Explanation and determination

Gijsbers, V.A.

Chapter 3

Against Inference to the Best Explanation

3.1 Introduction

Inference to the Best Explanation (IBE) – a term introduced by Harman 1965 [38], though the idea can be retraced to Peirce, Whewell and even Aristotle – is a contentious topic in philosophy. According to proponents of IBE, the observation of a theory's success in explaining a phenomenon is a reason to raise our estimate of the probability of the theory's being true: in the words of Lipton ([72], 56), "explanatory considerations [are] an important guide to judgments of likeliness".[1] Its opponents, on the other hand, deny that explanatory considerations are relevant for determinations of likeliness.[2] In this chapter, I will give a concise overview of the current state of the debate, and I will present new criticisms of several arguments in favour of IBE.

In a first phase of the debate, proponents of IBE defended the thesis that there is a correct inference scheme of the form "if E, and if H is the best explanation of E, then H". Several arguments were raised against this thesis, and the consensus is now that this thesis cannot be maintained. Recent proponents of IBE have therefore defended somewhat weaker claims. They claim no longer that IBE is a formal inference scheme on a par with deduction, but only that explanatory considerations point us to truth, and are thus epistemically relevant. My arguments specifically target this more modest

[1]For other defences of IBE, see Lipton (1991 [69], 2001a [70], 2001b [71], 2007 [73]); [38]; Psillos (1999 [93], 2007 [95]); Lycan 2002 [75]; Day and Kincaid 1994 [20]. The latter article also contains an extensive overview of how IBE is used in different areas of philosophy.

[2]See Van Fraassen 1989 [29]; Achinstein 1992 [1]; Salmon (2001a [114], 2001b [115]); Barnes 1995 [6].

version of IBE, and I will attempt to show that it, too, is too strong. Explanation is not an epistemic category at all – where I use the term 'epistemic', as I will do everywhere in this chapter, in the narrow sense of 'relevant to judgments of likeliness or truth'.

The arguments with which the modern, more moderate version of IBE is defended fall into three broad groups.[3] First, there are attempts to link explanatory virtues like simplicity and scope to the epistemic virtue of high probability. Second, there are case studies of scientific research where explanatory considerations allegedly do epistemic work. Third, there are arguments to the effect that IBE fits well with and therefore draws support from Bayesianism.

Against the first class of arguments, I will argue that none of the so-called explanatory virtues is conducive both to explanatory loveliness and to high probability. For instance, an explanation which gives a detailed causal mechanism is *ceteris paribus* more explanatory than one which is vague and non-committal, but also and for the very same reason *ceteris paribus* less likely to be true. In the reverse direction, rigorously applying the distinction between loveliness and likeliness shows us that a virtue like simplicity is not conducive to explanatory loveliness, even though it might be an epistemic virtue.

The second class of arguments cannot be dismissed *in toto*: I have to limit myself to discussing a single example. This will be Lipton's (2004) study of the research of Ignaz Semmelweis, which is the most convincing and thorough study in the literature. I will argue that although the study shows that explanatory considerations played an important role in Semmelweis's research, they did not play the role envisaged by Lipton. The case study licenses only a much weaker explanationist position.

As for the third class of arguments, I will attempt to show that the possibilities of reconciliation presented by Lipton 2004 [72] are not, in fact, very promising.

I will conclude that we do not have good reasons to accept the form of IBE defended by its recent proponents. If this conclusion is correct, we have to rethink the role of explanation in science, and reassess a large set of arguments for scientific realism that depend on IBE.

[3] I will be mainly concerned with the arguments given in Lipton 2004 [72], since his is by far the most extensive defence of IBE yet given.

3.2 IBE as an inference scheme

It is unfortunate that IBE is called IBE, because this gives us the idea that its defenders posit an inference scheme like those of deductive logic; an inference scheme – let us call it S – that would roughly be "if E, and if H is the best explanation of E, then infer H".

However, there are at least three reasons for thinking that S cannot be valid. **First**, we may not yet have thought of a good enough explanation of E: if we have only unconvincing explanations of E, the rational thing to do is to withhold our assent from them all, including the best of them. For instance, string theory may be our *best* explanation of certain physical phenomena, while not being a *good* explanation of those phenomena. That it is the best is by itself too weak a ground to believe it to be true.

This argument is known in the literature as 'the argument from a bad lot', and there is consensus that it shows that S cannot be valid – see, for instance, Lipton (2004 [72], p. 63). Nevertheless, there has been a debate in the recent literature about this argument. This debate was kicked off by Van Fraassen (1989 [29], pp. 142-150), who presents the argument from a bad lot as a decisive argument against Inference to the Best Explanation. Both Psillos (1999, [93], pp. 215-222) and Lipton (2004 [72], pp. 151-163) attack Van Fraassen's argument, which might give the impression that they try to save S from Van Fraassen's criticism.

However, this is not the case. Lipton and Psillos attack Van Fraassen in order to save scientific realism; and in their defence of realism, they affirm the invalidity of S. Thus, Psillos writes (1999 [93], p. 216): "[I]t is logically possible that our best theory is the best of a bad lot. Clearly, any sensible model of abduction must allow for this possibility." He then proceeds to defend the idea that under some circumstances, often realised in science, we may nevertheless have good reasons to believe that our theories form a 'good' lot. But a scheme of inference that is valid only if certain unstated background conditions are true is evidently not itself a correct scheme of inference. Lipton takes the same approach.

Second, there may be too many good explanations available, and in such a case, the difference between the best explanation and the other explanations might be too small to warrant inferring to it. It may be irrational to choose even among a good lot if we don't have strong reasons to prefer one alternative over all the others.

Third, S would commit us to an epistemic form of the Principle of Sufficient Reason. If it is valid, then for *any* phenomenon E, if there is at least one potential explanation H of E, it is rational to conclude that there is a true explanation of E. But some things – the decay at a particular time

(rather than some other time) of a radioactive atom, the existence of something (rather than nothing) – may not have true explanations at all, though they do have potential explanations (hidden variables, the necessary existence of God).[4] Since S would have us infer to the best explanation of these phenomena, it precludes the possibility that they cannot be explained.

There is consensus in the current literature that no scheme like S is valid. Salmon (2001a [114], p. 62) speaks of IBE as "a slogan, not [...] an accurate account of any form of nondemonstrative inference"; Lipton (2004 [72], p. 57) says that "it still remains more of a slogan than an articulated account of induction", after which he goes on to qualify the slogan in various far-reaching ways that dash any hopes that IBE might be a simple scheme of inference; Day & Kincaid (1994 [20], p. 275) state that "IBE is not a special, foundational inference scheme", by which they mean that explanatory considerations give us only defeasible warrant; Psillos (2007 [95], p. 442) says that IBE is "an instance of [...] defeasible reasoning". All these authors except Salmon are proponents of IBE.

Recent defenders of IBE have thus abandoned the idea of IBE as a scheme of inference. Instead, they claim that explanatory considerations do not determine but only inform our judgments of likeliness. Henceforth, we will be concerned only with this weaker form of IBE, which will be formulated in more detail in section 3.4.

3.3 Loveliness and likeliness

In the previous section, we saw that Inference to the Best Explanation is not really inference; in this section, we will look at the second half of formula. We must eliminate a major ambiguity in the term 'best explanation', namely, that between what Peter Lipton has dubbed the loveliest explanation and the likeliest explanation – or rather that between loveliness and likeliness, since there is little reason to continue speaking in superlatives once the idea of scheme S has been renounced.

A **lovely** explanation of a phenomenon F is an explanation which, if it *were* true, *would* offer a satisfying explanation of F; or, in other words, an explanation which, if it *were* true, *would* give us the great understanding of F. (What makes an explanation satisfying? We would of course like to have a detailed account of this, but the difference between loveliness and likeliness can be understood without it, and a purely intuitive notion of

[4] Pruss 2006 [92] argues for the necessary truth of the Principle of Sufficient Reason on the basis of the correctness of Inference to the Best Explanation; but I suspect that most philosophers will see this as a *reductio ad absurdum* of this kind of IBE.

which explanations are satisfying and which are not will prove enough for the arguments in this chapter.) Loveliness is an explanatory virtue, or rather, the *summum bonum* of explanation.

A **likely** explanation of a phenomenon F is an explanation which, considering all the available evidence, is likely to be true. Likeliness is thus equal to probability, and an *epistemic* good (in the restricted sense of this term established in the introduction).

An explanation can be likely without being lovely, and at least *prima facie* also lovely without being likely. An example of the former is the explanation of why people fall asleep after they have smoked opium by saying that opium has a 'dormitive virtue'. This is almost guaranteed to be true (if we construe the term 'dormitive virtue' broadly enough), but isn't very satisfying. Whether and under what conditions we find the latter, that is, an explanation that is lovely but not likely, is the central question of the debate about IBE; and it will therefore be addressed throughout this chapter.

When the proponent of IBE tells us that explanatory goodness inform our judgments of likeliness, what do they mean by explanatory goodness? Do they mean loveliness or likeliness? Lipton (2004) argues persuasively that if IBE is to be an informative doctrine, it must be the doctrine that loveliness informs our judgments of likeliness. If being a good explanation simply means being a likely explanation, then of course we should infer to good explanations – because likeliness is an epistemic good, prior to and independent of any explanatory considerations.[5] Explanatory considerations become relevant to epistemology only if they allow us to make inferences we wouldn't have been allowed to make without those considerations. Thus, the defenders of IBE must postulate a link between loveliness and inference; their doctrine must be that sometimes we can infer to an explanation *because* it is a lovely explanation.

On the other hand, it would be absurd if the proponent of IBE were to ask us to infer to an explanation that we judge less likely to be true (even after we take into account explanatory considerations) than some of its competitors. We should always infer to that theory which is most likely to be true. So the defender of IBE has to claim that lovely explanations can also be judged (after we take into account explanatory considerations) to be likely explanations; and in order to do so, he must forge a link between loveliness and likeliness. This link must be such that we sometimes find out about likeliness by considering loveliness.

[5]Harman 1965 [38] was not clear about this, glossing "best explanation" as "most plausible explanation". A sensitivity to the likely/lovely distinction is one of the virtues of modern defences of IBE.

3.4 The guiding thesis

Before we look at the arguments given in favour of IBE, let us formulate the doctrine somewhat more rigorously. Lipton introduces a formulation of IBE which he calls the 'Guiding Thesis'. The Guiding Thesis states that when we are looking for true theories (2004 [72], p. 56) "[explanatory] considerations tell us not only what to look for, but also whether we have found it." This is not to say that explanatory considerations alone can tell us whether a theory is true; but it is to say that they are relevant to making such judgments. Loveliness does not equal likeliness, but loveliness is a *guide* to likeliness.[6]

The guiding thesis is *not* equal to the claim that sometimes, the (subjective) probability of a theory T increases when we find out that some phenomenon E of which T furnishes a lovely explanation actually occurred. This latter claim is obviously true, and can be accepted by those who deny the guiding thesis. We can often judge E to be evidence for T without taking into account T loveliness as an explanation of E, and the vast majority of confirmation theories do so. Even theories that furnish only unlovely explanations can increase in (subject) probability when we get new evidence that they explain (in their unlovely way). "Opium has a dormitive virtue" becomes (subjectively) more likely once we see some people fall asleep after smoking opium, even though the theory gives us only an unlovely explanation of the evidence.

The guiding thesis, then, is more specific. In saying that loveliness is a guide to likeliness, it claims that sometimes, explanatory considerations play an *essential* role in determining whether or not a theory is to be believed. Insights in loveliness allow us to make claims that we would not be licensed to make if we paid no attention to explanation. (On a Bayesian approach, where there is generally no clear threshold for acceptance, we should reformulate this as follows: insights in loveliness make a difference to our posterior probabilities.)

The initial plausibility of the guiding thesis comes from examples where we do infer to the best explanation of a phenomenon we have witnessed. Salmon (2001a [114], p. 73) gives the example of two people who go to the woods, find, collect, cook and eat some mushrooms, and then, several hours later, suffer severe gastrointestinal distress. The best explanation of this is

[6] Barnes 1995 [6] calls this claim "Lipton's Central Thesis". His attack on it and mine complement each other well: he discusses some arguments that I will leave aside, such as the claim that Mill's method is a form of IBE; while I discuss the Semmelweis case study and Bayesian arguments, which he does not touch on. The only overlap is in our discussions of the explanatory virtues: here we are in substantial agreement, though the details of our arguments vary considerably.

that the mushrooms they collected were poisonous; we do in fact infer this hypothesis; and it is tempting to believe that we do the latter because of the former.

But is this true? Do we infer the hypothesis because it is the best explanation, or does that get the order wrong, and is the mushroom story the best explanation because we are allowed (on other grounds) to infer it? Salmon continues: "But how can we determine that consumption of poisonous mushrooms is the best explanation? It seems to me that we judge it to be best because it is the most likely, among those available, to be true. ... [the] explanatory beauty does not enter into our choice of the hypothesis as most likely." (Explanatory beauty is another term for loveliness.)

Consider the following alternative explanation of the gastrointestinal distress: although all the mushrooms were perfectly harmless, the mushroom gatherers drank a bottle of red wine when they came back from their trip, and this wine had been adulterated by CIA agents with a poison that causes gastrointestinal distress. Is this a good explanation? Let us disambiguate that question. First, is it a likely explanation? No, it is very unlikely to be true (given the facts at our disposal and our background beliefs). Second, is it a lovely explanation? That is, would it be a good explanation if it *were* true? Well – yes. If it were true, it would perfectly explain the suffering of the gatherers. Indeed, it appears as lovely as the explanation using the poisonous mushrooms. And yet we do not infer to its truth.

We can now see that many supposedly obvious examples of IBE in action are in fact no such thing. When the detective tries to solve the murder – the naive defender of IBE may say – surely he infers to that hypothesis that best explains the available evidence? True. But does he infer to that hypothesis *because* it is the loveliest explanation, that is, because it would be the most satisfying explanation if it were true? Or is the explanation accepted as that which best explains the available evidence *because* the evidence makes it likelier than any of its competitors? In the latter case, explanatory considerations do no epistemic work at all, and IBE does not take place.[7]

These considerations diminish the initial plausibility of the guiding thesis; so before we accept it, we will need arguments. The task of a defender of IBE, then, is to show that we do sometimes infer a hypothesis H because

[7]We have assumed throughout that explanatory and epistemic factors can be pried apart. This is probably not true; it is perhaps the case, for instance, that H's probabilistic relevance for E is a necessary condition both for H explaining E and for E being evidence for H. But we follow the literature in assuming that explanation is something *more*, something not reducible to simple confirmation theories. If this assumption is wrong, so much the worse for IBE; for it then would reduce to these simpler confirmation theories that did not mention explanatory factors.

a phenomenon E would be beautifully (that is, lovelily) explained by H if H were true; and that it is not instead the case that witnessing E makes H sufficiently likely to justify inferring it quite apart from the question of whether it affords a beautiful explanation of E.

3.5 First argument: virtues

We turn to the first argument that the proponents of the Guiding Thesis offer.[8] This argument has the following form. Explanations that are lovely, are lovely because they have certain properties. We call these properties explanatory virtues. If it turns out that these explanatory virtues are *also* epistemic virtues, then loveliness is a guide to likeliness. To make the argument more concrete, one has to identify the explanatory virtues, and one has to show that these virtues are also epistemic virtues, that is, that these virtues also point to likeliness.

One immediate worry with this strategy is that the role of explanation appears to be very small. If the explanatory virtues are also epistemic virtues, why not just call them epistemic virtues and leave explanation out of the story? Still, it might be the case that it is harder for us to judge which virtues are instantiated in a hypothesis H, and to what extent, than it is for us to judge how good an explanation of the evidence H is. If so, explanation would be psychologically necessary though analytically superfluous.

The real problem with the first argument is much deeper: there are in fact no explanatory virtues that are also epistemic virtues. Simplicity, scope, unification, precision, exhibiting a mechanism – all of these are either explanatory or epistemic virtues, but none of them is both. And if that is the case, the sought-for link between loveliness and likeliness is not established. So what we need to do now, is to look at all the supposed virtues in turn and see whether they can bridge the gap from the explanatory to the epistemic.

It is strange that the defenders of IBE have taken little trouble to defend the idea that the explanatory virtues are also epistemic virtues. Lipton 2004 [72] spends about half a page (122) of a 200-page book on this topic, giving a list of virtues rather than substantive arguments. Lycan 2002 [75] discusses a number of virtues, but he makes no clear distinction between loveliness, likeliness, and pragmatic reasons for theory acceptance, and argues only that, all things considered, we prefer simpler (and so on) theories. We do – but this does not establish a link between loveliness and likeliness.

[8] For another critical discussion of the explanatory virtues, see Barnes 1995 [6], pp. 257-266.

If we wish to judge whether, say, simplicity is an explanatory virtue, we need to look at examples where a simpler and a less simple explanation are competing with each other. Similarly for the other virtues; and thus, we need some examples on hand before we continue our investigations. I will give two.

1. I am carrying a bunch of red roses. Why am I carrying a bunch of red roses (rather than something else)? Explanation A: I want to surprise my girlfriend. Explanation B: I wanted to surprise my girlfriend, so I went to the theatre to buy tickets for the performance of *Medea* tonight. But then, when I walked home, I remembered that my girlfriend loathes Euripides (because she agrees with Nietzsche that he destroyed the Dionysian strength of Greek tragedy). So I went back, but they wouldn't refund me my money, which left me with only a couple of euros in cash. All I could do then was buy a cheap bunch of flowers, and these roses – which are already a bit past their prime – were the cheapest they had at the flower shop. Neither I nor my girlfriend like roses all that much, but it's the best I could do.

   Of these explanations, A is clearly the more likely, since B entails A. However, B is the lovelier: if it is true, it would be a better explanation of the phenomenon than A. This is not because B gives us more details about the causal history of the explanandum, but because it more specifically singles out the explanandum from its contrast class: explanation A leaves open the possibility that I want to surprise my girlfriend with tickets for *Medea*, or a gold ring, or an expensive bouquet, while explanation B excludes all those possibilities.

2. Heavy objects fall down, lighter-than-air objects fall up. Why is this the case? Because of the law of gravity. If we add some details, we have a good explanation for each occasion where something heavy falls down or something light falls up.

   Assume, however, that in my basement heavy objects fall up instead of down, while lighter-than-air objects such as helium balloons fall down instead of up. My basement is the only place in the world where this happens. One potential explanation of these phenomena is that in my basement, the law of gravity does not hold, but a reverse law of gravity holds that involves a repulsive force rather than an attractive one.

   We would be loath to accept this explanation, since it seems overwhelmingly likely to be false. We would need an immense amount of evidence before we would even consider believing it. But leave the epistemic

> considerations aside and suppose the hypothesis *were* true – would it then offer us a good explanation of a brick's in my basement falling up? It would.
>
> An explanation of something in my basement falling up in terms of a reverse law of gravity is just as lovely as an explanation of something elsewhere falling down because of the normal laws of gravity; it is just as lovely, the only problem is that it is less likely.

With these two examples in hand, we will be in a better position to judge the virtues.

Whether **simplicity** is linked to truth is a controversial issue. Salmon points out (2001b [115], p. 129) that scientists often regard simple explanations with suspicion and even reject them for being too simple. Thus historians may well feel, and rightfully so, that any simple theory about the decline and fall of the Roman Empire is bound to be wrong. A political scientist who wishes to explain that there was a surge in the number of votes received by populist parties in the Netherlands after 2000 may well dismiss out of hand candidate explanations like "other parties didn't face up to the problems of immigration" – the proposed cause being too simple to be, on its own, an explanation of the phenomenon.

However, let us grant for the sake of argument that simplicity *is* an epistemic virtue, and that simpler theories are *ceteris paribus* more likely to be true than complex ones. The question remains whether simplicity is also an explanatory virtue. Do simpler theories generally generate lovelier explanations?

Our first example, about why I am carrying a bunch of red roses, casts doubt on this proposal. It seems that as we get more details of a certain kind, namely, details that help us to discard additional elements of the contrast class, we are also able to understand the situation better. But more details make a hypothesis more complex, not more simple; so there is at least one aspect of simplicity – lack of detail – that harms rather than aids loveliness. Good explanations are often very specific stories that appeal to contingent features of the particular situation.

Is there another aspect of simplicity that does improve the loveliness of the explanation? A simpler explanation can be used in more circumstances – it is potentially more unifying than a complex one; but we will look at unification as a separate virtue next, and see that it does not improve loveliness. A simpler explanation is easier to grasp than a complex one; but this seems to be a merely pragmatic advantage, not something that we can tie to loveliness. A simple explanation of the fall of the Roman Empire may be easier to grasp, but a more complex one that details the interaction between a host of causal

factors will grant us more understanding once we've finally worked through the details and understood it.

There may be other aspects of simplicity that do increase loveliness, but it is up to the defenders of IBE to make them explicit. I conclude that as far as our analysis has gone, simplicity seems to be an explanatory vice rather than an explanatory virtue.

We now turn to **unification** and **scope**. Let us notice first that an explanation of a particular phenomenon is not the kind of thing that can unify, or that can have a scope. So the appeal to unification or scope must be understood in this sense: if an explanation of a phenomenon P involves theories that also explain many, or a great variety of, other phenomena, then by virtue of this it is a lovelier explanation of P.

This is false, as can be seen from our second example. The reverse gravity hypothesis is much less unificatory than the standard laws of gravity, and it also has a much smaller scope. Yet if it is true, it gives a lovely explanation of the strange behaviour of the objects in my basement. This explanation would not become lovelier if more objects in the Universe were to behave in accordance with this law – the behaviour of these other objects is perfectly irrelevant for how much understanding the explanation gives us about what is happening here in my basement. (It would, of course, become easier to believe that the explanation were *true* if it applied to more objects; but this is by definition irrelevant to how lovely it is.)

Suppose we see an explanation as the instantiation of a more general explanatory pattern (as we must do to talk about unification and scope). Then how lovely this explanation is can depend only on two things: the properties of the pattern itself, and whether the pattern is indeed instantiated in this case. The *number* of instantiations of the patterns is irrelevant. The dormitive virtue explanation of opium putting people to sleep does not become any lovelier when more people start smoking opium so that it can be applied more often; nor does the explanatory loveliness of a Newtonian explanation of the movement of our planets decrease if it turns out that the law of gravity holds only in our local part of the Universe. So we have reason to doubt that unification and scope are explanatory virtues.[9]

There is a sense in which unification and scope are 'explanatory virtues', but this is not a sense that is useful to the defender of IBE. That sense is this: if a theory is unifying, or has great scope, we can use it to construct more explanations than if it were not unifying or had limited scope. So in a *quantitative* sense, theories that unify and have great scope are better if we

[9]There is a substantial literature linking explanation with unification; for criticism of this tradition, I point the reader to chapter 2 above.

want explanations; but loveliness is a *qualitative* matter. We might want to say that a unifying theory is explains more, and that this is a virtue of the theory; but it does not follow that the individual explanations it furnishes are any lovelier. Thus, Lipton (2004 [72], p. 122) is wrong when he states (without argument): "An explanation that explains more phenomena is for that reason a lovelier explanation".[10] In the absence of an argument or an example, my examples are good reasons to believe that Lipton is mistaken and the unification and scope are not explanatory virtues.

A final point we have to make here is that lack of unification or scope can sometimes be an indication that a more lovely explanation is available. For instance, the fact that Galileo's law of free fall is valid near the surface of the earth but invalid in many other parts of the universe, can be taken to indicate that a deeper explanation is available. And indeed, there is the explanation of free fall given by Newton's theory of gravitation, which is valid everywhere in the universe. Newton's explanation is more lovely than Galileo's; it is also more unifying and has greater scope. In addition, it was the failure of Galileo's law to have a great scope that led us to think that there might be a lovelier explanation available. Does this mean that scope generates loveliness after all?

It does not, because the lack of scope is not itself a feature that makes the explanation less lovely. Rather, it is an *indication* that we haven't arrived at the true causes yet, that a deeper explanation is available. In certain domains, such as that of physics, we believe (for empirical, and perhaps partly metaphysical, reaons) that the basic causal structures are uniform over large areas of the universe. In such domains, we take a lack of scope to be an indication that we haven't arrived at those basic causal structures yet. But what makes the scope-lacking explanations less than perfectly lovely is that they don't get at these true causes; it is this failure, not the lack of scope, that lowers their loveliness.

Scope has an additional problem: it is not an epistemic virtue. Theories with greater scope have greater content than theories with a more limited scope; and they are therefore more likely to be false. It is more probable that Newton's Laws hold in my attic than that they hold in my attic *and* my basement. We have excellent reasons to desire theories with a large scope, but these are not and cannot be *epistemic* reasons (in the sense of that word used in this chapter).

Finally, let us look at **precision** and **giving a mechanism**. It is clear

[10] It is of course not an explanation but an explanation pattern that can explain more or fewer phenomena. A single explanation explains only one thing, its explanandum. But we can assume that Lipton was thinking about patterns here.

that these are explanatory virtues. But it is unclear that they are epistemic virtues, or rather, it is (at least in the case of precision) blatantly obvious that they are not. When I make an explanation more precise, the probability of its being true decreases. Our first example shows this to be the case: explanation B much more precisely picks out the explanandum from the contrast class than does explanation A; it is also, precisely because of this increased precision, less likely to be true.

When two explanations compete and one of them gives us a causal mechanism leading up to the phenomenon while the other does not, the former may well be lovelier than the latter. I do not know whether this is a general rule, but let us assume that it is. Is the explanation which gives a causal mechanism also more likely? The opium example shows that this is not the case: the 'dormitive virtue' explanation is very likely indeed, although it does not give a mechanism; while any explanation that *does* give a mechanism would be more precise than the dormitive virtue explanation, and hence less probable.

But let us look at two explanations, one of which gives a mechanism and the other of which doesn't, which have an equal precision. Can we then say anything in general about their respective probabilities? Is it more or less likely that the plane crashed because a certain Al Qaida leader wanted it to (no mechanism), than that it crashed because seven screws in the left wing were rusted, broke in mid-flight, and caused the left wing to fall off (causal mechanism)? The only way to answer this question is empirical-statistical research; an appeal to the fact that the second explanation contains a mechanism is powerless. I admit to seeing no general way to link mechanisms with probability; and that would mean that giving a mechanism is not an epistemic virtue.

An appeal to explanatory virtues therefore does not help us bridge the gap between loveliness and likeliness.

3.6 Second argument: case study

Another way to bridge the gap between loveliness and likeliness is to show that explanatory considerations are used as guides to inference in actual and successful scientific practice. While this would not necessarily decide the normative issue of whether IBE is epistemically justified, it would certainly be a step in the right direction. This second type of argument, which must take the form of a case study, is used extensively by Lipton (2004). The example of actual scientific practice he discusses is Ignaz Semmelweis's research on childbed fever between 1844 and 1848. I will first summarise Lipton's

account, and then discuss its flaws and merits. (For a more complete discussion, see Lipton 2004 [72], pp. 74-82; Hempel 1966 [42], pp. 3-8; Semmelweis 1861 [123].)

Ignaz Semmelweis worked at a Viennese hospital that had two maternity wards. It was observed that significantly more women in First Obstetrical Clinic than in the Second Obstetrical Clinic contracted childbed fever. Since childbed fever was the most significant cause of death in these wards, the mortality rate in the first ward was much higher than in the second ward.

Semmelweis set out to find an explanation of this contrast. After many false starts, he devised the theory that medical students, who examined women in the first but not in the second ward and often did so after dissecting dead bodies, infected the women with 'cadaveric matter'. Making the students disinfect their hands before examining the women led to an immediate and highly significant decrease of the rate of childbed fever in the first ward.

This was undoubtedly a case of successful science. If Semmelweis used IBE, and had to use IBE, in order to reach his conclusions, we will have a strong argument in favour of IBE. So the question to be answered is: did Semmelweis infer to his final theory (at least partially) because that theory was lovelier than any of its competitors? Let us look at the case in more detail, and let us pay special attention to the role of explanatory considerations.

The first thing we notice is that Semmelweis starts out by looking for an explanation: he needs to know *why* the mortality rates in the two wards are different. This means that he is looking for a *contrastive explanation*, as described by Lipton (2004 [72], pp. 30-54) and Hitchcock (1996 [46]). Semmelweis is looking for a causal variable A with the following properties: (1) A takes the value a_1 in the first ward, (2) A takes the value $a_2 \neq a_1$ in the second ward, (3) changing A in the first ward from a_1 to a_2 lowers the mortality rate there to (approximately) the level in the second ward.

Given that this is the case, how does Semmelweis construct and test hypotheses? He first notes a causal variable X that takes different values in the two wards. He then constructs the hypothesis that the difference in X explains the difference in mortality. This hypothesis must pass three tests before it can be accepted. The first test – a kind of filter, rather – is to reject the hypothesis if background knowledge tells us that it cannot possibly be true. The second test is to verify that the variable assumes different values in the two wards. If it doesn't, it cannot be the explanation of the difference in mortality. The third test is to actually change a_1 to a_2 in the first ward: if this results in the hoped for decrease in mortality, the hypothesis is accepted, and if it does not, the hypothesis is rejected.

Let us see how this works in practice. The first test rules out all hypotheses that violate our background knowledge, such as "the different names of the wards explain the difference in mortality rate". These hypotheses barely cross the scientist's mind, and are never written down. This first test leaves us with those hypotheses that are not patently absurd. One of these is the idea that perhaps medical students, who examine patients in the first but not in the second ward, do so in a much rougher way than the regular nurses, and that this explains the difference in mortality rate. However, this hypothesis does not pass the second test: it turns out to be the case that the medical students are not significantly rougher in their examinations than the nurses. Hence, roughness cannot explain the difference in mortality rates.

This leaves us with those hypotheses that involve actual causal differences (that are not obviously irrelevant to the difference in mortality). An example of these is "seeing the local priest pass through the ward, when he is on his way to give the last sacraments to a dying woman, is a cause of childbed fever". It was in fact the case that the priest, when he visited the hospital, passed through the first but not through the second ward; and it was not inconceivable that his presence depressed or agitated the women and thus caused physical illness. However, this hypothesis did not pass the third test: Semmelweis got the priest to take another route, but this did not change the mortality rates. Hence, the presence of the priest did not explain the difference in mortality.

Finally, Semmelweis hit on the hypothesis that medical students infected the women with cadaveric matter, since they usually dissected corpses before they came to the maternity ward. This hypothesis was not obviously absurd; it described a real difference; and taking away the difference (by having the medical students disinfect their hands with chlorinated lime before they entered the ward) did in fact take away the difference in mortality rate. Having passed all three tests, the hypothesis was accepted. Semmelweis concluded that he had found the explanation of the difference in mortality rate; and his policies of disinfection, later broadened to include instruments, led to an almost complete disappearance of childbed fever from his hospital.[11]

This description of the case, which I have reproduced faithfully from Lipton 2004 [72], is reasonable enough. But does it show IBE in action? As far as we have seen, explanatory considerations helped Semmelweis to

[11] A brief historical note: Semmelweis's theory met with much resistance from the Austrian medical establishment, due to both scientific and political reasons. He was fired from his Viennese hospital, and within a couple of years, the mortality rates there were back to their previous levels. In the meantime, Semmelweis had gone to Hungary, where his method of disinfection was quickly adopted with good results.

formulate and test hypotheses; but it was the outcomes of the tests, not considerations of explanatory beauty, that decided whether a hypothesis was rejected or accepted. In fact, explanatory loveliness did not once enter the picture. At no point were we forced to postulate that Semmelweis rejected or accepted a hypothesis because of its (lack of) loveliness. Apparently, then, Semmelweis did not use IBE.

Of course Lipton (2004 [72], p. 136) anticipates this remark. He concedes that "[i]n effect, Semmelweis converted the question of which is the best explanation of the original data into the question of which is the only explanation of the richer set. We often decide between competing hypotheses by looking for additional data that will discriminate between them." This is *prima facie* a problem for IBE. If loveliness could be used to decide between the hypotheses, why would we need additional data? And if the additional data rule out all but one hypothesis, why do we then need explanatory loveliness? If there is, in the final situation, only one possible explanation, then its comparative loveliness can no longer be an issue.

Lipton, however, qualifies his last claim (ibid.): "Perhaps in some extreme cases that discrimination works through the refutation of one of the hypotheses; but what seems far more common is that the additional evidence, though logically compatible with both hypotheses, can only be explained by one of them". This brings us back to a familiar dilemma. Is the less fortunate hypothesis discarded because it does not *explain* the additional evidence and is therefore improbable, or is it discarded because the additional evidence makes it less probable and therefore a bad explanation? I think that the facts of the Semmelweis case leave the door wide open for the second option; and if that is so, we do not need to postulate that he used IBE.[12]

If a priest passing through the ward is the main cause of childbed fever, it is very probable that intervening on this variable will change the rate of childbed fever. Semmelweis carries out the intervention and notices that the expected change does not occur. This means that either (1) the priest is not the causal variable that explains the difference, or (2) a highly improbable statistical fluke occurred, or (3) the situation is akin to the famous philosophical example of the two assassins, where only an intervention on two distinct variables will make a difference in the outcome of the causal process (pre-emption). The third possibility is relatively unlikely. The second possibility is unlikely by definition. Thus, the first option is very likely: the new evidence makes it very likely – quite apart from considerations of explana-

[12]Here one is tempted to quote the (probably apocryphal) remark of Laplace: "But where is Inference to the Best Explanation in your scheme?" "Je n'ai pas eu besoin de cette hypothèse."

tory loveliness – that the priest is not the causal variable that Semmelweis is looking for. The new evidence alone, without IBE, is enough to reasonably drop the hypothesis. This possibility of probabilistic disconfirmation is not taken seriously enough by Lipton when he formulates his dichotomy between deductive refutation and IBE.

One strategy that remains open to the defender of IBE is to argue that although in *this* example IBE seems merely to duplicate the results of independent probabilistic reasoning, there are nevertheless cases where it allows us to refute hypotheses that could not be refuted in another way; that there are cases, to slightly change Lipton's words, where the additional evidence, though not unlikely given either hypothesis, can be only explained by one of them. In such a case, IBE would tell us to drop the hypothesis which cannot explain the additional evidence.

But this criterion is far too strong. Recall the two potential explanations of the illness of the mushroom gatherers. First: the mushrooms they collected and ate were poisonous. Second: the wine they drank had been poisoned by a CIA agent. Now, assume that the following additional evidence is gathered: there are fingerprints on the bottle of wine that do not belong to either of the two gatherers. This new evidence can be explained by the second hypothesis, but not by the first hypothesis – hence, the criterion under consideration would say that we have to drop the first hypothesis.

This is obviously absurd. It is true that the mushroom hypothesis does not explain the fingerprints, because it doesn't say anything about fingerprints and bottles of wine at all. But since it is not particularly unlikely, given the mushroom hypothesis, that there are fingerprints on the bottle not left there by one of the gatherers (someone in the wine shop must have handled the bottle, after all), the fingerprints give us no reason at all to ditch the mushroom hypothesis. The fact that hypothesis H does not explain new evidence E is not in itself a reason to drop hypothesis H. What, in addition, is needed? Precisely, I would say, that E makes H unlikely. But if that is the case, IBE again appears to be unnecessary.

At this point the defender of IBE may wish to retreat even further. He may wish to suggest that my arguments hold water only in those situations where we can make informed judgments about probabilities; but that in situations where we cannot make such judgments, we may still be able to make judgments about explanatory loveliness. If this is the case, then perhaps in such situations IBE can play an epistemic role, as a substitute for general probabilistic reasoning.

This kind of proposal has been made within the context of reconciling IBE and Bayesianism, and I will take it up in the next section.

3.7 Third argument: Bayesianism

We have seen that what may appear at first sight to be Inference to the Best Explanation can turn out to be describable with nothing more than the standard theories of probabilistic reasoning. It is no surprise, then, that some defenders of IBE have tried to set up an alliance between IBE and probabilistic reasoning. The form this takes in Lipton (2004 [72], pp. 103-120) is a set of arguments that set out to establish that IBE is a useful addition to Bayesian reasoning. Day and Kincaid (1994 [20], p. 286) also state this possibility: "IBE can be embedded in determining the priors and likelihoods that make up Bayesian calculation". Lipton goes into more detail, so it is his version I will discuss.[13]

At first sight, Bayesianism leaves no room for IBE. We start out with a probability distribution over all propositions. We get some new evidence and, using conditional probabilities, we update the probability distribution. This is our new state of belief. Explanatory considerations do not play a role.

But let us not be too hasty. Lipton proposes three ways in which explanatory considerations might be hidden at the heart of the Bayesian machinery (2004 [72], p. 114). First, explanatory considerations might figure in the determination of likelihoods. Second, explanatory considerations might figure in the determination of priors. Third, explanatory considerations might figure in the determination of relevant evidence. We need to look at these three suggestions in turn.

The first suggestion is that explanatory loveliness is correlated with **likelihood**. Likelihood is a technical term that designates the conditional probability of the evidence on the hypothesis. Thus, if the hypothesis H is that the diners ate poisonous mushrooms, and the evidence E is that they both fell ill, then the likelihood of the evidence is $P(E|H)$, which is the probability of falling ill when you have eaten poisonous mushrooms. Lipton (2004 [72], p. 114) suggests that "although likelihood is not to be equated with loveliness, it might yet be that one way we judge how likely E is on H is by considering how well H would explain E." This would be pragmatically useful "if in fact loveliness is reasonably well correlated with likelihood, and we find it easier in practice to judge loveliness than likelihood."

[13] See Hitchcock 2007 [48] and Psillos 2007 [95] for other appraisals of Lipton's attempt to reconcile IBE and Bayesianism. Hitchcock argues that, although such a reconciliation is probably possible, it cannot go through the explanatory virtue of simplicity. (However, I argue above that simplicity is not an explanatory virtue.) Psillos argues that defenders of IBE should be the enemies rather than the friends of Bayesianism, and suggests that Lipton yields too much ground to the Bayesians. See Lipton 2007 [73] for replies.

This proposal might be a good argument for IBE, if it were to be found correct. In order to justify IBE, we must show that $P(H|E)$ (the probability of H on E) is positively correlated with how lovely an explanation of E is given by H. Now, suppose we can show – as Lipton hopes – that loveliness is positively correlated with the likelihood $P(E|H)$. Then, using the rules of probability and the two further assumptions that loveliness is not negatively correlated with $P(H)$ or positively correlated with $P(E)$ (which would have to be independently argued for), we can show that loveliness is positively correlated with $P(H|E)$. This is exactly what we need to prove in order to vindicate IBE.

But *do* we base our estimates of likelihood on the explanatory loveliness of H for E? What is it that we do when we desire to know the probability that, say, people fall ill after eating poisonous mushrooms? We attempt to gather the relevant empirical evidence. Experience, not explanatory considerations, teaches us whether in 95% or 30% or 4% of the cases eating poisonous mushrooms leads to illness.[14]

Let me elucidate this by an example where it is clear that (1) we know nothing at all about likelihoods, yet (2) we can easily judge loveliness. The following explanation of the illness of our mushroom gatherers is put forth: all the mushroom they gathered were edible, but as they walked through the woods, they were made the unwitting test subjects of an experimental non-lethal weapon mounted on a CIA satellite. This weapon sends out a strongly focused gravitational wave that, with probability p, causes sudden movements in the digestive system that allow more gastric acid to enter the duodenum than is normal. After a couple of hours, this leads to pains and illness.

This explanation is highly unlikely. It is also quite lovely: if it *were* true, it would be a good explanation of the illness of the mushroom gatherers. But is there anything, anything at all except that it is non-zero, that we can say about p once we have seen that this explanation is quite lovely? Does the loveliness of the explanation allow us to draw any conclusions about the success rate of a fictional CIA weapon? It is unlikely.

Thus, our examples cast doubt on the idea that there is a useful link between loveliness and likelihood. This does not prove that there is no such link; but since we have not been given any concrete arguments to believe that there is, it casts the ball back to the defenders of IBE.

One final note. There are some cases in which likelihood and loveliness *are*

[14]Unless we define 'poisonous' in terms of the percentage of cases that lead to illness, in which case it is not experience, but linguistic analysis or stipulation that gives us the correct percentage.

correlated: namely, those cases where we add detail to an explanation with the express purpose of pointing out the scenario with the greatest likelihood. For instance, we can simultaneously increase the loveliness and the likelihood of the hypothesis in the mushroom example by adding that the mushrooms weren't just poisonous, they were in fact devil's boletes, and that devil's boletes contain a poison called bolesatine which causes gastrointestinal distress. But such cases are no comfort to the defenders of IBE. The probability of the (more detailed) devil's boletes hypothesis is necessarily less than the probability of the (less detailed) poisonous mushroom hypothesis. Thus, in these cases loveliness, although positively correlated with likelihood, is negatively correlated with probability – which defeats the purpose of the proposal.

A second place where explanatory considerations might play a role is in the **determination of priors**. Lipton proposes one mechanism by which this could take place: the prior probabilities of hypotheses might be decided on the basis of explanatory virtues like simplicity. That brings us right back to the argument from explanatory virtues, which we have already discussed in section 3.5.

The third possibility is that explanatory considerations are used to **determine the relevant evidence**. Lipton (2004 [72], p. 116) writes: "Bayes's theorem ... does not, however, say *which* evidence one ought to conditionalise on. In principle, perhaps, non-demonstrative inference should be based on 'total evidence', indeed on everything that is believed. In practice, however, investigators must think about which bits of what they know really bear on their question, and they also need to decide which further observations would be particularly relevant. ... [T]his seems yet another area where the explanationist may contribute. ... [W]e sometimes come to see that a datum is epistemically relevant to a hypothesis precisely by seeing that the hypothesis would explain it."

Lipton is onto something here – thinking about explanation *does* help us when we decide which evidence to take into account and what further experiments to perform. But notice two things. First, we are not talking about explanatory loveliness, but about whether a hypothesis explains something or not. This is an absolute rather than a comparative judgment. Second, on this proposal explanatory considerations do not influence our judgments of likeliness directly, but only by telling us what evidence we ought to gather and consider. This role for explanation in scientific method is *not* the role postulated by IBE; so I can (and do) grant Lipton's claim here without granting that it constitutes a defence of IBE.

What we have seen, then, is that the attempt to tack IBE onto Bayesianism fails; or at least, that we do not at this moment have any reason to believe that it succeeds. Since this was the final class of arguments to be considered,

we can conclude that the current defence of IBE is not successful.

But I do not wish to end on such a negative note. In the next section, I will give a more positive account of Lipton's results; because although I do not believe that they establish the validity of IBE, I do think that they allow us to see that explanation plays a more important (if non-epistemic) role in scientific method than the famous methodological systems – falsificationism, hypothetico-deductivism, Bayesianism – have allowed us to see.

3.8 Modest explanationism

Not all methodological roles are epistemic roles. Thus, it is possible that explanation plays an important role in scientific method even if IBE is wrong. Lipton (2004 [72], p. 116) hints at one possible non-epistemic role explanation could play when he says that "we sometimes come to see that a datum is epistemically relevant to a hypothesis precisely by seeing that the hypothesis would explain it." Even more can be gleaned from his descriptions of the Semmelweis case.

We saw that the very *aim* that Semmelweis had is to find an explanation of the difference in mortality rate between the two maternity wards. (Why does he look for an explanation? If we accept an interventionist theory, such as that defended by Woodward 2003 [142], this is easy to understand: explanations are precisely the kind of knowledge that allows us to intervene successfully on a state of affairs.) Thus, we will understand his scientific methodology better when we understand what an explanation is.

Following Lipton 2004 [72], Hitchcock 1996 [46] and Woodward 2003 [142], we can say that explanations – at least explanations of the kind Semmelweis is looking for – are causal stories that show us why A_1 rather than the incompatible A_2 is the case, because (1) B_1 rather than the incompatible B_2 is the case and (2) there is a causal law (mechanism, scenario) leading from B_1 to A_1 and from B_2 to A_2. So what Semmelweis is looking for is some causal factor that takes different values in the two maternity wards, and which exhibits the relevant causal characteristics.

This observation helps us understand why Semmelweis discards some hypotheses that might well be true. Thus, when he considers the hypothesis that childbed fever is caused by overcrowding, he discards it because both wards are equally crowded. This is not an instance of falsification or disconfirmation or having too low a probability after Bayesian updating: there is no evidence at all that overcrowding is not a cause of childbed fever. For all Semmelweis knows, it might account for all cases of childbed fever in the second ward and a small number of cases in the first ward. But he rejects

this hypothesis because it cannot give him the explanation he is looking for.

When it comes to testing the hypotheses that are deemed worthy of testing, that is, the hypotheses that could explain the difference between the wards if they were true, it is the structure of explanation that tells Semmelweis how to proceed. He changes the value of causal variable B in the first ward, and observes whether the difference in mortality rates is removed. If it is, B is a cause of childbed fever. Again, there is much more potential evidence that is relevant for testing the hypothesis under consideration; but by paying attention to the fact that Semmelweis formulates the hypothesis in order to explain a specific phenomenon, we can understand why he tests his hypothesis in the way he does.

When he finally finds a causal factor that makes the difference between the two mortality rates, Semmelweis concludes that this factor must furnish the explanation of the difference. This concludes his research.[15]

None of this is particularly surprising; it is methodological common sense. But we do see that insights into the structure of explanations, and judgments on whether a given hypothesis would explain a certain phenomenon if it were true, inform science. We decide which hypotheses to test, and which experiments to carry out, based on such considerations. This is not an *epistemic* role for explanation; but it is an important methodological role nonetheless. Seeing explanation at work in these common sense contexts gives us reason to believe that developing a better, more detailed theory of explanation will give us a better, more detailed understanding of scientific practice. Thus, my attack on IBE notwithstanding, I believe that explanation is an important topic in the philosophy of scientific methodology.

I do not agree, then, with Peter Lipton's implicit suggestion that we are forced to choose between IBE and a trivial role for explanation (2004 [72], 62): "I want to insist that [IBE] makes out explanatory considerations to be an important guide to judgments of likeliness, that [it does] not reduce to the true but very weak claim that scientists are in the habit of explaining the phenomena they observe." There is a middle course between the Scylla of triviality and the Charybdis of making explanatory goodness an epistemic category; and it is a virtue of Lipton's own work that it shows us this course.

[15] What we see here is, more or less, what Bird 2007 [8] calls 'Inference to the Only Explanation'. I don't think Bird's insistence that all alternatives are actually falsified – rather than just made unlikely – by the experiment is particularly helpful, though.

3.9 Conclusion

We have seen three arguments for IBE: the first claims that there are explanatory virtues which are also epistemic virtues, thus linking loveliness to likeliness; the second claims that we can simply see IBE at work in specific instances of scientific research; and the third claims that IBE has a role to play within the framework of Bayesianism. We have also seen that all these three arguments fail – or at least, that they have not yet been adequately defended. From this, we can conclude that IBE itself has not been adequately defended. The many doubts that have been raised in the course of this chapter lend credence to the idea that a successful defence of IBE cannot be found; but only the future can tell whether this assessment is correct.

This does not mean that explanation plays no role in scientific method. Lipton's case study on the research of Semmelweis does show that explanatory considerations play an essential role in the methodological decisions of scientists. No theory of scientific method which does not take account of both the importance of explanation and the structure of explanations will be able to make sense of what scientists actually do. By bringing this fact out in the open, the work of the defenders of IBE has borne fruit, even if it was not the fruit they were after.

IBE is used in many areas of philosophy, as Day and Kincaid 1994 [20] show. One notable example is the realism debate, where realists often propose arguments that are variants of the following: theory T makes excellent predictions; the best explanation thereof is that T is (approximately) true; therefore, T is (approximately) true. If we reject IBE, these arguments may also have to be rejected – but this warrants a more detailed consideration than we can give it here. Much research on the topic of IBE remains to be done.