the loveliest. Secondly, accommodating audience-relativity is a purely

descriptive benefit. Maybe we do infer on grounds of loveliness, but Hungerford's objection raises the question: is loveliness a *good* criterion of inference? Perhaps its subjectivity is holding us back; *ought* we to infer on grounds of loveliness or would inferential agreement be likelier with different criteria? Lipton fails to answer this normative part of the objection.

1.2. Voltaire's objection

Lipton's response to Voltaire's objection consists in defending a further two claims (ibid.: 144-145; 151-163):

(iii) Voltaire's objection reduces to Hume's problem, which affects all accounts of inductive inference
(iv) if not an instance of Hume's problem, Voltaire's objection paraphrases van Fraassen's 'best of a bad lot' argument, against which IBE can be defended.

Claim (iii) is very important. Voltaire's objection asks the defender of IBE to justify the claim that inferring the loveliest explanation means inferring a true explanation. It asks for some independent reason to think that loveliness is correlated with truth. This, Lipton notes, is unreasonable. Since Hume, we have been aware of the impossibility of finding independent reasons to think that *any* inductive method will generate true conclusions (see chapter 1, section 4). Voltaire's objection merely highlights that IBE cannot be *shown* to be truth-conducive, which is just to say that IBE is not deductive. We cannot demonstrate a link between loveliness and truth a priori, since there is nothing in the description or application of IBE that guarantees it will preserve truth.[1] An a posteriori argument is out of the question, on pain of circularity. Thus IBE cannot be formally justified. But we knew this already; these points apply to any account of inductive inference. Lipton concludes that Voltaire's objection raises a serious problem for IBE – Hume's problem is a serious problem! – but an equivalent problem affects all of its competitors.[2]

Lipton makes two related points that put Voltaire's objection in perspective. First, he notes that, like any inductive method, IBE does not guarantee to preserve truth – we may infer falsely, from true premises – so a response to the objection need not establish a necessary connection between loveliness and truth. Second, he notes

that psychological research suggests that our inductive practices are, to some extent, systematically unreliable. Studies such as those by Nisbett and Ross and Kahneman *et al* (Lipton discusses these studies several times throughout his book, e.g. 128-132) are taken to demonstrate such things as our tendency to ignore base rate information when undertaking statistical reasoning, leading us to infer incorrectly.[3] Lipton reminds us that IBE, like any account of induction, should be able to accommodate such data; Voltaire's objection should not ask IBE to "make us out to be more reliable than we actually are" (ibid.: 145).[4]

Lipton is right, but right also to note that Voltaire's objection retains some force in spite of these points – the opponent of IBE (like the opponent of any inductive method) is owed an explanation of how inferential success is even possible on that account. But this is where my agreement with Lipton ends. His response to Voltaire's objection diverts attention from the real worry. Justifications of inductive methods may be circular and thus unpersuasive to sceptics, but they may explain and clarify those methods to those already disposed to use them. Those with inductive inclinations will see the merit in circular arguments for inductive principles (see chapter 5, section 2.1). Lipton shows that Voltaire's objection is wrong to ask us to justify IBE in a way Hume showed to be impossible, but this does not mean the defender of IBE is spared the burden of arguing for his view. Even a circular argument for a reliable link between loveliness and truth would do. Here, Lipton falls back on his descriptive approach: evidence suggests we do in fact reason according to IBE, and it serves us pretty well. His arguments are persuasive, but the thought behind Voltaire's objection lingers: what's so good about loveliness?

Looked at in another way, Voltaire's objection is motivated by the idea that IBE makes inductive success more miraculous than rival accounts. You don't need to be an inductive sceptic to think that the link between lovely explanations and true inferences needs clarification in a way that the link between, say, past and future instances does not. Hence Voltaire's objection: our aim in inference is truth, and in the absence of a reason to think we live in the loveliest of all possible worlds, or even just "the loveliest of those worlds where our observations hold" (ibid.: 145), we have no reason to infer on grounds of loveliness (rather than anything else). The objection asks for a reason to think it rational to use loveliness as a guide to inference, which Lipton does not provide. As with Hungerford's objection, he fails to rise to the normative challenge.

1.2.1. Lipton on van Fraassen's 'best of a bad lot' argument

So much for claim (iii), what about claim (iv), that Voltaire's objection reformulates van Fraassen's 'best of a bad lot' argument, against which IBE can be defended? This is the claim that, if not an expression of radical inductive scepticism, Voltaire's objection is an expression of more moderate inductive scepticism. Unlike Hume, van Fraassen grants that our inductive practices (as captured by IBE – the argument is directed specifically at IBE) are justified, but only up to a point.[5] His argument is this. According to IBE, inference depends on generating a pool of candidate hypotheses. For reasons already discussed (chapter 2, section 2), we never generate a full range of potential explanations, only a limited subset. In order rationally to infer one of them, we must believe that the truth is probably among them: "for me to take it that the best of set X will be more likely to be true than not, requires a prior belief that the truth is already more likely to be found in X than not" (van Fraassen 1989: 143). We have no reason to believe this, van Fraassen claims. Thus we can only ever be sure that we infer the best of the range we happen to have formulated, and "our selection may well be the best of a bad lot" (ibid.).

In short, the complaint is that explanatory quality cannot compel belief, since at any time all we know about our best explanation is that it's the best *available* explanation; this is the extent of our inductive abilities. Note that this is so even when we *should* believe our best available explanation because as a matter of fact it's true. Lipton considers this an instance of Voltaire's objection because it tries to drive a wedge between loveliness and truth; we can tell which available hypothesis is loveliest, but we're not justified in taking that to indicate its truth, for the truth might lie beyond – perhaps far beyond – the candidates we considered. A response to van Fraassen needs to show that we tend to generate potential explanations in the ballpark of truth, and thus that when we choose the loveliest, we thereby choose the (likely) truth. Given that both generation and selection stages of IBE are governed by loveliness, this means arguing that loveliness is a guide to likeliness, which is just what Voltaire's objection demands.[6]

Lipton claims van Fraassen's argument boils down to two premises. "The *ranking* premise states that the testing of theories yields only a comparative warrant... 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" (Lipton 2004: 152). The conclusion is that scientists can know which theory is the best in a range, indeed they can know that, of all those in the range, *that* theory is the likeliest to be true, but they cannot know how likely that theory is overall. Lipton's argument is then that "given an uncontroversial feature of the way scientists rank theories, the two premises of the argument... are incompatible" (ibid.: 157). The feature Lipton adduces concerns the role of background theories. We've already noted how background beliefs determine loveliness, but background theories help scientists in other ways. As Lipton notes, they also "influence the scientists' understanding of the instruments they use in their tests, the way the data themselves are to be characterized... and bearing of the data on the theory" (ibid.).

In order for scientists' comparative ranking to be reliable, as the ranking premise states, these background theories must be approximately true. Were they not, "they would skew the ranking... leading generally to true theories, when generated, being ranked below falsehoods" (ibid.). But (and here's Lipton's crucial observation) the background theories were once hypotheses, themselves the subject of comparative ranking; likewise, hypotheses ranked topmost in present comparisons will form part of the future background. This suggests ranking isn't comparative, but absolute. If scientists are capable of finding approximately true theories in order to form a background conducive to reliable ranking, then the hypotheses ranked highest in their pools are likely to be approximately true, not just likelier to be approximately true than their competitors.[7] Of course this means the approximate truth tends to be among the hypotheses scientists generate, and scientists may know this (scientists know they are reliable comparative rankers, thus they may know their background is approximately true, which means they may know that reliable comparative ranking is reliable absolute ranking). This contradicts the no-privilege premise. Lipton concludes that if the ranking premise is true, the no-privilege premise must be false, and the 'best of a bad lot' argument "self-destructs" (ibid.: 158).[8]

Lipton's general point is that "the level of reliability a background confers depends upon 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" (ibid.: 158-159). Thus even if the ranking premise conceded only a moderate reliability of scientists to rank (and this might be truer to life – the original ranking premise is

surprisingly generous) there would still be a connection between reliable comparative and reliable absolute ranking.[9] Any facility that scientists have for ranking must be due to an ability to locate the truth: "even moderately reliable ranking requires moderate privilege" (ibid.: 159). For van Fraassen to change the 'best of a bad lot' argument and claim that scientists' ranking is *completely* unreliable would turn it into an expression of Humean scepticism, and the interesting thing about the argument was that it was different from Hume's problem. Instead of arguing that our entire inductive practice is unjustified, it conceded that we are justified up to a point (reliable comparative ranking), and even *that*'s not enough to give us reason to think we have a nose for the truth. But as Lipton notes, such an intermediate scepticism turns out to be incoherent. The role of the background in theory evaluation means that we cannot have "inductive powers without inductive achievements" (ibid.).[10]

Let's evaluate Lipton's argument. Comparative ranking, as Lipton suggests, will be based partially on empirical testing. A hypothesis that does sufficiently badly in tests will be rejected altogether; of the remainder, the more empirically successful will tend to be ranked above the less. But this is not the whole story. Perhaps tomorrow's tests would elevate today's under-achiever above the present star pupil. Or perhaps they would merely tie. Perhaps more than one hypothesis is already top of the class even after today's tests. But testing cannot go on forever and we cannot wait until all possible results are in. At some point we have to take the plunge and *judge* which candidate is likeliest to be correct. This is all the more obvious when there are no available tests that will distinguish between two or more empirically equivalent competitors. So in cases with and without what Lipton calls 'inductive ties', we must make a judgement about which hypothesis to accept. According to IBE this judgement is made on grounds of explanatory loveliness. Thus although empirical testing is important, comparative ranking depends mainly on loveliness. The hypothesis ranked highest in any given range is the loveliest of that range. Given the link between comparative and absolute ranking, the loveliest available hypothesis is also likely to be true. Thus it may seem as if Lipton has established that loveliness is a guide to truth, and has responded to Voltaire's objection.

But matters aren't so straightforward. For one thing, Lipton's response to van Fraassen barely mentions loveliness. Empirical testing and likeliness are mentioned frequently, but (as with Lipton's account of the Semmelweis example – see chapter 2, section 2.1.2) loveliness falls out of the picture somewhat. The previous paragraph is

a mere sketch of how it may be reinstated, consistent with what Lipton says, but emphasising loveliness where he suppresses it. Perhaps as a consequence of that suppression, Lipton fails to give us an account of how loveliness, qua standard or set of criteria, favours approximately true hypotheses. He *has* given us reason to think that the loveliest available hypothesis is also likely to be true, which is no small achievement. But he has told us nothing about *how* loveliness affects hypothesis generation and selection such that we reliably generate approximate truths and then select them from among others that (presumably, given the influence of a truthlike background) are similarly close to the truth.

If fit with background were all there is to loveliness, Lipton's story would be satisfactory. We would generate hypotheses compatible with that background, and then select the one that (after testing, elimination of competitors and so forth) fitted most comfortably. It's obvious that the hypothesis that fits best with a truthlike background is likeliest to be true. But fit with background is just one criterion of loveliness; we saw that mechanism, unification and other (even less elaborated) factors are also influential. How do these other criteria affect generation and selection such that the loveliest hypothesis by their lights also comes out as likely true? I argued that the content and weighting of these criteria is determined by the background, but the background is not *identical* with the standard of loveliness. A truthlike background informs the standard of loveliness such that it becomes truth-tropic. A full response to Voltaire's objection must include an account of how this happens. The picture of feedback between truthlike background beliefs and loveliness is attractive, but at the moment we only know about one side of the partnership.

It's all the more important to develop such an account given that it's the increasing truthlikeness of background belief that makes loveliness a good guide to truth. The thought is plausible for everyday life: as we learn and experience more, we are able to explain better the world around us. But it's essential for science as construed by the defender of IBE. Science makes progress, evolving an increasingly truthlike theoretical account of the world. This is facilitated by an increasingly truth-tropic standard of loveliness. In order to account for this phenomenon (consistently, given what else we've said about loveliness) we must explain how loveliness develops in response to new theoretical discoveries. Again, although Lipton shows us *that* background beliefs must be approximately true in order for likeliness and loveliness to coincide in a given range (and thus that they don't just coincide in a given range),

what he doesn't show is *how* those beliefs fix a standard of loveliness such that it channels their (increasing) truthlikeness into the generation and selection stages of IBE. Showing that there is privilege is not enough; we must show that it's loveliness that brings it about. This would really get to the heart of the matter, in that it would acknowledge Voltaire's objection's normative force, which as we know, Lipton has neglected. Why *ought* we to use loveliness as our criterion of inference, rather than anything else? This is what Voltaire's objection really asks.

In the remainder of this chapter I show how an increasingly truthlike theoretical background furnishes scientists with an increasingly truth-tropic standard of loveliness, thus answering these questions. Since that background is shared, the variability of loveliness judgements is minimised; thus we also have a response to Hungerford's objection. The theoretical background in question is part of Thomas Kuhn's account of science; thus the next section will summarise the relevant features of his work.

2. Kuhn's account of science

Kuhn's landmark book *The Structure of Scientific Revolutions* (1962, 3rd edn. 1996) identified a pattern in the histories of mature sciences: extended periods of *normal science* punctuated by occasional *scientific revolutions*. Kuhn characterised normal science by the operation of a *paradigm* or tradition of scientific work, which defines the phenomena in which the science is interested, the methods and instruments it uses to investigate them, and the theory it accepts as background to its research. The business of normal science is *puzzle-solving*. Kuhn explains that paradigms not only define the puzzles that should occupy scientists but also provide the resources for their solution. He chooses the term 'puzzle' (rather than, say, 'problem') because although the solution of puzzles may be of little intrinsic interest, it is part of the fact that a puzzle is defined in the first place that there is an assured solution and there are certain "rules that limit both the nature of acceptable solutions and the steps by which they are to be obtained" (Kuhn 1996: 38) (though Kuhn sometimes substitutes 'problem' for 'puzzle'). Were we able to step outside the paradigm, puzzles would not make sense (though this doesn't make them trivial). The commitments that a paradigm imposes on scientists – "conceptual, theoretical, instrumental, and methodological" – are the source of these rules, and they enable scientists to "concentrate with assurance

upon the esoteric problems that these rules and existing knowledge define" (ibid.: 42). Kuhn admits that during normal science, some puzzles are dismissed as distractions, the business of other branches of science, or simply too time-consuming. A paradigm can even "insulate the community from those socially important problems that are not reducible to the puzzle form, because they cannot be stated in terms of the conceptual and instrumental tools the paradigm supplies" (ibid.: 37). But Kuhn thinks that, far from being a weakness of paradigms, this insulation is highly beneficial: "one of the reasons why normal science seems to progress so rapidly is that its practitioners concentrate on problems that only their own lack of ingenuity should keep them from solving" (ibid.).

Central to paradigms and puzzle-solving are what Kuhn calls *exemplars*, "the concrete problem-solutions that students encounter from the start of their scientific education [and] at least some of the technical problem-solutions found in the periodical literature that scientists encounter during their post-educational research careers and that also show them by example how their job is to be done" (ibid.: 187). They are called 'exemplars' because they provide exemplary puzzle-solutions: "scientists solve puzzles by modelling them on previous puzzle-solutions" (ibid.: 189). Thus exemplars are crucial to productive normal scientific work, but this is not their only function. Exemplars are at the root of paradigms' ability to define – both conceptually and in the sense of demarcation – the subject matter of a particular science during a period of normal science. Further, they tell the relevant scientific community how to investigate it. They do this by focussing attention on certain phenomena, certain kinds of observation, certain instruments set up in a certain way, and so on.

Exemplars constitute the theoretical background of normal science; scientists must be familiar with them in order to understand the field in which they work. Kuhn claims that mere statements of laws are not sufficient to furnish scientists with the requisite understanding. During their training, scientists are shown exemplars in action, solving puzzles. This exposure to exemplary puzzle-solving is a necessary accompaniment to simple verbal statements of relevant laws. Once familiarity is achieved, new scientists join the rest of their community in accepting exemplars unreflectively. On Kuhn's account, it is this widespread dogmatism that allows productive puzzle-solving to take place; normal scientists do not constantly question the basis of the science in which they work. In short, familiarity with relevant

exemplars is essential in order for scientists to (a) understand the normal science that has gone before and (b) be capable of contributing to normal science in the future.[11]

The above is a summary of the features of Kuhnian normal science relevant to the present project. Normal science, in Kuhn's own work and in the secondary literature, is much more complex. Scientific revolutions, the interruptions to normal science, are more complex still. Naturally, issues related to scientific revolutions appear in the course of the following discussion, but I introduce the relevant detail as we go along. I do not discuss revolutions here, partly for reasons of space, and partly because I want to keep attention focused on the way in which Kuhnian exemplars help respond to Hungerford's and Voltaire's objections. It is to the first of these responses that we now turn. (In what follows I call Hungerford's objection 'Hungerford's objection (I)'. This is because my response to it generates a new form of the objection, 'Hungerford's objection (II)', with which I deal in section 5.)

3. A Kuhnian response to Hungerford's objection (I)

Hungerford's objection (I) made the claim that "beauty is in the eye of the beholder": loveliness is too subjective to guide inference. A proper answer to the objection must show that judgements of loveliness are not highly contingent and that scientists will converge on a single explanation when selecting the loveliest from a range of competitors. One way of doing this is to show how a shared cognitive background can minimise the variability of loveliness judgements. Kuhn's account of science offers structure to this approach.

The first thing to notice is that the exemplars of Kuhnian normal science have an especially relevant feature with respect to Hungerford's objection (I): within the community to which they apply, they are the source of all scientific understanding. Further, any puzzle-solving success achieved under exemplars' guidance (and this should be considerable, given that paradigms are set up to be effective puzzle-solving mechanisms) may be indirectly attributed to them. Thus exemplars hold an elevated status within the relevant community. As the name suggests, they are seen as the very definition of a good theory. Now recall that according to IBE the loveliest explanation is that which offers the most understanding. My claim is that exemplars function by providing scientists with a standard of loveliness against which to assess potential explanations. We've seen that normal-scientific inference is a matter of

evaluating candidate puzzle-solutions by reference to exemplars, so if puzzle-solving means explanation-giving, as it often does, then that inference is governed by *exemplars of loveliness.*[12] The potential explanation/puzzle-solution that best conforms to the standards set by the relevant exemplar is the one that provides the loveliest explanation of the evidence, and should thus be inferred.[13]

Scientists may be unaware that their loveliness judgements are guided by exemplars, and may be unaware of the value to science of so judging. Similarly, the concept of loveliness that scientists form through familiarisation with exemplars may never be fully articulated. Nevertheless, they will have such a concept – it's simply a matter of resemblance. It's well-motivated too: exemplars are the source of all theoretical and practical understanding in normal science and the key to puzzle-solving success. Thus they are held in high esteem. A concept of loveliness is an entirely natural, perhaps inevitable, consequence of training and working within a normal science tradition.[14] This gives us our response to Hungerford's objection (I): loveliness not relative to individuals or groups, but shared among all members of a scientific community. Consequently, using loveliness as the desideratum of good inference does not mean disagreement about what to infer. Beauty is not in the eye of the beholder; loveliness is not too subjective to guide inference.

(A possible example of an exemplar determining loveliness is given by Psillos (1996: 38-39). It involves Fresnel and Arago's work on the phenomena of the polarisation of light. Two explanatory hypotheses were considered: one stated that light-waves are uniquely transversal, another that light-waves are transversal and longitudinal (their experimental results entailed that at least some light-waves are transversal). The former was chosen as the better explanation because "it explained the phenomenon of polarization more simply, more completely and without needing any *ad hoc* mano euvre" (Psillos 1996: 39). The explanation then became part of background theory, constraining subsequent attempts to explain light phenomena. Psillos notes that Fresnel and Arago's initial background contained the wave theory of light, which had superseded the emission theory. Thus Huygens' wave theory is a possible exemplar of loveliness. It was clearly operative at the generation stage – both hypotheses assume it. But it's not so clear that the wave theory determines Fresnel's criteria of simplicity, completeness and non-adhocness at the selection stage. The example serves Psillos' twin purposes of showing how background knowledge narrows down the space of potential explanations, and how explanatory virtues are

used to distinguish equally successful competing hypotheses. Thus it is also supportive of the view of IBE I want to emphasise. But more work needs to be done to show that the example is fully supportive of the Kuhnian account developed here.)

Before considering some obvious objections to my response, let me stress something important, and quickly note an advantage of my view. What I want to stress is that it is a concept of *loveliness* that scientists form in response to exemplars. Where puzzle-solving means explanation-giving, exemplars govern that process by prompting in scientists a clear idea of the features a puzzle-solution should have in order to create understanding of a phenomenon. Exemplars do not inculcate a concept of likeliness, though they may influence scientists' ideas about which features a puzzle-solution should have in order to achieve a high degree of likeliness, and rightly so. One would expect the empirical success that exemplars enjoy to be related in some way to their structural, loveliness-making, features. But it is *these* features scientists seek to reproduce, not the exemplars' likeliness directly; likeliness cannot be faithfully reproduced in this way. At bottom, likeliness is a matter of probability, or warrant given the evidence. Exemplars may instantiate such relations in a certain way, but the extent to which resulting puzzle-solutions also do so is a separate, empirical/epistemic (though not purely logical), matter. Exemplars do not decide the likeliness of the puzzle-solutions they inspire (though the relationship just noted, between exemplars' structural features and likeliness, suggests the kind of link between likeliness and loveliness developed below, in my response to Voltaire's objection). They are exemplars of loveliness, and scientists form a corresponding concept of loveliness.

The advantage of my view that I wish to note here is that it need not commit itself to a particular account of loveliness. For my purposes loveliness can be defined simply as 'those loveliness-making properties exhibited by the relevant exemplar'. The notoriously vexed issue of which properties make for lovelier explanations can thus be left to one side. My account is even silent about whether or not any properties contribute continuously to loveliness across all exemplars. Maybe all lovely explanations instantiate some particular loveliness-making property (in some way), maybe they don't. Maybe there's some loveliness-making property that, when instantiated (in some way), always makes an explanation lovely, maybe there isn't. These issues won't be investigated here; I make no attempt to ascertain whether any property is necessary or sufficient for lovely explanation. Judgements about the

plausibility of my account are not dependent on loveliness being analysed in a particular way; nor even are they dependent on the possibility of loveliness being analysable in any illuminating way. None of which is to say that I don't think such issues are interesting and worth investigating. On the contrary, as I note in section 6, I think my account opens up potentially fruitful avenues for further work on the nature of loveliness itself.

Now for the obvious criticisms. Firstly, on my view loveliness is no longer subjective, but it is relative to scientific communities during periods of normal science. Thus Hungerford strikes again: beauty is still in the eye of the beholder, but the beholder is a normal-scientific community rather than an individual scientist. Where there's relativity, there's the possibility of disagreement, and this counts against the idea of explanatory loveliness as an inferential guide. Secondly, exemplars may generate some consensus about loveliness, but it's another matter whether that consensus is truth-tropic. Indeed, it might look unlikely given the anti-realist associations of paradigm-based science. This second charge is basically Voltaire's objection, which I shortly argue can be undermined by adopting a reliabilist reading of Kuhnian science. I then tackle the first charge, Hungerford's objection (II); my response to that is informed by my comments on Voltaire. For now I restrict myself to the following remarks.

The best way to make the above criticisms vivid is to think about them in terms of understanding. According to IBE, understanding is the goal of inference, and we endorsed Lipton's definition of loveliness as that which would generate understanding. My opponent is objecting that on my account, loveliness cannot fulfil its function. To increase understanding of the world, standards of loveliness must be applied more or less consistently across all inferences, not relativised to individual sciences at individual times, the y claim. Further, from the point of view of the psychology of the individual scientist, loveliness must remain constant, or at the very least develop continuously, within a discipline. The argument here is that my account of loveliness is psychologically implausible. It leaves room for quite different standards of loveliness to succeed one another in a single science; if such a possibility were realised, loveliness would bring not understanding but confusion for the scientists that must make the transition. My opponent's last claim is that the possibility of wildly variant standards of loveliness also tells against any attempt to show that loveliness is truth-tropic.

The first thing to say in response is that there's nothing obviously objectionable about loveliness being localised to certain communities. The fact that many domains of research have understanding as their goal does not entail that they should all seek it in the same way. Indeed, the various branches of science display exactly the kind of localised standards of loveliness we would expect from disciplines investigating quite different subject matter. The search for empirical success encourages explanations in the sciences to take various forms, but as long as that success is achieved, there is no tension between localised loveliness and understanding. This suggests my second point: loveliness need not develop continuously in order to satisfy its inferential purpose. Indeed, if at any time the search for empirical success suggests a certain science must adopt a new exemplar, then the standards of loveliness for that science *must* be updated accordingly, even if this means a radical overhaul of scientists' shared concept. It seems that if we accept that a science uses IBE, i.e. treats loveliness as indicative of truth, we must be prepared to accept that that science may alter its concept of loveliness whenever it finds new ways of achieving empirical success. If we don't accept this, we may as well give up on understanding as a goal of science altogether. The important point here is that a new exemplar *need not* disrupt the smooth development of a certain concept of loveliness, but if it does, this does not mean that loveliness has become less *refined*, where 'refined' means 'able to track the truth'. On the contrary, on the above reading, scientists' concept of loveliness becomes increasingly refined even if the contents of the concept undergo considerable change.

These remarks have already introduced several issues related to Voltaire's objection, that we've no reason to think we live in the loveliest of all possible worlds. I now respond to that objection.

4. A Kuhnian response to Voltaire's objection

Kuhn associated the structural features he discerned in the histories of mature sciences with the philosophical position of relativism. His relativism is milder in form than many people assume (see e.g. Bird 2000: 271), but the balance of his work is clearly motivated by a rejection of the idea that successive paradigms approach the truth. For Kuhn, no single paradigm affords a more truthlike perspective on the world than any other. I reject Kuhn's relativism. This is legitimate: Kuhn's account of

the structure of science does not entail relativism. Although I make use of Kuhn's descriptive-historical account of how science works, this can be divorced from the normative-philosophical relativism with which he interprets it.

Despite being a relativist, Kuhn made room in his account of science for an idea of progress. He claimed that science makes puzzle-solving progress; successive paradigms display increased puzzle-solving ability. Puzzle-solving progress is possible thanks to the organisation of science into communities specialising in areas of specific interest, governed by their prevailing paradigm: "the nature of [scientific] communities provides a virtual guarantee that both the list of problems solved by science and the precision of individual problem-solutions will grow and grow" (Kuhn 1996: 170). Such admissions did not dent Kuhn's relativism. He was clear that although "later scientific theories are better than earlier ones for solving puzzles in the often quite different environments to which they are applied" (ibid.: 206), this does not justify the belief that science approaches truth. This stance is in tension with certain other claims he makes, such as that "normal science… is a highly cumulative enterprise, eminently successful in its aim, the steady extension of the scope and precision of scientific knowledge" (ibid.: 52). But consistency issues to one side, we must ask: what sense can be made of puzzle-solving progress in the absence of relativism?

Since "a scientific community is an immensely efficient instrument for solving… puzzles" (ibid.: 166), and since the resources needed for puzzle-solving are provided by exemplars, exemplars are part of an effective puzzle-solving method. Given puzzle-solving progress, we know that method gets more effective with the passing of time. Why does the method of Kuhnian scientists improve in this way? The obvious answer is the reliabilist answer: it improves because it gets better at tracking the truth.

Let's look more closely at this idea. Kuhn thought that the rejection of truth as the aim of science followed from the view of science as paradigm-governed. He is right if we need independent reasons for such a belief, i.e. independent reasons for believing that each paradigm-shift cements the relationship between theory and truth. However, to the epistemological externalist, such independent reasons are not necessary. Externalism about justification is the view that "some of the facts that make a true belief into knowledge may be unknown – indeed, unknowable – to the knower" (Bernecker and Dretske 2000: 65). This contrasts with internalism, according to which all factors relevant to justification must be internally available to the knower,

that is, they must be "cognitively accessible to – already known or experienced by – the subject" (ibid.). Reliabilism is the most popular form of externalism, and holds that "what qualifies a belief as knowledge or as justified is its reliable linkage to the facts that make the belief true. What makes this view externalist is the absence of any requirement that the knower have any sort of cognitive access to, any appreciation of, the relation of reliability that makes her true belief knowledge" (ibid.). Accordingly, reliabilism focuses on belief-forming methods, as these are the most obvious candidates for the job of linking beliefs with facts. According to reliabilism, justification and knowledge are a matter of being formed by an appropriate method; "so long as one's belief is brought about by a method which as a matter of fact is reliable, then one's belief is justified" (Bird 1998: 232). This is so no matter whether or not one has any beliefs about the method in question or its reliability.

Reliabilism is standardly illustrated with a perceptual example. I believe that there is a glass of orange juice on my desk in front of me. Whether or not my belief is justified is not a question of having made an appropriate inference from perceptual experience to the presence of a glass of orange juice, for example one invoking a belief about the general reliability of my eyesight. My justification depends rather on features of my visual system. If I am justified it is because the conditions are such that my eyes are at the centre of a mechanism that enabled me to form a true belief. What matters are facts about the reliability of my eyesight; justification does not depend on my having beliefs about my eyesight, in particular the belief that it is reliable.

The inferential method of science we're considering is that of assessing potential explanations against exemplars of loveliness. According to reliabilism, this method is reliable. If it weren't reliably connecting beliefs with facts, it wouldn't be a successful puzzle-solving tool. And if it weren't growing *more* reliable, successive paradigms wouldn't *increase* in puzzle-solving ability. The facts of puzzle-solving progress mean that scientists work with an inferential method that reliably delivers true beliefs and gets better at doing so. Thus to the reliabilist, puzzle-solving progress is evidence that successive exemplars approach the truth. If exemplars weren't approaching the truth, the standards of loveliness they provide wouldn't be part of an increasingly reliable inferential method. To put it differently, the increased puzzle-solving ability of successive paradigms means that if any paradigm has some non-trivial truth content then its successors are guaranteed to build on it. There is evidence that paradigms have such content: the method they prescribe leads to reliable beliefs. Thus we have

reason to believe that successive exemplars approach the truth; truth begets truth. Retention of puzzle-solving ability means retention of truth content, and increases in puzzle-solving ability mean increases in truth content.[15]

Note that on this account, normal-scientific IBE is not *guaranteed* to produce true beliefs, but to demand such a guarantee is unreasonable. Firstly, such a guarantee would be tantamount to having internal justification for puzzle-solving inference; such justification is unnecessary on the reliabilist view. Secondly, as Lipton reminds us, guarantees of truth-preservation are unavailable to all inductive methods. What is available is evidence that distinguishes between good and bad inductive methods, and the puzzle-solving evidence suggests that normal-scientific IBE is a good inductive method. We have *good reason* to think that it will generate true beliefs. Note though that there's no need for scientists themselves to have independent reasons to think their paradigm truthlike (or more truthlike than its predecessor). Indeed, there's no need for them to reflect on their method in any way. *They* don't have to have evidence for the reliability of IBE in order to use it justifiably. Reliabilism doesn't prohibit scientists from gathering such evidence much as we have done; rather, it maintains that such evidence will not make a difference to the justification of their inferences. As long as IBE continues to solve problems, scientists need not have access to any of the above information. This is an advantage because it insures against scientists being distracted by considerations of likeliness when inferring. Normal-scientific IBE is wholly governed by loveliness and hence avoids the trivialisation of IBE that Lipton warns us about.

I briefly considered above the following objections: (a) loveliness cannot be conducive to understanding if standards are local and diverse; (b) my account of loveliness is not psychologically plausible; (c) loveliness must develop continuously in order to stand a chance of being truth-tropic. I say more about all three during my response to Hungerford's objection (II), but further consideration of (a) I leave wholly until then. As for (c), the foregoing discussion of puzzle-solving progress shows it to be misconceived. Loveliness is just those properties exhibited by the latest and most successful exemplar. That exemplar may share numerous loveliness-making properties with its predecessor, only some, or none at all; it doesn't matter. What matters is that it's better at solving scientific problems (the reliabilist account of Kuhnian science doesn't take the puzzle/problem distinction seriously). It's precisely the fact that loveliness *need not* develop continuously that allows it to track the truth.

Scientists' concept of loveliness is refined through successive paradigms; it gets better at tracking the truth. This refinement may or may not coincide with the linear development of that concept.

This makes (b) all the more pressing: are scientists so adaptable that they can allow their concept of loveliness to change, perhaps radically, if a new exemplar dictates? Given that any such change represents a problem-solving advance – a step towards truth – then yes. This doesn't mean scientists treat their acquired concept of loveliness as purely instrumental, an easily replaceable means to the end of true inferences. For sure, scientists develop a certain attachment to (a broadly aesthetic preference for) the accepted notion of loveliness that's tough to relinquish. But this reluctance to surrender the traditions that form the background to scientific work is nothing new. It is merely a restatement of the familiar Kuhnian point that the ease with which scientists adopt a new paradigm is inversely proportional to their attachment to its predecessor. Kuhn has much to say about the sociology and psychology of paradigm-shifts; his remarks about how progressive factions, open-minded scientists and the recruitment of young trainees secure commitment to a new paradigm will be familiar. The doubts expressed in (b) are simply doubts about the plausibility of Kuhn's story. There's no doubt that, *if* a new exemplar demands that standards of loveliness be drastically overhauled, then scientists will find it difficult to make the transition. But they will be encouraged by the promise of puzzle-solving progress, and importantly, it will be no more difficult than the equivalent transition in Kuhn's original account.

Initially then, new exemplars may be thought unlovely by most of the scientific community. They achieve their status by solving the problems with which science is most concerned; this empirical success is enough to outweigh scientists' dislike of their loveliness-making properties. Gradually, as the new exemplar displays more general puzzle-solving prowess, the community develops a new concept of loveliness. Consider briefly an example: the replacement of Aristotelian with Newtonian mechanics. Newton's theory relied on the possibility of action-at-a-distance, a deeply unlovely concept to anyone wedded to the Aristotelian tradition. But as the Newtonian approach solved more problems, theories that assumed action-at-a-distance became highly desirable. It was taken as given that they would be successful; scientists modelled their theories according to the concept of loveliness that the Newtonian system supplied. Their inferences were truth-tropic because they modelled

new explanations on Newton's, which had been shown to have high truth content. This is a case of empirical success promoting the acquisition of a new concept of loveliness.

Voltaire's objection is answered: I have justified IBE's claim that loveliness is a guide to truth. Scientists assess potential explanations against the standard of loveliness supplied by their exemplars. Those that most closely meet that standard are likeliest to be true. The loveliest explanation is not guaranteed to be true – nor should it be (recall Lipton's point about Hume's problem) – but there is good reason to think it will be approximately true. Further, inferences under later paradigms are more truth-tropic than earlier ones. But in order to achieve all this, we have had to make loveliness relative to periods of normal science within individual sciences. Hungerford's objection (II) claims that this kind of relativity means that loveliness pulls inference in all directions, most likely away from understanding and truth. Thus loveliness is a poor guide to inference. This is the objection I turn to now.

5. A Kuhnian response to Hungerford's objection (II)

The relativity of loveliness is uncontroversial in relation to distinct sciences. Geology, for example, will have a different standard from particle physics, or evolutionary biology. There seems no reason to expect that disciplines dealing with such diverse subject matter should cleave to a single standard of loveliness derived from their various exemplars. What this tells us is that there are different standards of loveliness appropriate to different *puzzle-solving contexts.* Sciences are differentiated according to such contexts, and where a single science fosters several such contexts, there are appropriate exemplars for each one (there is nothing in Kuhn, or in my interpretation of Kuhn, that states that any science is governed by only one exemplar). I claim that loveliness isn't relative to individual sciences during periods of normal science; rather, loveliness is relative to puzzle-solving context. Within a single science, if a new exemplar demands a radical overhaul of loveliness, it's because it introduces a new puzzle-solving context into that science (the science in question may later separate into distinct sciences along puzzle-solving lines). Thus there is another reason why we should not expect the standard of loveliness within a single science to develop continuously in order to be truth-tropic. Loveliness changes as the various sciences begin to deal with new kinds of problem.

Before we go any further, let's look at an example. Consider the transition from Newtonian into Einsteinian physics. It is well-known that even after the revolution the Newtonian system continues to provide a simpler and more accurate way to account for the physical behaviour of medium-sized objects. Newton's theory was designed to account for the physics of objects such as billiard balls and planets; for these objects it continues to provide superior exemplars. Whenever scientists seek to explain phenomena of this familiar kind they look to Newtonian physics for guidance. To the extent that the Einsteinian paradigm applies itself to problems of this old sort, the standard of loveliness appropriate to that context – the Newtonian standard – remains, and remains truth-tropic.

In the new puzzle-solving context brought in by relativity theory, dealing with massive objects, large distances, high speeds and so on, new exemplars introduced a new standard of loveliness appropriate to those kinds of problem. Because this sort of loveliness has not evolved through paradigms, it may not be as truth-tropic as its more refined counterpart. But as long as the relevant exemplars continue to solve problems, there are grounds for thinking that they have some truth content, which will lead, over time, to a more truth-tropic concept of loveliness. This very same puzzle-solving ability will cause scientists to become as attached to the new kind of loveliness as they are to any other (I say more about this attachment shortly).

Note that relativity to puzzle-solving context is not another layer of relativism. Rather, it is simply the proper way to analyse the relativism that Hungerford's objection (II) worries about. The reference to puzzle-solving context does introduce a modest *pluralism* about loveliness, but the moral of the above example is that this is entirely palatable. In fact, it is to be expected, given the way the sciences discover, frame, and try to solve the problems to which they are devoted. Different kinds of problem demand different kinds of solution, and any one science will generate exemplars of loveliness appropriate to each. As noted above, sometimes this may lead to the science splitting into two or more separate sciences, each dedicated to certain kinds of problem, each with corresponding exemplars of loveliness. Alternatively, once a new puzzle-solving context is introduced into a science it may take over, becoming the sole focus of attention, the old context and exemplars being found to be redundant or fruitless. In such cases, individual sciences foster a single standard of loveliness in the same uncontroversial way remarked on above. But note that the physics case falls somewhere between the two: the new Einsteinian problem-solving

context dominates research in the field, but there is no significant splitting, and old Newtonian exemplars are appealed to where necessary. It is an example of a single science tolerating different kinds of loveliness simultaneously. True, sociologically or psychologically speaking, Newtonian loveliness is relegated to a subordinate, functional and relatively mundane role, but it is still fruitfully applied in the appropriate puzzle-solving context. That is, it is still scientifically successful. When physics encounters problems that can be solved in the Newtonian way, they may not be particularly exciting, but there is an effective standard of loveliness available to which physicists refer when judging which of their proposed explanations is likeliest to be true.

The Einsteinian concept of loveliness, meanwhile, governs most inference in physics, since the associated problem-context is the dominant source of interest. Newer standards of loveliness may be relatively unrefined, but as long as they continue to solve problems, there is reason to believe they are truth-tropic and will get more so over time. Science takes a risk: when it chooses to focus on a new puzzle-solving context, it chooses to make a new concept of loveliness the basis of inference. Scientists don't know how likely it is that their inferences will go wrong; they are governed by a concept of loveliness with uncertain refinement. But this is old news. Such risks are part and parcel of the scientific endeavour; in the attempt to push back the boundaries of ignorance scientists must concern themselves with new problems, and the new ways of working that they promote. As long as inference allows them to solve problems, scientists have no reason to think the move a misguided one. Further, the risk can be seen as merely yet another manifestation of Hume's problem: we can never be sure that our inductive inferences won't go wrong, so the best we can do is use methods that reliably produce true beliefs. As long as new concepts of loveliness help scientists solve problems in the appropriate contexts, there's no reason to mistrust them. One final point on this score: scientists' absorption in new kinds of problem is a risk, but also a benefit. The sooner and more widely they apply a new concept of loveliness, the greater the opportunity for it to grow more refined.

We now have something new to say about the psychological plausibility of this account of loveliness. The worry was that loveliness has been gerrymandered; what I'm calling loveliness could never count as such because scientists could never form the requisite attachment to it, such that they both trusted it as an inferential guide and found it in some broad sense aesthetically pleasing. Were a new exemplar to demand

radical changes to a concept, the worry continues, scientists would either fail to make the changes, or if they did, the lack of attachment thus expressed would show that whatever it was that changed isn't loveliness. This worry expresses the worst-case scenario for scientists strongly attached to certain exemplars of loveliness: a new puzzle-solving context *takes over* a science, replacing the old context rather than complementing it or forcing the science to split. The new standard of loveliness may indeed be hard to adopt. Chances are it is markedly different from the standard it replaces; scientists must abandon their tried-and-tested concept of loveliness and grow to love something at best alien, at worst downright unlovely (by old lights). Either that or be forced out of the scientific community altogether.

To answer this concern, two points may be reiterated. First, scientists are encouraged by the knowledge that they're now dealing with a new kind of problem which *merits* the new kind of loveliness. Sometimes, science encounters new kinds of puzzle; in order to continue to achieve puzzle-solving success it must adapt its inferential methods accordingly. It may take some effort to alter a concept of loveliness, but scientists have every motivation to do so. Relativity to puzzle-solving context rather than periods of normal science makes the adoption of new standards of loveliness by the collective scientific mind entirely plausible. The second point to reiterate is this: there *is* a psychological difficulty involved in altering one's concept of loveliness in this way, but it is just the familiar Kuhnian difficulty of growing accustomed to a new post-revolutionary puzzle-solving context. The difficulty of internalising a new concept of loveliness, and the dangers associated with not doing so, are mere restatements of familiar features of Kuhnian science. Kuhn argued that scientists strongly attached to old ways of doing science may never be able to work on new kinds of problem. Those who do make the transition may be persuaded by puzzle-solving promise, but they may equally be persuaded by factors such as the fear of exclusion from the scientific community, the need to win research grants, or the desire to follow a charismatic leader. Critics of my view would be right to point out that to some scientists, new and radically different standards of loveliness may seem unappealing, counterintuitive and even counterproductive. But if Kuhnian science is plausible, such features are not especially problematic.

I have argued that loveliness is not relative to individual sciences during periods of normal science, but instead relative to puzzle-solving context. Sciences may foster one or more puzzle-solving contexts. Within each context exemplars provide an

appropriate standard of loveliness, effective at solving its puzzles. Loveliness develops continuously, refining its connection with truth. If a new exemplar dictates this development must halt, it's because it introduces a new puzzle-solving context into the science. Scientists may not have to adapt, but if they do they will struggle to internalise the new standard. However, once they've got used to the new puzzle-solving context and realised that the new standard effects puzzle-solving success, their concept will begin to form accordingly. In time, they will trust it as a guide to inference and regard it as lovely in the requisite 'aesthetic' sense, just as they did the concept appropriate to the old context. This coincides with the increasing refinement of the new standard. I conclude that Hungerford's objection (II) is answered. Neither uniqueness of standard nor uniformity of development is necessary for loveliness to provide understanding or be truth-tropic. In fact, allowing loveliness to change according to puzzle-solving context enables scientific inference better to track the truth.

6. Kuhnian science and IBE

Hopefully the marriage of IBE and Kuhnian science presented above seems natural enough. However, this section presents further reasons why Kuhnian science makes sense with my account of IBE as its inferential engine.

The first reason to endorse Kuhnian IBE is that my account of loveliness fits with the sociological thread in Kuhn's work. Sociologists sometimes observe that widespread subjectivity in matters of taste, the kind Hungerford's objection (I) worries about, is often ironed out by cultural pressures, and those who fail to conform to the relevant norms are considered deviant. Likewise, on Kuhn's view, any subjectivity about how to practice science is removed during training; those who fail to adopt the standards supplied by exemplars are outcasts from the scientific community. Scientists are only inducted into the scientific community when they are familiar enough with exemplars to understand the goals of normal science and how to achieve them. On my view, this is explained in terms of loveliness: failure to form the right concept of loveliness – that provided by exemplars – means inability to work within a science. Those with the wrong 'taste' in loveliness are rejected by the prevailing culture.

The second reason to support Kuhnian IBE is that my account makes sense of progress in Kuhnian science. On Kuhn's account, the kind of progress he claims science makes – puzzle-solving progress – is something of an anomaly. Given his relativism, it's not clear why we should expect successive paradigms to get better at solving scientific problems. However, my account explains why paradigms increase in puzzle-solving ability: because the puzzle-solving/belief-forming method of normal science becomes increasingly good at linking up with the facts. Exemplars are the key to that method. Exemplars facilitate puzzle-solving progress – they instruct scientists how to solve puzzles, as Kuhn describes. Given that they perform this function increasingly well, on a reliabilist understanding this shows that exemplars approach the truth. On my view, scientific inference is IBE and exemplars instruct scientists in puzzle-solving by providing a standard of loveliness against which to assess potential explanations. With loveliness defined in this way, scientific inference is truth-tropic. Naturally, this results in puzzle-solving progress.

The third way in which my account fits with Kuhnian science involves the nature of loveliness itself. Kuhn's work suggests at least one way in which the project begun here could be developed. It involves his five values of theory choice – accuracy, consistency, scope, simplicity, and fruitfulness – that apply across all paradigms.[16] Kuhn (1977) explains that independently, all five are imprecise and may conflict with one another, but within a paradigm shared exemplars serve to fix their interpretation and weighting.[17] This is just the conception of exemplar-driven inference expressed in my account of IBE; but not only this, the five values are arguably constitutive of loveliness. Let's look at them more closely.

Accuracy, Kuhn claims, is a matter of a theory's deducible consequences being in quantitative and qualitative agreement with relevant observations, the results of previous experiments, and so on. Consistency is a basic requirement: theories should be internally consistent and consistent with other accepted theories. Theories' scope should be broad, says Kuhn, meaning they should have consequences beyond what they were designed to explain. They should also be simple, which Kuhn defines as bringing order to otherwise disparate and confusing phenomena. Fruitfulness is a matter of disclosing new phenomena or new relationships among those already known (Kuhn claims this is the most overlooked theoretical virtue).

All five values can be seen as explanatory virtues and thus as part of IBE (cf. Bird 1998: 282); certainly they promote understanding and puzzle-solving ability, and

so count as criteria of loveliness on my account. Maybe it is not too optimistic to suggest that evidence for their consistent, paradigm-neutral usage could be translated into the continuous, linear development of loveliness within a puzzle-solving context. Likewise, maybe the evidence supporting Kuhn's view – that the usage of the five values changes between paradigms – could be seen to support my claim that new standards of loveliness are introduced along with new puzzle-solving contexts. I claim that this aspect of Kuhn's work shows that the prospects for a Kuhnian account of loveliness are promising.

The final reason to endorse Kuhnian IBE is provided by Bird (1999). He claims that IBE *explains* the pattern of Kuhnian science; we should expect science to move in cycles (normal science, anomaly, crisis, revolution, new paradigm, normal science, and so on) if IBE is its core inferential method. Bird argues that normal science uses IBE alongside a principle of Least Disruption, which preserves existing theory by preventing each application of IBE from taking into account all possible alternative explanations and trying to explain all relevant evidence (doing so would inhibit progress). This emulates full IBE by preserving the explanatory virtue of the overall system. (We may understand Least Disruption as part and parcel of loveliness at the generation stage of IBE.) Sometimes though, Bird claims, it becomes apparent that this strategy has not been working: "that science has been led astray by Least Disruption would be suggested by a decreasing success in solving problems, by mounting anomalies and increasing need for unsupported ad hoc hypotheses. All these suggest that current theory, as revised, is no longer well-supported by total evidence as its predecessors were" (Bird 1999: 38).

In this case, Bird argues it is rational for scientists to abandon Least Disruption and consider an IBE that accounts for the total evidence. This may result in the significant changes to theory that characterise a scientific revolution. Once scientists have settled on the best explanation of current evidence, they will reinstate Least Disruption, and the Kuhnian cycle begins again. Bird claims that "on this picture, which describes theory change in terms of (an explicable alternation of) rational principles, we would expect periods of conservative change to be followed, in the face of accumulating anomalies, by radical change, which is itself followed by another period of conservative change" (ibid.: 42). We may conclude that we should see IBE as the Kuhnian method because IBE generates the Kuhnian pattern (Bird himself concludes that IBE shows that Kuhnian science is rational).

Thus Kuhnian IBE is a plausible and sensible idea. But notice something about Bird's argument. His claim was that Least Disruption, as an independent principle, allows scientists to approximate IBE. Usually, IBE would recommend the explanation that preserves the loveliness of existing theory; when it becomes clear that it wouldn't, Least Disruption may be dropped and we can see what IBE really would recommend. This implies that on his view there is a standard to which IBE can appeal even when what's at stake is the theoretical background of science. On our view, Least Disruption is an integral part of loveliness; the picture of loveliness as defined by background theory (specifically exemplars) already contains the kind of conservative instruction Least Disruption gives. There is no IBE without such instruction because there is no loveliness independent of exemplars. What we gain in being able to say that IBE (not just an approximation) is the inferential engine of normal science, we lose by being unable to say the same thing about revolutionary science. In order to replace exemplars, scientists must assess potential alternatives for loveliness, but the only things determining loveliness are the very exemplars they seek to replace. This is the problem of revolutionary inference.

To bring this problem out, return to what Kuhn says about the five values. Independent of a paradigm, the values are vaguely defined and may recommend different theories. Kuhn introduces them as a basis for the rational comparison of theories from different paradigms but maintains that they cannot determine theory choice outside a paradigm. Specifically, the five values cannot determine a choice *between* paradigms; the values mean that proponents of different paradigms may understand their counterpart's preference, but they do not mean that they will regard it as, from their perspective, fully justified. For Kuhn, paradigms have no common measure that dictates a choice between them; paradigms are incommensurable. By noting how the five values correspond to features of loveliness and how exemplars fix their interpretation, haven't we just agreed with Kuhn that there's no standard of theory choice for scientific revolutions?

On the contrary, my account of IBE gives reasons to think that paradigms are not incommensurable and that loveliness may determine theory choice even in revolutions (this is mainly because Hungerford's objection, in both forms, and Voltaire's objection expressed the concerns that motivate incommensurability).[18] Standards of loveliness are determined by exemplars, but the standards are not identical with those exemplars. The standards may remain even when the exemplars

that created them are being replaced. They may legitimately inform the inference of new exemplars because they are truth-tropic. Standards of loveliness may not remain fully intact in the new paradigm, since the process of ensuring that new exemplars are empirically successful may force them to alter. I claim that the only situation in which loveliness changes radically is when a new puzzle-solving context fully takes over a science. In this case, either the new standard grows gradually out of the old one following a slow succession of alterations, or it enters the science from outside – it's no coincidence that new exemplars of this kind are usually generated by scientists of genius whose affiliation to the scientific community does not involve them being fettered by existing standards of loveliness. In all other cases, there's no reason to expect that the standards of one paradigm can't govern the institution of another, nor that they shouldn't.

This picture, in which radical changes to a concept of loveliness are rare, is supported by Bird's account of Kuhnian science. Bird (2000: chapter 2) stresses that the simple two-stage model of normal science/revolutionary science does justice neither to the historical facts nor Kuhn's careful investigation of them. Not all revolutions are prompted by a serious anomaly (or an accumulation thereof) and not all result in radical revisions to a paradigm. Anomalies themselves may be minor or serious, and do not always precipitate crisis. Assuming we retain the word 'revolution' to describe any alteration to a paradigm, revolutions may be largely conservative, requiring only partial revision to paradigm beliefs. And revolutions do not exhaust the possible responses to anomaly. Even serious anomalies may be solved without revision to a paradigm; otherwise, non-paradigm beliefs may be changed instead, or the anomaly simply ignored. "Instead of a clear dichotomy between normal and revolutionary science there is a range of different historical episodes within each such that the most innovative normal science looks just as significant as the least radical moment of revolutionary science" (Bird 2000: 62).[19] This suggests the difference between normal science and revolutionary science is one of degree not of kind. Bird proposes that revolutions are distinguished not by being revisionary or caused by crisis, but by their effect on subsequent science.

"The impression that revolutions are major revisions brought on by crisis and so are very unlike normal science is fostered by focusing on a few revolutions such as those of Einstein, Dalton, Darwin, Newton and Lavoisier, that change an entire science and dominate it for many decades or centuries. Yet there is no reason to think

that such a focus is justified by our pre-theoretical conception of a revolution, nor is it endorsed by the detail of Kuhn's discussion" (ibid.: 60-61). Revolutions are not just those (rare) occasions when an entire science undergoes a radical shift in focus. There are revolutionary discoveries that do not affect whole fields but only parts of them; Bird cites examples from chemistry including Kekulé's discovery of the structure of benzene, which "transformed organic chemistry but not inorganic or physical chemistry" (ibid.: 61). Bird claims it's natural to think of the proliferation of branches of science as resulting from such conservative or mildly revisionary revolutions. These branches then undergo their own 'mini-revolutions' (Bird notes that this stretches Kuhn's picture, but is not inconsistent with it). Elsewhere, Bird (1999: 40-41) argues that revolutions at any level, being variable in scale, may occur in stages. This picture – different degrees of revolutionary change, revolutions not different in kind from normal science, radical revolutions occurring gradually, revolutions introducing new sub-disciplines – supports the view of loveliness I have developed above. Revolutions inevitably force loveliness to adjust, but not every revolution involves scientists in dramatic adjustment of their concept.

Our use of certain aspects of the structure of Kuhnian science to illuminate IBE does not commit us to Kuhn's interpretation of paradigms, the five values and so on. Kuhn claimed that the values cannot constitute a standard against which exemplars may be assessed because they rely on exemplars for their content. We have agreed that exemplars fix the interpretation of the values, but we need not follow Kuhn in thinking that one paradigm's interpretation cannot be carried over into another's. On the contrary, we should maintain that it usually is, and that this is explained and justified by the fact that successive paradigms approach the truth. This in turn is explained and justified by the reliabilist view of scientific progress. Without Kuhn's views about paradigms as non-progressive, irreconcilably different ways of seeing the world, we may regard the five values as fully-fledged criteria, whose progressive use under successive exemplars is conducive to truth. To reiterate a claim made above, loveliness is *refined* by revolutions.

Of course, Kuhn's opposing view has historical examples in its support (though the value of any example may be disputed). But as we see in the next section, there is historical support for the view advocated here too. Ernan McMullin uses the case of the Copernican revolution to argue that values such as those Kuhn mentions do govern the inference of new exemplars *and* that they are invoked as explanatory and

epistemic virtues, just as IBE claims. History does not commit us to radical variance in standards for theory choice – standards of loveliness – between paradigms.[20]

6.1. McMullin on the rationality of revolutions

McMullin ([1993] 1998) distinguishes two groups of values in science. The first group of values can be seen as the goals of science, ends in themselves: "predictive accuracy (empirical adequacy) and explanatory power are the most obvious candidates" ([1993] 1998: 129). Call these 'ends-values'. The second group of values are means to these ends. "Some of these are quite general and would apply to any epistemic activity. Logical consistency (absence of contradiction) and compatibility with other accepted knowledge claims would be among these... Other values are more specific to science, for example, fertility, unifying power, and coherence" (ibid.). McMullin notes that these values "are obviously not goals in themselves; they would not motivate us to carry on an activity in the first place" (ibid.). Call these 'means-values'. Ends-values are self-justifying as they serve to define the activity of science: "if, as Kuhn notes, one relinquishes the goal of producing an accurate account of natural regularity, the activity one is engaged in may be worthwhile, but it is not science" (ibid.: 130). Means-values may be justified if it can be shown that they are means to achieving the ends specified by the ends-values.

McMullin believes that this can be shown by using historical examples. He claims that in science, means-values are in fact justified by their being conducive to ends-values. Further, he claims that we can only make sense of their use in historical episodes if we assume realism; that is, we need *both* of the ends-values McMullin mentions – predictive accuracy *and* explanatory power – if we are to justify the means-values. This is a problem for Kuhn, as he advocates the use of the means-values as a way to achieve the ends-values, but denies realism. If truth is absent as an aim of science, explanatory power lacks justification in terms of those aims. It may be justified on pragmatic grounds; indeed anti-realists often prefer explanatory theories for just this reason. But explanatory power is not justified in terms of constitutive anti-realist aims such as predictive accuracy or empirical adequacy; observational evidence suffices to determine theories' status on that score (a claim that a theory is, e.g., empirically adequate 'goes beyond' the evidence, but still the only evidence the anti-realist considers is observational).

IBE helps get a grip on this idea. Suppose an anti-realist accepts an explanatory theory. She sees its explanatory power as pragmatically useful, but her epistemic attitude towards it is one of acceptance rather than belief. By accepting the theory as empirically or predictively adequate rather than believing it to be true, she allows that whatever the explanation posits as doing the explaining may not exist, and hence that the explanation is not really an explanation at all. "For the facts F to be the explanation of [evidence] E, F *must* exist," says Bird. "Someone who employs Inference to the Best Explanation cannot but take a realist attitude to a theory which is preferred on these grounds" (Bird 1998: 146; see also chapter 4, section 1.1 here). For the realist, explanatory power is evidential; it gives a reason to believe a theory is true. Hence the anti-realist will struggle to account for any historical instance of theory choice in which explanatory power is used in an evidential way. Such examples will count against anti-realism. In particular, they will count against any anti-realism that tries to maintain, unusually, that explanatory criteria *are* essential to theory choice. As we are about to see, this appears to be Kuhn's position.

Famously, after the publication of *The Structure of Scientific Revolutions*, Kuhn was determined to respond to the charge that on his view, science was irrational. The basis of this charge was his claim, central to the idea of incommensurability, that paradigm choice cannot be independently justified. Consequently, the choice is not wholly a matter of argument, but rather a matter of conversion or 'gestalt switch', for which a full justification can never be given (we may explain scientists' choice, but such explanations will make essential reference to non-epistemic, broadly sociological factors). Kuhn's critics' claim was that if the proponent of one paradigm cannot justify it in terms acceptable to the proponent of another, then science is irrational. Thus Kuhn introduced the five values as a basis for rational comparison of paradigms. He insisted that by themselves they could not determine paradigm-choice, but as McMullin notes, "the fact that neither side can persuade the other does not undermine the claim each can make to have good reasons for what they assert" ([1993] 1998.: 120). Nevertheless, Kuhn's anti-realism holds firm. As we know, Kuhn thought that successive paradigms make puzzle-solving progress, or as McMullin says, "(to put this in a different idiom) they predict better" (ibid.: 131). But he denied that this represents an approach to truth. Paradigm change may be a rational affair, but Kuhn "rejects in a most emphatic way the traditional realist view that the explanatory

success of a theory gives reason to believe that entities like those postulated by the theory exist, i.e., that the theory is at least approximately true" (ibid.).

But McMullin argues that Kuhn cannot have it both ways. Kuhn cannot use the history of science to support both the five values and his anti-realism, since history suggests that where those values are put to use, it is with realist intentions. Theory choice *is* guided by the five values, but their rational application only makes sense if we assume realism. Briefly, McMullin's argument is this. Recall the two groups of values: ends-values and means-values. Four of Kuhn's five values (all except accuracy) are means-values, but Kuhn doesn't think they need justification from ends-values. He sees them as aesthetic or pragmatic factors, grounds for rational preference and (perhaps indirectly) conducive to puzzle-solving progress, but justified on their own terms. McMullin argues that proper attention to relevant episodes in the history of science reveals that they are, on the contrary, applied in order to serve the ends-value of explanatory power. Means-values are used to describe the way in which one theory *explains* the phenomena better than another, and hence why it's worthier of acceptance. Kuhn's values do not bring any predictive advantage, so their use does not make sense if acceptance depends only on the ends-value of predictive accuracy. McMullin concludes that explanatory power must be seen as a goal – an ends-value – of science too, if the five values are to have the rational role Kuhn desires. This means Kuhn's values imply a realist perspective.

McMullin argues this by looking at an example of a scientific revolution: the move from Ptolemy's geocentric theory of planetary motion to Copernicus' heliocentric model. Kuhn accepts the example: the details McMullin uses are extracted from Kuhn's own book, *The Copernican Revolution*. McMullin examines the reasons given by astronomers of the time for accepting Copernicus' theory over Ptolemy's, prior to Galileo's work to distinguish them. McMullin reports that in *The Copernican Revolution* Kuhn argues that there was little between the two accounts in terms of predictive accuracy, and that what persuaded the eminent astronomers of the day was the Copernican system's aesthetic appeal. McMullin disagrees, arguing that their reasons were epistemic, not aesthetic. He refers back to Copernicus' original arguments, and states that "what Copernicus claims to discover in the new way of ordering the planets is a 'clear bond of harmony,' 'an admirable symmetry'" (ibid.: 133). McMullin then asks "why should this carry conviction [as a mere aesthetic preference], especially since (as Kuhn emphasises) Copernicus in the end had to

retain an inelegant and far from harmonious-seeming tangle of epicycles?" (ibid.). These are aesthetic reasons to *disfavour* the theory. Copernicus' claims about harmony and symmetry make better sense when seen as epistemic reasons to favour his theory.

McMullin continues: "the heliocentric model could *explain*, that is, provide the *cause* of, a whole series of features of the planetary motions that Ptolemy simply had to postulate as given, as inexplicable in their own right" (ibid.), for example the orbits of Venus and Mercury, and the increased brightness of Mars, Jupiter and Saturn when rising or setting together. Kuhn argues that the Ptolemaic system explains these phenomena just as well, if only through an ad hoc addition to the theory. Again, McMullin disagrees: "the Ptolemaic system does not explain the phenomena mentioned above at all. Ptolemy is forced to postulate that the center of the epicycle for both Venus and Mercury always lies on the line joining the earth and the sun" (ibid.). This allows Ptolemy to account for the phenomena, "but this is surely not *accounting for* in the sense of explaining" (ibid.). The fact that the phenomena can only be accounted for in an ad hoc way prevents Ptolemy's theory from being an explanation good enough to be accepted for realist or anti-realist reasons.

Kuhn admits that the Copernican system is "more natural" than the Ptolemaic, since it can discard such ad hoc elements. Again McMullin argues this is not an aesthetic claim, but an epistemic one: "Copernicus himself makes the genre to which [this argument] belongs quite clear. He says that [he] is able to assign the cause of these features to the planetary motions, whereas Ptolemy is not. There is no reason in Ptolemy's system for them, other than the mere need to get the predictions right" (ibid.: 134). Concluding his argument, McMullin discusses Kepler's development of Copernicus' arguments to do with the retrograde motions, noting "their relative size and frequency from one planet to another and the lack of any such motions on the part of the sun and moon are exactly what one would be led to expect in a system where we are observing the motions from the third planet and the moon is not a true planet but a satellite of earth" (ibid.). Kepler also noted that, on either the deferent or the epicycle circle of the Ptolemaic model, the period of rotation for all planets is exactly one year. This seemed an extraordinary outcome, given Ptolemy's assumption that the planets' motions are independent of one another. McMullin claims that "Kepler is clear that the issue here is one of causal explanation; one of the systems can provide such an explanation, the other cannot. He is also clear that the criterion of prediction alone will not be enough to decide in all cases between two rival

accounts of the planetary motions and thus that a different genre of argument (he calls it 'physical') is needed" (ibid.).

The arguments in favour of the Copernican model invoke the kind of criteria Kuhn specifies as definitive of theory choice, but in an explicitly epistemic form. Rational sense can only be made of this episode if we regard explanatory power as a goal of science, alongside predictive or descriptive adequacy. As McMullin says,

> "Copernicus's criterion of 'naturalness,' the elimination of ad hoc features, the virtue that might today be called coherence, is not aesthetic; it is epistemic. He is not just appealing to his reader's taste, or sense of elegance... He is saying that a theory that makes causal sense of a whole series of features of the planetary motions is more likely to be true than one that leaves these features unexplained... Besides coherence, one could make similar cases for fertility and unifying power. It is hard to make sense of the role played by these values if one adopts the instrumentalist standpoint that Kuhn feels compelled to advocate" (ibid.: 134-135).

The coherence of Copernicus' theory is cited as a source of explanatory power, that is, as a reason to believe the theory as making true claims about worldly stuff. (Coherence isn't in Kuhn's original list of five values, but it is close to the value of simplicity, and anyway he admits that the original list is not exhaustive.) Note that on McMullin's account, explanatory power is not a means-value conducive to the ends-value of truth. Truth is not a constitutive goal of science; explanatory power is the constitutive goal, which brings truth with it. Science aims at a full explanatory account of the world, and if it achieves such an account, it thereby achieves truth. Thus the Copernican example supports the view that Kuhn's five values are used in a realist way. Truth-based goals, explanatory power specifically, are required to account for such episodes in a way that maintains the rationality of revolutions. McMullin claims the example is not unique and that other historical episodes show the same thing for Kuhn's other values. By introducing the five values, Kuhn opens himself up to this kind of historical attack.

Thus McMullin's argument supports our view that the five values guide inference as explanatory/epistemic virtues and that this applies even to scientific revolutions. Although he doesn't mention IBE, McMullin's work clearly supports the view that scientific inference is best understood as IBE, especially in the claim that explanatory power is a constitutive aim of science. So the historical perspective McMullin brings is likewise supportive of the claim that IBE governs inference during scientific revolutions: new exemplars are inferred using IBE. In those exceptional cases where a revolution causes a single standard of loveliness to change radically, it

may be hard to see how old exemplary standards of loveliness can play a role, but McMullin points us in the right direction. Attention to historical examples will provide the answer. There are independent grounds for endorsing IBE as the correct account of scientific inference, so with cases that exhibit discontinuity in standards of loveliness, the correct approach is to investigate what it was about the use of IBE that brought it about. Perhaps the Copernican example is one such case of discontinuity. McMullin doesn't discuss the extent to which the explanatory standards applied in that case were formed by existing theoretical commitments. But there is enough evidence that IBE is in use to encourage research into the influence of prior scientific knowledge on Copernicus' standard of loveliness and that of his supporters.

This discussion of my account of IBE has thus far focused on the connection between loveliness and understanding, but loveliness also has an aesthetic dimension. I wrote above about how scientists develop what might be called an aesthetic attachment to a certain standard of loveliness, which would make it difficult to relinquish were a revolution to demand it. Indeed, there is an aesthetic implication to the term 'loveliness' that Lipton occasionally exploits, e.g. when talking about how IBE might explain scientists' reports that they inferred a theory because it was elegant or simple (2004: 66). I do not want to drag loveliness away from understanding and towards aesthetics; I think an emphasis on the aesthetic implications of the terms 'loveliness' and 'explanatory virtue' leads to a dangerous misunderstanding of IBE. But it would be interesting to see whether inference on aesthetic grounds is compatible with Kuhnian IBE, and if so, what that tells us about the aesthetic aspects of exemplars of loveliness. To do this, we will take a look at James McAllister's account of aesthetic considerations in Kuhnian science.

6.2. McAllister on aesthetics in Kuhnian science

In *Beauty and Revolution in Science* (1996), McAllister argues for a Kuhnian account of science, but inverts Kuhn's view of empirical and aesthetic criteria. Kuhn thinks that aesthetic values are incidental to normal science, theory choice being determined on purely empirical grounds. Once science is in crisis, however, Kuhn thinks empirical criteria no longer hold firm, and at such times aesthetic criteria may be decisive.[21] McAllister argues that, on the contrary, normal science is characterised by consensus over aesthetic criteria, which the community regards as decisive in theory

choice. Scientific revolutions are periods during which the community replaces one set of aesthetic criteria with another, and empirical criteria govern theory choice. McAllister's aim is to maintain the rationality of science while explaining both the occurrence of revolutions and scientists' reports of using aesthetic criteria in much of their work. We should note that McAllister endorses constructive empiricism, under which the aim of science is empirical adequacy rather than truth (see chapter 4, section 1). But just as with Kuhn, we may draw helpful morals from McAllister's account of the structure of science without worrying too much about his background philosophical views.

The normal science consensus over aesthetic criteria is expressed in what McAllister calls the 'aesthetic canon' (1996: 34-35). An aesthetic canon is the set of all properties to which the scientific community might have an aesthetic response, with weightings attached to each property determining the degree to which it should be pursued.[22] Weightings are updated whenever the community performs what he calls the 'aesthetic induction' (ibid.: 77-81), which proceeds as follows. First, empirical criteria for theory choice are clarified. McAllister argues this is a straightforward task, involving simple analysis of the goals of science. This might be easier for a community of constructive empiricists than a community of realists, but even then it may not be straightforward. Perhaps more confusingly, McAllister (ibid.: 76) gives explanatory power as an example of an acceptable empirical criterion; as McMullin has just shown, this is a realist claim. But let's assume that a scientific community of any kind can agree on empirical goals. Then the community reviews current and recent theories that have fulfilled those criteria to a high degree. They examine these empirically successful theories to discern the aesthetic properties correlated with their success. Those properties then receive an increased weighting in the aesthetic canon. In the finalised aesthetic canon each property has a weighting proportional to the empirical success of the current and recent theories in which it's found. The canon is then applied to all theory choices until the aesthetic induction is next performed; McAllister explains that it may be performed as often as the community finds necessary. Theory choice based on the aesthetic canon is rational because the properties it recommends are correlated with empirical success.[23]

McAllister argues that scientists come to regard properties recommended by the aesthetic induction as beautiful. His most persuasive support for this claim is an analogy with the applied arts (ibid.: chapter 9). He gives several examples of aesthetic

canons in architecture and civil engineering which changed in response to the success of new materials involved in audacious projects, also highlighting how that success moulded aesthetic tastes. As new materials and new designs evolved to suit one another, people came to have positive aesthetic responses to projects that some years before would have been considered ugly. McAllister claims that scientific theory choice exhibits the same phenomenon. For McAllister the constructive empiricist, the purpose of the aesthetic canon is to enable scientists to isolate empirically successful theories more easily; theories do not benefit empirically from having a given aesthetic aspect. As we will see later, this part of McAllister's view can be safely jettisoned without losing the insights he offers into IBE.

McAllister (ibid.: 81-85) makes clear that aesthetic canons are inherently conservative. The scientific community seeks theories that resemble those of the recent past. The success of those theories in turn means that the aesthetic properties they feature, also present in the first set of theories, are again given priority when future theories are chosen. This also means aesthetic canons tend to retrench. The correlation between aesthetically pleasing properties and empirical success receives increasing support from the growing body of theories, so the weightings of those properties in the aesthetic canon increase each time scientists perform the aesthetic induction. Thus those properties are sought with still higher priority in the next instance of theory choice.

However, because the primary goals of science are to do with empirical success, scientists' objective is to accommodate the phenomena, not to propose aesthetically pleasing theories. Thus from time to time, usually in response to an anomaly, aesthetic properties considered unattractive, and not conducive to empirical success, will start to appear in new theories. Because of the conservatism and retrenchment of aesthetic canons, it will take the aesthetic induction some time to react to this. McAllister regards such a time lag as unproblematic, saying that "in a changing environment, an evolving system is unlikely to be able to readjust all its properties so that they are always optimal for the prevailing circumstances... human cultures invariably contain both elements for which there is a current justification on utilitarian grounds and elements that can be justified only by appeal to tradition or heritage" (ibid.: 82). In fact, McAllister treats the aesthetic induction's ability to incorporate such a time lag as an advantage of his view. He sees it as partly explaining the

displeasure regularly reported by scientists who encounter an aesthetically unorthodox theory.

After the interval during which the aesthetic canon and empirical criteria are at odds, the weighting attached to the property newly correlated with empirical success will start to increase, as the induction over past theories takes into account successful theories bearing that property. It is then valued to a greater extent by the community, will start to be sought in new theories, and in time will be regarded as aesthetically appealing. In this case, the anomaly has been resolved. Sometimes however, scientists may only be able to accommodate the anomalous data with theories radically different in aesthetic properties from those they replace. McAllister (1996: 128-133) says that such deep conflict between empirical and aesthetic criteria divides the scientific community into two factions who disagree about the correct way to solve the anomaly. One faction, which McAllister calls 'conservative', commits itself to the prevailing aesthetic canon; hitherto, it has been an excellent indicator of empirical success and it is therefore irrational to adopt aesthetically unappealing theories. The other faction, which McAllister calls 'progressive', abandons the aesthetic canon and concerns itself only with empirical criteria. Decisive for them is the fact that the empirical performance of aesthetically pleasing theories with respect to the anomaly is poor compared to that of new, aesthetically unorthodox proposals.

If the conservative faction cannot reconcile aesthetic and empirical criteria, and if the progressive faction continues to widen the empirical gap between its theories and theirs, then the conservative faction has only one option: it must relax its commitment to the aesthetic canon and come to favour the progressive faction's theories despite their aesthetic drawbacks. This gradual abandonment by the entire scientific community of an old aesthetic canon in favour of a new one, based around the properties of a new style of empirically successful theory, is McAllister's description of a scientific revolution. He terms it "revolution as aesthetic rupture" (ibid.: 125).

Normal science resumes as the correlations between aesthetic properties and empirical success grow stronger, and the community reapplies the aesthetic induction in order to stabilise the aesthetic canon. Over time, new theories come to be regarded as aesthetically attractive. Some open-minded scientists, McAllister claims, will allow their aesthetic taste to be shaped by the correlation with empirical success, perhaps because of the sense of aptness invoked when aesthetics and empirical success are in

harmony. Otherwise, the community's positive response to the new aesthetic properties will derive from pre-revolutionary scientists leaving the profession and by new recruits, trained under the post-revolutionary theories and without prior aesthetic commitment, joining up.

McAllister offers several examples in support of his account. An extended discussion of Copernicus' theory in astronomy and relativity theory in physics results in neither being seen revolutionary: "Copernicanism is most appropriately seen as the culmination of Ptolemaic-style astronomy, and relativity theory should be seen as the culmination of classical physics" (ibid.: 164). This is because both adhered to the prevailing aesthetic canons.[24] In the former case this meant fulfilling "a long-standing requirement placed on theories in mathematical astronomy, which may be described in terms of simplicity, symmetry, or metaphysical allegiance: the requirement that the motion of celestial bodies should be interpreted as uniform motion along circles or combinations of circles" (ibid.).[25] In the latter case, relativity theory "satisfied requirements that had become established in physics in the nineteenth century, that theories should be deterministic and show particular symmetries" (ibid.). On the other hand, McAllister finds that Kepler's theory was revolutionary because it abandoned the aesthetic canon in astronomy, resulting in it being accepted only with difficulty; likewise quantum theory in physics, which had to overcome commitments to visualisation and, more notably, determinism. McAllister notes that he departs from tradition in not calling all four examples revolutionary. I do not discuss further whether McAllister's case is historically plausible, except to note that he presents a wide range of evidence in its support, and that, further, his account and the traditional view may be reconciled via Bird's account of degrees of revolution (see above).[26] My aim is to use McAllister to support my claim that loveliness qua aesthetic standard guides rational theory choice in science, whether or not the choices in question are properly described as revolutionary.

Here though it may seem that McAllister's view and mine are in tension. His claim is that revolutions involve the abandonment of one set of aesthetic preferences in favour of a different one. I maintain that all theory choices, even the ones McAllister calls revolutionary, are governed by loveliness (where for the purposes of this discussion, loveliness is seen from an aesthetic angle). There are two things to say here. Firstly, McAllister is more moderate than this sketch suggests. On his account the reception of a revolutionary theory has three stages (see e.g. ibid.: 180): resistance

due to aesthetic displeasure, recognition of empirical superiority over competitors, and recognition of sufficient empirical success to warrant the community to "relax" or "de-emphasise" its aesthetic preferences. The last stage, recall, involves an adjustment of the weightings attached to properties in the aesthetic canon. The canon includes a list of all the aesthetic factors to which the community might have an aesthetic response. At the last stage, the community will step back from those with a (high) weighting attached and take a general overview of the entire list. The new theory may leave some weightings in place (though these will be in the minority), and cause others to be raised or lowered; properties that were heavily favoured before may be totally de-valued, and others that had a weighting of zero may end up highly sought-after. Empirical success motivates this readjustment, but nothing in this picture suggests that revolutions are necessarily attended by the kind of wholesale replacement of standards that would be a worry for my account. Further, it doesn't imply that an entire scientific community is left languishing without its trusted aesthetic standard at any point. All it suggests is that those who wish to develop (what McAllister considers) a revolutionary theory must relinquish their attachment to the prevailing aesthetic canon, and that if they are successful, others will be motivated to follow suit.

The second thing to say is that, in any case, we need not endorse any part of McAllister's view that suggests that a given standard of aesthetic preference/loveliness must be wholly abandoned and (some time later) replaced with another (though it may be sensible to do so if history suggests it has, at any point, actually happened). Kuhn's account of science was the starting point for my account of IBE, not McAllister's, and I've already argued against the equivalent problems on Kuhn's account. The purpose of discussing McAllister was rather to lend plausibility to the claim that loveliness qua aesthetic preference may be general to a scientific community and rationally applied in theory choice. To this end, numerous features of McAllister's account of science agree with my account of IBE. There is an aesthetic standard for theory choice in normal science, which is inherently conservative, and shared in virtue of being generated by exemplars (McAllister doesn't talk about exemplars, but it's clear this is what he means when he refers to "current successful theories" (ibid.: 79) as evidence for the aesthetic induction, and to the revolutionary replacement of theories that determine aesthetic standards). The standard generates empirical success because of its inductive connection with exemplars (on my account

however, there's no equivalent of the aesthetic induction – no explicit reasoning goes into forming standards of loveliness; rather, they are tacitly introduced by habituation to exemplars).[27] McAllister further agrees that aesthetic criteria may undergo adjustments during normal science. For him, this happens via periodical reapplications of the aesthetic induction; for me, it occurs through the kind of minor revolution discussed by Bird. Lastly, McAllister's account strongly endorses my claim that empirical success generates aesthetic attachment; he suggests this may have an evolutionary explanation (ibid.: 81).

I close by noting a possible advantage of my account over McAllister's. On McAllister's view, the aesthetic canon is merely a methodological expedient; it points scientists towards empirical adequacy, but it is not essential to science's goals (empirical adequacy may be achieved with empirical criteria alone). Consequently, the properties the canon promotes bring about minimal epistemic satisfaction, and the aesthetic satisfaction they produce is merely a happy coincidence. On my view however, standards of loveliness are essential to meeting the goals of science; loveliness is a guide to truth in a way that empirical criteria alone cannot be. Moreover, on my account, loveliness isn't applied to theory choice because it is the latest inductively justified tick-list of properties; it is applied because scientists are habituated to making judgements according to their acquired concept. Altogether, this means the epistemic and aesthetic satisfaction brought about by loveliness is both significant and easily explicable. I have a clear and credible motivation for the use of loveliness which McAllister's account of aesthetics in science cannot replicate.

7. Summary

This chapter has tackled the two most serious objections to IBE and defended it via an account of IBE in Kuhnian science. The objections were Hungerford's objection (I), that loveliness is too subjective to guide inference, and Voltaire's objection, that loveliness is not truth-tropic. Lipton's responses to these objections were helpful. In particular, his response to Hungerford's objection (I) pointed out that inference is to an extent audience-relative, so for it to be guided by something similarly audience-relative is entirely appropriate, and his response to Voltaire's objection drew attention to the fact that it is, at bottom, Hume's problem. But Lipton did not do enough to show that loveliness is normatively desirable as an inferential

guide. In sections 3 and 4 I showed how Kuhn's account of science offers better responses to the supporter of IBE. Exemplars of loveliness guide normal scientists towards explanatory puzzle-solutions. Because they are shared, judgements of loveliness converge, and because (under reliabilism) they approach the truth, the belief that loveliness is truth-tropic is justified. Kuhnian IBE then generated Hungerford's objection (II), that loveliness is relative to paradigms. Section 5 argued that in fact, loveliness is relative to puzzle-solving context, which is unobjectionable.

Section 6 presented several reasons to endorse Kuhnian IBE. Notable here were claims that we might expect IBE to generate the patterns Kuhn discerns in the history of science, and that Kuhn's five values work to show that science uses IBE and undergoes rational revolutions (McMullin provided historical evidence to this effect). Bird's picture of degrees of revolution also helped to reinforce the claim that revolutions needn't mean radical changes in standards of loveliness. Lastly, McAllister's account of aesthetics in Kuhnian science showed that normal science is characterised by consensus over loveliness and revolutions by aesthetic disruption. Thus the aesthetic dimension to Kuhnian loveliness should not be considered unscientific.

In the next chapter, we turn our attention to another kind of case for IBE. A certain kind of philosophical argument for scientific realism proceeds via a defence of IBE as the inferential method of science. That's partly been the aim of this chapter: I've wanted to convince you that IBE can answer two crucial objections, but I've also wanted to convince you that in doing so, I haven't presented an artificial account. Chapter 4 notes a number of ways in which Kuhnian IBE, as presented here, works in concert with the defence of scientific realism.

Endnotes

[1] Though see Ladyman (2005): 334.

[2] Lipton (ibid.: 145-147) adds that the only accounts of inference less affected by Hume's problem are deductivist (e.g. Karl Popper's falsificationism) or in some way more parsimonious with their inductions (e.g. van Fraassen's constructive empiricism). He rejects the first very swiftly, since deductivism cannot offer a descriptively adequate account of induction. He takes longer over constructive empiricism, but comes to the same conclusion (also claiming that it's inconsistent).

[3] Lipton needs to be careful here, since the psychological studies of which he is fond may be used against his descriptive approach to Hungerford and Voltaire. It could be argued that if our inductive practices are systematically unreliable then whatever is driving them is at fault. Since Lipton claims that loveliness guides inference, this might reinforce Voltaire's objection.

[4] As part of his response to Voltaire's objection, Lipton (ibid.: 147-148) reviews some features of IBE that make it coherent and allow it to capture certain normatively attractive features of other accounts of induction (Mill's methods and Bayesianism) and of inference in general (the role of background

belief). This is the weakest part of his response, perhaps because he's done such a good job of convincing us of these things already that they no longer seem especially relevant.

[5] Van Fraassen's sophisticated critique of IBE has justifiably received considerable attention. The criticisms so often dealt with in isolation – as indeed they are here – are more potent in unison, even more so when their relationship to van Fraassen's constructive empiricism is properly understood (as Ladyman et al. (1997) urge). I have mentioned parts of van Fraassen's critique as they have become relevant to present concerns. I acknowledge this does not do it justice, but hope that my project isn't excessively weakened by the omission.

[6] This presentation has a flavour of what Psillos calls van Fraassen's 'argument from indifference' (1996: 43-46). To the extent that Lipton's response to 'bad lot' is not a response to 'indifference', the latter is not discussed here.

[7] Ladyman (2005) argues that Lipton's argument fails when we rank for approximate truth rather than truth. Lipton (2005) responds.

[8] For a brief criticism of this response, see Okasha (2000: 696). Psillos (1996) offers a similar defence of IBE against van Fraassen; Ladyman et al. (1997) respond. Day and Kincaid (1994: 286) anticipate a response to van Fraassen of this kind (Lipton originally published his argument in 1993, but it's plausible that Day and Kincaid's paper was already forthcoming).

[9] Lipton (ibid.: 162) notes that scientific realists should not maintain that scientists are absolutely reliable rankers, as that would entail a completely true background This is incompatible with diachronic changes and excessively optimistic.

[10] Lipton goes on (ibid.: 159-161) to make analogous claims against the well-known argument from underdetermination: the fact that any theory we actually choose is just one of those in principle compatible with the evidence does not show that that theory is arbitrarily far from the truth. An interesting part of this discussion is Lipton's emphasis that methodological principle and substantive belief are not independent. This claim is echoed in my Kuhnian defence of IBE.

[11] Bird (2000: chapter 3) emphasises the importance of exemplars to Kuhn's paradigm-based explanation of scientific change.

[12] I refer those who doubt that puzzle-solving often means explanation-giving to section 6 below.

[13] Psillos notes that Boyd (e.g. 1984) holds the view that "the virtues which constitute explanatory power become evidential precisely because they are present in theories which enjoy theoretical plausibility and evidential support" (1999: 172). Psillos (ibid.) also remarks that Salmon held a similar view, but additionally showed how theoretical virtues could bear on (Bayesian) confirmation via an inductive assessment of past theories.

[14] Bird (2000: 71-79, 90-96) notices that Kuhn's view on exemplars – exposure during training generating shared intuitions about puzzle-solving via the inculcation of learned similarity relations – is supported by the theory of connectionism in cognitive science.

[15] This is intended to be a very basic version of the sophisticated reliabilist interpretation of Kuhn given by Bird (2000: chapter 6, especially 245-266).

[16] Kuhn does not claim this list of values is exhaustive.

[17] It is not explicit in Kuhn that exemplars determine the values in this way. Bird (2000: 78) says it is a "plausible answer" to the question of how scientists acquire a set of values. To this I would add only that if the five values were otherwise fixed, they may recommend puzzle-solutions different from those recommended by exemplars. If exemplars and value-standards must give a univocal decision in inference, it makes sense for the former to determine the latter.

[18] For Kuhn, there are two sources of incommensurability: the paradigm-relativity of standards for theory choice, and the paradigm-relativity of the meaning of theoretical terms. The first is dealt with here; the second I do not tackle, except to say that I broadly endorse the arguments against it given by Bird (2000: chapter 5). I co-opt only the Kuhnian structure of science, not the philosophical lessons Kuhn drew from it. Of course, the problem of meaning variance is still serious to the extent that it is supported by historical examples.

[19] I have removed the subscript 'K' from the word 'revolutionary' in this quotation. Bird attaches this to signify the Kuhnian meaning of the term: revolutions as any change to paradigm (ibid.: 42).

[20] In the next chapter, we see that naturalism has something further to say about IBE during scientific revolutions.

[21] The extent to which Kuhn's five shared values are empirical or aesthetic is a matter of debate. If they are aesthetic, normal science is not as empirically-motivated as Kuhn thinks.

[22] McAllister is a projectivist about aesthetics.

[23] It's worth noting again Psillos' remarks on Salmon's view that "given two theories T and T' which have the same observational consequences but are differentiated in respect of some theoretical virtues, one should regard T as more plausible than T' if, given the past record, theories which exhibit the

virtues of T are more likely to be true than are theories like T'" (1999: 172). Arguably, this amounts to a realist aesthetic induction.

[24] McAllister argues that "Kuhn's own finding that astronomers switched from Ptolemy's to Copernicus's theory primarily under the impulse of aesthetic factors ought to persuade him that this episode constituted no revolution" (ibid.: 176).

[25] McAllister (ibid.: 40) classifies the aesthetic properties present in scientific theories under five headings: form of symmetry, invocation of a model, visualisability/abstractness, metaphysical allegiance, and simplicity. He admits these classes are not exhaustive, and that any aesthetic property may not fit uniquely into any one class. He argues that the properties are distinctively aesthetic in virtue of being described in aesthetic language by scientists, and stimulating a sense of 'aptness', which he argues is definitive of beauty.

[26] Neither do I discuss whether McAllister's account of normal and revolutionary science offers a better account of history than Kuhn's.

[27] McAllister notes that the aesthetic induction may be carried out unselfconsciously (ibid.: 79).

Chapter 4

The no miracles argument

1. Introduction: realism and anti-realism

We've been discussing IBE, especially IBE in science, and at various points we've remarked on issues to do with scientific realism and anti-realism. Crudely, scientific realism is the view that science aims at truth; scientific anti-realism denies this. Realism and anti-realism come in various forms, some of which will be discussed in this chapter. Our main concern will be the role of IBE in an important argument for realism: the 'no miracles' argument. Realism is the view I have championed thus far, and indeed the view to which IBE is closely attached; this connection has been noted already, but will be discussed again shortly. First, let's take a look at anti-realism.

Why be anti-realist? There are various motivations for the view, but perhaps the most common is the 'pessimistic meta-induction' (cf. Laudan [1981] 1996). The intuitive idea can be expressed as follows. On many occasions in the past, scientists have been convinced they were right about the world. Subsequently, they turned out to be wrong, and radically so. Current scientists may think they're right, or at least on the right track, but there's nothing to say that they're any better situated than their mistaken predecessors. Thus the view that science progresses towards truth is unjustified. Those persuaded by this argument think that instead, we should construe science as having a more modest aim, such as accurate prediction. Such progress can be measured; we can be *certain* that a later theory makes more accurate predictions than an earlier one. Indeed science does increase the accuracy of its predictions, but anti-realists claim this tells us nothing about the truth-value of its theories. Even theories with a 100% track-record of accurate prediction may still be false.

But anti-realists aren't simply inductive sceptics; on the contrary, they often suggest that we judge whether to believe our latest theories by looking at the evidence of past science. If history shows that science is largely cumulative then it's rational to believe current theories to be true, nearly true or approaching the truth. But as we just saw, this is just what the anti-realist claims history doesn't show: the history of science

is a graveyard of false theories, and reminds us that later theories frequently paint a picture of the world radically different from earlier ones (themselves fervently championed in their time). Thus the anti-realist concludes that the only reasonable stance is to deny that science aims at truth, in favour of a more epistemically accessible aim such as accurate prediction.

A popular modern form of anti-realism is constructive empiricism, first defended by van Fraassen (1980). Constructive empiricism accepts much of what traditional anti-realism denies, notably the claim that scientific theories are to be 'taken at face value', i.e. interpreted as making all the claims about the world they seem to make, usually about the existence and behaviour of certain worldly entities or mechanisms. Thus constructive empiricism allows that theories are capable of being true or false, and that if true, they would explain the observable phenomena (traditional anti-realism holds that scientific theories could not be true or false, just better or worse tools for accommodating and predicting data). What constructive empiricists deny is that science *aims* at true theories and that theories' empirical success is *evidence* for their truth (cf. Bird 1998: 123-125).

Constructive empiricists take the same view as realists with respect to inferences involving only observable phenomena; the dispute is over what we can't observe. Realists think we can have good grounds for inferring the existence of unobservable stuff, while constructive empiricists think that inferences to unobservables are never well-founded. For them, theories involving unobservables can never properly be called true. Rather, they can only be known to be *empirically adequate*. A theory is empirically adequate if all its observable consequences are true. The motivating thought is that true theories are indistinguishable from false but empirically adequate ones. The constructive empiricist thinks that only empirical evidence can tell theories apart, and empirical evidence can only tell us which theories are empirically adequate, not which are true. Of course constructive empiricists cannot deny that science does sometimes deal in unobservables, but they regard them as useful only to the extent that they enable a theory to have true observable consequences. Constructive empiricists are *agnostic* about the existence of unobservables, and thus about the truth of theories that mention them. These concerns about the limitations of evidence mean constructive empiricists insist that the aim of science is empirically adequate theories, not true ones.[1]

Van Fraassen defines realism, the position he opposes, thus: "science aims to give us, in its theories, a literally true story of what the world is like; and acceptance of a scientific theory involves the belief that it is true" (1980: 8, italics removed). This definition needs urgent qualification. Most realists would blanche at the idea that acceptance of a theory involves the belief that it is true. Instead, they insist that theories may be accepted as nearly true, approximately true, or (in the philosophical jargon) truthlike.[2] For acceptance to involve belief in truth is too bold; it suggests that when a scientist accepts a theory, they believe they have the matter 'all wrapped up' and all investigations may cease. This is implausible. It's one thing for science to aim at truth, entirely another for it to achieve it. It's doubtful that science has ever discovered the whole truth about anything, and some doubt it ever will. Consequently, scientific realism cannot be the view that whenever scientists accept theories, they believe their work is done and they can pack up their instruments and go home. Rather, acceptance of theories involves belief in their truthlikeness; accepted theories are taken to be close to the truth, or at least along the right lines.[3] Thus Stathis Psillos defines scientific realism as "the view that mature and genuinely successful scientific theories should be accepted as nearly true" (Psillos 1999: xvii).[4] This definition will be assumed here. Psillos distinguishes realism by its attitude of "epistemic optimism". Epistemic optimism holds that "science can and does attain theoretical truth no less than it can and does attain observational truth, where by 'theoretical truth' we understand the truth of what scientific theories say about unobservable entities and processes... there is some kind of *justification* for the belief that theoretical assertions are true (or nearly true)" (xx-xxi). Realists hold that we can have theoretical knowledge; this is just what constructive empiricists deny, holding instead that science can only furnish us with truths about the observable realm.

1.1. Constructive empiricism and IBE

Having sketched the contemporary debate between realism (as qualified by Psillos) and anti-realism (as constructive empiricism), we may ask: can constructive empiricism make use of IBE? Some suppose that it can. For example, Lipton (e.g. 2004: 153, 200-203) considers a constructive empiricist version of IBE under which we infer that the best explanation is empirically adequate, which rivals the realist account he develops. In fact, he attributes such a version to van Fraassen. This is not

uncommon, but as Okasha (2000: fn. 2) and especially Ladyman et al. (1997: 311-314) observe, van Fraassen did not endorse a constructive empiricist version of IBE. He introduced the idea as part of his critique of IBE, but merely as a rhetorical device.[5] This is just as well, since I take it that such a model would be incoherent.

To see why, we must recall something about the nature of explanation. According to IBE, scientists infer that the loveliest potential explanation in any pool of competitors is an actual explanation. The potential/actual distinction helps to clarify the two-stage process of IBE, but it is also misleading. Really the distinction is between potential explanation and explanation simpliciter. If something is actually an explanation then it's an explanation; calling something an "actual explanation" is a bit like calling something "really real". We've already noted that in order for an explanation to explain, whatever it cites as doing the explaining must exist; otherwise it's merely a potential explanation or 'explanatory hypothesis'. If we make room for the non-existence of the explanatory stuff, we claim only that *it is as if* whatever purports to do the explaining is really there. In this case the alleged explanation doesn't explain anything, whatever the proponents of such a view might think. Thus IBE is necessarily realist: it commits us to the *truth* of the best explanation. This is not a quirk of Lipton's or anyone else's interpretation, but a result of the fact that under IBE we infer *explanations*.

In cases where explanations mention unobservables, constructive empiricist users of IBE would allow that the 'explanation' they infer may be false. Their claim would be that it is only as if the unobservable entities and processes it cites are as claimed. But if the world may be otherwise, then what they have is a mere hypothesis, a story about the world whose explanatory qualities are nothing but a happy coincidence.[6] It may be epistemically attractive, pragmatically valuable even, but it is questionable whether the hypothesis provides *understanding*. For this, it would surely have to reveal something about the world, and whether or not it does is precisely what's at issue between the two alleged versions of IBE. If hypotheses do not provide understanding, they can hardly be called explanations, and if what's inferred is not an explanation, it cannot be a version of IBE.

Constructive empiricists limit their truth-claims to observable truths, so their version of IBE would only deviate from the realist's when explanations refer to unobservables. But there are at least three reasons not to limit IBE only to observables. First, the distinction between the observable and unobservable is

dubious. The anti-realist would need to know exactly where the line was drawn in order not to go beyond it, and it seems that any proposal leads straightaway to a boundary dispute (many think this counts against constructive empiricism itself). Secondly, IBE would lose much of its distinctiveness as an account of inference. One of the advantages IBE has over other models of induction is that it gives a unified account of the legitimacy of inferences to observables and unobservables – both are justified just in case they provide the best explanation. From a descriptive angle, even those who might favour constructive empiricist IBE would admit that science sometimes makes inferences to unobservables. An account of scientific inference should accommodate these. The constructive empiricist would have to bring in some other story about inference in order to do so; in which case, why adopt IBE for only the observable inferences? This suggests the third reason why IBE should not be limited to inferences within the observable domain. The constructive empiricist would have to provide a principled reason to think that loveliness is indicative *only* of observable truth, rather than truth simpliciter. A crucial part of the anti-realist attack on IBE is the denial that explanatory virtue is indicative of truth and thus gives reasons to believe. If anti-realists are to avail themselves of such virtues in inference, they need to explain why it's legitimate for them, but not for realists (cf. Psillos 1996: 46).

It's worth mentioning that in some cases where unobservables are concerned, realists shouldn't infer the truth of the best explanation, but rather its approximate truth, as per the above definition of realism. This is unproblematic. The inference to the truthlikeness of the best explanation does not violate the explanatory commitments IBE implies. Inference to truthlikeness allows that an explanation might be incorrect in some finer details, but does not allow that the explanation might be false. Thus there is a commitment to the stuff doing the explaining, even if the stuff might, on further investigation, turn out to be (in some suitably limited sense) otherwise. Such commitments mean that truthlike explanations provide genuine understanding of the observable phenomena; that understanding may be incomplete, but this doesn't mean there's no explanation.

If we're convinced that IBE is the correct account of scientific inference, then we agree that science is inherently truth-directed. But ought we to be scientific realists? We might agree that science, as a matter of fact, aims to discover the truth, but are we justified in having such confidence in its methods? An argument for

realism is thus an argument for the reliability of IBE. A popular attempt to show that IBE takes us to truth is the no miracles argument, the articulation of which will occupy us for the remainder of this chapter. Using Psillos (1999: chapter 4) as our guide, I introduce the argument and consider its influential first formulation, due to Maxwell (1962, 1970). Then I move on to the definitive version proposed by Psillos and Richard Boyd.

2. The no miracles argument

The no miracles argument (henceforth 'NMA') earned its name after Hilary Putnam called scientific realism "the only philosophy of science that does not make the success of science a miracle" (Putnam 1975: 73). Crudely put, the NMA is this: scientific theories are empirically successful, and later theories are more successful than earlier ones; this success would be best explained by the hypothesis that scientific inference, IBE, delivers truthlike theories; therefore IBE is reliable. The NMA is itself an instance of IBE: the ability of realism (about the results of IBE) to explain scientific success better than its rivals is offered as reason to accept it. The NMA argues that that success is to some extent puzzling unless we adopt realism. (Note here that the success of science is accepted by realist and anti-realist alike. Even supporters of the pessimistic meta-induction agree that science enjoys empirical success; they just don't think it indicates an approach to truth.)

Psillos' discussion of the NMA begins with an analysis of two historical precursors to Putnam's argument, due to J.J.C. Smart (1963) and Grover Maxwell (1962, 1970). I endorse Psillos' (1999: 72-73) account of Smart's NMA as a plausibility argument (defined below) rather than an IBE, though I do think that Psillos doesn't draw the distinction quite as clearly as he might, for example when he notes that judgements of plausibility might be grounded in judgements of explanatory power (ibid.: 73). Further, Smart comes closer to offering an IBE than Psillos acknowledges, at one point noting that he is aware of "the importance for the problem of the reality of theoretical entities of C.S. Peirce's notion of 'abduction'" (Smart 1963: 39). Indeed, shortly afterwards, he offers an argument analogous to his NMA, involving a detective finding evidence of criminal activity. The detective correctly predicts that further evidence will follow, and concludes that "if there really were a criminal [rather than a theoretical fiction] then these predictions would no

longer be surprising" (ibid.: 47). This is strikingly similar to the examples often given in support of IBE (e.g. Harman 1965). Nevertheless, we should agree with Psillos that Smart offers a plausibility argument, mainly because of Smart's own intentions (cf. Smart 1963: 8-12) and the fact that his argument compares realism and anti-realism on explanatory grounds but does not use the former's superiority as logically compelling grounds to infer it.

However, I disagree with Psillos' account of Maxwell's NMA. I think Maxwell's argument *is* an IBE, as I now show.

2.1. Maxwell's historical precedent

We may represent Psillos' (1999: 73-74) reconstruction of Maxwell's argument as follows. (Note that Maxwell's target, instrumentalism, is a kind of anti-realism different from constructive empiricism. It sees scientific theories as merely as instruments used to organise and unify data that would otherwise be thought mutually irrelevant. Science accumulates these theoretical tools because they facilitate accurate prediction. Instrumentalism denies that theories are capable of being true or false, and usually seeks ways to eliminate purely theoretical terms from scientific discourse.)

P1: The empirical success of science demands explanation.

P2: Instrumentalism views theories as inscrutable 'black boxes', mere tools which, if fed true observational premises, yield true observational conclusions; this provides no explanation of those black boxes' contribution to the success in P1.

P3: Realism comprehensively explains the success in P1, and it does so simply and without ad hoc modifications.

C: Realism is more plausible than instrumentalism; hence it should be accepted.

As the conclusion reveals, Psillos thinks Maxwell's NMA, like Smart's, is a plausibility argument. Philosophers use plausibility arguments to "clarify conceptual disputes, i.e. disputes which are not amenable to empirical tests" (ibid.: 73). Arguments for two opposing positions are offered and we identify, a priori, the

plausibility or arbitrariness of each. Plausibility arguments rely not on a logically compelling form, but on an intuitive judgement, or as Psillos puts it, "anyone with an open mind and good sense... [finding] the conclusion of the argument intuitively plausible, persuasive and rational to accept" (ibid.). According to Psillos, Maxwell argues that realism is more plausible than instrumentalism because its ability to make sense of scientific success renders it intuitively more appealing. Psillos expresses his view merely by saying that Maxwell's argument is "analogous" (ibid.) to Smart's plausibility argument, and that Maxwell uses explanatory virtues (see P3) to "ground the plausibility judgements that are required for a defence of realism" (ibid.: 74). In effect, Psillos merely asserts that Maxwell's NMA is a plausibility argument; on Psillos' own presentation, Maxwell's argument could easily be an IBE; "more plausible" in C above could easily be replaced by "a better explanation". I argue that we should make this replacement; Maxwell's NMA is indeed an IBE. Maxwell had a commitment to explaining the success of science, rather than a commitment to highlighting the plausibility of realism.

2.1.1. Explanation, comparison and explanatory virtue

I will highlight three features of Maxwell's NMA that reveal it as an IBE: its explanatory intention, its comparison of competing explanations, and its use of explanatory virtues to urge acceptance (further evidence is provided by Maxwell's Bayesian version of his argument: see sections 2.1.3 and 2.1.4).

The first IBE-related feature of Maxwell's NMA is that it promotes realism on *explanatory* grounds (see P1). It emphasises that the success of scientific theories is explained by the existence of the unobservable entities they mention: "the only reasonable explanation for the success of theories of which I am aware is that well-confirmed theories are conjunctions of well-confirmed, genuine statements and that the entities to which they refer, in all probability, exist" (Maxwell 1962: 18). Elsewhere, Maxwell claims that "the reality of theoretical entities provide[s] an explanation of the occurrence of the observational events which they predict. And – equally important – an explanation for the fact that theories 'work' as well as they do is, as already noted, also forthcoming; it is simply that the entities to which they refer exist" (ibid.: 20). Maxwell is not arguing that realism's ability to explain enhances its plausibility – there is no mention of plausibility here – rather he is arguing straight

from explanatory power to acceptance: realism explains success; *therefore* it should be accepted. Such a method is the basis of IBE.

The second important feature of Maxwell's argument is that it *compares* the realist and instrumentalist explanations of the success of science. He says this of the instrumentalist explanation (see P2):

> "to say that theories are *designed* to accomplish this task [explaining and predicting observations] is no reply... the thesis that theoretical entities are 'really' just 'bundles' of observable objects or of sense data would, if true, provide an explanation; but it is not taken very seriously by most philosophers today – for the very good reason that it seems to be false" (ibid.: 18).

Maxwell finds the instrumentalist explanation wanting. In fact, he finds that it's no explanation at all, because it's false. Later, he is clearer still: "instrumentalism... cannot provide an explanation as to why its 'calculating devices' (theories) are so successful. Realism provides the very simple and cogent explanation that the entities referred to by well-confirmed theories exist" (ibid.: 22). Comparison of competing potential explanations is a key feature of IBE. The instrumentalist proposal is eliminated because it's false, while the realist's proposal is not only better supported but also virtuous.

Thus we find the third IBE-related feature of Maxwell's NMA: it compares explanations on grounds of *explanatory virtue* (see P3). Above he describes the realist explanation of success as "simple and cogent". In the later presentation of his argument he says that "realism explains [success] very simply" and that "the competitors of realism become more and more convoluted and ad hoc and explain less than realism" (Maxwell 1970: 12). As noted, Psillos sees these explanatory virtues as grounding Maxwell's judgement that realism more plausible. But given Maxwell's avowed desire to *explain* the empirical achievements of science, it's fairly clear he's invoking simplicity, comprehensiveness and non-adhocness as explanatory virtues that recommend realism directly, rather than via a plausibility judgement.

Moreover, Maxwell (ibid.) argues that as the success of science increases, realism retains its virtue, while instrumentalism loses whatever virtue it had and cannot regain it. He is talking about the features of two hypotheses and the way in which their virtuousness changes in relation to new evidence. He is suggesting that this relationship is objective, or at least in some sense evidential; degree of simplicity, comprehensiveness and non-adhocness has a direct bearing on the support realism and instrumentalism enjoy as hypotheses about the nature of science. This is

inconsistent with the virtues being the basis for a plausibility judgement. The virtues are not being used philosophically, to determine degree of plausibility, but rather *scientifically*, to determine degree of support. Given Maxwell's interest in explanation, the natural interpretation of this is that he sees them as explanatory. If the argument is about evidential support and cites explanatory virtues, then there is none of the wriggle-room afforded by plausibility arguments; explanatory virtue compels the inference.

Maxwell has compared the realist explanation of success with the instrumentalist's and found that the former is more virtuous and thus better supported. He has made no reference to plausibility, but has used explanatory power to argue directly for acceptance. I conclude that Maxwell's NMA is an instance of IBE.

2.1.2. Reconstructing Maxwell's argument

Psillos claims that Maxwell's use of explanatory virtues shows that the plausibility judgement he motivates between realism and instrumentalism is not "distinctively philosophical". He then claims that "Maxwell's argument is the 'bridge' between Smart's a priori argument and the subsequent Putnam–Boyd *naturalistic* version" (Psillos 1999: 74). As we shall see in more detail later, naturalism is the view that philosophy is continuous with the natural sciences, particularly in the sense that it should treat its subject matter as material for scientific investigation. Under naturalism, nothing is distinctively philosophical. Thus Psillos acknowledges the influence of scientific methods on Maxwell's thinking: Maxwell's NMA is not distinctively philosophical, like Smart's plausibility argument, but neither is it full-bloodedly naturalistic, like the Putnam–Boyd version (see section 2.2). Psillos is right, yet by noting that Maxwell has imported the considerations of simplicity, comprehensiveness and non-adhocness from science, and that doing so has changed the character of his argument to something other than "distinctively philosophical", Psillos concedes that my interpretation of Maxwell is correct. He must think of Maxwell as seeing the scientific use of explanatory virtues as different from their philosophical use; otherwise how could they stop Maxwell's argument from being distinctively philosophical like Smart's? It is odd, not to say inconsistent, for Psillos to

maintain that Maxwell offers a plausibility argument while acknowledging that he saw explanatory virtues as evidential.

We can begin to explain Psillos' error by noting the following. It is possible that science might use simplicity, comprehensiveness and non-adhocness in a broadly evidential sense, and even call them explanatory virtues, without using those considerations as the basis for inference. In arguing that Maxwell's NMA is an IBE I have not committed him to the view that science uses IBE. All I have committed him to is what's already implicit in Psillos' account, viz. that he saw the considerations he mentions as explanatory and evidential. But Maxwell's lack of commitment to IBE in science does not stop his own argument from being an instance of IBE. It's not the case that a philosophical argument can be an IBE only if its author thinks that science uses IBE and for that reason uses it himself. We'll see later why a naturalistic approach is an advantage for anyone proposing a version of the NMA, but what's important here is that a philosophical argument can be an IBE in the absence of naturalistic intent. Nor is it a condition on being a version of the NMA that it provide a direct defence of IBE; all an NMA needs to do is explain the success of science. Perhaps Psillos thinks that because Maxwell didn't see science as using IBE, we can't interpret him as proposing an IBE-based NMA. But in fact, we can view him as calling upon explanatory virtues in a non-distinctively philosophical way without assuming that he thought scientific inference is IBE. I have shown that Maxwell's argument has (a) a commitment to explaining the evidence of scientific success, (b) identified realism as the best explanation of that evidence, and (c) shown realism to be compelling on those grounds. This is what's needed to be an instance of IBE; views about scientific inference, and the relationship between it and philosophical method, are unnecessary.

Perhaps that is a bit quick. Maxwell doesn't explicitly say that explanatory virtue rationally compels the inference to realism. Maybe *this* is why Psillos is reluctant to call Maxwell's NMA an IBE. But this can't be right. What this really amounts to is saying that Maxwell doesn't give his argument the structure of an IBE. But just because there isn't a clear-cut reproduction of the logic of IBE doesn't mean an argument can't count as an instance of it. Such a requirement would be far too strict: only a handful of arguments from the history of philosophy – or, for that matter, the history of science – would be reconstructible in terms of IBE. It's too much to ask that they have the rigid form of IBE in their original sources.

Looked at from the other side, numerous historical arguments, scientific ones in particular, have been reconstructed as IBEs. In many cases this is highly illuminating, so why can't it be done with Maxwell? There may be a reason why he himself resists calling his argument an IBE (or in 1962, an abduction), or resists putting it in that form more explicitly. But it's incumbent upon neither philosophers nor scientists always to announce the form their argument will take, or to formulate them in terms that make them easily identifiable as type *a*, *b* or *c*. One of the assumptions of exegesis is that it's a non-trivial exercise to examine arguments and decide their form; one of the benefits of the exercise is that this is often revealing to audience and author alike. Looking again at the original evidence, I have argued that this is Maxwell's claim: because realism is better at explaining the success of science it should be accepted. One might ask: what more do we need for IBE?

2.1.3. Maxwell's Bayesian NMA

Psillos' distinction between Maxwell's NMA, IBE, and the Putnam–Boyd NMA starts to look even more strained when we look at his discussion of Maxwell's next move. Maxwell (1970) gives his argument for realism a Bayesian interpretation. In the original text, the argument is given informally in only a few lines; Psillos reconstructs it for his discussion as follows. If we take realism and instrumentalism as hypotheses about science, we may take it that both entail that science is successful. In accordance with Bayes' theorem (see chapter 1, section 2.3), it follows that the likelihood of the evidence (the success of science) given either realism or instrumentalism, is 1. The prior probability of the evidence will be the same for both hypotheses – they both account for the same success. Thus any difference in posterior probability between realism and instrumentalism will come down to a difference in the prior probabilities of the hypotheses. Realism and instrumentalism are inconsistent – they cannot both be true – so if we want Bayes' theorem to distinguish them evidentially, they must be given different priors. Given that realism is a simpler, more comprehensive and less ad hoc hypothesis than instrumentalism, Maxwell claims that its prior should be much greater than its rival's.[7] Thus if either realism or instrumentalism is better confirmed by the evidence of scientific success, realism is.

The purpose of discussing Maxwell's Bayesian NMA is to bring out his sympathy for the scientific approach. It is not a new argument; Psillos correctly

discusses it as the earlier NMA recast in Bayesian form. Thus it is especially hard to grasp why he doesn't see Maxwell as straightforwardly naturalistic in the first place; that is, why he sees Maxwell's NMA as "the 'bridge' between Smart's a priori argument and the subsequent Putnam–Boyd *naturalistic* version" rather than simply akin to the Putnam–Boyd argument. Psillos notes Maxwell's view that Bayesianism and the virtues of different hypotheses play a role in scientific inference. He also notes Maxwell's view that the realism-instrumentalism debate is much like a scientific problem in which evidence can't distinguish between two competing hypotheses. These considerations lead Maxwell to bring scientific method to bear on philosophy, as Psillos realises when quoting the following passage: "my reasons for accepting realism are of the same kind as those for accepting any scientific theory over others which also explain current evidence" (Maxwell 1970: 18).

Thus Psillos acknowledges that, in this case at least, Maxwell supports a naturalistic approach. He describes how Maxwell uses a scientific outlook to deflate the "grand conceptual dispute" – philosophical positions are seen as hypotheses trying to account for evidence – and then how he solves the problem by bringing in a scientific method – Bayesianism – to distinguish the two sides. Why then does Psillos not call this argument naturalistic? The only reason I can see is that Maxwell himself didn't call his argument naturalistic. But as we've seen already, how philosophers label themselves or their work should not constrain our interpretation. Maxwell's Bayesian NMA *is* naturalistic, and this is yet another respect in which his work anticipates that of Putnam, Boyd and Psillos himself.

Clearly, Maxwell's Bayesian NMA is not an IBE – it's the original IBE couched in Bayesian terms. But of course Psillos thinks of it as a Bayesian representation of a plausibility argument. Once again this is mistaken. Like the original, the Bayesian argument has key explanationist components. Psillos thinks of the priors in Maxwell's argument as "initial plausibility rankings" (Psillos 1999: 74) of realism and instrumentalism, but once again he can't avoid talking in terms of explanation. When noting how the virtues influence the priors, he says "the realist explanation of the success of science is simpler, more comprehensive and less ad hoc than any instrumentalist attempt at such an explanation" (ibid.: 75). Given that Maxwell's own approach to the problem is in terms of explanation (this is evident again in the quotation in the previous paragraph, which Psillos himself chooses to illustrate Maxwell's Bayesian-scientific intent) this cannot reasonably be judged a matter of

plausibility. Maxwell's emphasis on explanation and his desire to employ scientific methods mean his NMA must be construed as arguing that realism explains better the success of science, and that *on those grounds* it should be accepted (because better explanation issues in greater confirmation).

This interpretation of Maxwell would be tempting in any case, but it is even more plausible in the light of Lipton's discussion of 'Bayesian abduction' (2004: chapter 7; see chapter 1, section 3.3 here).[8] According to Lipton, IBE provides a heuristic that enables us better to realise the Bayesian calculation. Explanatory considerations guide us towards sensible values for the priors and likelihoods. The only value relevant to the Bayesian NMA is the prior of the hypothesis, and sure enough, explanatory considerations step in to help. Lipton reiterates Maxwell's comments almost exactly: "this is where considerations of unification, simplicity and their ilk would naturally come into play" (Lipton 2004: 115).[9] But Lipton rightly notes that Bayesians have long been happy with explanatory considerations playing a role in fixing priors. What's supposed to be distinctive about Lipton's account is that explanationist thinking expresses the Bayesian calculation, not vice versa. So why is Maxwell's NMA a Bayesian abduction, rather than straightforwardly Bayesian?

The question is misconstrued: Maxwell's NMA is a Bayesian abduction *and* straightforwardly Bayesian. Lipton's point isn't to deny the possibility of Bayesian argumentation; he doesn't claim that every time we see a Bayesian argument what we really see (and must be able to argue for) is an IBE. Rather, he claims that wherever there's a Bayesian argument, there's an argument that's likely to have had explanationist inputs: "it may be possible to see at least one heuristic, Inference to the Best Explanation, in part as a way of helping us respect the constraints of Bayes' theorem" (ibid.: 112); Lipton's aim is to show that "Bayesianism is compatible with the governing idea of Inference to the Best Explanation…, the idea that explanatory considerations are a guide to likeliness" (ibid.: 107). Arguments may be laid out in Bayesian language, even in Bayesian form, but still be at bottom explanationist; likewise IBEs can be given a Bayesian twist, just as I claim Maxwell did. One does not have to trump the other; the two are not in competition. My claim is that since Maxwell's Bayesian NMA fits Lipton's description of Bayesian abduction, we may regard him as using IBE as a heuristic to help fix the Bayesian values. This lends further support to my claim that the argument is a Bayesian reconstruction of an IBE.

We should remember that Maxwell's Bayesian NMA is given in just a few lines. Psillos does a lot of exegetical work to get his points across, and while he doesn't misrepresent Maxwell, the limitations of the source material mean some interpretation must be entered into. I have argued that my interpretation is better supported by the background to Maxwell's discussion and the language he uses to express his reasons for supporting realism.

2.1.4. Maxwell on explanation and science

My IBE-based interpretation of Maxwell is further supported by the context of his Bayesian argument. Maxwell (1970) is arguing for a structural realism about the external world on the basis of what scientific theories tell us about sense perception. Part of that argument is a rejection of what he calls 'strict inductivism', the view that "confirmation can only be accomplished by inductive arguments of a very simple kind such as induction by simple enumeration… and Mill's methods" (Maxwell 1970: 5). More generally, Maxwell rejects 'strict confirmationism', the view that "the confirmatory relationship between evidence and hypothesis is of a purely logical nature, where 'logical' means purely formal as well as rational in all possible worlds" (ibid.: 6, italics removed). Such views, he claims, are instances of the 'fallacy of epistemologism': "the confusion of meaning with evidence or confusion of what a proposition asserts with how it comes to be known" (ibid.: 16). Instead, Maxwell adopts a view of confirmation he calls 'hypothetico-inferential reasoning', the view that "theories or hypotheses are proposed as a result of yet poorly understood creative acts of the mind. When a satisfactory hypothesis is conjoined with certain other knowledge claims… the evidence may be inferred from this conjunction, usually deductively" (ibid.: 6).[10] He claims that confirmation by this method, "both in scientific inquiry and everyday reasoning,… is frequently employed" and when we do this "our main reason for holding that a certain hypothesis is true or quite likely to be true is that it *explains* the facts so well" (ibid.: 7).[11]

The purpose of Maxwell's paper is not to argue for hypothetico-inferential reasoning or discuss it in great detail. Rather it is to show how, by giving us an alternative to "strict inductivism and other varieties of the fallacy of epistemologism" (ibid.: 29), that method, in some suitable form (again not discussed), allows us to know structural properties of the external world, scientific theories, other minds etc.

With regard to the latter, Maxwell says "the grounds for belief in the existence of 'other minds' – of the existence in others of thoughts and feelings very similar to our own – become clear and unproblematic. Such beliefs are the hypothetico-inferentially best confirmed explanations of the relevant observable evidence, which, in its turn, is causally produced in our own sense experience by the others' behaviour and by other relevant events" (ibid.).

Maxwell is not endorsing IBE as we know it. But he very clearly *is* endorsing the influence of explanation on inference and confirmation. Specifically, he's exploiting the symmetry that Hempel noticed between the hypothetico-deductive model of confirmation and the deductive-nomological model of explanation: the evidence that confirms a hypothesis is the evidence it would explain. On Maxwell's account, we gather evidence, formulate an explanatory hypothesis, join it with certain factual conditions, and deduce what we've observed, thus confirming the explanation.[12] While this isn't IBE, it does emphasise Maxwell's naturalistic tendency and his concomitant intention that the realism debate be decided by explanatory power rather than plausibility. This supports my view that Maxwell's argument is best reconstructed as an IBE.

Having presented his Bayesian NMA, Maxwell summarises his view as follows:

> "That a philosophical position such as realism turns out to be a contingent theory and one, moreover, for which scientific theory and observational evidence are relevant should not surprise us once we have freed ourselves of the mistaken views about the nature of philosophy that arise from strict confirmationism… What holds for these issues regarding realism holds also, I believe, for most – perhaps not all – of the important, interesting problems of epistemology and metaphysics. These problems differ only in degree and not in kind from more run-of-the-mill scientific problems" (ibid.: 18).

Maxwell's approval of (some form of) hypothetico-inferential confirmation imported from science, for which explanatory hypotheses are the starting point, drives home the point that for him, the realism debate can be settled on grounds of explanation because it's basically no different from a scientific dispute.

Perhaps, had it been more fully-developed at the time, Maxwell would have endorsed IBE more explicitly. His brief comments on hypothetico-inferential reasoning and the Bayesian NMA show that he would have advocated the use of explanatory considerations at the generation stage of IBE, while his earlier comments about explanatory virtues show that he would have endorsed their use at the selection stage. I have argued that Maxwell's original NMA is not a plausibility argument but an

IBE, and the later reformulation is not a Bayesian reconstruction of a plausibility argument but a Bayesian abduction. I have argued this on the grounds that Maxwell's dual concern was to support realism because it better explains the success of science, and (to some extent in 1962 but definitely so in 1970) to argue that the realism debate is akin to a scientific problem, and should be resolved accordingly. On this issue at least, Maxwell's views are broadly naturalistic. He didn't use that word to describe his arguments, in the same way he didn't call them IBEs. To this extent they differ from the Putnam-Boyd NMA; but this is a difference of terminology, or historical accident, alone. I conclude that Maxwell's NMA is not, as Psillos claims, "the 'bridge' between Smart's a priori argument and the subsequent Putnam–Boyd naturalistic version", but rather a fully-fledged IBE-based precedent to that argument.

2.2. Boyd and the NMA

Smart's and Maxwell's versions of the NMA are presented in a matter of a few pages, in works that don't examine critically the strategy of defending realism by appeal to scientific success. By contrast, Boyd's version of the NMA (I do not consider Putnam's views any further) has been developed through several publications whose collective aim has been a thoroughgoing explanationist defence of realism, a defence which brings with it a distinctive approach to epistemology in general. The key features of this approach are its *naturalism* and *reliabilism*. We find out below what benefits these bring to the NMA. This section will also examine the relationship between the Boyd–Psillos defence of realism and my defence of IBE in chapter 3, showing how they stand in a relation of mutual support. But let's start by noting one way in which their strategy improves upon that of Maxwell.

Maxwell's NMA is effective only against instrumentalism; it is ineffective against van Fraassen's constructive empiricism. The reason is this. Maxwell establishes that the empirical success of scientific theories supports a realist construal of those theories. Van Fraassen agrees that theoretical claims should be interpreted realistically but denies that if such claims (so interpreted) entail well-confirmed predictions then they are themselves well-confirmed. For Maxwell, this thesis about confirmation follows from the semantic realism he has argued for elsewhere. But van Fraassen's position is precisely that successful prediction is *not* a reason to believe even realistically-construed theories to be true (at least insofar as they concern

unobservables). He holds that we should remain agnostic about the truth of theoretical claims; predictive success, no matter how impressive, establishes nothing more than empirical adequacy.

Crucially, van Fraassen's views on confirmation and the inference from empirical success to scientific realism are two effects of a common cause: his denial that IBE is truth-conducive. Were IBE to be shown to produce rationally compelling conclusions, claims about unobservables would be justified and the inference from success to realism would be legitimate. So to mount an argument against constructive empiricism, the realist needs to show that scientists' first-order IBEs – inferences to (the truth of) scientific theories – are justified. Then by the same token, the realist can show that his second-order IBE about those IBEs – the NMA – is justified. A defence of IBE is both a defence of approximate truth as the best explanation of success and a defence of the inference to that conclusion.[13] Unlike Maxwell, Boyd defends realism by providing a defence of IBE; hence, if his argument is successful at all, it's successful even against constructive empiricism.

Boyd expresses his argument in several different ways (e.g. 1984, [1990] 1996), and in each case he emphasises modifications and restrictions to the crude form of the NMA we've so far been discussing.[14] However, the following is a reasonable reconstruction which echoes that found in Psillos (1999), and which Psillos himself endorses and extends:

P1': Scientific methodology is theory-laden: previously accepted background theory informs all parts of scientific practice.

P2': Mature sciences achieve success: (non-background) theories deliver correct predictions; the methodology of mature sciences is instrumentally reliable.

P3': The best explanation of the success in P2' is that any background theory asserting a causal connection, or other mechanism in virtue of which a (non-background) theory yields a correct prediction, is approximately true.

C': In virtue of P3', realism (about any background theory used to design experiments whose results agree with the prediction under test) should be accepted.

The starting point for Boyd's argument is not the success of science simpliciter, or even the predictive success of certain theories, but the *instrumental reliability of mature scientific methodology*. Boyd defines this as "the extent to which [its] practice is conducive to the acceptance of instrumentally reliable theories", and the instrumental reliability of a theory as "the extent of its capacity to make approximately true observational predictions about observable phenomena" (Boyd [1990] 1996: 221). Crucially, "scientific realists and their opponents largely agree that the methods of actual recent scientific practice are significantly instrumentally reliable" (ibid.).

Boyd sees it as a key philosophical task to offer an adequate explanation of why scientific methods have this feature. His insight is that those methods depend heavily on background theory. It is often remarked that background theory influences such things as the design of experiments and the interpretation of evidence: observation is theory-laden. But Boyd's claim is different: background theory determines "judgements of projectability and degrees of confirmation"; that is, it provides "the standards by which scientists determine which general conclusions are even real candidates for acceptance given an (always finite) body of available data" (Boyd 1984: 57). According to Boyd, scientists have no theory-independent way to gauge the relative confirmation of competing hypotheses (much less a single hypothesis in isolation), and no theory-independent way to tell which hypotheses carry terms that have 'latched onto' objective features of reality (i.e. terms that refer – at least partially – to real entities, properties, relations etc). Consider a simple example.[15] I have two competing hypotheses about the pH level of a liquid: H1 says it is acidic (has a pH less than 7); H2 that it is alkaline (has a pH greater than 7). I dip some litmus paper into the liquid; it turns red. Which hypothesis is confirmed? An answer depends on background theory: ionic equilibrium, conjugation theory, LeChatlier's principle, even theories of the transmission and absorption of light, collectively tell me that litmus paper turns red in contact with acid. Thus H1 is confirmed. Confirmation depends on theories defining the relevant parts of the world – acids, pH levels, litmus etc – and telling me when an observation confirms a hypothesis that mentions those worldly items.

If this sounds a bit Kuhnian, it's no accident. Boyd agrees with Kuhn about the profound influence of the prevailing theoretical tradition on the practice of science. He also agrees that such traditions change over time. Where Kuhn and Boyd differ is over the explanation of the success that such theory-bound science achieves. Boyd

phrases the key question thus: "why should so theory-dependent a methodology be reliable at producing knowledge about (largely theory-independent) observable phenomena?" (ibid.). Kuhn thought that any such success could be explained by saying that the world is somehow defined by the theoretical tradition. If scientific methodology and the world it investigates are both defined by the prevailing paradigm, it's no surprise that the former should be in tune with the latter, and generate 'success' accordingly.[16] In reply Boyd notes that, amongst other things, anomalies, which Kuhn defines as observations the relevant paradigm cannot explain, show that the "instrumental reliability of particular scientific theories cannot be an artifact of the social construction of reality" (ibid.: 60).[17] Boyd's point is that if reality were so constructed, then there would be no anomalies.[18]

2.2.1. Naturalism and reliabilism

We noted that there are two key philosophical theses upon which the Boyd–Psillos defence of realism depends: naturalism and reliabilism. Having sketched Boyd's approach, we now look at these views. Reliabilism is familiar; it was discussed as part of my interpretation of Kuhnian science in chapter 3. We'll return to it shortly, but let's look at naturalism first.

Ronald Giere, a leading naturalist in the philosophy of science, says that naturalism "may be characterized only by the most general ontological and epistemological principles, and then more by what it opposes than by what it proposes" (Giere 2000: 308). Maybe so, but hopefully it's not too controversial to say that the core naturalist belief is that philosophy is continuous with the natural sciences (this was the characterisation I used when discussing Maxwell). Usually this means a belief that the methodology of science (appropriately construed) is the correct way to investigate the subject matter of philosophy. Two motivations are often cited. First, there's the failure of certain alternative philosophical approaches. Here naturalists usually mention traditional a priori methods such as conceptual analysis and programmes such as that of logical empiricism. The former is taken to have furnished us with no uncontested conclusions despite centuries' effort (cf. Boyd [1990] 1996: 226-227), the latter to have shown that knowledge cannot be reduced ultimately to observation statements, thus refuting empiricist foundationalism (the naturalist programme is arguably most advanced in epistemology, thanks to the work

of W.V.O. Quine). The second customary motivation for naturalism is the progress exhibited by science. Crudely, this is the thought that science is highly successful with respect to its goals, certainly more so than philosophy, so philosophy could learn a lot by taking its methods seriously.

Underlying these two motivations is the scientific precept that whatever phenomena philosophy chooses to examine are 'out there in the world', any dispute about them being in principle decidable by empirical investigation. The naturalist holds that scientific methods are our best means of contact with the world, so the best way of investigating, say, knowledge, is to treat it as a 'natural kind', apt to be found in certain worldly locations, displaying certain investigable properties, having been generated by certain investigable causal processes, and so on. In Quine's terms, naturalists believe there is no 'first philosophy', no perspective on the world prior to the scientific one. That which is called 'philosophical' and that which is called 'scientific' are both just part of our overall theory of the natural world.

Now recall that reliabilism is the thesis that one's beliefs are justified just in case they are formed by a reliable method, where 'reliable' means something like 'tracks the truth'.[19] Reliabilism requires only that the method is *in fact* reliable; for the reliabilist, justification has nothing to do with what, if anything, the believer herself believes about the method in question. If we wish to investigate the reliability of our methods (reliabilism doesn't prohibit such research), the investigation is an empirical matter. It's contingent whether a method is reliable, so we must determine a posteriori whether certain inferences, or certain types of inference, successfully hook up with truth. Justification is not some mysterious property or relation accessible (if at all) only to introspection; we must examine evidence external to the knower. Thus reliabilism makes justification a worldly matter, compatible with the naturalist approach; in the jargon, reliabilism is an obvious way to *naturalise* justification.[20]

Naturalism also motivates reliabilism, not only in the obvious sense that naturalism promotes the general approach of which reliabilism is an instance (naturalism entails externalism about justification, of which reliabilism is the best-motivated and best-developed kind), but also in the sense that the naturalist analysis of science finds evidence in favour of reliabilism. As Boyd observes, it seems as though justification in science really is a matter of being generated by reliable methods. An appropriate relationship with certain methods of inference and testing, namely those involved in the cycle of dialectical improvement (see 2.2.2), is all there is

to the justification of scientific theories. What else might scientists appeal to? It's because methodology does the justificatory work that science has developed such a sophisticated mechanism for its improvement. If we think that findings about science furnish morals about knowledge-gaining per se (which many philosophers do, naturalist or not), then the evidence of scientific practice speaks in favour of reliabilism as the correct account of justification (and even if it's not the correct account of justification simpliciter, it might be the correct account of justification in science).

In science, the evidence of reliability is demonstrable success or progress; for example, we saw in the discussion of Kuhn that puzzle-solving success is evidence that the Kuhnian puzzle-solving method – on my account, IBE – is reliable. The evidence that concerns Boyd is the *instrumental* reliability of scientific methods, where 'instrumental reliability' is defined in terms of the tendency to promote acceptance of empirically adequate theories. Scientific methods are instrumentally reliable, and this is evidence that they track the truth. Boyd's claim is not just that his argument only makes sense from the naturalist-reliabilist perspective; it's that if one buys into the naturalist-reliabilist package, then scientific realism follows. Realism about the theories that determine the reliability of scientific methods is a consequence of acting scientifically and seeking to explain the evidence of instrumental reliability in scientific methodology – evidence admitted by all sides of the realism debate. The facts of instrumental reliability support realism and realism alone.

As this suggests, under naturalism philosophical theories are empirical hypotheses, testable and potentially false. Naturalist philosophers' foremost concern is that their theories accommodate the evidence, in our case the evidence of scientific practice. This evidence is not questioned, as it might be by traditional philosophers, who may allow their a priori analysis of the concept 'science' to colour their assessment of what science does and should be doing. Instead, the naturalist's aim is primarily descriptive: to arrive at theories that accommodate and explain the evidence of real science. Sceptical questions are ignored; after all, the evidence tells against sceptical hypotheses, and sceptical doubts about that evidence are relics of failed philosophy. Normative issues are postponed; naturalists usually claim that good theoretical explanations of science will prescribe an instrumental rationality, but we need to find them first.[21] Such explanations are the naturalist goal; we need to understand how science really goes about its business. Thus naturalist philosophy of

science is guided by the claim that no descriptively inadequate theory can hope to be the correct account of science.

For Boyd, the evidence suggests that science develops along Kuhnian lines, yet that same evidence also reveals that scientific methodology is instrumentally reliable, and increasingly so. Non-naturalist philosophy might perceive a tension here and see its task as reconciling Kuhnian discontinuities with methodological progress (we encountered this alleged tension while developing Kuhnian IBE). But Boyd takes the evidence at face value: given that science is like this, we need to explain why it exhibits these two features and why they are in fact compatible. As we know, Boyd and Psillos think that realism, specifically about the background theories upon which methodology depends, is the only way to understand this science. Boyd's claims about the dependence of methodology on background, and the 'radical contingency' of the successful method that results, are grounded in naturalism; they are findings of a scientific investigation of science. Thus realism is defended as a scientific hypothesis. This is the basis of the Boyd–Psillos NMA, to which we return below.[22] First we shall look more closely at the complementary relation between Boyd's realist take on paradigm-based science and my account of Kuhnian IBE.

2.2.2. Boydian realism and Kuhnian IBE

As outlined schematically above, Boyd argues that, contra Kuhn,

> "[the realist] conception of the enterprise of science provides the only scientifically plausible explanation for the instrumental reliability of the scientific method. In particular, I argue that the reliability of theory-dependent judgments of projectability and degrees of confirmation can only be satisfactorily explained on the assumption that the theoretical claims embodied in the background theories which determine those judgements are relevantly approximately true, and that scientific methodology acts dialectically so as to produce in the long run an increasingly accurate theoretical picture of the world" (Boyd 1984: 59).

This "dialectical" improvement of scientific theory is an interesting consequence of Boyd's argument, especially with respect to the picture of scientific IBE developed in previous chapters. Boyd outlines it thus (using experimental measurement as an example): "the approximate truth of current theories explains why our existing measurement procedures are (approximately) reliable. That reliability, in turn, helps to explain why our experimental or observational investigations are successful in uncovering new theoretical knowledge, which, in turn, may produce improvements in

measurement techniques, etc" (Boyd [1990] 1996: 222, italics removed). This makes possible a development of scientific knowledge that is "cumulative by successive (but not necessarily convergent) approximations to the truth" (ibid., italics removed). My account of IBE had exactly the same message. My claims were that exemplar-based standards of loveliness are truth-tropic just in case exemplars are successful at puzzle-solving, and that standards of loveliness grow more truth-tropic to the extent that later exemplars are better at puzzle-solving than earlier ones. On Boyd's picture, this is just what happens in science, through the dialectical improvement of theory and methodology. The approximate truth of prevailing exemplars explains why IBE, governed by loveliness, is a reliable puzzle-solving method. That reliability, in turn, explains why IBE contributes to the solution of further problems. This success, in turn, improves (or in our terminology, refines) the standard of loveliness governing inference, enabling it to solve still more problems.

Boyd's realist explanation of the instrumental reliability of scientific methodology gives an account (indeed, he claims, the *only* account) of how such feedback between methods and results facilitates progress in Kuhnian science:

> "Scientific method provides a paradigm-dependent paradigm-modification strategy: a strategy for modifying or amending our existing theories and methods in the light of further research that is such that methodological principles at any given time will themselves depend upon the theoretical picture provided by the currently accepted theories. If the body of accepted theories is itself relevantly sufficiently approximately true, then this methodology operates to produce a subsequent dialectical improvement both in our knowledge of the world and in our methodology itself" (ibid.: 223).

What my account of IBE adds to Boyd's picture is *an explanation of how this process works for a specific inferential method* – importantly, the method Boyd himself thinks scientists use. Scientific methods are engaged in a process of paradigm-dependent dialectical improvement because IBE is governed by ever-improving exemplars of loveliness. IBE facilitates puzzle-solving progress through paradigms, which in turn furnishes scientists with standards of loveliness progressively better suited to solving puzzles. This contribution that IBE makes to the increasing reliability of scientific methods can only be explained if loveliness is truth-tropic, i.e. if the exemplars that define loveliness approach the truth.[23] Just as Boyd says, the need to explain the instrumental reliability of scientific (in this case inferential) methods compels us to see the background theories of science as approximately true.

Thus my account of IBE and Boyd's defence of realism are mutually supportive. My account provides a specific mechanism, based chiefly on exemplars, through which we can understand how scientists conceive of loveliness and put it to work in inference. This mechanism fully admits – indeed, encourages – the kind of feedback between inferential method and problem-solving success that Boyd calls 'dialectical improvement'. My account of IBE is thus an articulation of a particular aspect of his explanationist defence of realism. Boyd's defence then supports my account of IBE by telling a story about the structure of science into which my account fits naturally. For Boyd, science is both Kuhnian and truth-tropic, with theory-dependent and self-improving methods. This is just the view of science that I support, but where my defence of it was minimal, Boyd's is sophisticated and compelling. One notable advantage of this happy marriage is that it reconciles Lipton's account of IBE and Boyd's view of science. To the extent that he discusses it at all, Lipton at least equivocal about whether loveliness can be identified with the theory-dependent plurality of standards Boyd and I advocate. At various points however, he reveals a sympathy with the general thrust of Boyd's approach (e.g. 2004: 157), so it's satisfying to account for loveliness in a way that harmonises the two projects.

One final point of mutual appreciation. We already know that for Boyd, "the methods for identifying inductively appropriate empirical generalizations… [are] profoundly theory dependent" (Boyd 1984: 77). But what we haven't yet seen is that Boyd realises this includes (what we've been calling) standards of loveliness. For example, he notes that "theory dependence of methods and the consequent dialectical interaction of theory and method are entirely general features of all aspects of scientific methodology [such as] standards… for assessing the quality and methodological import of explanations [and] principles governing theory choice" (Boyd [1990] 1996: 222). This is just the conception of loveliness I maintain is correct. Part of the purpose of my account is to show how loveliness can absorb the changes recommended by new puzzle-solutions *that it has itself provided*, and to show how, far from being a hindrance, this approach contributes to scientific progress. This view, anticipated by Boyd, strikes me as the only way in which IBE can hope to play a part in an increasingly reliable scientific methodology.

2.2.3. Radical contingency and revolutionary inference

Boyd draws out two important consequences of his view, both of which reinforce the notion of scientific IBE as guided by exemplars of loveliness. The first consequence is that the reliability of scientific inference is entirely contingent on the approximate truth of the background theories upon which inference depends. I argue that scientific inference is IBE and that scientists infer on grounds of explanatory loveliness, but just what loveliness amounts to is determined by prevailing exemplars. The result is that the basic justificatory work is being done by those background theories, not by the rule of IBE itself. As Boyd says, rules of scientific inference such as IBE "are not reducible to some more basic rules whose reliability as a guide to the truth is independent of the truth of background theories" (ibid.: 227). We cannot read off from a mere description of the rule of IBE that it's reliably truth-tropic (recall Lipton's remarks about Hume's problem: see chapter 1, section 4; chapter 3, section 1.2). Were the relevant background theories – exemplars of loveliness – not approximately true, IBE would not be a reliable method of inference. Boyd calls this the "radical contingency" of scientific methods: "since it is a contingent empirical matter which background theories are approximately true, the justifiability of scientific principles of inference rests ultimately on a contingent matter of empirical fact... [T]here are no a priori justifiable rules of non-deductive inference, and it is an a posteriori question about any such inference whether or not it is justifiable" (ibid.).

The second consequence Boyd stresses is an extension of radical contingency. If scientific inference depends for its reliability on truthlike background theories, then there's no story – specifically, no inferential story – to be told about how that reliability came to be. Boyd's point is that successful methodology *begins* with the first successful paradigm in a particular field; there are no reliable inferential principles logically or temporally prior to such a development. About this, he is emphatic:

> "The emergence of successful modern scientific methodology as we know it depended upon the logically, epistemically, and historically contingent emergence of a relevantly approximately true theoretical tradition. It is not possible to understand the initial emergence of such a tradition as the consequence of some more abstractly conceived scientific or rational methodology that itself is theory-independent. There is no such methodology." (ibid.).

But these comments don't just dovetail with my account of IBE; they provide a response to the problem of revolutionary inference (see chapter 3, section 6). How can scientists infer new exemplars when their grounds for inference are provided by

the very exemplars they seek to replace? Boyd's remarks tell us two things about the problem. The first thing we learn is that a response is not going to consist in providing an account of inference markedly different from that of normal science. There's no such thing as a theory-independent method of inference, so however the inference of new exemplars works, it must be dependent in some way on the existing theoretical commitments of the science concerned (recall that the dialectical relationship between methods and results allows IBE to facilitate progress by self-correcting). Following Boyd on this issue mitigates against a trivialisation of IBE: it rules out the idea of a special account of revolutionary inference that makes IBE exclusive to the mundane inferences of normal science and inapplicable to the exciting inferences of revolutions.

The second thing Boyd reveals about the problem is that it's general to all Boydian realists, whatever their preferred account of scientific inference.[24] Anyone who endorses the Kuhnian view of science, and the view that inference is theory-dependent, has a problem: to explain how inference licenses those radical changes to background theory that revolutions precipitate. However the Boydian realist characterises inference, and whatever account of revolutions that generates, the challenge is to show how inference can work when what justifies it are the very theoretical commitments (at least some of which) it sets out to replace. In our terms: where an exemplar defines a standard of loveliness different from that of its predecessor (recall that not all revolutions institute standards of loveliness discontinuous with earlier ones), how does the antecedent standard justify the inference of the subsequent one?

Boyd tells us that a response will be a matter of empirical investigation. The theory-dependence of scientific methods means that empirical investigation is the only way to find out what justifies any inference, revolutionary ones included. Recall Boyd's approach. Evidence from the history of science shows us three things: there are occasional revolutions that considerably change the theoretical background of science; scientific methodology depends heavily on that background; and through the dialectical relationship, methodology improves over time. This last point is crucial since it means we can ask the question: what are the features of scientific inference that generate methodological progress? This question can and should be answered by *empirical investigation of the relevant inferences.* Specifically, we need to look at which parts of background theory were active in justifying revolutionary conclusions.

It is an advantage of uniting Boyd, Lipton and Kuhn that we now have a programme for this investigation. We need to examine how IBE operates during those revolutions that brought about a discontinuity in standards of loveliness.[25] It's a matter of fact that science sometimes infers new and unusual exemplars; what we need to do is work out exactly which loveliness-making features of their predecessors determined the inference. Whatever they were, we know that they helped to ensure methodological improvement; this is already established by the instrumental reliability of IBE under the new exemplars. Discovering the features of loveliness that were at work need not deliver overarching morals about good inference, but it will corroborate the view that IBE is reliable. Returning to the problem of revolutionary inference, we can now say: agreed, it's certainly curious that different standards of loveliness should collaborate in this way. But given that it happens (we have independent grounds for thinking IBE is a correct account of inference) and it contributes to progress (increasing instrumental reliability), the proper is response is not to challenge IBE, but to figure out empirically how it works in such circumstances.[26]

2.2.4. The Boyd–Psillos NMA

With the Boydian explanationist approach clarified and endorsed, we may look again at the Boyd–Psillos version of the NMA. Boyd and Psillos hold that the inferential method of science is IBE (sometimes preferring the term 'abduction'), so that's the method they use in their defence of realism:

> "Neither the empiricist nor the constructivist can explain the most striking feature of the recent history of science, that is, the instrumental reliability of its methods. Only scientific realism provides the resources for explaining this crucial historical phenomenon. It is for this reason that realism is to be preferred to rival accounts of scientific knowledge" (Boyd 1984: 79).

The best, indeed the only, explanation for the reliability of scientific methodology is that the background theories upon which it depends are approximately true; on these grounds it should be accepted. Psillos calls this "a kind of *meta*-abduction", saying that successful instances of IBE in science "provide the basis (and the initial *rationale*) for this more general abductive argument" (Psillos 1999: 79). Yet the NMA is "not just a generalisation over scientists' abductive inferences"; rather, it aims "to defend the thesis that Inference to the Best Explanation… is reliable" (ibid.). The success of IBE

in science shows that it's reasonable to accept the theories involved as relevantly approximately true, and the NMA argues that science *can indeed* deliver approximately true theories, i.e. that IBE *is* reliable. This argument itself proceeds via IBE. In Psillos' words, the "NMA asserts that the best explanation of why scientific methodology has the contingent feature of yielding correct predictions is that the theories which are implicated in this methodology are relevantly approximately true" (ibid.). Just as scientists accept their theories as approximately true because they best explain the evidence, philosophers should accept realism for precisely the same reason.

Psillos urges two qualifications to the NMA. First, he stresses that part-and-parcel of any sensible realism is an acknowledgement of theoretical failures in the history of science. Psillos frequently emphasises that any defence of realism should take the pessimistic meta-induction (see section 1) seriously. The best way to stop the anti-realist mounting an argument from historic failures is to accommodate those failures within the realist position. This is not self-defeating: scientific methodology is not vitiated by the fact that it sometimes misleads, just as "the fact that I have occasionally failed to find my lost keys does not entail that a thorough search of the places where they could have been left is not a reliable method for finding lost keys" (ibid.: 80). Amid the failures, science has achieved many successes, and it is these upon which realists should concentrate. It could have been that scientific methodology scored no successes at all, "so to ask how it is possible at all that scientific theories yield correct predictions, especially novel ones, and to offer explanations of this contingent feature of scientific methodology are essential for understanding science" (ibid.).[27] The realist must face up to failure but draw attention to success, and argue that there is enough of the latter to arouse curiosity.[28]

The second of Psillos' qualifications is to localise the NMA's scope. He restricts its conclusion to asserting the approximate truth of certain parts of successful background theories, rather than theories as a whole. There may be an explanatory connection between empirical success and a theory's making some correct claims about the unobservable world, but "it is far too optimistic... to claim that *everything* the theory asserts about the world is thereby vindicated" (ibid.: 80). The NMA is made more convincing by being less optimistic – it should assert that success is best explained by successful theories' having "*truth-like theoretical constituents* (i.e. truth-like descriptions of causal mechanisms, entities and laws)" (ibid.: 80-81). Psillos explains further:

> "The theoretical constituents whose truth-likeness can best explain empirical successes are precisely those which are essentially and ineliminably involved in the generation of the predictions and the design of the methodology which brought those predictions about. From the fact that not every theoretical constituent of a successful theory does and should get credit from the successes of the theory it certainly does *not* follow that none do (or should) get some credit. If, on top of that, it is shown that, far from being abandoned, the theoretical constituents of past theories which did essentially contribute to their successes were retained in subsequent theories of the same domain, then the realist case is as strong as it can be" (ibid.: 81).

Realism is not diluted by Psillos' qualifications. He merely notes that science doesn't always get it right first time, and when it does get it right, success is explained only by those parts of a theory which definitely brought it about. It makes sense to modify the NMA to represent better the kind of success science achieves. Boyd realised this. He claimed that "it is no part of my thesis that [the development of theoretical traditions embodying approximate truths] was progressive in all particular instances, or uniform with respect to different disciplines, sub-disciplines, or even problem areas within sub-disciplines" (Boyd 1984: 76). Boyd also claimed that the NMA should conclude that inferential methods are reliable "only because, and to the extent that, the relevant background theories are *relevantly* approximately true" (Boyd [1990] 1996: 227, my italics). But Psillos is right to stress these qualifications, especially given the way they dovetail with his closely-related argument against the pessimistic meta-induction (1999: chapter 5). This argument is not discussed here but should be noted as complementary to his defence of the NMA.

3. Summary

The Boyd–Psillos NMA represents the strongest positive argument available for scientific realism. The main purpose of this chapter has been to describe what it is. Section 1 began by introducing the realism/anti-realism debate and sketching constructive empiricism as the most important modern form of anti-realism. We then noted the important point that constructive empiricists (and indeed anti-realists in general) cannot make use of IBE; because of what it means to infer an explanation, IBE is an inherently realist method. Section 2 introduced the NMA and argued that Maxwell's version of it is an IBE, and thus a fully-fledged precedent of the modern version. That version has been most thoroughly articulated by Boyd and Psillos. We have discussed its naturalist-reliabilist context and made some essential qualifications, and also seen how my views on IBE, developed in the previous chapters, offer

support to, and gain support from, the Boyd–Psillos approach. In the next chapter we look at the defence of the NMA against the objections that it's circular, and that it fails to establish that realism is the best explanation of the success of science.

Endnotes

[1] The debate in this chapter and the next will take constructive empiricists to be representative of modern anti-realists, but other versions of anti-realism, most notably the various forms of instrumentalism, are available. These will also enter the discussion at various points, mainly in virtue of the fact they are mentioned by other authors I discuss. I hope any confusion caused is benign.
[2] These terms are used interchangeably here.
[3] I offer no rigorous definition of truthlikeness, but Psillos' 'intuitive' definition seems about right (1999: chapter 11). Not that this isn't controversial: Bueno (2001) offers one criticism; other definitions have generated other criticisms.
[4] Bueno (2001) criticises Psillos, claiming he does not adequately define maturity (cf. Psillos 1999: 110). Psillos (2001) responds satisfactorily.
[5] See chapter 1, section 3.3; chapter 3, section 1.2.1.
[6] See chapter 3, section 6.2.
[7] Psillos (1999: 75-76) has further arguments for why realism should be judged to have an objectively higher prior probability than instrumentalism.
[8] Psillos (1999) didn't have access to this chapter of Lipton, though Bird (1998) had already noted the value of explanatory virtues to establishing Bayesian priors.
[9] It could be argued that Maxwell and Psillos provide evidence in favour of Lipton's position. They both present explanation as paramount in fixing the Bayesian values.
[10] Maxwell claims that "'hypothetico-inferential' is a better term than 'hypothetico-deductive', for sometimes inferences from hypotheses to evidence may be nondeductive, e.g., statistical" (ibid.: fn. 5).
[11] Maxwell goes on to identify explanation with the deductive-nomological model, but this is unimportant; also unimportant are various more detailed claims he makes about the nature of confirmation.
[12] This is not quite Hempel's view.
[13] Boyd's defence of IBE is different from mine and closer to Lipton's. It tries to establish that, given the successful use of IBE in science, IBE is reliably truth-conducive. As we shall see though, Boyd's defence and mine are complementary.
[14] Boyd's treatment of realism is sophisticated and complex. It deserves a more thorough analysis than it can be given here.
[15] I owe this example to Joe Morrison's paper 'Just How Controversial is Evidential Holism?' (forthcoming in *Synthese*).
[16] The inverted commas are not meant to imply the triviality of puzzle-solving success on Kuhn's construal.
[17] Clearly, Boyd thinks that unlike confirmation, observation is not heavily theory-laden.
[18] This is not the full extent of Boyd's attack on the constructivist explanation of success, which we return to briefly in the next chapter. See also Boyd ([1990] 1996: 231-234, 252-253).
[19] It's worth noting again Lipton's innovative attempt to use reliabilism, specifically in a form incorporating a Nozick-style tracking condition, to solve the raven paradox (Lipton 2007a). If reliabilism can help to account for the distinctions between confirming, disconfirming, and irrelevant evidence, this is a major reason to favour it.
[20] It remains a concern with reliabilism that it's hard to define 'belief-forming method' and individuate them 'in the world'.
[21] There is still some dispute over the extent to which naturalism can answer normative questions.
[22] Arguably, a hardcore naturalist would never think to offer the NMA, since it challenges the results of science. As it happens, the naturalist NMA endorses those results, but it could have been otherwise. This isn't far from Boyd's attitude; he talks about explaining the success of science, but never talks about the NMA as a formal IBE about IBEs. Instead, he thinks that realism and the reliability of IBE simply 'fall out' of the naturalist attitude.
[23] Boyd doesn't define instrumental reliability in terms of puzzle-solving but in this context the blurring of his and my definitions of this term is unproblematic.

[24] This assumes that it is possible to be a Boydian realist without endorsing my account of IBE as the inferential method of science!

[25] This problem has already been diminished by my response to Hungerford's objection (II): chapter 3, section 5.

[26] I acknowledge a slight hypocrisy here: I criticised Lipton for similarly-motivated responses to Hungerford's and Voltaire's objections.

[27] Psillos suggests that this might proceed in a piecemeal fashion.

[28] Here there's evidence of Psillos' view that realists should focus on successful novel prediction when proposing the NMA, as it stands in greater need of explanation than successful prediction simpliciter (ibid.: 105-108; see also Lipton 2004: chapter 10). Boyd doesn't single out novel prediction for special attention, but this has no special significance for the present discussion.

Chapter 5

The NMA defended

1. Introduction: the circularity and poor explanation objections

The last chapter described the NMA (henceforth, 'NMA' will be taken to refer to the Boyd–Psillos version of the argument). This chapter will defend it, against two principal objections: that it is circular (section 2), and that it does not establish that realism offers the best explanation of the success of science (section 3). The 'poor explanation' objection is simple enough. It consists in the anti-realist, in our case the instrumentalist, claiming that their position can offer an explanation of scientific success as least as good as the realist's. If she can force the realist to agree, the realist can no longer infer his explanation as the best. The 'circularity' objection is slightly more complex, but the basic claim is obvious: the NMA tries to defend IBE but is itself an instance of IBE. That the first-order IBEs it defends are 'scientific' and the second-order NMA is 'philosophical' is not a real distinction. There is no difference between the methods; indeed, the naturalistic defence of IBE depends on their being the same. Thus proponents of the NMA must assume that IBE is a legitimate form of inference in order to show that IBE is a legitimate form of inference.

Section 2 will conclude that Psillos' defence against circularity is the right way to proceed, but that we must be fully aware of what the NMA does and doesn't show. Lipton's critique of the NMA helps bring out the fact that the NMA is really a formal statement of the naturalist-reliabilist attitude towards science, rather than a full-blown argument for it. The NMA is still important, but should not be used to persuade anti-realists. Section 3 will find that the poor explanation objection never really gets going, and that its value is to remind the realist of the need to justify the inference to realism on grounds of explanatory virtue. This returns us to the issues of loveliness discussed in chapters 2 and 3, and I close by briefly noting some interesting features of the naturalistic use of loveliness.

2. The circularity objection

Arthur Fine (various publications, e.g. 1984) offers the definitive version of the circularity objection. Fine calls the NMA viciously circular, since it employs "the very type of argument whose cogency is the question under discussion" (quoted in Psillos 1999: 81). The NMA argues that the best (only?) explanation of the success of scientific methods is that the background theories that determine those methods are approximately true. Since those theories were arrived at via IBE, IBE is justified. Scientific realism falls out: we should be realists about the relevant background theories, otherwise their contribution to success is inexplicable (having defended IBE, we should be fairly confident about more recent, 'foreground' theories too). Thus the NMA is itself an IBE; it requires its audience to take explanatory power as a reason for belief. Crucially, it requires this prior to its articulation; in order to justify IBE, it must assume IBE is justified. Thus the NMA is circular. In Fine's words, "no support accrues to realism by showing that realism is a good hypothesis for explaining scientific practice. If we are open-minded about realism to begin with, then such a demonstration (even if successful) merely begs the question that we have left open ('need we take good explanatory hypotheses as true?')" (Fine 1984: 86). Fine concludes that the NMA preaches to the converted: it only convinces those who are already disposed to use IBE, i.e. realists (and even then it doesn't really *convince* them of anything).

To get a flavour of how a response to Fine gets going, it's worth noting that philosophical arguments can have two distinct purposes:

1. They can aim to persuade an open-minded audience of a particular view, or persuade an audience that holds one view of the correctness of an opposing view.

2. They can aim to provide an audience with justification for some view which, consciously or unconsciously, they already hold.

Both aims are worthwhile. Fine establishes that the NMA cannot achieve 1, but not that it can't achieve 2. Realising this is essential to understanding any good response to the circularity objection.

Also essential to bear in mind as we discuss circularity is the philosophical position that generates the NMA: the naturalist-reliabilist package. On this view, there's no possibility of an inferential method receiving philosophical justification prior to its use. Justification can come only from an a posteriori investigation of successful inferences. That investigation must itself use the kind of method whose justification it sets out to check; to get any empirical investigation going, we *must* take some justification for granted. Thus to the naturalist, circularity is not an obstacle but simply a non-eliminable feature of the successful justification of inferential methods.[1] But rather than simply bang the table, let's look more closely at how Psillos uses the naturalist-reliabilist package to deal with the circularity objection.

2.1. Psillos on rule-circularity, reliabilism and the justification of induction

Psillos notes a distinction, introduced by Richard Braithwaite (1953), between 'premise-circular' and 'rule-circular' arguments. Premise-circular arguments include some proposition *p* in their premises, or include a premise presupposing *p*, while purporting to offer an argument for *p*. Rule-circular arguments are instances of the rule vindicated by their conclusion. In rule-circular arguments, the conclusion is not a premise; rather, such arguments apply some inference rule *r* to premises *p1–pn* to give a conclusion *q*, but *q* asserts that *r* is reliable, hence the circularity. *Viciously* circular arguments purport to offer reasons for accepting a proposition, one of those reasons being the proposition itself. Thus Psillos follows Braithwaite in identifying vicious circularity with premise-circularity. Rule-circular arguments are not viciously circular: "the conclusion of a rule-circular argument is *not* one of the premises. Nor is the argument such that one of the *reasons* offered for the truth of the conclusion is the conclusion itself" (Psillos 1999: 82).

The conclusion of the NMA – that the background theories of science are approximately true – is not one of its premises; nor is it even assumed. Thus the NMA is not premise-circular. But clearly it is rule-circular: "the truth of the conclusion of the NMA is (part of) a sufficient condition for accepting that IBE is reliable" (ibid.: 83). Psillos brings this out by noting that there's no a priori guarantee that the conclusion of the NMA will be the approximate truth of relevant scientific theories, which there would be if the argument were premise-circular. If the conclusion of the NMA is true, it is true on the basis that it happens to provide the

best explanation of the evidence. Psillos points out that anti-realists implicitly concede this contingency when they argue that their explanations of scientific success are better than the realist's (see section 3).

Of course, rule-circular arguments are still circular, and hence potentially suspicious. Worse, viciousness still threatens. Opponents may argue that assuming the reliability of IBE requires prior acceptance of the conclusion of the NMA. Their claim is that realists must establish the conclusion in order to justify the use of the rule, by first accepting the reliability of the rule in order to establish the conclusion. But as we know, under reliabilism, this is not necessary: we do not need to know that our inferential methods are reliable in order to be justified in using them. Nothing about the reliability of a method need be assumed by its instances; nor does its reliability need to be defended in order for its conclusions to be compelling. Thus the conclusion of the NMA is justified as long as IBE is in fact reliable; the NMA is rule-circular but not viciously circular. Looked at from the other end, the NMA is only as viciously circular as the first-order scientific applications of IBE; they are not viciously circular (their conclusions aren't among their premises), so neither is the NMA.

Psillos is emphatic about the importance to the NMA of externalist views such as reliabilism:

> "the issue of whether rule-circular arguments are vicious turns on the theory of justification one adopts. Realists should have to be externalists if they take NMA seriously. And their critics will have to argue for internalism, if the charge of vicious circularity is to go through. Given an externalist perspective, NMA does not have to assume anything about the reliability of IBE" (ibid.: 85).

Psillos is aware that externalism remains controversial, but notes that those who agree with his analysis are "in good company" (ibid.: 83). To be sure, if it turned out all forms of externalism were false, the NMA would be in trouble, but absent such an outcome, whatever circularity the argument displays can be shown not to undermine it formally.[2] Thus Psillos has shown that the NMA's circularity is not worrying.

Fine's objection gives new expression to an old philosophical chestnut: Hume's problem. Hume showed that any justification for induction must derive ultimately from experience and would thus, sooner or later, have to employ induction, assuming what it sets out to defend. Fine's objection makes an analogous claim: if the NMA defends the reliability of IBE, then it lacks rational force, since it must first assume IBE, which as yet is undefended.[3] Viewed in this way, rule-circularity seems

unacceptable, and rightly so; if it didn't, Hume's problem wouldn't be a problem at all, a result that 250 years of philosophy tell against.

We've already seen one kind of response to Humean worries about IBE (chapter 3, section 1.2). Lipton's response to Voltaire's objection consisted mainly in the claim that the objection reduces to Hume's problem, which affects all accounts of inductive inference, so IBE is no worse off than its competitors. Psillos pursues a similar line with the NMA. If justifications for inferential rules such as induction and IBE are necessarily rule-circular, as Hume and Fine argue, then either we offer a rule-circular justification or no justification at all. Psillos notes that the situation is the same for all rules of inference, including deductive ones, no matter whether we're internalist or externalist about justification: "if the rule-circularity of a defence is taken to be an outright vice, then we should simply have to forgo any attempt to explain or defend any of our *basic* inferential practices" (ibid.: 86). If rule-circularity is inevitable, then the NMA's rule-circularity *is* a problem, but only to the extent that Hume's problem is a problem *and no further*.[4] The NMA is simply an example of the best kind of defence we can offer for any rule of inference, deductive or inductive.

Psillos illustrates the deductive situation with modus ponens: *p*; if *p* then *q*; therefore *q*. Deductive inferences preserve truth; if we put true premises in, we will get a true conclusion out (in the strongest possible sense of 'will').[5] This seems 'obvious' from the schema just given; but *proving* that modus ponens is truth-preserving is more difficult. Such a proof would have to follow some rule of inference, and this rule had better be truth-preserving too, otherwise the proof won't prove anything. But what rule can we use to prove that *that* rule is truth-preserving? A further rule, whose truth-preservation would have to be proven using a further rule, and so on. Either we face a regress, or more realistically, we acknowledge that the justification of modus ponens must ultimately make use of modus ponens. (More technically, we can only prove that modus ponens is truth-preserving 'in a language' if the language we use to talk about that language also has modus ponens as a rule. Either that or it must feature other deductive rules, for which the same problem will ultimately arise.) It cannot be proven that deductive rules preserve truth, and thus they cannot be formally justified; intuitively, we would have to justify all such rules before we could justify any of them: "no justification of *modus ponens* is possible which does not rest on some presuppositions" (ibid.: 87), such presuppositions being the

meanings of the logical operators and the truth-tables that demonstrate how they respond to the truth-values of their inputs.

Psillos notes that any such justification for a deductive rule is not justification in the formal sense. We don't supply independent grounds for trusting the rules; rather, they are justified 'from within' the deductive system. They are reinforced to anyone who already knows what the rules mean and how to use them by a process of "*explanation* and *defence*". Psillos characterises this approach as follows: "by reflecting on *modus ponens* (and other deductive rules we use), we aim to systematise it, to explain to ourselves the ways in which we should use it, and to show that, *given* the meaning of the logical connectives and the truth-tables, it delivers its goods – it is truth-preserving" (ibid.).

Returning to induction, Psillos describes how Carnap came to a similar conclusion about Hume's problem.[6] Reconstructing Carnap's argument, Psillos says "reasoners are either 'inductively blind' – where 'inductively blind' refers to reasoners who make no inductive inferences and who are not disposed to make any – or they are not. If the reasoners are inductively blind, then we cannot possibly show them when an argument is inductively valid and when it is not" (ibid.: 88). Because defences of induction are necessarily inductive, recognising good inductive arguments requires an 'inductive disposition', which cannot be brought about by rational argument. This disposition is simply the disposition to learn from experience; for those who have it, rule-circular arguments to the effect that induction is reliable are not problematic, and the fact that there are no non-rule-circular arguments to persuade those who don't is irrelevant. Our rule-circular attempts to persuade the inductively blind are thus "indispensable and harmless" (ibid.); indispensable because no other arguments are available, and harmless because nothing can persuade them anyway. When we attempt to persuade those *with* the inductive disposition, inductive justifications of induction are similarly indispensable (no non-inductive argument is available) and harmless (such arguments provide self-clarification). Carnap realises this situation has a deductive analogue, similar to the modus ponens example. Psillos describes the position Carnap adopts:

> "In one sense, no inferential rule carries an absolute rational compulsion, unless it rests on a framework of intuitions and dispositions which take for granted the presuppositions of this rule (truth preservation in the case of deductive reasoning, learning from experience in the case of inductive reasoning, searching for explanations in the case of abductive reasoning). When we attempt to vindicate or defend certain rules of inference... this is not because we want either to justify them without any assumptions, or to prove that they are rationally compelling for any

> sentient being. It is because we want to evaluate our existing inferential practices... Such evaluations cannot be made from a neutral epistemological standpoint" (ibid.: 88-89).

Carnap's view is that the justification of all inferential practices is a matter of seeing to what extent they are reliable and explaining why they are reliable to that extent. Instead of using inferential rules uncritically we thus find out something about them and how we use them. Justification is *not* a matter of trying to persuade the sceptic. So for Carnap, while Hume's problem isn't to be dismissed, it's also nothing special. If all justifications are rule-circular, the moral is not that we should ignore Hume's problem, but that we should learn to construe our responses to it in the appropriate way.

As with inductive justifications of induction, so with the NMA as a justification of IBE. Those with the 'abductive disposition' should not find it any more problematic than any other attempt to defend an inferential rule. Further, Psillos claims the class of the abductively disposed is highly populated, extending far beyond those who would defend IBE with the NMA. Significantly, within it are those opponents of realism who would use IBE but disagree with the conclusion of the NMA. As Psillos remarks, that there are such critics "follows from that fact that at least some critics of the realist NMA try to show that there are better potential explanations of the success of science than the realist one. If sound, the NMA can have rational force for all of them" (ibid.: 89). Psillos suggests that the development of the realism debate shows that more people find the NMA compelling than would care to admit it.

2.2. Lipton and Psillos on the case for IBE

Although an advocate of IBE and a scientific realist, Lipton disagrees with Psillos about the NMA. He thinks that the circularity objection and the poor explanation objection (see below) are enough to show that the NMA is "almost entirely without probative force" (Lipton 2004: 206). However, he admits that the guiding intuition remains: a theory's long-term and varied explanatory and predictive success is, in the absence of reasonable competitors, a reason to infer its approximate truth. Lipton saves this intuition by finding arguments for realism that do not depend on a second-order IBE about IBEs, but rather on structural features of the first-order scientific instances of IBE (cf. ibid.: 200-206). These arguments are not discussed

here, but in a nutshell, Lipton takes for granted that scientific inference is IBE, but instead of arguing from scientific success to the reliability of IBE, he argues that only realism can account for certain important features of the way IBE is used in science.[7] Thus he avoids the circularity and poor explanation objections (though he readily admits his arguments are far from decisive against the anti-realist), but Lipton's main motivation is that the NMA "appears to introduce no new evidence for the truth of successful theories and so no new evidence for the reliability of inference to the best explanation as a route to true theories" (Lipton 2001: 350). For Lipton, the relationship between the *scientific* evidence, *scientific* IBEs and the success of the resulting *scientific* theories is all the case for realism needs: "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 explanation of the phenomena that the evidence describes" (Lipton 2004: 206). To Lipton, the NMA is a superfluous argument, and a bad one at that.

Lipton agrees that there are legitimate inductive justifications of inductive practices, but thinks that legitimate ones find further evidence for the reliability of those practices. He gives the example of checking the reliability of barometers. Lipton shares Psillos' externalism, so this consists in checking the track record of barometers at predicting storms. If we find a correlation between predictions (certain barometer readings) and storms, we may infer that barometers are reliable. But Lipton claims this inference from past to future (or general) reliability is only legitimate since the process of checking gave us new evidence for the claim, viz. the observed correlations. Correspondingly, the justification the NMA purports to offer for IBE is illegitimate, for the reason that it provides no new evidence of IBE's reliability. The successes of a scientific theory are evidence for its truth, "but when the philosopher comes on the scene *she* does not gather further evidence of the track record of particular theories. Rather, *she* simply says that when theories are successful they are probably approximately true" (Lipton 2001: 350). Lipton argues that the evidence for the connection between IBE and truth is exhausted by the scientific evidence that supports scientific theories, even when the NMA focuses on background theory. Scientists may not notice the support that background theories get when predictions are fulfilled, but the support is still part of the scientific evidence; it doesn't count as new evidence just because the NMA draws attention to it.

In response, Psillos asks us to consider instrumentally successful theories *T1,... Tn*. Although the fact that they were delivered by IBE offers no new evidence that they are probably true (above and beyond the first-order scientific evidence) this fact might provide new evidence for the probable truth of a new theory *Tn+1*, also delivered by IBE. "The successes of *T1,... Tn*, and the fact that they were arrived at by IBE, supports, via NMA, the view that IBE is reliable, and this works in addition to the first-order evidence for *Tn+1* to make *Tn+1* more credible" (Psillos 2001: 367). The fact of being generated by a reliable method adds support to a theory, "for its truth will also be supported (indirectly) by all the (first-order) evidence that has led scientists to accept the method as reliable" (ibid.).

Psillos' example suggests this: we formulate theories *T1,... Tn*, they enjoy success, we run the NMA, and the conclusion – that IBE is reliable – then enhances the case for *Tn+1*. But as Psillos says, this is new evidence for the approximate truth of *Tn+1*; it is *not* new evidence for the reliability of IBE, which is what's needed to counter Lipton's argument. Beyond the scientific case that it best explains certain phenomena, there's no evidence for the approximate truth of *Tn+1* other than the fact that it was inferred using IBE. This by itself cannot be new evidence that IBE is reliable. The mere fact that a reliable method *is used* to infer theories isn't evidence in favour of that method's reliability. The claim cannot be: IBE is reliable, so whenever we use it to infer theories, those theories necessarily provide new evidence for IBE's reliability. Surely, there would only be new evidence on the assumption that *Tn+1* goes on to enjoy success. But if the realist then runs the NMA a second time to try and take advantage of that success, *Tn+1* is in the same position as *T1,... Tn*: the fact of being delivered by a reliable method is part of the scientific case and the second iteration of the NMA would offer no new evidence in favour of IBE. Consider Lipton's barometer. Whilst it's true that his latest barometer reading is more likely to be correct because barometers are reliable, the mere fact that it's a (reliable) barometer that generates the reading is *not* new evidence to support the reliability of barometers. It only counts as new evidence that barometers are reliable if the reading turns out to be correct, but in that case it becomes part of the evidence for the next inductive assessment of barometers. Generalisations about known instances of successful barometers plus new barometer readings (of uncertain accuracy) do not add up to *new* evidence that barometers are reliable.

Part of what makes Lipton's argument strong is that as soon as the NMA is run the first time, the conclusion – that IBE is reliable – becomes part of the scientific evidence. But the very fact that, in Psillos' example, the NMA has to have been run once already, prior to any new evidence in favour of IBE that *Tn+1* might provide, shows that the NMA has been run in the absence of new evidence of the reliability of IBE, i.e. (by Lipton's lights) illegitimately. Again, consider Lipton's barometer. In that example, evidence of the reliability of barometers is gathered and that supports the subsequent inference to the claim that barometers are reliable. In Psillos' example, there is no gathering of evidence in favour of IBE *beyond the scientific evidence*, prior to running the NMA for the first time. *Tn+1* can only support the claim that IBE is reliable on the assumption that the NMA has already illegitimately established IBE's reliability *and Tn+1* goes on to achieve success (that can only count as part of the scientific case for IBE on the next iteration of the NMA). Thus Psillos' response to Lipton fails.

Lipton shows that even though Psillos establishes that the NMA isn't compromised by circularity, neither is it much of an argument, giving no new reason to think scientific theories are approximately true. The next section argues that the NMA does have a legitimate function, albeit one slightly different from that envisaged by Psillos.

2.3. The real status of the NMA

Psillos expresses his stance on the NMA as follows:

> "That the world is such that – as a contingent matter of fact – IBE tends to yield (approximately) true theories is a new general claim about the world which is not entailed by the (scientists') first-order IBEs… [T]he actual track-record of successful applications of IBE does offer genuine evidence for the reliability of IBE. In particular, successful novel predictions issued by first-order theories arrived at by IBE do lend extra credence to the claim that IBE is reliable" (ibid.).

Psillos' claim is that the NMA establishes something about IBE over and above what can be gathered from successful scientific inferences, viz. that "there must be a feature of the world that answers to IBE's reliability" (ibid.). All evidence of the connection between IBE and truth is scientific, but the NMA takes a feature of those inferences – their success – and concludes (since it's the best explanation) that the method used is reliable. But as Lipton notes, "this insistence may not be on new evidence. It may rather be on the point that the conclusion is different in the two

cases, the philosophers' inferred from the scientists'" (Lipton 2001: 351). Lipton summarises the situation thus: "scientists tell us that particular theories are true, and philosophers use this result to go on to show that the scientists' methods are reliable" (ibid.). This means the NMA not only uses the form of inference the realist seeks to justify, but also uses as a premise the conclusion of the relevant scientific inference, namely that the theories thus inferred are (approximately) true. Lipton calls this version of the NMA "unexceptionable", but claims that it only serves to emphasise "that the realist here is in no way introducing new evidence or testing the scientist's methods" (ibid.). In this form, he claims, the NMA is simply the argument that IBE has worked well in the past; the best explanation of this is that it works well in general; therefore IBE works well in general.

2.3.1. The NMA as a statement of naturalism

If the NMA is at bottom a generalisation about the reliability of IBE, structured so as to appeal only to realists, then, however "unexceptionable", it is hardly potent. Yet the discussion of Boyd and Psillos in chapter 4 certainly made the NMA seem like more than "the drawing of a general moral from the prior commitment to the truth of specific theories" (ibid.). So why do Boyd and Psillos think there's more to it, and where has this extra content gone? Further, can the NMA be restored to its position at the forefront of the realist attack? My answer to the second question is nothing short of a resounding "perhaps"! It should become clear why once I've presented a more informative answer to the first.

Lipton is right that the NMA is not much of an *argument* for realism. Rather, it is a formal expression of a certain attitude – the naturalist-reliabilist attitude – towards scientific evidence and inference. (From now on I will speak of the 'naturalist attitude', since epistemological naturalism entails externalism about justification, of which the most plausible form is reliabilism.) In effect, the NMA says: here's what naturalists think about science, IBE and justification. To adopt naturalism just is to think that IBE is justified, if at all, by its successful scientific use. Taking scientific success for granted, we carry out an investigation of science that itself involves IBE, but since all a posteriori investigation is necessarily rule-circular, such Humean worries are no obstacle. If science uses IBE, infers the approximate truth of its theories, and so on, then naturalists are realists. Non-naturalists may not find this

convincing, but the NMA still has merit. It allows naturalist realists to describe their position to other realists and to non-realists, such that they may learn its details and understand why the position is internally justified, i.e. justified to those with all the relevant presuppositions (this is analogous to Carnap's point about inductive justification).

Note that the NMA does not reduce to Lipton-style arguments that appeal directly to the structure of scientific IBE. It expresses something more: a general outlook on scientific inference; a diagnosis of when science works well and in what respects; a research programme for investigating scientific inference and justification. In short, what it gives the realist over and above a claim about first-order IBEs is Boyd's position, which is perhaps why Boyd (1984, [1990] 1996) barely mentions the NMA qua independent, central argument for realism. The NMA just is Boyd's position.

As Lipton observed, the NMA does take the scientists' claim that their best theories are approximately true as a premise. But we now see that its justification is independent of its contribution to the general conclusion about the reliability of IBE. It is justified by the naturalist's views on the continuity between philosophy and science. Although sympathetic to naturalism, Lipton's language can drive a wedge between scientific and philosophical activity. Remarks such as "the conclusion is different in the two cases, the philosophers' inferred from the scientists'" represent Lipton's rhetorical need to differentiate two arguments. This is respectable enough, but it also imposes an artificial distinction between 'scientific' and 'philosophical' reasons for thinking that IBE is reliable. Naturalists see both arguments as parts of a single scientific-philosophical case for the reliability of IBE. Scientists assess their methods philosophically and philosophers offer realism as a scientific hypothesis (I continue to use the terms 'scientist', 'philosopher' and their cognates for brevity and convenience, recognising that the two sets of terms do not represent a sharp distinction).

It's certainly not implausible to think that scientists proclaim the approximate truth of their theories because, perhaps implicitly, they are persuaded by NMA-style considerations. Perhaps sometimes philosophers do nothing but generalise over all the successful instances of IBE and conclude that it's a reliable method. But it's not the case that naturalism would see scientists doing all the philosophy and philosophers only making boring generalisations about exciting science. Rather, the

naturalist thinks that what Lipton calls 'the scientific case' for the reliability of IBE *is* the philosophical case, and vice versa. It doesn't matter who runs the NMA, the scientist or the philosopher. What matters is that there's a strong case presented by scientist-philosophers, or philosopher-scientists (where these hyphenated terms co-refer). I suggest that in practice, scientists do some basic, probably implicit, philosophy, which philosophers then articulate as a full-blown defence of realism. Philosophers take the evidence of scientific practice and explain it back to scientists; thus both parties better understand their position, and are better able to defend it. This *doesn't* mean scientists are prevented from criticising their own methods (or indeed philosophical methods), or that philosophy necessarily stands in judgement of science. It just means that while scientists are busy getting on with their work, philosophers have the resources to reflect on what they're doing.

It seems Lipton would agree here, but as naturalists in the Boydian tradition, we are not driven to abandon the NMA in favour of Lipton's 'structural' arguments based on first-order IBEs. We've seen that the NMA isn't a powerful argument, but it still has value for reasons already rehearsed: it clarifies and explains the naturalist attitude, justifies it to those who hold it, and allows others to understand why those who hold it do so. Thus the NMA is still an argument for realism, albeit one with no probative force for non-naturalists. Hopefully this discussion has shown that what matters is not who the argument belongs to or which discipline generates the key inferences, but what the argument is such that scientists and philosophers may use it fruitfully.

2.3.2. The NMA from within and without

Naturalism has a few final things to say about the NMA and circularity. Under naturalism, science is something to be investigated, not questioned. To this end, it promotes the employment of scientific methods; these are designed to investigate worldly phenomena, and are successful at it too. As noted, this implies externalism about justification: if we're to put scientific methods to use, we cannot question whether they are 'really' useful, seeking independent a priori reasons to regard their success as 'genuine'. We must take their prima facie success as a reason to trust them.[8] This outlook means that, providing science continues to produce successful IBE-derived theories, the NMA may be iterated. The conclusion of the NMA – that IBE

is reliable – would be reinforced on each iteration. Thus realism would be reinforced too. The NMA may not be able to persuade others round to the naturalist-realist way of thinking, but the process of iteration makes the internal justification of that position progressively stronger. That science has developed a methodology with which it achieves success is radically contingent, as we know from Boyd. But under the kind of naturalism he espouses and which we've been discussing, this is not something we need question. What matters is that we develop a philosophy of science that explains why success happens. The NMA and the philosophical position it expresses are the basis of that explanation.

Thus we may agree with Lipton that the NMA isn't an argument in any thick sense but disagree that the NMA should therefore be abandoned. Other arguments, such as Lipton's own, might help the case for realism, but the NMA's value resides in its formalisation of a certain philosophical viewpoint, namely that viewpoint from which realism may most fruitfully be defended. Naturalists are comfortable with the idea that there's not much to add to their case for IBE over and above certain generalisations about scientific inference. But as we saw in our discussion of Boyd, such generalisations can be quite sophisticated, and on the iterative picture just described, they can grow increasingly potent. Psillos argues that the NMA, being rule-circular, won't convince inductive sceptics that IBE is reliable. It turns out that the NMA is even less persuasive: all it can do is articulate a certain philosophical position and defend it to those who hold it. But this audience needn't be small; if the arguments of Boyd and Psillos are sound, it should include all scientific realists.

Nevertheless, it may seem a high price to pay to see the NMA retained. Two closing points should ease the anxiety. First, consider anti-realism. There are various arguments for various forms of anti-realism; indeed some of them may set about trying to explain, or explain away, the success of science. The realist may not be convinced by those arguments, and this may well be because she doesn't share the presuppositions upon which they are based. Yet she would not reject those arguments; on the contrary, they would enable her to understand such opposing positions and their internal justification. This is just the value that the NMA has for the realist.[9] Secondly, we should insist that the NMA be appraised on its own terms. It never was a decisive argument for realism; many attempts have been made to salvage it from its evident circularity, and all have failed. Boyd's and Psillos' work on this matter has shown us that the NMA is at its best when set against a background of

naturalism, but the real moral of their discussion is that the NMA is a *consequence of* that view rather than an *argument for* it. As long as the arguments for naturalism itself are convincing, the NMA arrives as part of the philosophical package, with all the benefits of self-clarification, explanation and (limited) justification already noted.

These are concessions only if we retain unachievable ambitions for the NMA. Of course, all parties in the realism debate should continue the search for arguments that might persuade their opponents – it would be an end to philosophy if we were content merely with arguments that reinforce our prejudices – but in the meantime the NMA is better than no argument at all.[10]

3. The poor explanation objection

With the circularity objection dealt with, the second key objection to the NMA is that realism does not provide the best explanation of the success of science. Once again, Fine is at the forefront of the attack, but this time Psillos argues convincingly against him, the guiding claim being that the instrumentalist is generally on very shaky ground when the realism debate is phrased in terms of explanation.

Fine argues that the NMA fails because the instrumentalist has a better explanation of the success of science than the realist, which should be inferred instead. Psillos notes that in order to make the argument, Fine would need an instrumentalist version of IBE. He cannot use the standard version, as this would be to accept that explanatory virtues are epistemic and thus that IBE generates true conclusions, which is exactly what he wants to deny. As we've seen, there is no non-realist version of IBE. This is the first and most serious problem with the poor explanation objection. But Fine gets his argument going again by trying to show that truth is redundant in the realist's explanation of success. Given that the instrumentalist's explanation would not contain such redundant parts, the realist cannot claim her explanation is better.

Fine's central point is that on the realist account, there must be some 'connection' between a theory's approximate truth and its practical successes. This connection, he claims, is the theory's pragmatic reliability; if a theory is true then it's also pragmatically reliable. Fine then claims that pragmatic reliability can do all the explanatory work that approximate truth does in the realist account of success. This explanation is superior to the realist's, claims Fine, on account of making no reference

to superfluous notions such as truth. Fine concludes that, by substituting pragmatic reliability for truth throughout the realist account, we turn the argument for realism into an argument for instrumentalism, based on the fact that it explains scientific success better than realism.

We should not regard Fine as arguing for instrumentalism on grounds of simplicity (somehow defined in terms of economy, for example) as an explanatory virtue. Rather, we should construe him as capitalising on the fairly uncontroversial claim that unnecessary components decrease a hypothesis' chances of empirical adequacy. He is arguing that instrumentalism is more likely to be empirically adequate than realism because it has no parts which do not contribute to explaining the data. Intuitively, this is justified by the Popperian idea that greater content in theories means greater likelihood of falsification. Any theory which makes unnecessary claims is more likely to be proven false; empirical adequacy is likelier if a theory can explain all data without such superfluous components. Hence, other things being equal, in a competition between two equally empirically adequate theories, the simpler of the pair may be judged more empirically adequate.[11]

Fine's argument is flawed, as Psillos shows. His first response concerns Fine's notion of a 'pragmatic intermediary', operating between a theory's practical success and its approximate truth. Psillos assumes that the likeliest candidates to play this role are "methods, auxiliary assumptions, approximations, idealisations, models and probably other things" (Psillos 1999: 92) derived from background theory. If this is the kind of pragmatic intermediary Fine has in mind, then as Psillos notes, it does no explaining at all. All it does is prompt a change in explanandum, from the success of some theory to the instrumental reliability of the background theory which helped bring it about. Why are *those* models, assumptions, approximations and so on better than any others which might have been used? Approximate truth must be brought in to explain the reliability of the pragmatic intermediary. If Fine did mean to refer to this kind of stuff, then approximate truth is not superfluous to explanation.

Psillos's second response is more general. To provide a better explanation of scientific success than the realist, the instrumentalist must replace the realist's approximate truth with some notion of instrumental reliability. Psillos argues that the explanatory value of any such notion is extremely doubtful: "Instrumental reliability is nothing but a summary statement of the fact that the theory performs successfully practical tasks. If we then try to explain the theory's empirical success by saying that

background theories are instrumentally reliable, we simply paraphrase what needs to be explained" (ibid.: 92-93). Psillos draws the useful analogy of trying to explain the success of hammers at nail-driving by saying that hammers are instrumentally reliable at nail-driving. Instrumentalism turns out to offer no explanation at all of the success of science, let alone an explanation better than the realist's. More generally, instrumentalism is based on a commitment to instrumental reliability and nothing more. Given that such reliability is just what needs explaining, instrumentalism ultimately lacks the resources to do anything except paraphrase the success it tries to explain.

But perhaps the realist has skewed the debate in his favour. The instrumentalist may insist we 'deflate' the search for explanation, and Psillos considers one way in which this might be done, which he calls the 'induction about abduction'. Providing the instrumentalist accepts rule-circular justifications of inductive methods, and has the usual inductive dispositions, she should recognise at least a prima facie need to explain scientific success. But then, as Psillos says, "instead of accepting the realist's explanation, [s]he identifies explanation with retrodiction and prediction" (ibid.: 94). She then offers the induction about abduction: since IBE has delivered empirically successful theories in the past, it will continue to do so in the future. By doing this, she is "equipped with inductive generalisations about the instrumental reliability of abductive scientific methodology on the basis of which one can predict or retrodict the instrumental reliability of scientific methodology in particular cases" (ibid.). There's no big explanation here, but the instrumentalist's point is that she doesn't need one. A general statement of a link between IBE and success is enough: IBE is good at generating empirically successful theories, thus success was/will be enjoyed because IBE was/will be used. The major advantage of this deflationary move is supposedly that, since the generalisations are inductive rather than abductive, they "do not commit one to the existence of unobservable entities, nor do they entail that abductive reasoning is a reliable guide to theoretical truth" (ibid.).[12]

But as Psillos argues, inductive generalisations about past successful theories are not free of theoretical commitments. Before the instrumentalist makes the induction about abduction, she must accept that science achieves success, and that when theories contribute to that success, the inductive generalisations they make about observables are thereby supported. This judgement about inductive support is theoretically loaded. As Psillos notes, "from myriad generalisations that involve

observables, scientists pick only some as genuinely empirically supported" (ibid.: 95). This choice depends on background theory, which determines how scientists differentiate natural kinds, which predicates they choose as projectible, and so forth. Thus the instrumentalist does not accept induction as theory-independent. Psillos follows this up by noting that, once again, the instrumentalist does not offer a genuine explanation of success; rather she paraphrases what needs to be explained and passes it off as an explanation. Via the induction about abduction, the instrumentalist obtains the generalisation that IBE reliably generates empirically successful theories. This is equivalent to the claim that IBE has generated empirically successful theories in the past, does so in the present and will continue to do so in the future. But to say that IBE was, is and will be instrumentally reliable in the past, present and future is just to say that science is empirically successful. Empirical success is just what needs to be explained; thus the instrumentalist's generalisation paraphrases rather than explains it.

With the induction about abduction, the instrumentalist shows that explanation cannot be deflated; inductive generalisations that enable prediction and retrodiction do not explain. For example, suppose I note numerous instances of big-selling records by boring singer-songwriters. I might make the generalisation that boring singer-songwriters record successful albums. This allows me to predict that the next album by a boring singer-songwriter will sell well, and retrodict that some such album from the past will have sold well too. The generalisation that boring singer-songwriters record successful albums does not, however, explain why such albums sell in large quantities, nor does it explain why any individual album achieves high sales. Generalising about successful methods does not provide any new information about why those methods deliver success (this is what we had to face up to when discussing circularity). We may conclude that the instrumentalist cannot provide an explanation of scientific success, much less one that could challenge the realist.

3.1. Is the realist explanation good enough?

Psillos shows that the poor explanation objection never really gets off the ground, despite Fine's best efforts. But Lipton (2001) offers a different version of the poor explanation objection, which although similarly handicapped, reminds the realist of a chink in the NMA's armour. Lipton's criticism is that in order to claim that

realism should be inferred because it best explains the success of science, the realist must justify his account of explanatory virtues and specify which of them the realist explanation has. This is not because other explanations threaten the realist's, but because the realist explanation must be good enough by itself to be inferred. It turns out that explanatory virtues do not need an independent justification, but the realist must still work to point out which of them his explanation has.

Lipton notes two alternative explanations of the success of science that Psillos does not consider. The first of these is the 'underdetermined alternative theories' explanation, based on the idea that science could have developed differently and yet been equally empirically successful (as Lipton realises, this is really a class of explanations, one for each different possible science). For each theory science has actually inferred, there are theories it could have inferred that would have scored the same successes. The approximate truth of the theories in each of these possible sciences would explain the success of actual science. Lipton's second alternative explanation is the 'Kuhnian-Kantian' explanation. This appeals to the idea that science is successful not because its theories are approximately true, but because they are true in the world as constituted by current scientific activity. The realist needs to establish that the approximate truth of actual scientific theories explains success better than their Kuhnian-Kantian truth, and better than the approximate truth of any underdetermined alternatives.

The realist view is that, to the extent that a hypothesis explains the evidence better than a rival, it is more worthy of belief. The realist accepts the realist explanation of scientific success largely on such grounds, but here the spectre of circularity returns to haunt him. Before the realist can argue for realism on explanatory grounds, he must first establish that explanatory virtues really are epistemic virtues. For Psillos the justification of the explanatory/epistemic virtue identity is part and parcel of the NMA – just another aspect of the fact that that argument must assume IBE. But Lipton's new worry is that even if Psillos' rivals – proponents of the underdetermined alternative theories and Kuhnian-Kantian explanations – overlook *that* circularity, still he has not established that realism offers the better explanation of scientific success.

Looked at one way, the appeal to explanatory virtues is equivalent to an assumption that scientific inference is truth-tropic, which means that "the foils to realism are being excluded without reason" (Lipton 2001: 353). Lipton is arguing that

the NMA gerrymanders the explanandum – scientific success – such that the realist explanans always comes out on top. If the realist doesn't assume that scientific inference is truth-tropic, then it's "unclear that the foil explanations are worse than the realist theory" (ibid.). Refusing to prejudge the explanandum would alleviate some of the NMA's difficulties, but would also allow rival explanations back in. If science doesn't in fact infer the approximate truth of its theories, then its success may be given a Kuhnian-Kantian spin, and if science doesn't in fact treat explanatory virtues as epistemic (i.e. use IBE), then there's no justification for supporting realism because it better explains success.

Once again, it seems the realist ought to concede that the NMA is an expression of naturalism about science: science does use IBE, and IBE is (inherently) truth-tropic. But even from inside the naturalist perspective, the realist still needs an account of why his explanation is, by his own lights, superior. At the very least he needs to show that it's good enough for him to infer it. "Even if we take it that the epistemic virtues are justified by the predictive success of the theories that have them, there remains the different question of whether the realist theory itself enjoys these virtues" (ibid.). Lipton considers two such virtues that the NMA relies upon: indispensability in prediction of the evidence, and issuing of novel predictions. Realism cannot claim superiority in either case. The success of science's best theories is entailed by their approximate truth, but it's also entailed by their Kuhnian-Kantian truth and the truth of any underdetermined alternatives. Hence approximate truth is not indispensable to the prediction of success. Lipton then asks, "particular scientific theories sometimes make novel predictions, but what novel prediction follows from the claim that scientific practices can attain the truth?" (ibid.). It seems the realist explanation lacks the virtue of novel prediction too. The alternatives to realism do not have these virtues either, so Lipton's claim is not that they explain success better than realism. Rather his claim is that the realist must establish (a) which virtues his explanation of success has that its competitors lack, and (b) that those virtues are sufficient for it to be inferred. Without this, the NMA has not established what it set out to show, even on its own terms.

Lipton concentrates on fairly uncontroversial epistemic virtues rather than those that have a distinctively explanatory flavour, but his criticism is general. It reveals that a pressing task for the realist is to give an account of explanatory virtues and show that the realist explanation of success has them, and in sufficient quantity to make it

inferable by anyone who accepts IBE. Note that this is not the same as arguing that explanatory virtues are epistemic; that would be part of an attempt to win round opponents of realism. One message of this discussion is that the realist should concentrate on a good internal defence of his position, the lack of which is more embarrassing than the much-discussed inability to convince non-realists. Boyd talks about approximate truth being the *only* explanation of scientific success, which dissuades him from talking in terms of virtues, comparatively or otherwise. He does mention epistemic and explanatory virtues at various points, but not in a way that establishes that realism is more virtuous than the sort of competitor Lipton considers (though Boyd does have something to say about the Kuhnian-Kantian explanation: see section 3.3). Psillos says similarly little about virtues, except insofar as he endorses remarks made by Maxwell in support of his version of the NMA.

I suggest Maxwell's work is a good starting point for this part of the realist project. He cited the virtues of simplicity, comprehensiveness and non-adhocness on behalf of realism, and hinted at the virtue of unification. The realist explanation of success does indeed unify many disparate phenomena – the various instances of success – citing a common cause: approximate truth. It's comprehensive in that it covers all instances of success, and simple in that it uses no mysterious concepts (providing we can make sense of approximate truth). Realism also explains success without making any ad hoc modifications in response to quirks of the explanandum (as we've just seen, the realist does construe the explanandum in a certain way before trying to explain it, but that's a separate issue).

Sadly, the underdetermined alternative theories and Kuhnian-Kantian explanations have the same virtues. Kuhnian-Kantian truth does all the work of truth, unifying precisely the same instances under a common cause and without ad hoc modifications. The realist might get some purchase on grounds of simplicity, since Kuhnian-Kantian truth is arguably more mysterious than approximate truth. But the idea of the world being to some extent constituted by one's conceptual framework isn't all that hard to understand, and the realist must admit that the question of relative simplicity is moot until the debate over approximate truth is settled. In the case of the underdetermined alternatives, what we might call 'truth in a close possible world' explains the success of the theories in that world just as virtuously as truth does with the theories of the actual world. Once again we find the realist's explanation on a par with its rivals. It will be hard for the realist to show that his

explanation is more virtuous than its competitors (and virtuous enough to be accepted), but the challenge must be engaged, otherwise realists won't even have a complete defence of their position available to themselves, let alone one that could convince opponents.

3.2. The Darwinian explanation and explanatory depth

Psillos' final argument against anti-realist explanations, though successful, cannot help. The argument uses van Fraassen's Darwinian explanation of success to expose a structural feature of realist explanations. This feature appears to account for their superiority, and the inability of instrumentalism or constructive empiricism to do anything but paraphrase the success they try to explain. But it also reveals another way in which the realist skews the debate in his favour, and in any case it fails to banish Lipton's alternatives.

Van Fraassen says the following:

> "I claim that the success of current scientific theories is no miracle. It is not even surprising to the scientific (Darwinist) mind. For any scientific theory is born into a life of fierce competition, a jungle red in tooth and claw. Only the successful theories survive – the ones which *in fact* latched on to actual regularities in nature" (1980: 40).

This explanation, acceptable to the constructive empiricist (because it mentions nothing unobservable), says that the success of science is due to current theories surviving a process analogous to natural selection. They prevail because they describe real features of the world, while their competitors died out because they did not. Psillos calls this explanation *phenotypic* as it provides a mechanism by which all theories with some quality (empirical success) are selected. It claims that science is successful because it is organised in such a way that it selects empirically successful theories. The realist explanation Psillos calls *genotypic* as it cites an underlying feature (approximate truth) in virtue of which theories achieve empirical success. Psillos illustrates the distinction with Lipton's example of the group of red-haired people (Lipton 2004: 194). Each member's having red hair is phenotypically explained by their belonging to a club that only admits people with red hair. But this does not explain why any of the individuals in the group have red hair. In order to explain this, a different, genotypic, story must be told which deals not just with a selection mechanism, but with common

(in this case genetic) traits in virtue of which each individual came to have the quality for which they were selected.

Van Fraassen's phenotypic explanation and the realist's genotypic explanation are compatible, but Psillos claims that "the realist's is arguably preferable, because it is deeper" (1999: 96). One result of the Darwinian explanation's lack of depth is, as Psillos notes, that it only supports the claim that current theories have not yet been refuted. It does not provide a warrant for their success in the future. Psillos realises that such a warrant could come from the phenotypic explanation plus an inductive inference about scientific practice, but argues that the genotypic explanation "has this warrant up its sleeve: if a theory is empirically successful because it is true, then it will keep on being empirically successful" (ibid.: 97). Effectively, this is Psillos arguing that the realist's explanation has the Maxwellian virtues of simplicity, comprehensiveness and non-adhocness, and to a higher degree than van Fraassen's explanation. The realist explanation explains science's ongoing success, not just its past successes, so is more comprehensive than the Darwinian alternative, and arguably also more unifying. It's simpler too, in that it cites a single cause of success (approximate truth) rather than a range of causes (the various success-inducing features of scientific practice). And the realist's explanation is less ad hoc, since it does not require the additional inductive inference over past science, having all the required explanatory resources 'in house'. On Psillos' account, the realist explanation has these virtues because it is 'deeper': it descends to the genotypic level, inaccessible to the constructive empiricist.

This is the point that came out in our discussion of Fine: the added depth offered by an appeal to truth looks to be a realist trump card. As long as the realism debate is cast in terms of explanation, the realist can always 'go one better' by, in effect, asking an extra "why?". The various anti-realist explanations of success all make roughly the same fundamental claim: science is successful because it is set up in such a way that it ends up using empirically successful theories. Their explanations must stop there, but the realist can continue to ask *why* science has retained only those theories with the disposition to be instrumentally reliable, or *why* only those theories that have survived the Darwinian battle achieve ongoing empirical success. The instrumentalist and constructive empiricist cannot respond, since they have nothing consistent with their position that can fill the explanatory role of truth. If deeper

explanations are always better, it looks as though the realist has won the explanatory debate.

This isn't a new victory for the realist; this is simply an account of why, when the debate is cast in terms of explanation, the realist has the kind of advantage he's been seen to have throughout this discussion. Further, the realist's ability to cast the debate in terms of explanation is just what's at issue with the NMA. Anti-realists may not be able to answer the extra "why?" questions the realist persists in asking, but neither do they think such questions are legitimate. Anti-realists eliminate theoretical truth from their accounts of science for the reason that to claim truth rather than (say) empirical adequacy for scientific theories means straying too far beyond what the evidence supports. Anti-realist explanations of success are still explanations, but they are explanations that deny that an underlying, genotypic cause of theoretical success is warranted or explanatorily useful. For the realist to insist that deeper explanations are better starts to look like another facet of the question-begging that continues to dog the NMA. Anti-realists may retort that their explanations already go as deep as explanations can.

In any case, Lipton's rival explanations of scientific success – the underdetermined alternative theories and Kuhnian-Kantian explanations – are both genotypic. They cite more than a selection mechanism for empirically successful theories. The truth of underdetermined possible theories and the Kuhnian-Kantian truth of actual theories go as deep as the realist explanation. Lipton spots the reason why: "the former accepts a realist understanding of truth, but asks why we should suppose that it is the theories that scientists' practices take them to that are the true ones, rather than others that would have enjoyed the same sort of success. The Kuhnian alternative does not propose different theories, but rather a different conception of truth" (Lipton 2001: 352). Truth (of theories not actually inferred) and Kuhnian truth (of theories actually inferred) work on the same level as truth in the realist's explanation of success. This is why Lipton's alternatives share the Maxwellian virtues. Extra depth might've given the realist explanation an advantage over van Fraassen's, but these rivals are equally deep, so it's no surprise they're equally virtuous.

Lipton's alternative explanations are not artificial. A variety of anti-realist views may offer the underdetermined alternative theories explanation as a challenge to the realist. It may not reflect a particular anti-realist commitment, but it represents one of

the motivations general to most anti-realists: what reason is there to think that science is *in fact* approaching truth? The Kuhnian-Kantian explanation may be offered in some form by all constructivists (though as we noted with Fine, proponents of neither explanation may infer it on explanatory grounds if they wish consistently to defend anti-realism). So what, if anything, can the realist do now?

3.3. Against Lipton's alternatives

Take the Kuhnian-Kantian explanation first. The plausibility of that explanation rests on the plausibility of the constructivist position that's behind it. A positive case can certainly be made for constructivism based on Kantian or Kuhnian arguments (the extent to which Kuhn himself supported it is debatable), but it's open to some well-known objections, some of which are raised by Boyd ([1990] 1996). He sketches two main lines of attack. The first concerns the way in which the constructivist must deal with the causal powers and relations of the objects that science studies. The constructivist thinks that prevailing theories, conventions, institutions and so on contribute to the causal powers of objects, and not just to their identification. In other words, she thinks that the causal properties of objects are partially constituted by the framework within which they are studied.[13] The constructivist view on causation is independently quite implausible. As Boyd says, the realist view of causation and social institutions as separate "has very deep roots in quite diverse features of our understanding both of causation and of social phenomena. Thus any constructivist package will be prima facie vulnerable at any point at which it incorporates a distinctly constructivist conception of… causal relations" (Boyd [1990] 1996: 252-253). Given that such a conception is central to the constructivist account of scientific success, the constructivist epistemology of science cannot rival the realist's.

Boyd's second line of attack is more standard. In its simplest form it is just that anomalies provide counterexamples to the constructivist view. The fact that, historically, there have been features of the world with which scientific paradigms and theories have (initially at least) failed to cope suggests that the world is not itself constituted by those paradigms and theories, but is independent of them. Scientific theories and scientific methodology are forced to change in response to anomalies, which means that "the incorporation of a doctrine of social construction of the

reliability of scientific method seems hardly to strengthen the constructivist philosophical package" (ibid.: 253). Thus we see that in isolation, the Kuhnian-Kantian explanation rivals the realist's, but twinned with the position that generates it – constructivism that takes the success of science seriously, as something to be explained – it is substantially weaker. Boyd does not rule out a constructivist fight-back, but until the constructivist tells a plausible story about causation and anomalies, she cannot account for the success of science as well as the realist.

The underdetermined alternative theories explanation is more tricky. The heart of the problem is that although the underdetermined alternatives would have scored the same successes, they are *incompatible* with the theories science actually inferred. Since any actual theory and any one of its underdetermined alternatives cannot both be true, we need a reason to think that success is explained by the approximate truth of actual theories rather than the approximate truth of any of the ones scientists could've chosen. Thus what we face is the problem of underdetermination: given several equally empirically successful but incompatible theories, how do we choose the one that's (approximately) true? The plausibility of the underdetermined alternative theories explanation comes down to how seriously we take the problem of underdetermination.

There are a variety of responses to underdetermination, some of which deny that it's a real problem. Assuming that we take it seriously, to overcome it we must deny that empirical success is the only guide to truth. According to this response, epistemic virtues can distinguish between equally empirically successful theories, picking out those that are approximately true. Psillos (1999: chapter 8) argues this, and his proposals are typical of the realist. But we've already noted that strictly, the legitimacy of appeals to virtue is not yet settled. To claim that science may appeal to virtues to overcome underdetermination requires that we first defend the method that licenses such an appeal, viz. IBE. This is what the NMA seeks to do, by showing that approximate truth best explains theoretical success. But this in turn requires an appeal to virtues. This is circularity again, but it's not new circularity: the NMA aims to establish that IBE is legitimate, which *just is* to establish that explanatory virtues can solve cases of underdetermination.

The reminder of circularity shouldn't be depressing, especially since it reveals how to get rid of the underdetermined alternative theories explanation. The realist doesn't need to establish that his explanation is better in a way that may convince

opponents; he needs rather to establish that IBE is legitimate. This would solve the problem of underdetermination, and the rival explanation would vanish along with it. The realist's argument *against* underdetermination and *for* realism is one and the same thing: the NMA. Thus Psillos is right to see the justification of the virtues as part and parcel of the NMA; the argument kills two birds with one stone. The NMA cannot compel non-realists to accept realism; neither can it compel non-realists to accept that underdetermination isn't a problem. But for those who already accept realism, the argument has merit on two counts, providing we see it as an articulation and internally acceptable defence of the naturalist view of science.

We may conclude that the Kuhnian-Kantian and underdetermined alternative theories explanations can both be undermined on independent grounds. The realist does have the best explanation of the success of science. Nevertheless, the realist must work to spell out the explanatory virtues, not so that the NMA can be rationally compelling – we've seen that that won't work – but so that we may get a clearer idea of what the explanatory virtues at work in science actually are and see that the realist explanation has them. Consistent with my account of IBE, I suggest that this investigation should be conducted naturalistically, and that it should proceed by identifying and examining exemplars of loveliness. But is scientific loveliness amenable to naturalistic investigation? That is, is philosophical IBE continuous with scientific IBE? I close this chapter by considering the relationship between IBE in philosophy and science.

4. Explanatory loveliness in science and philosophy

If philosophy is continuous with the natural sciences, when philosophers use IBE they must use it in the same way as scientists. But can they do this, given the account of scientific loveliness I've developed elsewhere? It seems philosophy must generate exemplars of loveliness that determine admissible inferences. Philosophy must foster a certain puzzle-solving context, or more likely, several such contexts corresponding roughly to the different branches of philosophy (or the significant problems they consider). Philosophers must treat certain philosophical theories as providing a standard of loveliness appropriate for each kind of problem, against which solutions to the puzzles that those theories define are measured. The philosopher may leave her concept of loveliness implicit; all that matters is that her

conclusion does in fact resemble the relevant exemplar. On this picture, IBE would be structurally identical across science and philosophy, even though the details of loveliness may differ, just as they may differ between other sciences.

Is this picture plausible? One thing in its favour is that we may discern phases in the history of philosophy that emphasised certain kinds of problem and solution. Perhaps the most obvious example is positivism, with its anti-metaphysical doctrines of meaning, truth and so forth. Positivism rejected as irrelevant or meaningless many problems considered legitimate in pre- and post-positivist philosophy. Other problems were emphasised and certain standards applied to their solution. As noted in chapter 1, although scientific inference was a central positivist concern, explanation could not play a role in its characterisation for the reason that explanatory commitments are non-verifiable. (A problem with this example is that, in rejecting explanation, positivism strictly provided its practitioners with no standard of explanatory loveliness. But the example is only intended to be suggestive, and I do not want to discuss here the extent to which positivists really managed to expunge explanation from philosophy.)

To take a less secure example, perhaps Kuhn's account of science now provides an exemplary theory in philosophy of science. There seems to be fairly widespread consensus that Kuhn's historical researches show that science is cyclical, even if there is no consensus that this tells us anything about the realism debate. The ability to fit within, or explain elements of, the Kuhnian structure is generally taken to be a philosophical virtue.[14] A third possible example comes from epistemology. Arguably, over the last thirty years, externalism, and more specifically reliabilism, has become something of a research programme. It has been applied successfully to numerous problems of justification, and although it still has its detractors, it seems generally to be regarded as getting something right. A modern insistence on a thoroughgoing internalism is certainly unusual. Alleged contemporary exemplars such as this are difficult to assess since only history reveals their true significance; further, it's difficult to pin down an externalist theory that might play the exemplary role (though Alvin Goldman, David Armstrong and Robert Nozick have provided candidates). But the thought that externalism/reliabilism offers a standard against which modern epistemological puzzle-solutions are assessed is certainly plausible.

In all the above cases it should be clear that the standards dictated by the alleged exemplars are standards of loveliness. Fit with background belief is obviously part of

the story, but mechanistic constraints are also in evidence – explanations that invoke the Kuhnian mechanism are lovely in philosophy of science. Standards of simplicity and unification could also be drawn out of these examples. Puzzle-solutions that conform to reliabilist standards are taken to be simpler because they do not invoke internalist baggage such as the requirement that we know that we know – such claims are taken to be superfluous. Reliabilist explanations also have the virtue of unifying the various kinds of knowledge: perceptual knowledge, inferential knowledge, testimonial knowledge and so on are all accounted for in the same way – via the functioning of a reliable method.

But perhaps here is a disanalogy between IBE in science and philosophy. I've just been articulating the loveliness-making features of philosophical exemplars and corresponding inferences. It's one of the jobs of epistemology and philosophy of science to do so. Haven't I just shown that a dispensable part of scientific inference is an essential part of philosophy? I have argued that it might be beneficial to the functioning of normal science if standards of loveliness remain unarticulated; the same cannot be said of philosophy. The scrutiny of loveliness is an important philosophical activity; if we give up on it, we give up on a good account of inductive justification. If the articulation of loveliness is eliminable from science but not from philosophy, then philosophy doesn't use IBE in the same way as science, and the naturalist's alleged continuity of methods is undermined. Ben-Menahem (1990: 325-326) argues for a similar disanalogy between IBE in 'philosophical' and 'non-philosophical' contexts, due to different understandings of the status of IBE and of explanatory standards. She claims that philosophical IBEs have general structural features (of explanations, for example) as content and take timeless standards as given in their form. Scientific IBEs (for example) have empirical matters as content and use explanatory standards that evolve with increasing empirical knowledge in their form.

On the naturalist view, this argument is a non-sequitur. It does not follow from the fact that loveliness is an important philosophical subject that philosophers use IBE differently from scientists. Note that on the above account, instances of philosophical IBE are structurally identical to instances of scientific IBE. Both scientists and philosophers refer to exemplary standards of loveliness when inferring. These standards may differ because loveliness is relative to puzzle-solving context, but the inferential procedure is analogous. Neither philosopher nor scientist need know about their standard of loveliness in order for their inferences to be justified. In

either case, whether or not specific IBEs are justified depends on whether or not the hypotheses inferred in fact resemble relevant exemplars in appropriate ways. Investigation of philosophical loveliness, where it takes place, is a matter of locating philosophical exemplars and examining them, along with examples of typical inferences made in the relevant branches of philosophy. The same goes for investigation of scientific loveliness, whether carried out by scientists or philosophers. The fact that philosophy investigates loveliness more often than science, and the fact that science might benefit from not doing so (at least not too often), does not mean that on a case-by-case basis IBE is used differently in philosophy and science. Thus the naturalist view of inference is not threatened by the fact that inference, and hence loveliness, is part of the subject matter of philosophy. The view that exemplars of loveliness govern IBE in philosophy just as they do in science is poorly understood, but is at least prima facie plausible.

4.1. Naturalism and philosophical consensus

However, part of Ben-Menahem's point was that in science, loveliness is informed by increasing empirical knowledge. Within a given science, there is a body of theory over which there is very broad consensus (there may be some rogue scientists who reject parts of the received view, but as Kuhn has explained, they are usually excluded from the scientific community). I have argued this consensus is justified because science approaches the truth; science accumulates not just beliefs about but knowledge of the world, so scientists are right to agree on it.[15] In fact, their agreement is explained by increasing truthlikeness: increased puzzle-solving ability is indicative of increased verisimilitude, and were puzzle-solving to stall or decrease, scientists would no longer agree about the relevant parts of theory. Thus scientists are right to allow their increasing knowledge to influence their standards of loveliness. Doing so enables increasingly truth-tropic puzzle-solving to occur within the relevant context.

Is this picture reflected in philosophy? Well, yes and no. There *is* increased philosophical knowledge, often prompted by the sciences. For example, we know a lot more about the workings of the mind thanks to cognitive science, neuroscience and psychology. Their findings influence philosophical theories of mind, ruling out some ideas as implausible, revealing tensions between others, setting out ways in

which still others must be modified and developed. But philosophy can generate its own knowledge. We know now that induction cannot be simple enumerative, that knowledge is not justified true belief (where justification is taken to be a relation between known propositions), and that there are problems with the analytic/synthetic distinction. But however philosophy derives its knowledge, what these examples suggest is that there is knowledge about, and hence consensus over, what, e.g., justification isn't or cannot be, but not over what justification *is*. Reliabilism is a highly popular view, but unresolved problems bring forth a welter of criticism. Perhaps reliabilists are right to favour their view because some form of reliabilism is true. But while opponents are still significant in number there is no consensus on the matter, and as long as proponents of reliabilism take the problems of their view seriously, even they cannot be said to know that (some variant of) reliabilism is the correct account of justification, at least not in the same way that scientists can be said to know that heliocentrism is the correct account of planetary motion.

Thus even if philosophy achieves a broad consensus over a particular positive theory, it seems it does not become the unquestioned background to philosophical work that exemplars provide in the sciences. Traditionally, acceptance of a philosophical view has not been the same as gaining empirical knowledge of the world. Hence although we may be able to identify philosophical exemplars of loveliness, it is not clear that they represent the latest, best state of knowledge as it's clear they do in the sciences. This is why they do not generate consensus within the relevant philosophical community. And if there is not consensus, exemplars do not determine standards of loveliness in philosophy as they do in the sciences. This casts doubt on their claim to be exemplars of loveliness in the first place, and undermines the naturalist claim of continuity.

To illustrate, consider a particular criterion of loveliness: fit with background. Within a puzzle-solving context, scientists may agree on which hypothesis fits best with their background theory, because they have achieved consensus over some particular positive theory (on my account, this is because the theory is approximately true, and scientists realise this – the consensus has the status of knowledge). On the other hand, a large number of philosophers – reliabilists, say – may agree that a certain hypothesis fits best with their background belief in the positive theory of reliabilism, but since it's only belief and not knowledge, they have internalist colleagues operating within the same context (who are not rebels or outcasts from the

community), who favour a different hypothesis that fits with *their* background belief in the opposing positive theory of internalism.[16] Which provides the exemplar of loveliness? It depends who you ask. It seems philosophical IBE is subject to its own version of Hungerford's objection (II) (see chapter 3, section 5): good inference is relative to different philosophical communities, where communities correspond not to puzzle-solving contexts but to theoretical commitments.

How can philosophical debates such as that between reliabilism and internalism be resolved and Hungerford's objection (II) answered? Achieving consensus in philosophy over any issue, such that it may then achieve consensus about what kinds of inference to make, is a central motivation behind naturalism.[17] The idea is roughly that by turning philosophy into an empirical science we may test our way out of difficulty, converging on philosophical theories that give the best empirical account of the phenomena. Something about scientific methodology, and the corresponding view of the world as material for investigation, enables it to solve problems and engender almost universal agreement about concrete achievements. The naturalist hope is that by adopting the same approach, philosophy might enjoy similar successes and that consensus would naturally follow. Thus naturalism offers a way to overcome philosophy's version of Hungerford's objection (II). By adopting scientific methods throughout, philosophy may develop standards of loveliness appropriate to its various puzzle-solving contexts, eliminating the frustrating relativity of loveliness to theoretical commitments. Given time, exemplary theories would provide increasingly accurate accounts of philosophical subjects and their empirical success would engender agreement about loveliness across the relevant communities. Thus I suggest that naturalism is itself the key to bringing about the kind of continuity naturalism seeks between philosophical and scientific IBE.

5. Summary

This chapter has considered the two principal objections to the NMA: that it's circular and that it doesn't offer the best explanation of the success of science. Section 2 endorsed Psillos' defence against circularity, which pointed out that the NMA is only rule-circular, a form of circularity inherent in all inductive justification and, according to reliabilism, unproblematic. Lipton noted that even so, the NMA provides no new evidence for the reliability of IBE, and this led us to argue that the

NMA should be seen as a statement of, and internal justification for, the naturalist perspective that generates it, given that no-one outside that position will find it convincing (not a huge concession, since naturalists are great in number). We then turned to the poor explanation objection, agreeing with Psillos that it cannot be made in any potent form. Again, a criticism from Lipton encouraged realists to think about the status of their position, in this case its support from explanatory virtue. Section 4 closed by taking the idea of explanatory virtue (loveliness) in science, as developed in chapter 3, and applying it to naturalistic philosophy. It suggested that philosophical IBE is structurally continuous with scientific IBE, but awareness of, and consensus about, philosophical exemplars of loveliness has yet to be reached.

Endnotes

---

[1] Boyd ([1990] 1996: 249-250) sketches his own response to the circularity objection. He concludes that if, in order to do justice to the realism/anti-realism debate, it is essential to see realists as arguing for their position from a certain perspective, then any prima facie circularity generated by arguing from that perspective cannot be question-begging.

[2] Douven (2001) criticises Psillos' use of externalism. He argues that externalism doesn't imply a positive assessment of IBE; for example, an externalist persuaded by van Fraassen's 'no privilege' argument against IBE would be at least agnostic about its reliability. Psillos (2001) responds effectively: van Fraassen needs to give the externalist an additional reason to think we are in fact bad at generating actual explanations. The realist would here engage van Fraassen in a debate parallel to the one over the NMA, in which he already has the upper hand: cf. Lipton (2004: 151-163).

[3] Psillos notes that the NMA needn't be construed in this way. He says that, on the externalist view, the NMA "does not make IBE reliable. Nor does it add anything to its reliability, if it happens to be reliable. It merely generates a new belief *about* the reliability of IBE which is justified just in case IBE is reliable" (ibid.: 86). Externalists do not strictly speaking need 'new' arguments for the reliability of modes of inference. Thus for Psillos, using the NMA as a defence of IBE in this way is "optional".

[4] I use the word 'only' here rhetorically. I do not mean to belittle Hume's problem.

[5] Worries from chapter 1, section 1 here notwithstanding.

[6] Psillos attributes Carnap's argument to a 1968 paper 'Inductive Intuition and Inductive Logic', found in I. Lakatos (ed.) *The Problem of Inductive Logic* (Amsterdam: North Holland Publishing Company).

[7] Lipton's views on realism are presented in his (2004: chapter 11).

[8] This doesn't imply complacency about scientific methods. Naturalists may find those methods wanting in some respects and attempt with scientists to correct them. However, the relationship between naturalism and normativity cannot be discussed here.

[9] This picture is rather like the view of incommensurability in Kuhnian science developed in the chapter 3, section 6: scientists from different paradigms may disagree with their counterparts' choices but still understand why they made them.

[10] Whether my position is closer to Lipton's or Psillos' is moot.

[11] McAllister (1996: chapter 7) discusses arguments for the empirical superiority of simpler theories, and provides his own attractive account of the empirical value of a high degree of simplicity.

[12] Van Fraassen's explanation of scientific success adopts a similar deflationary tactic: see section 3.2.

[13] The realist admits that such things have some influence on causal powers, but that influence is itself causal: "the realist does not deny that the adoption of theories, etc., and the having of projects or interests, are themselves causal phenomena, and thus contribute *causally* to the establishment of, for example, those causal factors that are explanatory in, for example, the history, philosophy, and sociology of science and that, in consequence, the adoption of a theory in such a discipline could contribute causally to the causal powers and relations that are the subject-matter of the theory itself" (Boyd [1990] 1996: 232).

[14] I hope so anyway.

[15] The accumulation of knowledge may not be linear.

[16] We could see internalism and externalism as defining their own puzzle-solving contexts – internalists and externalists certainly have different agenda – but they are ultimately in direct competition over the same territory. To they extent that they define different problems for themselves, they do so only in order to achieve the same goal, viz. an account of justification.

[17] Other attempts have been made to bring about a programme of agreement in philosophy, notably pragmatism and positivism.

Conclusion

This thesis has considered two different defences of IBE. Chapters 1, 2 and 3 uncovered the structure of IBE and argued that for numerous reasons it made sense to construe inductive inference as guided by explanatory loveliness. Chapters 4 and 5 argued for the same thing, but this time via the idea that IBE could be used to support its own realist presuppositions. Thus we may see the two parts of this thesis as mirroring one another: the first three chapters started with the theory of IBE and argued that it could match inductive practice; the last two chapters started with the (successful) practice of science and argued that it matched IBE.

The argument central to the latter approach, the NMA, itself used IBE – the reliability of the scientific method was urged as the best explanation of scientific success. Thus the NMA is circular. For Lipton, this showed that the realist shouldn't bother with the NMA and should argue instead for a simple des criptive match between scientific inference and IBE. But chapter 5 argued that the NMA does represent a useful application of IBE, albeit one without probative force for non-naturalists. Having endorsed Psillos' arguments for rule-circularity, reliabilism and an appropriate justification of inductive practice, I argued that the NMA defends scientific realism by explaining, to realists and their opponents, why the position finds such support from the success of science. Thus it gives a quasi-normative justification of IBE that Lipton-style descriptive arguments do not. Prior to any belief about actual scientific practice, we ought to think that science is truth-tropic because otherwise its success is mysterious. This returns us to Putnam's original thought, offered before concerns about IBE properly entered the frame: other philosophies of science just won't do the job; the evidence of success supports realism. Now the way is cleared for us to claim that scientific inference is IBE; and sure enough, there's ample evidence that it is (much of the evidence for Kuhn's view of science is evidence for Kuhnian IBE, as chapter 3 argued). Realists thus find themselves in a position of greater internal strength; opponents of realism may not find the NMA convincing, but there's no better way for them to understand why realists are so confident of their view.

So if the circularity objection forces realists to retreat, the poor explanation objection allows them to stand their ground. Chapter 5 also noted that the realist gains the upper hand over the anti-realist simply by putting the debate in terms of

explanation. Psillos' debate with Fine began by hardly mentioning IBE or loveliness, and concentrated on whether or not (given an intuitive idea of what it means to explain) the anti-realist can be said to offer an explanation of science's success. Psillos argued that she cannot, but the challenge then is to convince her that success stands in need of explanation. As Psillos again shows, whenever she agrees it does, she concedes that we may argue for a position on its explanatory merit, which is to concede the legitimacy of IBE. Lipton's version of the poor explanation objection was similarly hamstrung by an inability to use IBE (though in this case it belonged to his rhetorical anti-realist opponents). But his argument was useful in that it reminded the realist that, no matter whether or not anti-realists may explain the success of science, he owes us an account of why his explanation is good enough to be inferred. I suggested this account should return to Maxwell's version of the NMA, which I argued is just as much an IBE as the naturalistic, Boyd–Psillos version. One of the main reasons for this was that Maxwell considers the virtues of the realist explanation, which tend to be overlooked.

The opposite approach to defending IBE, that of chapters 1, 2 and 3, began by arguing that IBE can account for fundamental features of our inductive practice that enumerative inductive, hypothetico-deductive, and Bayesian accounts cannot handle. Then, mainly through the work of Thagard and Ben-Menahem, chapter 1 began to articulate the idea of explanatory virtue and emphasise its dependence on background knowledge. This became the central theme of chapter 2, which introduced Lipton's account of IBE with its core notion of a two-stage process guided by loveliness. We saw that many of Lipton's original comments on loveliness, although scarce, endorsed the idea of background-dependence, but failed to distinguish properly between two ways in which a potential explanation might be said to have the virtue of 'fit with background'. Accordingly, we noted that a hypothesis may both virtuously cohere with background beliefs, and meet the other criteria of loveliness (unification, mechanism, simplicity, etc) as determined by that background. The latter sense of 'fit with background' is what enables IBE to have an all-important context-sensitivity: what loveliness consists in depends on what our background beliefs tell us is appropriate in the inferential context. It's this feature of loveliness that enables IBE to have many of the advantages over rival accounts of induction reviewed in chapter 1. It should also put an end to the a priori philosophical quest to find some sort of conceptual connection between rigidly-defined explanatory virtues and truth.

Responding to Barnes' criticisms brought out the reasons why background-dependent loveliness doesn't trivialise IBE, but chapter 3 began with two more serious objections. Background-dependence served to exacerbate the problems raised by Hungerford's objection, that loveliness is too subjective to guide inference, and Voltaire's objection, that loveliness is not a guide to truth. In response, I argued for a new account of IBE according to which loveliness in the sciences is determined by Kuhnian exemplars. In Kuhnian IBE, exemplars of loveliness provide standards against which explanatory puzzle-solutions are assessed. Because exemplars are endorsed by the entire scientific community, loveliness is not subjective. Further, because on a reliabilist reading successive exemplars approach the truth, the standards of loveliness they provide are correspondingly truth-tropic. These responses generated a new version of Hungerford's objection, which argued that loveliness cannot be a guide to likeliness because it's relative to paradigms. I argued that in fact, loveliness is relative to puzzle-solving contexts, and given context-sensitivity, this is to be expected.

The final section of chapter 3 argued for Kuhnian IBE on independent grounds. Bird's interpretation of Kuhnian science helped here, as, unwittingly, did Kuhn himself, in particular by introducing the five trans-paradigmatic values for theory choice. Kuhn's effort to show that paradigm-shifts are rational plays into the hands of supporters of IBE, and McMullin showed how the Copernican revolution cannot be understood without the five values being construed as explanatory virtues. My account also borrowed support from McAllister's account of aesthetics in Kuhnian science. He argues that normal science is characterised by the acceptance of an aesthetic standard for theory choice – an aesthetic canon – and revolutionary science by its overthrow and replacement. The arguments and evidence McAllister uses to support his view are effectively reasons to favour Kuhnian IBE, especially once we realise that revolutions needn't feature the kind of 'aesthetic rupture' he talks about.

One result of twinning naturalism with Kuhnian IBE, a partnership that chapter 4 argued was mutually beneficial, is that, taking success for granted, we investigate those features of any scientific inference that we identify as bringing that success about. There is enough evidence in favour of Kuhnian IBE that we may legitimately say that loveliness is a guide to likeliness (in scientists' heads *and* out in the world). Thus any point where a contextual standard of loveliness has changed is a point where we need to ascertain how the earlier standard justified the inference to the later

one (or rather the exemplars that define it). The idea that at least some of that philosophical investigation may proceed via IBE introduced the question of whether IBE in philosophy resembles IBE in science. The naturalist desires that it should, since naturalism owes its very name to a thesis about continuity of method between philosophy and natural science. I closed chapter 5 by suggesting that philosophical IBE is indeed structurally continuous with scientific IBE, and that naturalism's desire for philosophy to imitate science could be self-fulfilling.

At the end of *Inference to the Best Explanation*, Peter Lipton, whose work has been the concern of much of this thesis, says of himself, "endorsing a philosophy I cannot believe does not interest me" (2004: 206). I share this sentiment, and hope that this thesis has presented adequately the reasons why I believe IBE is the best account of inductive inference in general and scientific inference in particular. More specifically, I hope to have argued that naturalist, realist supporters of Kuhnian IBE, of whom I am one, are in an enviably strong position. As ever, there is more work to be done, but I flatter myself that those who do not endorse the view will struggle to find a better one.

Bibliography

Achinstein, P. 'Can There Be A Model of Explanation?' (1981). In D.-H. Ruben (ed.) *Explanation* (Oxford: Oxford University Press, 1993).

Achinstein, P. 'Inference to the Best Explanation: Or, Who Won the Mill-Whewell Debate?', *Studies in History and Philosophy of Science* **23**, 2 (1992), 349-364.

Barnes, E. 'Inference to the Loveliest Explanation', *Synthese* **103** (1995), 251-277.

Ben-Menahem, Y. 'The Inference to the Best Explanation', *Erkenntnis* **33** (1990), 319-344.

Bernecker, S. and Dretske, F. 'Externalism and Internalism: Introduction'. In S. Bernecker and F. Dretske (eds.) *Knowledge: Readings in Contemporary Epistemology* (Oxford: Oxford University Press, 2000).

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

Bird, A. 'Scientific Revolutions and Inference to the Best Explanation', *Danish Yearbook of Philosophy* **34** (1999), 25-42.

Bird, A. *Thomas Kuhn* (Chesham: Acumen, 2000).

Bird, A. 'Inference to the Only Explanation', *Philosophy and Phenomenological Research* **74**, 2 (2007), 424-432.

Boyd, R. 'The Current Status of Scientific Realism'. In J. Leplin (ed.) *Scientific Realism* (Berkeley: University of California Press, 1984).

Boyd, R. 'Realism, Approximate Truth, and Philosophical Method' (1990). In D. Papineau (ed.) *The Philosophy of Science* (Oxford: Oxford University Press, 1996).

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

Bueno, O. 'Review Symposia: Quests of a Realist', *Metascience* **10**, 3 (2001), 360-366.

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

Day, T. and Kincaid, H. 'Putting Inference to the Best Explanation in its Place', *Synthese* **98** (1994), 271-295.

Douven, I. 'Review Symposia: Quests of a Realist', *Metascience* **10**, 3 (2001), 354-359.

Douven, I. 'Wouldn't It Be Lovely: Explanation and Scientific Realism', *Metascience* **14** (2005), 338-343.

Ennis, R. 'Enumerative Induction and Best Explanation', *The Journal of Philosophy* **65** (1968), 523-529.

Fine, A. 'The Natural Ontological Attitude'. In J. Leplin (ed.) *Scientific Realism* (Berkeley: University of California Press, 1984).

Friedman, M. 'Explanation and Scientific Understanding' (1974). In J. C. Pitt (ed.) *Theories of Explanation* (Oxford: Oxford University Press, 1988).

Fumerton, R. A. 'Induction and Reasoning to the Best Explanation', *Philosophy of Science* **47** (1980), 589-600.

Gettier, E. L. 'Is Justified True Belief Knowledge?', *Analysis* **23** (1963), 121-123.

Giere, R. N. 'Naturalism'. In W. H. Newton-Smith (ed.) *A Companion to the Philosophy of Science* (Oxford: Blackwell, 2000).

Goodman, N. 'The New Riddle of Induction' (1983). In S. Bernecker and F. Dretske (eds.) *Knowledge: Readings in Contemporary Epistemology* (Oxford: Oxford University Press, 2000).

Grimm, S. 'Explanatory Inquiry and the Need for Explanation', *British Journal for the Philosophy of Science* **59**, 3 (2008), 481-497.

Hacking, I. *Representing and Intervening* (Cambridge: Cambridge University Press, 1983).

Hanson, N. R. *Patterns of Discovery* (Cambridge: Cambridge University Press, 1972).

Harman, G. 'The Inference to the Best Explanation', *Philosophical Review* **74** (1965), 88-95.
Harman, G. 'Enumerative Induction as Inference to the Best Explanation', *The Journal of Philosophy* **65** (1968), 529-533.
Harman, G. *Thought* (Princeton: Princeton University Press, 1973).
Hempel, C. G. and Oppenheim, P. 'Studies in the Logic of Explanation', *Philosophy of Science* **15**, 2 (1948), 135-175.
Hempel, C. G. *Aspects of Scientific Explanation* (New York: Free Press, 1965).
Hitchcock, C. 'The Lovely and the Probable', *Philosophy and Phenomenological Research* **74**, 2 (2007), 433-440.
Hume, D. *An Enquiry Concerning Human Understanding* (1748) (T. L. Beauchamp, ed.) (Oxford: Oxford University Press, 1999).
Kitcher, P. 'Explanatory Unification' (1981). In J. C. Pitt (ed.) *Theories of Explanation* (Oxford: Oxford University Press, 1988).
Kitcher, P. 'Explanatory Unification and the Causal Structure of the World' (1989). In Y. Balashov and A. Rosenberg (eds.) *Philosophy of Science: Contemporary Readings* (London: Routledge, 2002).
Kuhn, T. S. 'Objectivity, Value Judgment, and Theory Choice'. In *The Essential Tension* (Chicago: University of Chicago Press, 1977).
Kuhn, T. S. *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press, 1996).
Ladyman, J., Douven, I., Horsten, L. and van Fraassen, B. 'A Defence of van Fraassen's Critique of Abductive Inference: Reply to Psillos', *The Philosophical Quarterly* **47**, 188 (1997), 305-321.
Ladyman, J. 'Wouldn't It Be Lovely: Explanation and Scientific Realism', *Metascience* **14** (2005), 331-338.
Laudan, L. 'A Confutation of Convergent Realism' (1981). In D. Papineau (ed.) *The Philosophy of Science* (Oxford: Oxford University Press, 1996).
Lipton, P. 'Inference to the Best Explanation'. In W. H. Newton-Smith (ed.) *A Companion to the Philosophy of Science* (Oxford: Blackwell, 2000).
Lipton, P. 'Review Symposia: Quests of a Realist', *Metascience* **10**, 3 (2001), 347-353.
Lipton, P. *Inference to the Best Explanation* (London: Routledge, 2004).
Lipton, P. 'What Good Is An Explanation?'. In J. Cornwell (ed.) *Explanations: Styles of Explanation in Science* (Oxford: Oxford University Press, 2004) (Lipton 2004a).
Lipton, P. 'Wouldn't It Be Lovely: Explanation and Scientific Realism; Author's Response', *Metascience* **14** (2005), 353-361.
Lipton, P. 'Replies', *Philosophy and Phenomenological Research* **74**, 2 (2007), 449-462.
Lipton, P. 'The Ravens Revisited'. In A. O'Hear (ed.) *Philosophy of Science* (Cambridge: Cambridge University Press, 2007) (Lipton 2007a).
Maxwell, G. 'The Ontological Status of Theoretical Entities'. In H. Feigl and G. Maxwell (eds.) *Minnesota Studies in the Philosophy of Science Volume III: Scientific Explanation, Space and Time* (Minneapolis: University of Minnesota Press, 1962).
Maxwell, G. 'Theories, Perception, and Structural Realism'. In R. G. Colodny (ed.) *The Nature and Function of Scientific Theories* (Pittsburgh: University of Pittsburgh Press, 1970).
McAllister, J. W. *Beauty and Revolution in Science* (Ithaca: Cornell University Press, 1996).
McMullin, E. 'Rationality and Paradigm Change' (1993). In M. Curd and J. A. Cover (eds.) *Philosophy of Science: The Central Issues* (London: Norton, 1998).
Misak, C. 'Peirce'. In W. H. Newton-Smith (ed.) *A Companion to the Philosophy of Science* (Oxford: Blackwell, 2000).
Newton-Smith, W. H. *The Rationality of Science* (London: Routledge, 1981).

Newton-Smith, W. H. 'Explanation'. In W. H. Newton-Smith (ed.) *A Companion to the Philosophy of Science* (Oxford: Blackwell, 2000).
Norton, J. 'A Material Theory of Induction', *Philosophy of Science* **70** (2003), 647-670.
Okasha, S. 'Van Fraassen's Critique of Inference to the Best Explanation', *Studies in History and Philosophy of Science* **31**, 4 (2000), 691-710.
Peirce, C. S. *Collected Papers of Charles Sanders Peirce: Volume 5* (C. Hartshorne and P. Weiss, eds.) (Bristol: Thoemmes Press, 1998).
Psillos, S. 'On van Fraassen's Critique of Abductive Reasoning', *The Philosophical Quarterly* **46**, 182 (1996), 31-47.
Psillos, S. *Scientific Realism: How Science Tracks Truth* (London: Routledge, 1999).
Psillos, S. 'Review Symposia: Author's Response', *Metascience* **10**, 3 (2001), 366-371.
Psillos, S. 'The Fine Structure of Inference to the Best Explanation', *Philosophy and Phenomenological Research* **74**, 2 (2007), 441-448.
Putnam, H. *Philosophical Papers Vol. I: Mathematics, Matter and Method* (Cambridge: Cambridge University Press, 1975).
Rappaport, S. 'Inference to the Best Explanation: Is It Really Different From Mill's Methods?', *Philosophy of Science* **63** (1996), 65-80.
Ray, C. 'Logical Positivism'. In W. H. Newton-Smith (ed.) *A Companion to the Philosophy of Science* (Oxford: Blackwell, 2000).
Ruben, D.-H. *Explaining Explanation* (London: Routledge, 1990).
Ruben, D.-H. 'Introduction'. In D.-H. Ruben (ed.) *Explanation* (Oxford: Oxford University Press, 1993).
Salmon, W. 'Scientific Explanation: Causation *and* Unification' (1990). In Y. Balashov and A. Rosenberg (eds.) *Philosophy of Science: Contemporary Readings* (London: Routledge, 2002).
Smart, J.J.C. *Philosophy and Scientific Realism* (London: Routledge and Kegan Paul, 1963).
Thagard, P. 'The Best Explanation: Criteria for Theory Choice', *The Journal of Philosophy* **75** (1978), 76-92.
Van Fraassen, B. C. *The Scientific Image* (Oxford: Clarendon Press, 1980).
Van Fraassen, B. C. *Laws and Symmetry* (Oxford: Clarendon Press, 1989).