W.H. Newton-Smith (ed) *A Companion to the Philosophy of Science* (Blackwell, 2000) 184-193.

Inference to the Best Explanation

PETER LIPTON

Science depends on judgments of the bearing of evidence on theory. Scientists must judge whether an observation or the result of an experiment supports, disconfirms, or is simply irrelevant to a given hypothesis. Similarly, scientists may judge that, given all the available evidence, a hypothesis ought to be accepted as correct or nearly so, rejected as false, or neither. Occasionally, these evidential judgements can be make on deductive grounds. If an experimental result strictly contradicts an hypothesis, then the truth of the evidence deductively entails the falsity of the hypothesis. In the great majority of cases, however, the connection between evidence and hypothesis is non-demonstrative or inductive. In particular, this is so whenever a general hypothesis is inferred to be correct on the basis of the available data, since the truth of the data will not deductively entail the truth of the hypothesis. It always remains possible that the hypothesis is false even though the data are correct.

One of the central aims of the philosophy of science is to give a principled account of these judgements and inferences connecting evidence to theory. In the deductive case, this project is well-advanced, thanks to a productive stream of research into the structure of deductive argument that stretches back to antiquity. The same cannot be said for inductive inferences. Although some of the central problems were presented incisively by David HUME in the eighteenth century, our current understanding of inductive reasoning remains remarkably poor, in spite of the intense efforts of numerous epistemologists and philosophers of science.

The model of Inference to the Best Explanation is designed to give a partial account of many inductive inferences, both in science and in ordinary life. One version of the model was developed under the name `abduction' by Charles Sanders PIERCE early in this century, and the model has been considerably developed and discussed over the last twenty-five years. Its governing idea is that explanatory considerations are a guide to inference, that scientists infer from the available evidence to the hypothesis which would, if correct, best explain that evidence. Many

inferences are naturally described in this way. Darwin inferred the hypothesis of natural selection because, although it was not entailed by his biological evidence, natural selection would provide the best explanation of that evidence. When an astronomer infers that a star is receding from the earth with a specified velocity, she does this because the recession would be the best explanation of the observed red-shift of the star's characteristic spectrum. When a detective infers that it was Moriarty who committed the crime, he does so because this hypothesis would best explain the fingerprints, blood stains and other forensic evidence. Sherlock Holmes to the contrary, this is not a matter of deduction. The evidence will not entail that Moriarty is to blame, since it always remains possible that someone else was the perpetrator. Nevertheless, Holmes is right to make his inference, since Moriarty's guilt would provide a better explanation of the evidence than would anyone else's.

Inference to the Best Explanation can be seen as an extension of the idea of `self-evidencing' explanations, where the phenomenon that is explained in turn provides an essential part of the reason for believing the explanation is correct. For example, a star's speed of recession explains why its characteristic spectrum is red-shifted by a specified amount, but the observed red-shift may be an essential part of the reason the astronomer has for believing that the star is receding at that speed. Self-evidencing explanations exhibit a curious circularity, but this circularity is benign. The recession is used to explain the red-shift and the red-shift is used to confirm the recession, yet the recession hypothesis may be both explanatory and well-supported. According to Inference to the Best Explanation, this is a common situation in science: hypotheses are supported by the very observations they are supposed to explain. Moreover, on this model, the observations support the hypothesis precisely because it would explain them. Inference to the Best Explanation thus partially inverts an otherwise natural view of the relationship between inference and explanation. According to that natural view, inference is prior to explanation. First the scientist must decide which hypotheses to accept; then, when called upon to explain some observation, she will draw from her pool of accepted hypotheses. According to Inference to the Best Explanation, by contrast, it is by only by asking how well various hypotheses would explain the available evidence that she can determine which hypotheses merit acceptance. In this sense, Inference to the Best

Explanation has it that explanation is prior to inference.

There are two different problems that an account of induction in science might purport to solve. The problem of description is to give an account of the principles that govern the way scientists weigh evidence and make inferences. The problem of justification is to show that those principles are sound or rational, for example by showing that they tend to lead scientists to accept hypotheses that are true and to reject those that are false. Inference to the Best Explanation has been applied to both problems.

The difficulties of the descriptive problem are sometimes underrated, because it is supposed that inductive reasoning follows a simple pattern of extrapolation, with `More of the Same' as its fundamental principle. Thus we predict that the sun will rise tomorrow because it has risen every day in the past, or that all ravens are black because all observed ravens are black. This model of `enumerative induction' has however been shown to be strikingly inadequate as an account of inference in science. On the one hand, a series of formal arguments, most notably the so-called raven paradox and the new riddle of induction (see CONFIRMATION, PARADOXES OF), have shown that the enumerative model is wildly over-permissive, treating virtually any observation as if it were evidence for any hypothesis. On the other hand, the model is also much too restrictive to account for most scientific inferences. Scientific hypotheses typically appeal to entities and processes not mentioned in the evidence that supports them and often themselves unobservable and not merely unobserved, so the principle of More of the Same does not apply. For example, while the enumerative model might account for the inference that a scientist makes from the observation that the light from one star is red-shifted to the conclusion that the light from another star will be red-shifted as well, it will not account for the inference from observed red-shift to unobserved recession.

The best-known attempt to account for these `vertical' inferences that scientists make from observations to hypotheses about the often unobservable reality that stands behind them is the Deductive-Nomological model. According to it, scientists deduce predictions from a hypothesis (along with various other `auxiliary premises') and then determine whether those predictions are correct. If some of them are not, the hypothesis is disconfirmed; if all of them are, the hypothesis is confirmed and may eventually be inferred. Unfortunately, while this model does make room

for vertical inferences, it remains, like the enumerative model, far too permissive, counting data as confirming a hypothesis which are in fact totally irrelevant to it. For example, since a hypothesis (H) entails the disjunction of itself and any prediction whatever (H or P), and the truth of the prediction establishes the truth of the disjunction (since P also entails (H or P)), any successful prediction will count as confirming any hypothesis, even if P is the prediction that the sun will rise tomorrow and H the hypothesis that all ravens are black.

What is wanted is thus an account that permits vertical inference without permitting absolutely everything, and Inference to the Best Explanation promises to fill the bill. Inference to the Best Explanation sanctions vertical inferences, because an explanation of some observed phenomenon may appeal to entities and processes not themselves observed; but it does not sanction just any vertical inference, since obviously a particular scientific hypothesis would not, if true, explain just any observation. A hypothesis about raven coloration will not, for example, explain why the sun rises tomorrow. Moreover, Inference to the Best Explanation discriminates between different hypotheses all of which would explain the evidence, since the model only sanctions an inference to the hypothesis which would best explain it.

Inference to the Best Explanation thus has the advantages of giving a natural account of many inferences and of avoiding some of the limitations and excesses of other familiar accounts of non-demonstrative inference. If, however, it is to provide a serious model of induction, Inference to the Best Explanation needs to be developed and articulated, and this has not proven an easy thing to do. More needs to be said, for example, about the conditions under which a hypothesis explains an observation. Explanation is itself a major research topic in the philosophy of science, but the standard models of explanation yield disappointing results when they are plugged into Inference to the Best Explanation. For example, the best-known account of scientific explanation is the Deductive-Nomological model, according to which an event is explained when its description can be deduced from a set of premises that essentially includes at least one law. This model has many flaws (see EXPLANATION). Moreover, it is isomorphic to the Hypothetico-Deductive model of confirmation, so it would disappointingly reduce Inference to the Best Explanation to a version of hypothetico-deductivism.

The difficulty of articulating Inference to the Best Explanation is compounded when we turn to the question of what makes one explanation better than another. To begin with, the model suggests that inference is a matter of choosing the best from among those explanatory hypotheses that been proposed at a given time, but this seems to entail that at any time scientists will infer one and only one explanation for any set of data. Yet scientists are sometimes agnostic, unwilling to infer any of the available hypotheses, and they are also sometimes happy to infer more than one explanation, when the explanations are compatible. `Inference to the Best Explanation' must thus be glossed by the more accurate but less memorable phrase, `inference to the best of the available competing explanations, when the best one is sufficiently good'. But under what conditions is this complex condition is satisfied? How good is `sufficiently good'? Even more fundamentally, what are the factors that make one explanation better than another? Standard models of explanation are virtually silent on this point. This does not suggest that Inference to the Best Explanation is incorrect but, unless we can say more about explanation, the model will remain relatively uninformative.

Fortunately, some progress has been made in analyzing the relevant notion of the best explanation. We may begin by considering a basic question about the sense of `best' that the model requires. Does it mean the most probable explanation, or rather the explanation that would, if correct, provide the greatest degree of understanding? In short, should Inference to the Best Explanation be construed as inference to the likeliest explanation, or as inference to the loveliest explanation? A particular explanation may be both likely and lovely, but the notions are distinct. For example, if one says that smoking opium tends to put people to sleep because opium has a `dormative power', one is giving an explanation that is very likely to be correct but not at all lovely: it provides very little understanding. At first glance, it may appear that likeliness is the notion Inference to the Best Explanation ought to employ, since scientists presumably only infer the likeliest of the competing hypotheses they consider. This is, however, probably the wrong choice, since it would severely reduce the interest of the model by pushing it towards triviality. Scientists do infer what they judge to be the likeliest hypothesis, but the main point of a model of inference is precisely to say how these judgements are reached, to give what scientists take to be the symptoms of likeliness. To say that scientists

infer the likeliest explanations is perilously similar to saying that great chefs prepare the tastiest meals: true perhaps, but not very informative if one wants to know the secrets of their success. Like the dormative power explanation of the effects of opium, `Inference to the Likeliest Explanation' would itself be an explanation of scientific practice which provides only little understanding.

The model should thus be construed as `Inference to the Loveliest Explanation'. Its central claim is that scientists take loveliness as a guide to likeliness, that the explanation that would, if correct, provide the most understanding, is the explanation that is judged likeliest to be correct. This at least is not a trivial claim, but it raises at least three challenges. The first challenge is to identify the explanatory virtues, the features of explanations that contribute to the degree of understanding they provide. The second is to show that these aspects of loveliness match judgements of likeliness, that the loveliest explanations tend also to be those that are judged likeliest to be correct. The third challenge is to show that, granting the match between loveliness and judgments of likeliness, the former is in fact the scientists' guide to the latter.

To begin with the challenge of identification, there are a number of plausible candidates for the explanatory virtues, including scope, precision, mechanism, unification and simplicity. Better explanations explain more types of phenomena, explain them with greater precision, provide more information about underlying mechanisms, unify apparently disparate phenomena, or simplify our overall picture of the world. Some of these features, however, have proven surprisingly difficult to analyze. There is, for example, no uncontroversial analysis of unification or simplicity, and some have even questioned whether these are genuine features of scientific hypotheses, rather that mere artifacts of the way they happen to be formulated, so that the same explanation will count as simple if formulated in one way but complex if formulated in another.

A different but complementary approach to the problem of identifying some of the explanatory virtues focusses on the contrastive structure of many why-questions. A request for the explanation of some phenomenon often takes a contrastive form: one asks not simply `Why P?', but `Why P rather than Q?'. What counts as a good explanation depends not just on fact P but also on the foil Q. Thus the increase in temperature might be a good explanation of why the mercury in a thermometer rose

rather than fell, but not a good explanation of why it rose rather than breaking the glass. Accordingly, it is possible to develop a partial account of what makes one explanation of a given phenomenon better than another by specifying how the choice of foil determines the adequacy of contrastive explanations. Although many explanations both in science and in ordinary life specify some of the putative causes of the phenomenon in question, the structure of contrastive explanation shows why not just any causes will do. Roughly speaking, a good explanation requires a cause that `made the difference' between the fact and foil. Thus the fact that Smith had untreated syphilis may explain why he rather than Jones contracted paresis (a form of partial paralysis), if Jones did not have syphilis; but it will not explain why Smith rather that Doe contracted paresis, if Doe also had untreated syphilis. Not all causes provide lovely explanations, and an account of contrastive explanation helps to identify which do and which do not.

Assuming that a reasonable account of the explanatory virtues is forthcoming, the second challenge to Inference to the Best Explanation concerns the extent of the match between loveliness and judgments of likeliness. If Inference to the Best Explanation is along the right lines, then the lovelier explanations ought also in general to be judged likelier. Here the situation looks promising, since the features we have tentatively identified as explanatory virtues seem also to be inferential virtues, that is, features that lend support to a hypothesis. Hypotheses that explain many observed phenomena to a high degree of accuracy tend to be better supported that hypotheses that do not. The same seems to hold for hypotheses that specify a mechanism, that unify, and that are simple. The overlap between explanatory and inferential virtues is certainly not perfect, but at least some cases of hypotheses that are likely but not lovely, or conversely, do not pose a particular threat to Inference to the Best Explanation. As we have already seen, the dormative power explanation of opium's soporific effect is very likely but not at all lovely; but this is not threat to the model, properly construed. There surely are deeper explanations for the effect of smoking opium, in terms of molecular structure and neurophysiology, but these explanations will not compete with the banal account, so the scientist may infer both without violating the precepts of Inference to the Best Explanation.

The structure of contrastive explanation also helps to meet this

matching challenge, because contrasts in why-questions often correspond to contrasts in the available evidence. A good illustration of this is provided by Ignaz Semmelweis's nineteenth-century investigation into the causes of childbed fever, an often fatal disease contracted by women who gave birth in the hospital where Semmelweis did his research. Semmelweis considered many possible explanations. Perhaps the fever was caused by `epidemic influences' affecting the districts around the hospital, or perhaps it was caused by some condition in the hospital itself, such as overcrowding, poor diet, or rough treatment. What Semmelweis noticed, however, was that almost all of the women who contracted the fever were in one of the hospital's two maternity wards, and this led him to ask the obvious contrastive question and then to rule out those hypotheses which, though logically compatible with his evidence, did not mark a difference between the wards. It also lead him to infer an explanation that would explain the contrast between the wards, namely that women were inadvertently being infected by medical students who went directly from performing autopsies to obstetrical examinations, but only examined women in the first ward. This hypothesis was confirmed by a further contrastive procedure, when Semmelweis had the medics disinfect their hands before entering the ward: the infection hypothesis was now seen also to explain not just why women in the first rather than in the second ward contracted childbed fever, but also why women in the first ward contracted the fever before but not after the regime of disinfection was introduced. This general pattern of argument, which seeks explanations that not only would account for a given effect, but also for particular contrasts between cases where the effect occurs and cases where it is absent, is very common in science, for example wherever use is made of controlled experiments.

This leaves the challenge of guiding. Even if it is possible to give an account of explanatory loveliness (the challenge of identification) and to show that the explanatory and inferential virtues coincide (the challenge of matching), it remains to be argued that scientists judge that an hypothesis is likely to be correct <u>because</u> it is lovely, as Inference to the Best Explanation claims. Thus a critic of the model might concede that likely explanations tend also to be lovely, but argue that inference is based on other considerations, having nothing to do with explanation. For example, one might argue that inferences from contrastive data are really applications of Mill's method of difference (see MILL), which makes no

explicit appeal to explanation, or that precision is a virtue because more precise predictions have a lower prior probability and so provide stronger support as an elementary consequence of the probability calculus (see PROBABILITY).

The defender of Inference to the Best Explanation is here in a delicate position. In the course of showing that explanatory and inferential virtues match up, he will also inevitably show that explanatory virtues match some of those other feature that competing accounts of inference cite as the real guides to inference. The defender thus exposes himself to the charge that it is those other features rather than the explanatory virtues that do the real inferential work. Meeting the matching challenge will thus exacerbate the guiding challenge. The situation is not hopeless, however, since there are at least two ways to argue that loveliness is a guide to judgments of likeliness. As we have seen, at least some of the competing accounts of inference are fraught with difficulties, inapplicable to many scientific inferences and incorrect about others. If it is shown that Inference to the Best Explanation would give a better account of more inferences than any other available account, this is a powerful reason for supposing that loveliness is indeed a guide to likeliness. Secondly, if there is a good match between loveliness and likeliness, as the guiding challenge grants, this is presumably not a coincidence and so itself calls for an explanation. Why should it be that the hypotheses that scientists judge likeliest to be correct are also those that would provide the most understanding if they were correct? Inference to the Best Explanation gives a very natural answer to this question, similar in structure to the Darwinian explanation for the fact that organisms tend to be well-suited to their environments. If scientists select hypotheses on the basis of their explanatory virtues, the match between loveliness and judgments of likeliness follows as a matter of course. Unless the opponents of the model can give a better account of the match, the challenge has been met.

We have been considering the prospects of Inference to the Best Explanation as a partial solution to the problem of describing the structure of scientific inferences, but the model has also been applied to problems of justification. The most fundamental problem of inductive justification is due to David HUME, who argued that there can be no good reason to believe that our inductive practices are even moderately reliable, tending to take us from true observations to true hypotheses or predictions.

According to Hume, to justify induction we would have to produce a cogent argument whose conclusion is that induction is generally reliable and whose premises are not themselves inductively based. The only such premises are reports of past observation and the demonstrative truths of logic and mathematics. All cogent arguments are either deductive or inductive. Now we face a dilemma. There can be no cogent deductive argument for the reliability of induction, since no number of past observations (along with demonstrative truths) deductively guarantees that induction is generally reliable. In particular, past observations will never entail that induction will be reliable in the future. Neither is there a cogent inductive argument for induction, since any such argument presupposes the very practice it is supposed to justify. For example, to argue that induction is likely to be reliable in future on the grounds that it has been reliable in the past would beg the question, even if it were granted that the past reliability of induction could itself be known on the basis of observation. Hence our inductive practices are unjustifiable.

If Hume's argument is sound, there is no reason whatever to believe any scientific claim that goes beyond what has been directly observed, which is at least to say that there is no reason to believe any scientific prediction, hypothesis or theory. This is incredible, but the sceptical argument has proven extraordinarily resilient and there is still no generally accepted answer to it. For all of Hume's sophistication in presenting the problem of justification, however, his solution to the problem of description is rather primitive. He seems to have accepted a version of the simple enumerative `More of the Same' model of induction discussed above. Consequently, one might hope that a more sophisticated and accurate account of inductive practice would make it possible to avoid or rebut Hume's sceptical argument. In particular, it is sometimes supposed that Inference to the Best Explanation provides such an account.

Unfortunately, Inference to the Best Explanation does not solve Hume's problem. The description he gave of induction was incorrect, but his sceptical argument does not depend on it. Indeed the argument seems to depend on little more than the undeniable fact that inductive arguments are not deductively valid. Reports of past observation will never entail that future inferences to the best explanations will in fact select true hypotheses; and any argument that the reliability of inference to the best explanation would itself be the best explanation of what we have observed

begs the question. It might even be claimed that Inference to the Best Explanation exacerbates the problem of justification, since it is quite unclear why the hypothesis that would, if correct, provide the deepest understanding is also in fact likeliest to be correct. Why should we suppose that ours is the loveliest of all possible worlds? This additional worry may, however, be an overreaction, since what Hume's sceptical argument suggests is that the success of any other method of induction would be equally mysterious.

Inference to the Best Explanation has also been invoked to solve more modest problems of inductive justification. Even if the model is of no avail against a complete inductive sceptic, it might have a role to play in the defence of scientific realism, according to which there are good reasons to believe that well-supported theories are likely to be at least approximately true, against positions such as constructive empiricism, according to which we can only have reason to believe that our best theories are empirically adequate, that their observable consequences are true. (Constructive empiricism has been developed in detail by Bas Van Fraassen, who is also a vigorous critic of Inference to the Best Explanation.) The constructive empiricist is no inductive sceptic, since to say that all the observable consequences of a theory are true is a much stronger claim than to say merely that its observed consequences are true; but the realist goes further by sanctioning in addition vertical inferences to the truth of a theory's claims about unobservable entities and processes.

Perhaps the best known example of this application of Inference to the Best Explanation in defence of scientific realism is the so-called `miracle argument', discussed by Hilary PUTNAM. He takes it that the model provides a good solution to the descriptive problem and proposes that philosophers may themselves make an inference to the best explanation in defence of scientific realism. Suppose that all the many and varied predictions derived from a particular scientific theory are found to be correct: what is the best explanation of this predictive success? According to Putnam, the best explanation is that the theory itself is true. If the theory were true, then the truth of its deductive consequences would follow as a matter of course; but if the hypothesis were false, it would be a `miracle' that all its observed consequences were found to be correct. So, by a philosophical application of Inference to the Best Explanation, we are entitled to infer that the theory is true, since the `truth-explanation' is the

best explanation of the theory's predictive success. This higher-level inference is supposed to be distinct from the first-order inferences scientists make, but of the same form.

This justificatory application of Inference to the Best Explanation has considerable intuitive appeal, but it faces three objections. The first is that the truth-explanation for the predictive success of a theory is not really distinct from the substantive scientific explanations that the theory provides and on the basis of which it was inferred by scientists in the first place. If this is so, then the miracle argument provides no additional reason to believe that the hypothesis is correct: it is merely a repetition of the scientific inference it was supposed to justify. This objection can be answered, however, by observing that the two sorts of explanation have a different structure. The scientific explanations a theory provides are typically causal, whereas the truth-explanation is logical. The truth of a theory does not physically cause its consequences to be true; the explanatory connection is rather that a valid argument with true premises must also have a true conclusion.

The second objection to the miracle argument is that, even if the truth explanation is distinct from the scientific explanations, the inference to the truth of the theory is vitiated by the same sort of circularity that Hume appealed to in his sceptical argument. In effect, the miracle argument is an attempt to use an inference to the best explanation to justify scientific inferences to the best explanation so, the objector will claim, such an argument must beg the question of the reliability of this form of inference. In particular, the constructive empiricist may insist that, although he will allow the legitimacy of some forms of induction, inferences to the truth of theories that traffic in unobservables are precisely those that are at issue. One possible response to the circularity objection is to argue that the circle is broken in virtue of the difference between inferences to causal and to logical explanations, but the objection has considerable force.

The third objection to the miracle argument is that truth is simply not the best explanation of predictive success, so the argument fails on its own terms. The obvious way to flesh out this objection is to give another explanation that is at least as good. For example, the constructive empiricist may claim that we can explain the predictive success of a theory by supposing that it is empirically adequate, that all its observable

consequences are true, whether or not the theory is true as a whole. In this case, however, the defender of the miracle argument has two ready replies. Firstly, it is far from clear that the explanation in terms of empirical adequacy is as lovely as the truth explanation, since it is dangerously close to saying that the consequences of the theory are true because they are true, an extremely unlovely explanation, reminiscent of the appeal to opium's dormative power. Moreover, even if, as in the opium case, we infer this explanation, it does not preclude an inference to the truth-explanation, since the two explanations are compatible: a theory may be both empirically adequate and true. The third objection to the miracle argument can however be made more pressing through a better choice of alternative explanations. For given any set of successful predictions, there are always in principle many theories incompatible with the original one which nevertheless share those consequences (see UNDERDETERMINATION). The truth of any of the competing theories would also explain the predictive success they share with the original theory and it is unclear that these alternative truth explanations would be any less lovely than the original. The inference to the truth of the original theory may thus be blocked.

Neither of the justificatory applications of Inference to the Best Explanation we have considered appears promising. If the model can help to solve problems of inductive justification, these are likely to concern more specific aspects of scientists' inductive practices. For example, the model has been plausibly applied in an argument to show why it is rational for scientists to put greater weight on data that an hypothesis correctly predicts than on data that was available when the hypothesis was formulated and which it was constructed to accommodate. Whatever the justificatory potential of Inference to the Best Explanation, however, the model may be counted a philosophical success if it can be shown to give an illuminating description of some of the general inferential principles that guide scientific practice.

Bibliography

Fraassen, B. van: The Scientific Image (Oxford: Oxford University Press, 1980).

Garfinkel, A.: Forms of Explanation (New Haven: Yale University Press, 1981).

Harman, G.: `The inference to the best explanation', The Philosophical Review, 74 (1965), 88-95.

Hempel, C.: Aspects of Scientific Explanation (New York: Free Press, 1965).

Hume, D.: An Enquiry Concerning Human Understanding (London:1777); (Oxford:1975), sections 4,5.

Lipton, P.: Inference to the Best Explanation (London: Routledge, 1991).

Pierce, C.S.: Collected Papers eds. C. Hartshorn and P. Weiss (Cambridge, MA: Harvard University Press, 1931), 5.180-5.189.

Putnam, H.: Meaning and the Moral Sciences (London: Hutchinson, 1978), pp. 18-22.

Thagard, P.: `The best explanation: criteria for theory choice', The Journal of Philosophy, 75 (1978), 76-92.