form $a = a$ and truths of the form $a = b$. Generally hostile to Frege, so it's especially important to read this if you feel persuaded by Frege.

SKYRMS, B. (1982), 'Intensional Aspects of Semantical Self-Reference', in Martin (1984). Mentioned for the development of the idea that the Liar paradox can be blocked by denying that substitution of identicals is valid in the relevant contexts. Probably will be followed up only by enthusiasts.

STRAWSON, P. F. (1950*a*), *'On Referring', *Mind*, 59: 269–86; repr. in *Logico-Linguistic Papers* (London, 1971). Essential reading. Famous attack on Russell's theory of descriptions. Important for distinctions between sentences, uses of sentences, and statements.

—— (1950*b*), 'Truth', *Proceedings of the Aristotelian Society*, supp. vol. 24: 129–56; repr. in Pitcher (1964). Classic paper on the correspondence theory; part of a symposium with Austin (1950).

—— (1965), 'Truth: A Reconsideration of Austin's Views', *Philosophical Quarterly*, 15: 289–301; repr. in *Logico-Linguistic Papers* (London, 1971). This concludes the story which begins with Austin (1950) and Strawson (1950).

—— (1967), 'Is Existence Never a Predicate?', *Critica*, 1: 5–15; repr. in *Freedom and Resentment and Other Essays* (London, 1974). Defends a Meinongian position, according to which there are some things, in particular fictional characters, which do not exist. Shows that in ordinary talk we are quite happy with the idea that there are things which do not exist.

TARSKI, A. (1937), 'The Concept of Truth in Formalized Languages'; repr. in *Logic, Semantics, Metamathematics* (Oxford, 1956). The canonical exposition of his definition of truth. Most of the article is extremely technical, but the early pages are informal and worth looking at.

—— (1969), 'Truth and Proof', *Scientific American*, 194: 63–77. Informal exposition of some of his main views. If you want to read some Tarski, this is the first piece to choose.

WIGGINS, D. (1976), 'Frege's Problem of the Morning Star and the Evening Star', in M. Schirn (ed.), *Studies on Frege* (Stuttgart). Useful detailed discussion of Frege's notion of sense. Argues that even expressions with the same sense (by intuitive standards) cannot be substituted *salve veritate* in every context in which they are used.

—— (1980*a*), *Sameness and Substance* (Oxford). Not really suitable for this subject area (though excellent for epistemology and metaphysics). Mentioned here for the use of the Barcan Marcus proof of the necessity of identity.

—— (1980*b*), 'What would be a Substantial Theory of Truth?', in Z. van Straaten (ed.), *Philosophical Subjects: Essays Presented to P. F. Strawson* (Oxford). Useful attempt to connect the notion of truth with other notions: assertion, convergence, etc.

WRIGHT, C. (1992), *Truth and Objectivity* (Cambridge, Mass.). Impressive recent book, arguing for a version of minimalism about truth. It's too detailed to be what the average beginner needs. But if you want to make a serious study of the realism–anti-realism debate, this book is for you.

3

METHODOLOGY: THE ELEMENTS OF THE PHILOSOPHY OF SCIENCE

David Papineau

INTRODUCTION

The subject-matter of *methodology* is best defined in opposition to that of *logic*. Logic is the study of deductively valid reasoning: in a deductively valid argument, the premisses provide conclusive reasons for the conclusion; it is quite impossible for the premisses to be true and the conclusion false. However, most of the reasoning that we actually engage in falls far short of this ideal. In both everyday life and in science the arguments we use do not provide conclusive reasons for their conclusions. They may in some sense give us good reason to believe their conclusions, but they do not compel us in the same absolute way as deductive arguments.

This discussion of methodology will be concerned with this kind of non-conclusive reasoning and with various philosophical issues that arise in trying to understand it. There will be five sections: (1) Induction and its Problems; (2) Laws of Nature; (3) Realism, Instrumentalism, and Underdetermination; (4) Confirmation and Probability; (5) Explanation.

1. INDUCTION AND ITS PROBLEMS

1.1. The Problem of Induction

Generally speaking, 'induction' refers to any form of inference in which we move from a finite set of observations or experimental results to a conclusion about how things generally behave. There are various forms of inductive inference, but we shall concentrate on simple *enumerative inductions*, which start from the premiss that one phenomenon has always followed another so far, and conclude that those phenomena will always occur together. So, for example, you might note that, every time you have seen red sky in the evening, there has been fine weather the next day, and conclude on that basis that red sky in the evening is *always* followed by fine weather. Or you might note that all the samples of sodium you have heated on a Bunsen burner have glowed bright orange, and conclude on this basis that in general *all* heated sodium glows bright orange. Schematically, the premiss to an enumerative induction is that '*n As* have all been observed to be *Bs*', and the conclusion is that 'All *As* are *Bs*'.

Note that these inductive inferences start with particular premisses about a *finite* number of past observations, yet end up with a general conclusion about how nature will *always* behave. This is the source of the notorious *problem of induction*. For it is unclear how any finite amount of information about what has happened in the past can guarantee that a natural pattern will continue for all time.

After all, what rules out the possibility that the course of nature might change, and that the patterns we have observed so far turn out to be a poor guide to the future? Even if all red-skyed evenings have been followed by fine weather so far, who is to say that they won't start being followed by rain in the next century? Even if all heated sodium has glowed orange up till now, who is to say it won't start glowing blue at some future date?

In this respect induction contrasts with deduction. In deductive inferences the premisses really do guarantee the conclusion. For example, if you know that 'Either this substance is sodium or it is potassium', and then learn further that 'It is not sodium', you can conclude with certainty that 'It is potassium'. The truth of the premisses leaves no room for the conclusion to be anything but true. But in an inductive inference this does not hold. If you are told that 'Each of the *As* observed so far has been *B*', this does not guarantee that 'All *As*, including future ones, are *Bs*'. It is perfectly possible that the former claim may be true, but the latter false.

I have illustrated the problem of induction with respect to enumerative inductions. There are other forms of induction apart from enumerative induction, as we shall see later. But the problem of induction is quite general. For what the different forms of induction have in common is that they take us from information about a finite number of instances to some general conclusion about a wider class of cases. Since nothing in logic seems to guarantee that the wider class will display the same behaviour as the finite instances, any such inference is for this reason equally problematic.

The problem of induction threatens both everyday and scientific knowledge. Most of the everyday knowledge we rely on consists of general principles like '*Whenever* you cut yourself, you bleed', or '*Whenever* the brakes are applied, cars stop'. Similarly, all scientific discoveries worth the name are in the form of general principles: Galileo's law of free fall says that '*All* bodies fall with constant acceleration'; Newton's law of gravitation says that '*All* bodies attract each other in proportion to their masses and in inverse proportion to the square of the distance between them'; Avogadro's law says that '*All* gases at the same temperature and pressure contain the same number of molecules per unit volume'; and so on. The problem of induction calls in question the authority of all these general claims. For if our evidence is simply that these generalizations have worked so far, then how can we be sure that they will not be disproved by future occurrences?

1.2. Initial Responses to the Problem

1.2.1. *A Principle of Induction*

One possible response to the problem of induction would be to appeal to some 'principle of induction' which asserts that, for some number *N*,

(P) For any α and β, whenever N αs are observed to be βs, then all αs are βs.

If such a principle were available, then we could add it to the original premiss of any enumerative induction—namely, that *N* (or more) *As* have been observed to be *Bs*—to conclude deductively that 'All *As* are *Bs*'. For once we add (P) as a premiss, then there is no longer any room for the premisses of the induction to be true and the conclusion to be false.

However, even if we leave to one side the question how big *N* needs to be to make (P) plausible, there is an obvious difficulty about the status of the proposed principle. Clearly (P) is not an analytic claim whose truth is guaranteed by its meaning: you could understand all the terms in it, yet not believe it. So it must be a synthetic claim, in need of support by empirical evidence. But since (P) is a generalization, this support would have to be some kind of inductive argument, taking as its premisses some finite body of instances where inductive inferences have worked in the past, and seeking to move to (P) as a conclusion. So in the present context of argument this would beg the question at issue, which is to defend inductive arguments against the challenge raised by the problem of induction.

1.2.2. *Inductive Arguments for Induction*

Suppose we abandon any principle of induction, and thereby accept that we cannot make inductive arguments deductive. Still, cannot we simply argue that inductive arguments are nevertheless acceptable because they *work*? Even if the premisses don't logically *guarantee* the conclusions, don't the conclusions normally turn out to be true anyway? After all, hasn't our experience shown us that patterns like red-sky–good-weather or sodium-heated–orange-flame continue to hold good in the future, once they have displayed themselves in the past?

But this suggestion runs into the same difficulty as the last one. We are arguing that inductions are generally acceptable because our experience has shown them to work so far. But this is itself an inductive argument. After all, even if observed patterns have tended to hold good so far, what guarantees that they will continue to do so? As Bertrand Russell once said, it is no help to observe that *past* futures have conformed to *past* pasts; what we want to know is whether *future* futures will conform to *future* pasts. Given that we are trying to vindicate induction against objections, an inductive argument for induction once more begs the question.

1.2.3. *Introducing Probability*

Another possible response to the problem of induction is to regard inductive inferences as merely generating *probable* conclusions, rather than certain ones.

Even if past evidence doesn't allow us to be sure about future patterns, might it not at least support conclusions about *probable* patterns?

Later on we shall see that the idea of probability is indeed important for our understanding of inductive arguments. But it is not difficult to show that on its own it is not enough to solve the problem of induction.

In fact, as we shall see later, there are really two notions of probability. Roughly, we need to distinguish probability in the sense of rational *degree of belief* from probability in the sense of *objective tendency*. When we say that it is 50 per cent probable that it will snow today, we might mean one of two things. First, we might be expressing a degree of belief: saying that we have an equal expectation both for its snowing and for its not snowing today. Alternatively, we might be making a claim about an objective tendency: saying that in general it snows on 50 per cent of days like today. Later on we shall look at these 'subjective' and 'objective' interpretations of probability in more detail. Here I merely want to show that neither helps with the problem of induction.

Suppose first that the conclusion of an inductive inference is a statement of *objective* probability, stating that in 90 per cent of cases, say, *As* turn out to be *Bs* (for example, that on 90 per cent of days following red-skyed evenings there is fine weather). The evidence for this claim will still be a finite body of observations, namely, that in our experience *so far* more or less 90 per cent of *As* have been *Bs*. So the problem of induction is still with us, for we still need to explain how a finite body of evidence can establish a general conclusion. For note that the probabilistic conclusion is still a claim, requiring not just that 90 per cent of *As* have been *Bs* in the *past*, but also that this will continue in the *future*. Even if the pattern we are now interested in is probabilistic, rather than exceptionless, we still face the same difficulty in explaining how past patterns can tell us about future ones.

Alternatively, we might take the conclusion of an inductive inference to be a statement of *subjective* probability, asserting that 'We should attach a high degree of belief to the proposition that all *As* are *Bs*'. (Note that we could also have a statement of subjective probability about a proposition of objective probability: for example, 'We should attach a high degree of belief to the proposition that 90 per cent of *As* are *Bs*'. The point which follows would apply just the same.) The difficulty once more is that our evidence for such a conclusion about subjective probability is simply that *As* have been observed to go with *Bs* *so far*. Yet the conclusion says that we should have a high expectation that *As* will go with *Bs* in the future as well as the past. So we still face the problem of explaining how facts about the past can tell us what to think about the future.

1.3. Popper's Alternative to Induction

A rather different line of response to the problem of induction is due to Karl Popper. Popper looks to the practice of science to show us how to deal with the problem. In Popper's view, science does not rest on induction in the first place. He denies that scientists start with observations and then infer a general theory. Rather, they first put forward a theory, as an initially uncorroborated conjecture, and then compare its predictions with observations to see whether it stands up to test. If such tests prove negative, then the theory is experimentally falsified and the scientists will seek some new alternative. If, on the other hand, the tests fit the theory, then scientists will continue to uphold it—not as proven truth, admittedly, but nevertheless as an undefeated conjecture.

If we look at science in this way, argues Popper, then we see that it does not need induction. According to Popper, the inferences which matter to science are *refutations*, which take some failed prediction as the premiss and conclude that the theory behind that prediction is false. These inferences are not inductive, but deductive. We see that some *A* is not-*B*, and conclude that it is not the case that all *As* are *Bs*. There is no room here for the premiss to be true and the conclusion false. If we discover that some piece of sodium does not glow orange when heated, then we know for sure that it is not the case that all heated sodium glows orange. The point here is that it is much easier to disprove theories than to prove them. A single contrary example suffices for a conclusive disproof, but no number of supporting examples will constitute a conclusive proof.

So, according to Popper, science is a sequence of conjectures. Scientific theories are put forward as hypotheses, and they are replaced by new hypotheses when they are falsified. However, this view of science raises an obvious question: if scientific theories are always conjectural, then what makes science better than astrology, or spirit-worship, or any other form of unwarranted superstition? A non-Popperian would answer this question by saying that real science *proves* its claims on the basis of observational evidence, whereas superstition is nothing but guesswork. But, on Popper's account, even scientific theories are guesswork—for they cannot be proved by the observations, but are themselves merely undefeated conjectures.

Popper calls this the 'problem of demarcation'—what is the difference between science and other forms of belief? His answer is that science, unlike superstition, is at least *falsifiable*, even if it is not provable. Scientific theories are framed in precise terms, and so issue in definite predictions. For example, Newton's laws tell us exactly where certain planets will appear at certain times. And this means that if such predictions fail, we can be sure that the theory behind them is false. By contrast, belief systems like astrology are irredeemably vague, in a way which prevents their ever being shown definitely wrong. Astrology may predict that Scorpios will prosper in their personal relationships on Thursdays, but when faced with a Scorpio whose spouse walks out on a Thursday, defenders of astrology are likely to respond that the end of the marriage was probably for the best, all things considered. Because of this, nothing will ever force astrologers to admit their theory is wrong. The theory is phrased in such imprecise terms that no actual observations can possibly falsify it.

Popper himself uses this criterion of *falsifiability* to distinguish genuine science, not just from traditional belief systems like astrology and spirit-worship, but also from Marxism, psychoanalysis, and various other modern disciplines he denigrates as 'pseudo-sciences'. According to Popper, the central claims of these theories are as unfalsifiable as those of astrology. Marxists predict that proletarian revolutions will be successful whenever capitalist regimes have been sufficiently weakened by their internal contradictions. But when faced with unsuccessful proletarian revolutions, they simply respond that the contradictions in those particular capitalist regimes have not yet weakened them sufficiently. Similarly, psychoanalytic theorists will claim that all adult neuroses are due to childhood traumas, but when faced by troubled adults with apparently undisturbed childhoods, they will say that those adults must nevertheless have undergone private psychological traumas when young. For Popper, such ploys are the antithesis of scientific seriousness. Genuine scientists will say beforehand what observational discoveries would make them change their minds, and will abandon their theories if these discoveries are made. But Marxists and psychoanalytic theorists frame their theories in such a way, argues Popper, that no possible observations need ever make them adjust their thinking.

1.4. The Failings of Falsificationism

At first sight Popper seems to offer an attractive way of dealing with the problem of induction. However, there is reason to doubt whether he really offers a solution.

The central objection to his account is that it only accounts for *negative* scientific knowledge, as opposed to *positive* knowledge. Popper points out that a single counter-example can show us that a scientific theory is wrong. But he says nothing about what can show us that a scientific theory is right. Yet it is positive knowledge of this latter kind that is supposed to follow from inductive inferences. What is more, it is this kind of positive knowledge that makes induction so important. We can cure diseases and send people to the moon because we know that certain causes *do* always have certain results, not because we know that they *don't*. If Popper cannot explain how we sometimes know that 'All *As* are *Bs*', rather than just 'It's false that all *As* are *Bs*', then he has surely failed to deal properly with the problem of induction.

Popper's usual answer to this objection is that he is concerned with the logic of pure scientific research, not with practical questions about technological applications. Scientific research requires only that we formulate falsifiable conjectures and reject them if we discover counter-examples. The further question whether we should *believe* those conjectures, and *rely* on their predictions when, say, we prescribe some drug or build a dam, Popper regards as an essentially practical issue, and as such not part of the analysis of rational scientific practice.

But this will not do, if Popper is supposed to be offering a solution to the problem of induction. The problem of induction is essentially the problem of how we can base judgements about the future on evidence about the past. In insisting that scientific theories are just conjectures, and that therefore we have no rational basis for *believing* their predictions, Popper is simply denying that we can make rational judgements about the future.

Consider these two predictions:

(A) When I jump from this tenth-floor window, I shall crash painfully into the ground.

(B) When I jump from the window, I will float like a feather to a gentle landing.

Intuitively, it is more rational to believe (A), which assumes that the future will be like the past, than (B), which does not. But Popper, since he rejects induction, is committed to the view that past evidence does not make any beliefs about the future more rational than any others, and therefore to the view that believing (B) is no less rational than believing (A).

Something has gone wrong. *Of course* believing (A) is more rational than believing (B). In saying this, I do not want to deny that there is a *problem* of induction. Indeed it is precisely *because* believing (A) is more rational than believing (B) that induction is problematic. Everybody, Popper aside, can see *that* believing (A) is more rational than believing (B). The problem is then to explain *why* believing (A) is more rational than believing (B), in the face of the fact that induction is not logically compelling. So Popper's denial of the rational superiority of (A) over (B) is not so much a *solution* to the problem of induction, but simply a refusal to recognize the problem in the first place. As a reviewer of one of Popper's books once said, Popper's attitude to induction is like that of someone who stands on the starting-line of a race and shouts, 'I've won, I've won'.

Even if it fails to deal with induction, it should be recognized that Popper's philosophy of science does have some strengths as a description of pure scientific research. For it is certainly true that many scientific theories start life as conjectures, in just the way Popper describes. When Einstein's general theory of relativity was first proposed, for example, very few scientists actually believed it. Instead they regarded it as an interesting hypothesis, and were curious to see whether it was true. At this initial stage of a theory's life, Popper's recommendations make eminent sense. Obviously, if you are curious to see whether a theory is true, the next step is to put it to the observational test. And for this purpose it is important that the theory is framed in precise enough terms for scientists to work out what it implies about the observable world—that is, in precise enough terms for it to be falsifiable. And of course if the new theory does get falsified, then scientists will reject it and seek some alternative, whereas if its predictions are borne out, then scientists will continue to investigate it.

Where Popper's philosophy of science goes wrong, however, is in holding that scientific theories never progress beyond the level of conjecture. As I have just agreed, theories are often mere conjectures when they are first put forward, and they may remain conjectures as the initial evidence first comes in. But in many cases the accumulation of evidence in favour of a theory will move it beyond the status of conjecture to that of established truth. The general theory of relativity started life as a conjecture, and many scientists still regarded it as hypothetical even after Sir Arthur Eddington's famous initial observations in 1919 of light apparently bending near the sun. But by now this initial evidence has been supplemented with evidence in the form of gravitational red-shifts, time-dilation, and black holes, and it would be an eccentric scientist who nowadays regarded the general theory as less than firmly established.

This example can be multiplied. The heliocentric theory of the solar system, the theory of evolution by natural selection, and the theory of continental drift all started life as intriguing conjectures, with little evidence to favour them over their competitors. But in the period since they were first proposed these theories have all accumulated a great wealth of supporting evidence, and nearly everybody who is acquainted with this evidence has no doubt that these theories are well-established truths.

1.5. Induction is Rational by Definition

I have just insisted, against Popper, that it is often rational to believe the conclusions of inductive inferences. However, as I said, this observation is by no means a solution to the problem of induction. For we still need to explain how inductive inferences can be rational, give that their conclusions are not logically guaranteed by their premisses.

Some philosophers have argued that we can solve the problem by focusing on the everyday meaning of the term 'rational'.[1] After all, they point out, in normal usage this term is by no means restricted to deductive reasoning. True, everybody recognizes that deductive reasoning is *one* species of rational argument. But at the same time nearly everybody also applies the term 'rational' to other kinds of reasoning, and in particular to inductive reasoning.

By way of illustration, consider three different ways of forecasting the weather. The first type of forecaster does not pay any attention to past weather patterns, but simply guesses at random at tomorrow's weather. A second type of forecaster does attend to past patterns, but predicts future weather on the

[1] See Paul Edwards, 'Russell's Doubts about Induction', *Mind*, 68 (1949), 141–63; and section 9 in P. F. Strawson, *Introduction to Logical Theory* (London, 1952).

basis of the assumption that future weather patterns are going to be different from past patterns: so, for example, on seeing red sky in the evening, this forecaster reasons that, since red sky has presaged fine weather in the past, tomorrow's weather will not be fine. The third forecaster works on the assumption that past weather patterns are indeed a guide to future patterns, and so, on the basis of past experience, takes red sky in the evening to be a sign that there will be fine weather tomorrow.

Now, if we ask people who understand the meaning of the word 'rational' which of these three weather forecasters is rational, there is no doubt that they will reply that the third forecaster is rational, and the other two are not. And there is no doubt that they would also say that in general people who anticipate the future on the basis of the past are rational, and those who merely guess, or expect the future to be unlike the past, are irrational.

Doesn't this now show that induction is rational? For what more could be needed to show this than that people who understand the meaning of the term 'rational' all agree that this term is applicable to inductive reasoning?

This form of argument is known as the 'paradigm case argument', and was very popular among British 'ordinary language philosophers' in the 1950s and 1960s. It was applied to other philosophical problems apart from the problem of induction. So, for example, in response to the thesis that human beings do not really have free will, ordinary language philosophers pointed out that anybody who understands the phrase 'acting of their own free will' will have no hesitation in applying it to a wide range of human actions. After all, the ordinary language philosophers argued, aren't such actions as drinking a cup of coffee or buying a new car paradigm cases of free actions, as we ordinarily use the term? And what more could be needed to show that free will exists than that people who understand the meaning of the term 'free will' all agree that it applies to this kind of human action?

However, this example also serves to bring out the weakness of paradigm case arguments. The only reason some philosophers doubt the existence of free will is because they think there is an underlying requirement for an action to be free, namely, that it not be determined by past causes, and because, moreover, they doubt that any human actions are not so determined. Any such philosopher will reply to the paradigm case argument for free will as follows: 'Maybe ordinary people are happy to apply the term "free will" to such actions as drinking a cup of coffee or buying a new car. But this is only because they are implicitly assuming that these actions are not determined by past causes. But in fact they are wrong in this assumption. All human actions are determined by past causes. So there is really no free will, and everyday people are just making a mistake when they apply the term "free will" as they do.'

The same point applies to the attempt to establish the rationality of induction by appeal to ordinary usage. For ordinary usage leaves it open that there may be some underlying requirement for a form of inference to count as rational. And inductive inferences may in fact fail to satisfy this requirement, despite the inclination of ordinary people to apply the term 'rational' to induction.

1.6. A Reliabilist Defence of Induction

Let us ask, then, whether there is some underlying requirement which a form of inference must satisfy if it is to qualify as rational. Well, a minimum requirement is surely that the conclusions of these inferences must generally be true, if the premisses are. The whole point of inferences is to increase our stock of knowledge. Inferences make new knowledge out of old: they take old knowledge as input, and generate new knowledge as output. But a form of inference will fail in this task if it issues in false conclusions even when provided with true premisses. For in such cases the inference will not be increasing our stock of knowledge, but rather leading us into error.

It is important to recognize that this requirement—that the conclusions of a form of inference should generally be true if its premisses are—does not necessarily amount to the requirement that the form of inference should be *deductively valid*. A form of inference is deductively valid if it is logically quite *impossible* for the conclusions to be false if the premisses are true. This is far stronger than the requirement that *as a matter of fact* the conclusions are never false when the premisses are true. By way of illustration, consider this form of inference.

$$\frac{X \text{ is a human}}{X \text{ is less than 200 years old}}$$

This is not deductively valid. It is logically *possible* for someone to be a human and to live for 200 years. But, as it happens, there are no such human beings, and so this form of inference will never in fact take us from a true premiss to a false conclusion. (Of course, this form of inference can be made deductive by adding the premiss that 'All humans are less than 200 years old'. But my point is that, even if we don't add this premiss, and so stop the inference being deductive, it still satisfies the requirement of never going from true premisses to false conclusions.)

Let us use the term 'reliable' for the requirement that true premisses should always yield true conclusions. Then deductively valid inferences can be thought of as inferences that are *necessarily* reliable. In the terminology of possible worlds, a deductively valid inference will generate true conclusions out of true premisses in *every* possible world. A reliable but non-deductive inference, by contrast, always generates true conclusions out of true premisses in the actual world, but would go astray in other possible worlds (such as worlds, say, in which humans live for more than 200 years).

Given this distinction, it seems clear that reliability is a minimal requirement for a form of inference to be acceptable. However, to ask in addition for deductive validity seems like overkill. If we have a form of inference which works in the actual world, why require in addition that it should also work in every other possible world, however unlikely or outlandish?

These points support the ordinary language philosophers in their insistence that deductively valid inferences do not exhaust the category of rational inferences. But they suggest a different kind of reason for recognizing some non-deductive forms of inference as rational. The ordinary language philosophers were prepared to count as rational any form of inference that normal speakers of English call 'rational'. The points made in this section, however, argue that a form of inference should only count as rational if it satisfies the underlying requirement of reliably transmitting truth from premisses to conclusion.

It should be said that it is a matter of active controversy whether reliability is sufficient for rationality. This issue is part of a widespread contemporary debate which involves not only the notion of rationality, but also such related notions as *knowledge* and *justification*. Few contemporary philosophers, I think, would still want to say that a belief is rational (knowledge, justified) only if it is arrived at in ways that are necessarily reliable (such as by deductive inference). But among the remainder there is a split, between those (let us call them 'reliabilists' henceforth) who think that a reliable source on its own suffices for a belief to be rational (knowledge, justified) and those who think that some further requirement, such as intuitive persuasiveness, also needs to be satisfied.

However, there is no question of resolving this wider dispute here. So in the remainder of this section I shall discuss the following conditional thesis: *if* you think the reliability of a form of inference is sufficient for its rationality, *then* you will have an answer to the problem of induction.

Note first that, if we do adopt the reliabilist point of view, the original argument against induction ceases to present a substantial problem. For the original argument was simply that the premisses of an inductive argument do not deductively imply its conclusion. But since we are no longer demanding that inductive arguments should be logically infallible, but only that they in fact reliably transmit truth, this is no argument against induction at all. For, as I have emphasized, a form of inference can be reliable without being deductively valid.

This is only part of a reliabilist defence of induction. For, even if the traditional argument fails to show that induction is not reliable, the reliabilist still needs to provide grounds for thinking that induction is reliable. Unlike the ordinary language philosopher, the reliabilist cannot simply defend induction on the grounds that most people regard it 'rational'. For, according to reliabilism, a form of inference is only rational if it satisfies the underlying requirement of reliably transmitting truth from premisses to conclusion.

But perhaps the reliabilist can answer this challenge. The issue is whether inductive inferences generally yield true conclusions if given true premisses. The reliabilist can point out that there is plenty of evidence that they do. When people have made inductions from true premisses in the past, the reliabilist can argue, their conclusions have turned out true. So we can infer, from this evidence, that inductive inferences are in general reliable transmitters of truth.

Of course this last step is itself an inductive inference, from the past success of inductions to their general reliability, and so this argument is simply a version of the inductive defence of induction I accused of begging the question in Section 1.2.2. However, at that stage we were assuming that the traditional argument raises a genuine problem for induction, and that therefore it would be illegitimate to use induction in further philosophical analysis. But the first point made by the reliabilist defence of induction was that the traditional argument, which merely points out that induction is not deduction, does nothing at all to discredit induction. So why shouldn't we use our normal inductive methods to ascertain whether induction is reliable? How else, the reliabilist can reasonably ask, are we expected to address the question?

Of course this kind of inductive defence of induction is not going to persuade somebody who does not already make inductions to start making them, for such a person will be disinclined to conclude, from the premiss that inductions have worked in the past, that they will do so in the future. But the reliabilist can respond that the inductive argument for induction is not supposed to cure any eccentrics who might reject induction. Rather, it is simply supposed to explain, to normal people like ourselves, how we are entitled to the view that induction is reliable, and hence rational.

Not all philosophers would agree that this reliabilist defence of induction avoids begging the question. But at this stage I propose to leave this issue and turn instead to a more direct objection. This defence assumes that inductions with true premisses have at least generated true conclusions so far, as the premiss to the inductive argument for induction. But is even this true? Aren't there plenty of cases where people have made inductions, and yet have arrived at false rather than true conclusions?

Clearly this is a challenge reliabilists need to answer. For even if we grant them the legitimacy of inductive arguments for induction, reliabilists aren't going to get anywhere if the past evidence indicates that induction is not reliable.

I shall examine two sorts of reason for thinking that induction is downright unreliable. One sort appeals to the history of science and notes that many inductively supported scientific theories, from Ptolemaic astronomy to Newtonian physics, have been shown by later evidence to be false. However, I shall postpone discussion of this sort of historical argument against induction until Section 3. First I want to examine a more abstract reason for thinking that induction, or at least enumerative induction, cannot possibly be generally reliable.

1.7. Goodman's New Problem of Induction

Suppose we define 'grue' as a term which applies to all and only those objects which are *first examined before* AD *2000 and found to be green* OR *which are not first examined before 2000 and are blue.*

Now imagine that we want to ascertain, by inductive means, which properties, if any, are possessed by all emeralds. Well, we can note that all the emeralds that we have observed so far have been green, and conclude on this basis, by an enumerative induction, that all emeralds are green. But we could also note that all the emeralds we have observed so far have been grue (since they have all been first examined before AD 2000 and found to be green) and so infer, by a quite analogous enumerative induction, that all emeralds are grue.

But now note that these two conclusions, that all emeralds are green and that all emeralds are grue, cannot both be true, given that some emeralds will only be first examined after AD 2000. For the first conclusion implies that these emeralds will be green, and the second conclusion implies that they will be blue, and so one of them must be wrong. Intuitively, of course, we are convinced that it is the 'grue hypothesis' that is wrong, and that emeralds will still be green after AD 2000. But this intuitive assumption is not needed to make Goodman's initial point, which is that both conclusions were reached by enumerative inductions of the form: 'A large number *n* of *A*s have all been seen to be *B*s', so 'All *A*s are *B*s'; yet at most one of these conclusions is true; so enumerative inductions cannot all reliably generate true conclusions.

Of course, grue is rather a funny property, and I'll come back to that in a moment. But the central moral of Goodman's argument is simply that, unless we put some restrictions on what *A*s and *B*s are allowed to enter into enumerative inductions, there are going to be far too many enumerative inductions for them all to have true conclusions. This is because, given any 'normal' predicate like 'green', we can easily cook up an infinity of funny grue-like predicates that will give rise to inductive conclusions that must be false, if 'normal' inductive conclusions are true.

The 'new problem' raised by Goodman is thus to distinguish, among all the complicated predicates that can be defined, that subclass which should be allowed to enter into inductive inferences. Goodman called this the problem of distinguishing 'projectible' from 'non-projectible' predicates.

Some philosophers have suggested that the problem can be dealt with fairly quickly, by simply banning any predicates whose definition makes reference to some particular time, in the way that the definition of 'grue' refers to AD 2000. But Goodman shows that the problem cannot be dealt with this easily. For suppose we define 'bleen' as 'first examined before AD 2000 and found to be blue OR not first examined before 2000 and green'. Then it is true that if we start with the predicates 'green' and 'blue', and define 'grue' and 'bleen' in terms of them, as above, then the definitions make mention of particular times. But suppose we instead started with 'grue' and 'bleen' as our primitive terms. Then we could define 'green' as 'first examined before AD 2000 and found to be grue OR not first examined before 2000 and bleen'; and we could define 'blue' as 'first examined before AD 2000 and found to be bleen OR not first examined before 2000 and grue'; and from this perspective it is then the definitions of 'green' and 'blue' that make mention of time. So in effect the appeal to time begs the question. For it is only because we start with the assumption that 'green' and 'blue' are respectable predicates, in terms of which 'grue' and 'bleen' need defining, rather than vice versa, that we deem 'grue' and 'bleen' not to be respectable.

Goodman's own solution is that the 'projectible' predicates are simply those which happen to be 'entrenched' in our inductive practices, in the sense that they are the ones which our community has used to make inductive inferences in the past. Other philosophers, however, have tried to devise less arbitrary ways of drawing the line, appealing to ideas of simplicity or of importance to science. It would be fair to say, I think, that there is no universally agreed solution to this question.

In any case, simply drawing a line between projectible predicates and the rest is arguably only half the problem. We would also like some explanation of why it is rational to make inductions with projectible predicates but not others. From the reliabilist perspective outlined in the last section, such an explanation would need to establish that inductions made using projectible predicates reliably produce true conclusions given true premisses.

There is the possibility of simply arguing once more that past inductions provide inductive evidence for induction's reliability, as was done at the end of the last section. But it can no longer be taken for granted that this move is available. For when we made this move in the last section, it was via an enumerative (meta-)induction. But we now know that enumerative induction is not always a satisfactory means of reasoning, and that at best some restricted category of such inductions is acceptable, namely, those that deal specifically with 'projectible' features of the world. Until we have some more detailed theory of 'projectibility', we cannot take it for granted that the success of past inductions is itself the kind of projectible pattern that provides inductive evidence for its own continuation.

At this stage, however, I propose to leave this topic. I shall return and consider it from a somewhat different perspective at the end of Section 3.6.

2. LAWS OF NATURE

2.1. Hume, Laws, and Accidents

In this section I want to consider a different puzzle raised by the existence of general truths about nature. Here the puzzle is not to do with our knowledge of such truths, but with the nature of the reality they describe: it is a problem in metaphysics, rather than epistemology. This problem is normally called the problem of distinguishing 'laws of nature' from 'accidental generalizations'.

A helpful way to approach this problem is to go back to David Hume's analysis of causation. Prior to Hume, philosophers assumed that when one thing caused another, this was because the cause possessed some kind of power which necessitated the occurrence of the effect. Moreover, they took it that we can know about these necessitating links a priori, in the sense that we can infer a priori that the effect will necessarily follow the cause, even if we have never had previous experience of their co-occurrence.

Hume argued against this account of causation. He pointed out that when we observe one event causing another (for example, the impact of one billiard-ball causing another to move), we never see any necessitating link. All we see is the initial event (the first ball's impact), and then the subsequent event (the second ball's motion), but never any third thing which might link them together. In addition, Hume argued that there is no a priori knowledge of the kind which such necessitating links would provide. People who have never observed billiard-balls cannot possibly tell, on the first occasion they see a moving ball approaching a stationary one, that the impact will make the stationary one move, rather than explode, or turn into a leprechaun.

Hume's own account of the link between a cause and its effect is simply that events like the cause are always followed by events like the effect. In Hume's view, there is nothing in a particular cause–effect sequence, other than that the first event occurs, and the second occurs after it. The link is simply that this sequence is an instance of a general pattern in which, to use Hume's terminology, events like the cause are 'constantly conjoined' with events like the effect.

One consequence of Hume's analysis of causation is the problem of induction discussed in the last section. Prior to Hume, it was assumed that we could know a priori that certain results would always follow certain causes. According to Hume, however, knowledge of causation is simply knowledge of constant conjunctions which are not the upshot of any a priori link between the cause and the effect. So our knowledge of causation can only derive from our experience of the cause being constantly conjoined with the effect. The problem of induction then emerges as the problem that our experience, which is always of a finite number of past cause–effect instances, is insufficient to guarantee what we need for causal knowledge, namely, knowledge that the cause will be constantly conjoined with the effect, not just in the past, but in the future too.

The problem of induction is a problem about our knowledge of general truths, a problem in epistemology. But Hume's analysis of causation also generates a problem about the nature of general truths, a problem in metaphysics. The problem is that Hume's analysis of causation makes it difficult to distinguish genuine laws of nature which state causal truths from accidental generalizations whose truth is a matter of mere happenstance.

According to Hume, a causal law is simply a statement of the form 'Whenever *A*, then *B*'. However, there are truths of this form which do not seem to express laws. Whenever I go to watch Arsenal play, the score is 0–0. That is a true statement of the form 'Whenever *A*, then *B*'. And it's going to stay true, because I'm not going to watch Arsenal any more. But it clearly isn't a causal law. Even though my attendance is in fact always followed by zero goals, my being at Highbury doesn't stop the players scoring.

But why not? If all that is required for a law is that *A*s are always followed by *B*s, then why isn't it a law that there are no goals when I watch Arsenal? After all, there is, by hypothesis, a perfect correlation between my being at Highbury and nobody scoring.

This is the problem of distinguishing laws from accidents. The Humean account of causation threatens to admit accidentally true generalizations into the category of laws. We need to find some way of keeping them out.

There are two general lines of response to this problem, which I shall call 'Humean' and 'non-Humean'. The Humeans stick to the basic Humean idea that causal laws state constant conjunctions, not necessary connections, and then try to explain why some constant conjunctions (the laws) are better than others (the accidents). Non-Humeans, by contrast, question this basic idea, and argue for a return to the pre-Humean view that the difference between laws and accidents is simply that laws, but not accidents, state necessary connections.

2.2. Counterfactual Conditionals

However, before exploring these two types of response, it will be useful to deal with a connected issue. One often-noted difference between laws and accidents is that laws but not accidents support *counterfactual conditionals*. A counterfactual conditional is an 'if . . . then . . .' statement with a false antecedent clause. So, for example, the claim 'If the temperature had fallen below 0°C, then there would have been ice on the road', made on an occasion where the temperature did not in fact fall below 0°C and the water did not freeze, is a counterfactual conditional. Indeed it is a counterfactual conditional that we intuitively accept as true, in virtue of the law that water always freezes at 0°C. But now consider the counterfactual conditional 'If I had gone to Arsenal, the score would have

been 0–0', made about a match to which I did not go and which was not a scoreless draw. Even though it is in fact true that on all occasions when I am present there are no goals, we do not accept this second counterfactual conditional as true on that account. Intuitively we feel that my presence would not have made any difference. Even if I had been there, the goals would still have been scored.

This is the sense in which laws but not accidents support counterfactuals. We intuitively project laws, but not accidents, into counterfactual situations. However, while this is certainly a good symptom of the difference between laws and accidents, it does not amount to an explanation of the difference.

The reason is that the meaning of counterfactuals is itself a matter that calls for philosophical explanation. We might start such an explanation by saying that counterfactuals state what happens in non-actual situations. But in what sense do non-actual situations exist? And if they don't exist, what makes counterfactual claims true?

One possible philosophical theory of counterfactuals is to say that counterfactuals are true just in case there is a law linking up the antecedent and consequent. But if we take this line on counterfactuals, then we obviously cannot turn round and use counterfactuals to explain the law–accident difference. For this theory of counterfactuals presupposes this difference.

As it happens, there are in any case well-known difficulties facing an explanation of counterfactuals in terms of laws. To mention just one, consider counterfactuals in which the antecedent is itself the denial of a law, such as 'If the force of gravity were inversely proportional to r, rather than r^2, then the universe would already be contracting'. This seems a perfectly cogent counterfactual assertion, but the notion of some further law linking antecedent and consequent does not seem to apply.

Because of this, contemporary philosophers have developed various other theories of counterfactual. One popular such theory, due to David Lewis,[2] appeals to the metaphysics of 'possible worlds', and says that the counterfactual 'If *A*, then *B*' is true if and only if the 'nearest' possible world in which *A* is true is also one in which *B* is true. This is an attractive theory of counterfactuals. But if we adopt this, or any other similar theory of counterfactuals, then we still cannot explain the law–accident difference in terms of counterfactuals. For, since we are now explaining counterfactuals in terms of possible worlds, rather than laws, we will need some further explanation of why laws, but not accidents, 'project' into nearby possible worlds. After all, on the Humean view, both laws and accidents simply state that *A*s are always followed by *B*s in the actual world. So why do laws, but not accidents, also tell us about other non-actual worlds?

A complete philosophy of these matters would combine an account of the law–accident distinction with an account of counterfactuals in order to yield an

[2] *Counterfactuals* (Oxford, 1973).

explanation why laws and not accidents support counterfactuals. But, until we have such a complete account, the counterfactual-supporting power of laws is part of the problem of explaining the difference between laws and accidents, not a solution.

2.3. Laws as Wide-Ranging Generalizations

The Humean strategy, remember, is to explain why some constant conjunctions (laws) are better than others (accidents). An obvious initial thought is that laws tend to be more general than accidents. The truth that water freezes at 0°C covers an indefinite, and perhaps infinite, number of instances. By contrast, the truth that there are never any goals when I am at Arsenal only applies to an odd half-dozen or so cases.

But this isn't in fact an invariable difference. There can well be laws with only a few instances. 'In any expanding universe, the rate of expansion decreases' presumably only has one instance, but it isn't any less a law for that. And it is even arguable that there are laws with no instances, such as 'A body subject to no forces will have zero acceleration'.

A related thought is that accidents are disqualified from lawlike status because they tend to be framed using terms which refer to particular spatio-temporal individuals, like 'David Papineau', and 'Arsenal football ground', rather than in purely qualitative terms like 'water', '0°C', and 'freezes'. Terms of the latter kind apply to any objects anywhere which have the right general properties, whereas non-qualitative terms like 'David Papineau' are restricted to specific individuals.

But this does not get to the heart of the matter either. Suppose we start with a true accidental generalization framed in non-qualitative terms, such as 'Whenever David Papineau goes to Arsenal, the score is 0–0', and simply replace the non-qualitative terms by qualitative descriptions detailed enough to pick out just the same individuals. That is, suppose we replaced 'David Papineau' by 'anybody with such-and-such an appearance' and 'Arsenal football ground' by 'any football ground with such-and-such-shaped stands', where the 'such-and-suches' were long descriptions which uniquely identified me and Arsenal football ground. Then 'Whenever somebody with such-and-such an appearance goes to a football ground with such-and-such-shaped stands' would be a true generalization framed in purely qualitative terms. But it would still be an accident.

2.4. Laws are Inductively Supported by their Instances

But still, despite the arguments of the last section, isn't there some sense in which accidents are too specific, too local, to function as general guides to the workings of the universe? J. L. Mackie has argued for a different way of capturing this intuition. The trouble with accidents, according to Mackie, is not that they have too few instances, as such, but rather that they aren't *inductively supported* by their instances. When we observe a number of cases of water freezing at 0°C, then this gives us good reason to suppose that all water freezes at 0°C. By contrast, that the teams failed to score on the first three or four occasions when I went to Arsenal seems a bad reason for supposing that my presence would preclude them from scoring the next time I went.

In effect, Mackie is suggesting that we explain the difference between laws and accidents in terms of the difference between projectible and non-projectible predicates.[3] Recall the discussion of Goodman's 'new problem of induction' in Section 1. Goodman shows that we need to recognize a distinction between patterns involving predicates like green, which can rationally be projected on to further unobserved cases, and patterns involving predicates like grue, which it is irrational to expect to continue. Mackie's suggestion, then, is simply that laws are those true generalizations that contain projectible predicates.

Note how this suggestion yields a natural explanation of why examples of accidents tend to be framed in non-qualitative terms and to have finite numbers of instances. According to Mackie, while laws can be asserted on the basis of subsets of their instances, accidents, which aren't inductively supported by their instances, can only be accepted as true when we know we have exhaustively checked through all the instances. (For example, we only knew the Arsenal generalization was true because I could promise you I wasn't going there any more.)

So it is a condition of an accident's being known to be true that it have a finite number of instances, for otherwise exhaustive examination would be impossible. And one natural way to ensure such finiteness is to frame examples of accidents in non-qualitative terms. (Which is not necessarily to rule out true accidents with an infinite number of instances. The point is only that such accidents cannot be known to be true, and so won't be available as examples for philosophical discussion.)

We can now see exactly why accidents are useless as guides to the workings of the universe. It's not that accidents are less true than laws, nor even that they are necessarily less general. It is just that we are never in a position to *use* them as guides, for we are never in a position to trust an accidentally true generalization, until we have already ascertained everything it might tell us by independent means.

[3] J. L. Mackie, *Truth, Probability and Paradox* (Oxford, 1973).

2.5. Laws and Systematization

I now want to look at a different Humean account of the law–accident difference. At the end of this section I shall compare it with Mackie's account. The central idea is that laws, but not accidents, are part of a scientific account of the ways the world works: the difference between 'Water freezes at 0°C' and 'There are no goals when David Papineau goes to Highbury' is that the former, but not the latter, is explainable in terms of basic scientific principles.

Of course, this suggestion needs to give some independent account of 'basic scientific principles', apart from their being basic *laws*. This is done by appealing to the idea of the *simplest systematization* of general truths. Imagine that from a God's-eye point of view, so to speak, there is a class of objectively true generalizations which includes both all the laws and all the accidents. Now think of the various ways these truths might be organized into a deductive system, based on a set of axioms. Some of these systematizations would have a greater degree of simplicity than others. (We can take it that the fewer the axioms, the simpler the system.) But simplicity will be bought at the cost of leaving some generalizations out of the systematization. (We could include *all* general truths in the system by simply taking them all as axioms. But this systematization would completely lack simplicity.) Arguably, there will be one systematization that optimally combines strength and simplicity, in that it has a small number of axioms, for simplicity, but nevertheless manages to include nearly all the original class of general truths as theorems which follow from these axioms. We can then distinguish laws from accidents by saying that the axioms and theorems in this optimal systematization are laws, while the general truths left dangling are the accidents.

In short, we say that laws are those general truths which follow from the axioms of science, and then use the simplicity-plus-strength argument to identify those axioms.

How does this idea, which was first put forward by F. P. Ramsey early this century, and later revived by David Lewis, relate to Mackie's suggestion? Let us assume, for the sake of this comparison, that the class of projectible predicates coincides with those which appear in the simplest-plus-strongest systematization. Even if we make this assumption, Mackie's theory differs from Ramsey's and Lewis's. For Mackie says that *any* true generalization framed in projectible predicates is a law; whereas Ramsey and Lewis require in addition that the generalization be deducible from the axioms of science.[4] So to decide between these two theories of lawhood, we need to consider the status of some generalization which is framed in projectible terms, but is not in fact deducible from the axioms of science.

[4] F. P. Ramsey, 'Universals of Law and Universals of Fact' (1928); repr. in *Foundations*, ed. D. H. Mellor (London, 1978); David Lewis, *Counterfactuals* (Oxford, 1973).

For example, imagine you are doing some research with some complicated electronic equipment, and you notice that, whenever the equipment and your radio are both on, your radio makes a funny noise. Suppose also that this is the only time this kind of complicated equipment will ever be constructed, because you dismantle it at the end of the experiment. Given that the properties of electronic equipment and radios are presumably projectible, if anything is, you infer that, whenever equipment of that kind is on, radios like yours make a funny noise. But suppose that in fact there is no real connection, and that your radio is making funny noises for some quite different reason. Then the generalization 'Whenever equipment of that kind is on, radios like yours make a funny noise' will be exceptionlessly true, and will contain projectible predicates. Yet it clearly isn't a law. This shows that Ramsey and Lewis are right about laws and Mackie is wrong, since the Ramsey–Lewis theory does not count this generalization as a law, while the Mackie theory does. (If you were the experimenter in the example, you would no doubt think that the pattern is a law, for you would no doubt think that it has some explanation in terms of basic science. But still, you will be wrong in thinking this, since it does not have such an explanation.)

2.6. The Non-Humean Alternative

One objection to the Ramsey–Lewis theory of laws is that its dependence on the notions 'strength' and 'simplicity' makes it vague and subjective. But even if we let that pass, and allow that the theory yields a reasonably precise way of distinguishing those true generalizations that qualify as laws, there is another objection, indeed an objection that can be levelled at all Humean theories. Namely, that the whole Humean approach to lawhood is highly counterintuitive.

Consider these two sequences: (1) The temperature falls below 0°C, and then the water freezes; (2) I go to Highbury, and then there are no goals. Humeans say that the only distinction between them is that, while they are both instances of true universal generalizations, the generalization covering (1) is somehow more significant than that covering (2). But this is surely counter to intuition. For it seems to leave out the idea that in (1) the first event made the second one happen, that it was because of the first event that the second event happened; whereas in (2) there is no such link between the two events. To say that this difference is a difference in the covering generalizations seems to put the difference in the wrong place, to make it a linguistic matter rather than an aspect of nature. Intuitively, the issue is whether there is a link in nature between the particular events, not whether the covering generalizations are sufficiently general, or inductively supported by their instances, or even part of the optimal systematization.

Of course, to side with intuition here is simply to reject Hume's analysis of causation. But a number of recent philosophers have argued that we should do just that. In the last two decades David Armstrong, Fred Dretske, and Michael Tooley[5] have all argued that causal laws are not simply statements of constant conjunction, but rather state *necessitating relationships* between the properties involved. They say that the way to represent the content of a causal law is not simply as 'All *A*s are (as it happens) followed by *B*s', but rather 'Nec(*A*, *B*)', where 'Nec' stands for the relationship of necessitation between the properties *A* and *B*. So, in the contrasting pair above, the low temperature necessitates the freezing, but my being at Highbury does not necessitate the absence of goals.

On the Armstrong–Dretske–Tooley view, a necessitating relationship between *A* and *B* certainly implies that all *A*s are *B*s. But the converse implication does not hold: there can be cases where all *A*s are *B*s even though it is not true that Nec(*A*, *B*)—namely, when it is an accident that all *A*s are *B*s.

So this non-Humean view offers an entirely straightforward explanation of the law–accident difference. The difference is simply that laws state something that accidentally true generalizations do not, namely, the existence of a necessitating relationship between properties.

Given the possibility of this simple solution, the obvious question to ask is why most philosophers in the 250 years since Hume have not availed themselves of it.

Hume had two arguments against the idea that causal laws involve necessitating links. First, we never see such links. Second, we cannot know laws of nature a priori, as would be possible if they stated necessities.

We need not dwell too long on Hume's first argument. The assumption that we cannot meaningfully talk about things we cannot observe has had few supporters in this century, even if it was generally accepted in Hume's time. The example of modern science, with its talk of atoms, electrons, and radio waves, has shown that meaningful reference is not restricted to observable phenomena. So the fact that we cannot see necessitating links does not automatically mean we cannot talk about them.

Hume's second argument deserves more attention. This argument assumes that if laws state necessities, then they must be knowable a priori (and so concludes that, since laws clearly cannot be known a priori, they cannot state necessities). The assumption that necessity implies aprioricity went unchallenged until very recently in the Western philosophical tradition. At the beginning of the 1970s, however, the American philosopher Saul Kripke argued that the *metaphysical* notion of necessity needs to be sharply separated from the *epistemological* notion of aprioricity. In particular Kripke argued that many statements of identity (for example, 'The Evening Star = the Morning Star') are necessary (for how *could* that planet not be itself?), even though it is only after a posteriori empirical discoveries that they can be *known* to be true.

[5] David Armstrong, *What is a Law of Nature?* (Cambridge, 1983); Fred Dretske, 'Laws of Nature', *Philosophy of Science*, 44 (1977), 248–68; Michael Tooley, 'The Nature of Laws', *Canadian Journal of Philosophy*, 7 (1977), 667–98.

It is striking that Armstrong, Dretske, and Tooley all put forward their non-Humean view of laws within five years of the publication of Kripke's ideas. This suggests that the key which allowed them to reject Hume's view of laws was the separation of necessity from aprioricity. For when they say that laws of nature state that *A* necessitates *B*, they certainly do not mean to imply that these laws can be known a priori. To this extent the kind of necessary connection they uphold is different from the kind Hume rejected. (It also means that their view of laws makes no difference to the problem of induction: since laws have to be derived from a posteriori evidence, we still need to explain how past evidence can tell us something that implies future patterns.)

While it seems highly plausible that Kripke's views about necessity prompted the re-emergence of non-Humean views of laws of nature, there are important differences between these two developments. Most centrally, the non-Humean necessitating connections are not in fact *necessary* in Kripke's sense. Kripkean necessities are supposed to obtain in all possible worlds. It is simply impossible that a planet should exist yet not be itself. But the modern non-Humeans do not require their laws of nature to be necessary in this sense. They allow that it is possible that the force of gravity might have been weaker than it is, that water might have frozen at a different temperature, and so on. Their idea of a necessitating connection is that of one property making another happen, not the Kripkean idea of a claim that *could not possibly* be false.

This difference points to a difficulty facing the non-Humean views of laws of nature. The non-Humeans say that necessitation involves something more than constant conjunction: if two events are related by necessitation, then it follows that they are constantly conjoined; but two events can be constantly conjoined without being related by necessitation, as when the constant conjunction is just a matter of accident. So necessitation is a stronger relationship than constant conjunction. However, the non-Humeans say very little about what the extra strength amounts to. We are told that it is not necessity in the Kripkean sense of truth in all possible worlds. But we are not given any positive characterization of this extra strength, except that it distinguishes laws from accidents. Critics of the non-Humean view argue that a satisfactory account of laws ought to cast more light than this on the nature of laws. They complain that the notion of necessitation simply restates the problem, rather than solving it.

So we can sum up our overall discussion of laws of nature with a choice. If you like explanations, and don't mind too much about intuitions, then you can go for a Humean strategy, with the Ramsey–Lewis theory the most promising version. But if you want an account of laws of nature that fits our pre-theoretical intuitions, and don't mind the complaint that it simply reifies the law–accident difference without explaining it, then you can take the modern non-Humean option.

3. REALISM, INSTRUMENTALISM, AND UNDERDETERMINATION

3.1. Instrumentalism versus Realism

In the first section I discussed the problem of induction. In this section I want to consider a different difficulty facing our knowledge of the natural world and scientific knowledge in particular. Much of science consists of claims about unobservable entities like viruses, radio waves, electrons, and quarks. But if these entities are unobservable, how are scientists supposed to have found out about them? If they cannot see or touch them, doesn't it follow that their claims about them are at best speculative guesses, rather than firm knowledge?

It is worth distinguishing this problem of unobservability from the problem of induction. Both problems can be viewed as difficulties facing theoretical knowledge in science. But where the problem of induction arises because scientific theories make general claims, the problem of unobservability is due to our lack of sensory access to the subject-matter of many scientific theories. (So the problem of induction arises for general claims even if they are not about unobservables, such as 'All sodium burns bright orange'. Conversely, the problem of unobservability arises for claims about unobservables even if they are not general, such as, 'One free electron is attached to this oil drop'. In this section and the next, however, it will be convenient to use the term 'theory' specifically for claims about unobservables, rather than for general claims of any kind.)

There are two schools of thought about the problem of unobservability. On the one hand are *realists*, who think that the problem can be solved. Realists argue that the observable facts provide good indirect evidence for the existence of unobservable entities, and so conclude that scientific theories can be regarded as accurate descriptions of the unobservable world. On the other hand are *instrumentalists*, who hold that we are in no position to make firm judgements about imperceptible mechanisms. Instrumentalists allow that theories about such mechanisms may be useful 'instruments' for simplifying our calculations and generating predictions. But they argue that these theories are no more true descriptions of the world than the 'theory' that all the matter in a stone is concentrated at its centre of mass (which is also an extremely useful assumption for doing certain calculations, but clearly false).

Earlier this century instrumentalists used to argue that we should not even interpret theoretical claims literally, on the grounds that we cannot so much as meaningfully talk about entities we have never directly observed. But as I said in the last section, the development of modern science, with its talk of atoms, electrons, and so on, has made this restriction on meaningful talk difficult to defend. So nowadays this kind of *semantic* instrumentalism is out of favour.

Contemporary instrumentalists allow that scientists can meaningfully postulate, say, that matter is made of tiny atoms containing nuclei orbited by electrons. But they then take a sceptical attitude to such postulates, saying that we have no entitlement to believe them (as opposed to using them as an instrument for calculations.)

3.2. Initial Arguments for Realism

An initial line of argument open to realism is to identify some feature of scientific practice and then argue that instrumentalism is unable to account for it. So, for example, realists have pointed to the fact that scientists characteristically seek to unify different kinds of scientific theory in pursuit of a single 'theory of everything'. In the nineteenth century, for instance, physicists working in thermodynamics developed the kinetic theory of gases, which explained variations in the temperature, pressure, and volume of gases by postulating that gases are made of swarms of tiny particles; at the same time chemists were developing the atomic theory of matter, which explained chemical combinations on the assumption that matter was made of atoms, one kind of atom for each element. An obvious question was to investigate the relation between the two theories: were the particles of the physicists combinations of atoms, and if so what kinds of combination? The resolution of this issue was not always easy, but in time a satisfactory conclusion was arrived at.

However, this whole procedure, the realist points out, only makes sense on the assumption that scientific theories are *true* descriptions of reality. After all, says the realist, if theories are simply convenient calculating-devices, then why expect different theories to be unifiable into one consistent story? Unification is clearly desirable if our theories all aim to contribute to the overall truth, but there seems to be no parallel reason why a bunch of instruments should be unifiable into one big 'instrument of everything'.

Other features of science appealed to by realists as arguments against instrumentalism include the use of theories to explain observable phenomena, and the reliance on theories to make novel predictions. I shall take these in turn. The topic of explanation will be discussed in detail in Section 5 below. But for the moment we need only note that scientists often explain the behaviour of observable phenomena in terms of unobservable mechanisms. Thus, to use one of the above examples, scientists explain why the pressure of an enclosed gas increases when its temperature does by referring to the behaviour of the tiny particles making up the gas. But surely, the realist urges, this only makes sense if these tiny particles really exist and the theory describing them is not just an instrument for making calculations. We surely cannot say that the pressure goes up because the tiny particles are moving faster if we don't believe in the existence of those particles.

Then there is the argument from prediction. Scientists often predict surprising and hitherto quite unknown observable phenomena on the basis of their theories. For example, Einstein predicted, on the basis of the general theory of relativity, that light would bend in the vicinity of the sun. Apart from his theory, there was no reason whatsoever to expect this. Yet this prediction was triumphantly confirmed by Sir Arthur Eddington's famous observations in West Africa during an eclipse of the sun in 1919. This provides another argument for realism. For the realist can insist that there would be no reason why such predictions should ever work if the theories behind them were not true.

These three arguments, from unification, explanation, and prediction, all give some support to realism. But none of them are conclusive. In each case, there are two possible lines of response open to instrumentalists. They can offer an instrumentalist account of the relevant feature of scientific practice. Alternatively, they can deny that this feature really is part of scientific practice in the first place. I shall go through the three cases in turn.

3.3. Initial Instrumentalist Responses

3.3.1. *Unification*

First, the argument from unification. The first possibility is for instrumentalists to offer an instrumentalist account of the scientific practice of unifying theories. They can do this by arguing that the unification of science is motivated, not by the pursuit of one underlying truth, but simply by the desirability of a single, all-purpose calculating instrument instead of a rag-bag of different instruments for different problems. If the aim of theories is convenience, rather than veracity, is it not more convenient to have one device that will deal with all problems, rather than having to worry which tool will be most effective for the problem at hand?

The second possibility for an instrumentalist faced by the argument from unification is to deny that unification is essential to science to start with. Thus Nancy Cartwright argues, in *How the Laws of Physics Lie*,[6] that science really *is* a rag-bag of different instruments. She maintains that scientists faced with a given kind of problem will standardly deploy simplifying techniques and rules of thumb which owe nothing to general theory, but have shown themselves to deliver the right answer to the kind of problem at hand. So, in Cartwright's view, unification is not central to science in the first place, and so not something instrumentalists need to account for.

[6] (Oxford, 1983).

3.3.2. *Explanation*

The same two lines of response can be made to the realist argument from explanation. Here the more normal line of response is the second, namely, to deny that explanation really is an essential feature of scientific practice. Instrumentalists can argue that the essential aim of science is to describe, not to explain. What we want of science, they will say, is an accurate account of how the observable world behaves. The further issue of why it behaves like that is a far more difficult question, which takes us beyond science, if it can be answered at all. (After all, the instrumentalist can observe, even realists have to stop explaining at some point. Maybe they can explain observables in terms of unobservables, and some unobservables in terms of others. But even realists will have to admit that at some point, perhaps with quarks or other fundamental particles, they run out of explanations, and can only describe the behaviour of the fundamental particles, without explaining it in terms of further mechanisms.)

As I said, this kind of denial that explanation is essential to scientific theorizing is the normal instrumentalist response to the argument from explanation. But a minority of instrumentalists try the opposite tack, and argue that there is nothing in scientific explanation that instrumentalism cannot account for. According to instrumentalists of this stripe, it is a mistake to think of scientific explanation as a matter of identifying genuine hidden causes for observable phenomena, as opposed to simply showing how these phenomena are part of some wider pattern. The scientist who 'explains' variations in the pressure of gases by the kinetic theory is not, from this perspective, specifying the true unobservable causes of those variations, but simply showing how they conform to the same underlying equations as other kinds of observable gas behaviour. (Perhaps this second response to the argument from explanation does little more than invent a new meaning for 'explanation'. But if this makes you uneasy, there is always the first response to fall back on.)

3.3.3. *Prediction*

There remains the realist argument from prediction. Here the two lines of instrumentalist response are again open. The more radical, and perhaps less plausible, would be to deny that the ability to make such predictions is a genuine feature of scientific practice. Instrumentalists who take this line will of course allow that scientists make 'predictions' in the sense that they draw observable consequences from their theories. But they can deny that this practice generates any more *true* predictions than random guessing would. After all, they can point out, the only predictions that we remember are those that succeed, like Einstein's prediction of light bending. But for every such successful prediction there are thousands of scientific experiments that do not produce the results that are expected. So what real reason do we have for thinking that theories about unobservables enable us to anticipate new observable phenomena? Maybe this is just an impression created by selective memory. If this is right, and science is not really predictively successful, then there is obviously no need for an instrumentalist explanation of this success.

However, as I said, this response is not entirely plausible. It seems unlikely that the ability of theories about unobservables sometimes to anticipate new observable phenomena is just a chance matter. However, even if we accept that science is predictively successful, there remains room for an instrumentalist account of this. The realist account, remember, was that theories about unobservables are themselves characteristically true, so it is no surprise that they issue in true predictions. Instrumentalists, who deny the truth of theories about unobservables, cannot say this. But they can say something else. They can accept that there is a well-established pattern, displayed in the history of science, of novel observable predictions suggested by theories about unobservables turning out to be true. And then they can simply insist, in line with their general instrumentalism, that there is no need to give any further explanation of this pattern, in terms of such hidden facts as the truth of the theories concerned. After all, instrumentalism is precisely the view that we do not need to explain manifest patterns in terms of hidden causes (or at most that we should 'explain' them by fitting them into larger manifest patterns). Given that instrumentalists start off by denying the need for unobservable explanations, it begs the question against them to insist that they should produce such an explanation for the predictive success of science.

3.4. The Underdetermination of Theory by Data

In the last section I argued that various arguments against instrumentalism can be resisted. I shall now allow instrumentalism to go on the offensive, and consider some positive arguments against realism. There are two strong lines of argument that instrumentalists can use to cast doubt on realism. In this section and the next I shall discuss 'the underdetermination of theory by evidence' and some related issues. In Section 3.6 I shall consider 'the pessimistic meta-induction from past falsity'. As it happens, I do not think that either of these arguments succeeds in discrediting realism. But they are both arguments which merit careful consideration.

The argument from underdetermination asserts that, given any theory about unobservables which fits the observable facts, there will be other incompatible theories which fit the same facts. And so, the argument concludes, we are never in a position to know that any one of these theories is the truth.

Why should we accept that there is always more than one theory which fits any set of observable facts? There are two routes to this conclusion. One stems

from the Duhem–Quine thesis, originally formulated by the French philosopher and historian Pierre Duhem at the turn of the century and later revived by the American logician W. V. O. Quine.[7] Duhem and Quine point out that a scientific theory T (such as the Newtonian theory of gravitation) does not normally imply predictions P on its own (about the motions of planets, say), but only in conjunction with auxiliary hypotheses H (concerning such things as the number of other planets, their masses, the mass of the sun, and so on).

$$T \,\&\, H \rightarrow P$$

Because of this, T can always be defended in the face of contrary observations (such as the well-known anomaly for Newtonian theory presented by the orbit of Mercury) by adjusting the auxiliary hypotheses H (by postulating a hitherto unobserved planet, say, or an inhomogeneous mass distribution in the sun). The point is that the observational refutation of P does not disprove T, but only the conjunction T & H.

$$\text{Not-}P \rightarrow \text{not-}(T \,\&\, H)$$

So T can be retained, and indeed still explain not-P, provided we replace H by some alternative, H', such that

$$T \,\&\, H' \rightarrow \text{not-}P.$$

This yields the Duhem–Quine thesis: Any theoretical claim T can consistently be retained in the face of contrary evidence by making adjustments elsewhere in our system of beliefs. The underdetermination of theory by evidences (UDTE) follows quickly. For the Duhem–Quine thesis seems to imply that the adherents of competing theories will always be able to maintain their respective positions in the face of any actual observational data. Imagine two competing theories T_1 and T_2. Whatever evidence accumulates, versions of T_1 and T_2, conjoined with greatly revised auxiliary hypotheses if necessary, will both survive, consistent with that evidence, but incompatible with each other.

The other route to UDTE, first put forward by physicists such as Henri Poincaré at the turn of the century, has a different starting-point.[8] It begins, not with two competing theories, but with some given theory, all of whose observational predictions are supposed to be accurate. Imagine that T_1 is the complete truth about physical reality, that it implies observational truths O. Then we can always construct some 'de-Occamized' T_2 which postulates some more complicated unobservable mechanism, but which nevertheless has precisely the same observational consequences.

For example, suppose we start with standard assumptions about the location

[7] P. Duhem, *The Aim and Structure of Physical Theory*, Eng. edn. (London, 1962); W. V. O. Quine, 'Two Dogmas of Empiricism', in *From a Logical Point of View* (Cambridge, Mass., 1953).

[8] H. Poincaré, *Science and Hypothesis*, Eng. edn. (New York, 1952).

of bodies in space-time and about the forces acting on them. A de-Occamized theory might then postulate that all bodies, including all measuring instruments, are accelerating by 1 ft/sec^2 in a given direction, and then add just the extra forces required to explain this. This theory would clearly have exactly the same observational consequences as the original one, even though it contradicted it at the unobservable level.

To bring out the difference between the two arguments for UDTE, note that the Duhem–Quine argument does not specify exactly which overall theories we will end up with, since it leaves open how T_1 and T_2's auxiliary hypotheses may need to be revised; the de-Occamization argument, by contrast, actually specifies T_1 and T_2 in full detail, including auxiliary hypotheses. In compensation, the Duhem–Quine argument promises us alternative theories whatever observational evidence may turn up in the future; whereas the de-Occamization argument assumes that all future observations are as T_1 predicts.

3.5. Simplicity and Elimination

My view is that the arguments of the last section give us good reason to accept UDTE, the thesis that there will always be incompatible theories to explain any given body of observational facts. I do not agree, however, that UDTE yields a good argument against realism. What UDTE shows is that more than one theory about unobservables will always fit any given set of observational data. But it is too quick to conclude, as many philosophers do, that this makes realism about unobservables untenable. For we should recognize that there is nothing in the arguments for alternative underdetermined theories to show that these alternative theories will always be equally well supported by the data. What the arguments show is that different theories will always be consistent with the data. But they do not rule out the possibility that, among these alternative theories, one is vastly more plausible than the others, and for that reason should be believed to be true. After all, 'flat-earthers' can make their view consistent with the evidence from geography, astronomy, and satellite photographs, by constructing far-fetched stories about conspiracies to hide the truth, the effects of empty space on cameras, and so on. But this does not show we need take their flat-earthism seriously. Similarly, even though Newtonian gravitational theory can in principle be made consistent with all the contrary evidence, by bringing in various hidden forces and other *ad hoc* devices, this is no reason not to believe general relativity theory.

Certainly practising scientists do not regard the UDTE as blocking their access to the theoretical truth. They recognize that we can always in principle concoct alternative explanations for any given body of data; but they simply discount as not worth taking seriously those complex alternatives that need to invoke hidden planets, or hidden forces, or other truth-hiding conspiracies. In

effect, scientists are taught, in the course of their scientific training, that only certain sorts of theory are possible candidates for the truth; and once they have data that rule out all but one of *these* theories, they quite happily ignore all the other conspiratorial theories that remain consistent with the data. (Perhaps the best way of describing this aspect of scientific practice is to say that scientists ignore all theories that are not sufficiently 'simple'; but if we do so we should not think of 'simplicity' as some innate or intuitive idea; rather, the relevant kind of simplicity is part of what scientists learn when they are trained as meteorologists, embryologists, physicists, or whatever.)

Still, even if scientists don't regard the UDTE as a serious obstacle, many philosophers, as I said, move quickly, from the premiss that different theories are consistent with the observational evidence, to the conclusion that none of them can be regarded as the truth. However, I think that they only make this move because they assume that the only good inferences from data to theories are deductively valid ones: they note that the data cannot deductively imply T, if they leave open the possibility that some inconsistent theory T' is true; and they conclude that this shows we are never entitled to believe such a T.

However, as we saw in our earlier discussion on induction in Section 1, there are good reasons for allowing that other inferences, apart from deductively valid ones, can be rational. In particular, in that discussion I suggested that the important underlying requirement might merely be that inferences should be reliable, not deductively valid.

In fact, the issue we are now addressing is closely related to our earlier discussion of induction. In Section 1 I focused on *enumerative* induction, in which we go from instances of a pattern to the theory that this pattern holds generally. The theory-choices we are now considering can be thought of as *eliminative* inductions, in which we assume that the truth lies among one of a limited number of theories (the reasonably 'simple' theories), and then use our observations to eliminate all but one of those theories.

The essential difference between these two forms of induction is that eliminative inductions consider only a limited number of theories to be candidates for truth. This might make it seem as if enumerative induction is a more general form of inference, since it rests on no such presupposition. But in fact our discussion of Goodman's 'new problem of induction' in Section 1 shows that even enumerative inductions rely on a similar presupposition: since there are so many possible ways of projecting observed patterns into the future, enumerative inductions are forced to restrict the generalizations they regard as candidates for the truth to the limited number which involve projectible predicates. For example, propositions of the sort 'All emeralds are green (yellow/red/etc.)' are reasonably 'simple', and so candidates for truth, but propositions of the sort 'All emeralds are grue (bleen/etc.)' are not. Someone investigating emeralds can then reach the natural conclusion by noting which of the candidates for truth is consistent with the observations made so far.

Given this, we may as well regard all inductions as in essence eliminative, rather than enumerative. Still, the question of the reliability arises in just the same way for eliminative induction as for enumerative induction. The fact that eliminative inductions are not logically valid does not itself mean that they are not reliable. But there remains the question whether they are reliable.

In Section 1 I suggested that it might be acceptable to answer this question for enumerative inductions by providing (enumeratively) meta-inductive evidence for their reliability. Perhaps we can try the same move again. That is, maybe we can take as evidence those occasions where scientists have chosen whichever 'simple' theory is consistent with the evidence, and then argue meta-inductively that the only 'simple' account of the success of these inferences is that such eliminative inductions are in general reliable guides to the truth. This move obviously involves some element of circularity, but, as I noted in Section 1, it is not clear that this kind of circularity is vicious.

It should be said that this is only one possible way in which we might try to defend the rationality of eliminative induction. The main point I want to make in this section is that the rationality of eliminative induction does not require that it be deductively valid. So the UDTE does not show that such inductions are never acceptable, and so does not discredit the realist view that well-attested theories about unobservables can be regarded as true descriptions of nature. How best to go beyond this, and show positively that eliminative inductions *are* rational, is perhaps too difficult a question to resolve here.

3.6. The Pessimistic Meta-induction from Past Falsity

Let me now turn to the other argument against realism mentioned earlier. This argument takes as its premiss that past scientific theories have generally turned out to be false, and then moves inductively to the pessimistic conclusion that our current theories are no doubt false too.

There are plenty of familiar examples to support this argument. Newton's theory of space and time, the phlogiston theory of combustion, and the theory that atoms are indivisible were all at one time widely accepted scientific theories, but have since been recognized to be false. So doesn't it seem likely, the pessimistic induction concludes, that all our current theories are false, and that we should therefore take an instrumentalist rather than a realist attitude to them?

This is an important and powerful argument, but it would be too quick to conclude that it discredits realism completely. It is important that the tendency to falsity is much more common in some areas of science than others. Thus it is relatively normal for theories to be overturned in cosmology, say, or fundamental particle physics, or the study of primate evolution. By contrast, theories

of the molecular composition of different chemical compounds (such as that water is made of hydrogen and oxygen), or the causes of infectious diseases (that chicken-pox is due to a herpes virus), or the nature of everyday physical phenomena (that heat is molecular motion), are characteristically retained once they are accepted.

Nor need we regard this differential success-rate of different kinds of theories as some kind of accident. Rather, it is the result of the necessary evidence being more easily available in some areas of science than others. Palaeoanthropologists want to know how many hominid species were present on earth three million years ago. But their evidence consists of a few pieces of teeth and bone. So it is scarcely surprising that discoveries of new fossil sites will often lead them to change their views. The same point applies on a larger scale in cosmology and particle physics. Scientists in these areas want to answer very general questions about the very small and the very distant. But their evidence derives from the limited range of technological instruments they have devised to probe these realms. So, once more, it is scarcely surprising that their theories should remain at the level of tentative hypotheses. By contrast, in those areas where adequate evidence is available, such as chemistry and medicine, there is no corresponding barrier to science's moving beyond tentative hypotheses to firm conclusions.

The moral is that realism is more defensible for some areas of science than others. In some scientific subjects firm evidence is available, and entitles us to view certain theories, like the theory that water is composed of H_2O molecules, as the literal truth about reality. In other areas the evidence is fragmentary and inconclusive, and then we do better to regard the best-supported theories, such as the theory that quarks and leptons are the ultimate building-blocks of matter, as useful instruments which accommodate the existing data, make interesting predictions, and suggest further lines for research.

At first sight this might look like a victory for instrumentalism over realism. For didn't instrumentalists always accept that we should be realists about observable things, and only urge instrumentalism for uncertain theories about unobservable phenomena? But our current position draws the line in a different place. Instrumentalism, as originally defined, takes it for granted that everything unobservable is inaccessible, and that all theories about unobservables are therefore uncertain. By contrast, the position we have arrived at places no special weight on the distinction between what is observable and what is not. In particular, it argues that the pessimistic meta-induction fails to show that falsity is the natural fate of all theories about unobservables, but only that there is a line within the category of theories about unobservables between those theories that can be expected to turn out false and those whose claims to truth are secure. So our current position is not a dogmatic instrumentalism about all unobservables, but merely the uncontentious view that we should be instrumentalists about that subclass of theories which are not supported by adequate evidence.

4. CONFIRMATION AND PROBABILITY

4.1. The Notion of Confirmation

At the end of the last section I argued that the history of science gives us reason to be cautious in our commitment to certain scientific theories. In at least some areas of science the evidence for even the best theories is often fragmentary and inconclusive, with the consequence that we should expect that such theories will turn out to be false.

It would be nice to be able to say more about the degree to which a given body of evidence supports a given theory. That is, it would be nice to have a quantitative account of the relationship between evidence and theory. Philosophers have sought to develop such accounts, under the name of 'confirmation theory'. They seek to understand the extent to which different bodies of evidence 'confirm' different theories. If a theory is highly confirmed by the available evidence, then we can be reasonably confident it is true; but if it has a lower degree of confirmation, then we should moderate our trust in it accordingly.

However, this intuitive notion of confirmation is less straightforward than it seems. I shall introduce some of the difficulties by describing two well-known paradoxes that any theory of confirmation must deal with.

4.2. The Paradox of the Ravens

Let us assume that there is a relationship of confirmation, according to which sometimes *E confirms T*, where *E* is some body of evidence and *T* some theory. Then it certainly seems natural to make the following two assumptions about confirmation:

(1) If $E = (Fa \,\&\, Ga)$ and $T =$ All *Fs* are *Gs*, *then E confirms T*.

(This first assumption simply says that generalizations are confirmed by their instances.)

(2) If *E* confirms *T*, and *T* is logically equivalent to *S*, *then E confirms S*.

As I said, these two assumptions seem highly uncontentious. But they can easily be shown to generate a puzzle.

Note first that the following two generalizations are logically equivalent:

(L) All ravens are black.

(M) All non-black things are non-ravens.

Now take as our evidence an observation that:

(I) The white thing over there is a shoe.

Since (I) is an instance of a non-black thing which is a non-raven, then assumption (1) tells us that (I) confirms (M).

But if we now put this together with the fact that (M) is logically equivalent to (L), then assumption (2) tells us that (I) confirms (L).

However, this seems absurd. For (L) is the claim that all ravens are black, and surely we cannot confirm *that* just by observing that some white thing is a shoe.

Something seems to have gone wrong somewhere. But it is difficult to see where. For there can scarcely be anything wrong with assumption (2)—logically equivalent propositions make exactly the same claims about the world, so it is hard to see how some piece of evidence could support one such proposition, without therewith supporting the other proposition. And assumption (1) seems almost as obvious—if anything is ever confirmed by anything, surely generalizations are confirmed by their instances.

(Some of you might think that the flaw in the reasoning lies with assumption (1). For isn't the lesson of Goodman's new problem of induction precisely that $Fa \,\&\, Ga$ cannot always confirm $(x)\,(Fx \rightarrow Gx)$? Goodman shows that, unless we restrict *F* and *G* to 'projectible' predicates, there are far too many *Fs* and *Gs* for all such generalizations to be confirmable by their instances. However, I do not think this helps with the raven paradox, given that there isn't anything particularly 'gruesome' about the predicates used to formulate it, namely, 'black', 'raven', 'non-black', and 'non-raven'. It is of course true that Goodman's argument shows that (1) is not acceptable as formulated without qualification above. But the paradox will still be generated even if (1) is restricted to apply only to 'projectible' predicates.)

4.3. The Tacking Paradox

Now for the second paradox. Here are two further assumptions that seem pretty obvious:

(3) If *T* entails *E*, then *E confirms T*.

(This is just the idea that a theory is confirmed if its consequences are observed to be true.)

(4) If *E* confirms *T*, and *T* entails *P*, then *E confirms P*.

(This is just the idea that, if some evidence entitles you to believe some theory, then it entitles you to believe what follows from it.)

But now take *any* theory N—Newtonian gravitational theory, say—and *any* consequence M it entails—the planets move in ellipses. And then consider *any* other proposition Q you like—the moon is made of green cheese. Since N (Newtonian theory) entails M (elliptical orbits), by hypothesis, N & Q (Newtonian theory *plus* the moon is made of green cheese) also entails M. So, by (3),

(*a*) M confirms N & Q.

But

(*b*) N & Q entails Q, trivially,

so by (4), applied to (*a*) and (*b*), it follows that M (elliptical orbits) *confirms* Q (the moon is green cheese). But this means that anything that follows from one theory—the planets move in ellipses—confirms any other theory you like—the moon is made of green cheese. And this is surely absurd.

Even so, this absurd conclusion follows from the apparently uncontentious assumptions (3) and (4). Once more, it is difficult to see where the fault in our reasoning lies.

This paradox is called the 'tacking' paradox, because it involves 'tacking' an arbitrary hypothesis (the moon is made of green cheese, in the above example) on to the theory you start with (Newtonian mechanics). A common initial reaction is that assumption (3) is at fault. Do the motions of the planets really confirm Newtonian-theory-and-that-the-moon-is-green-cheese? But I shall argue that this is in fact a perfectly sensible thing to suppose, and that it is assumption (4) that is really responsible for the tacking paradox. However, before I explain how I think the tacking paradox (and the paradox of the ravens) should be dealt with, it will be necessary to digress at some length, and explain some ideas about *probability*.

4.4. Interpretations of Probability

The notion of probability can be understood in a number of different ways. In particular, as I said in Section 1, there are both objective and subjective notions of probability. But there is one thing that ties together all the different notions of probability, namely, that they satisfy the *axioms of the probability calculus*.

These axioms are normally stated as follows:

(1) $0 \leq \text{Prob}(p) \leq 1$, for any proposition p.
(2) $\text{Prob}(p) = 1$, if p is a necessary truth.
(3) $\text{Prob}(p) = 0$, if p is impossible.
(4) $\text{Prob}(p \text{ or } q) = \text{Prob}(p) + \text{Prob}(q)$, if p and q are mutually exclusive.

Any way of assigning numbers to propositions so as to satisfy these axioms constitutes an interpretation of the probability calculus. We shall concentrate in

particular on the contrast between subjective and objective interpretations of probability.

The subjective interpretation takes the probability of p to be a measure of the strength with which you believe p. More specifically, for any person X, the subjective interpretation equates X's probability for p with the degree *to which X believes p*.

Some extreme subjectivists argue that this is the only notion of probability we need. But most philosophers who recognize subjective probabilities also recognize objective probabilities. Objective probabilities apply specifically to propositions which claim that a certain kind of result will occur on a certain kind of repeatable trial, such as that a certain kind of coin will come down heads when tossed. And in this kind of context a statement of objective probability specifies how much trials of that kind *tend* to produce the result in question. This kind of tendency is displayed by the frequency with which the result occurs—for example, by how often coins like this come down heads.

It should be clear that these subjective and objective interpretations give us different notions of probability. A degree of subjective belief is one thing, and an objective tendency is another. There is no guarantee that any particular person's subjective expectations should correspond to the objective tendencies; and there would still have been objective probabilities of atoms decaying, even if there had never been any human beings to form degrees of belief. Let us now look more closely at these two notions in turn.

4.5. Subjective Probabilities

The central assumption of the subjective interpretation is that belief comes in *degrees*. Normally we think of belief as something you either have or have not. But consider the attitude of someone who takes both umbrella and sunblock cream on a walk. Does this person believe it will rain or not? The natural answer is that the person has some expectation that this proposition is true, and some that it is not. Or consider the attitude of a company director who gives money to both the Labour Party and the Conservative Party before the election. Again it seems natural to say that the company director has a positive degree of belief that each outcome will happen. (Some people object to the idea of 'degrees of belief' because they think of beliefs as definite attitudes to propositions, for or against. If one preferred, one could think in terms of degrees of expectation rather than belief. This will make no difference to the points which follow.)

It is one thing to argue that beliefs come in degrees. It is another to show that we can attach definite numbers between 0 and 1 to these degrees. But the subjective theory needs to show this, since degrees of belief will have to equal such numbers if they are to have any chance of satisfying the probability axioms.

However, this is not necessarily as far-fetched as it seems at first sight. The obvious way to attach a number to someone's degree of belief is to see what minimum *odds* would induce them to bet on p. If you are only prepared to put up £N once your opponent offers £M or more, with the winner taking all if p turns out true, then this arguably shows that your degree of belief in p is $N/(N + M)$.

True, some people hate betting *per se*. And in such cases the odds that will induce them to bet will overestimate their degrees of belief. For example, you may be convinced that you are betting on a fair coin, and so attach a 50–50 probability to heads, but may be disinclined to risk your treasured £10 on heads until I am putting up £40 or more. The test suggested in the last paragraph would indicate that your degree of belief in heads is 0.2, rather than 0.5. But perhaps in this kind of case an investigator could still find out your true degree of belief by asking you to choose fair odds for bets on p, without telling you which way you are going to bet, or how much the bet is going to be. In this situation, any aversion to betting should cancel out and leave your chosen odds expressing your true degree of belief.

Despite these ingenious suggestions, you may still feel it is unrealistic to suppose that there are precise numerical degrees of belief for all propositions. Surely there is no fact of the matter of whether my degree of belief that X will win the next election is 0.3456 rather than 0.3457. But the defender of the subjective interpretation can reasonably respond that the postulation of exact numerical degrees of belief is a useful *idealization*, which facilitates our theorizing, and does no harm when understood as such. By way of comparison, consider the way that physicists suppose that physical objects, like stones and planets, have precise masses and sizes. This is never strictly true, since there are always some molecules dropping off or joining on to such objects. But the fiction of precise quantities is extremely useful in physics, and misleads no one.

Even if we grant the subjective interpretation that degrees of belief can be thought of as precise numbers, it still needs to be shown that these numbers actually conform to the probability axioms. It need of course be nothing more than convention that we assign 0 to the lowest possible degree of belief (namely, disbelief) and 1 to the highest (full belief), and so satisfy axiom (1). But it would not be a matter of convention that degrees of belief should satisfy the other axioms.

The normal way of establishing that degrees of belief conform to the probability axioms is via the 'Dutch book argument', which shows that somebody whose degrees of belief violate axioms (1)–(4) can be induced to make manifestly irrational bets.

Suppose that you attach a probability of 0.8 to 'It will rain today', and a degree of belief of 0.7 to 'It won't rain today'. Then your degrees of belief violate the probability axioms. (This is because the proposition that 'It will or it won't rain' is a necessary truth, and so, to satisfy axiom (2), needs to have degree of belief 1; but it is also the disjunction of the exclusive propositions 'It will rain today' and 'It won't rain today', and so, to satisfy axiom (4), needs to have a degree of belief equal to the sum of these propositions' separate degrees of belief; however, this sum is 1.5, not 1.)

Note also that, because you have these degrees of belief, you will be prepared to wager your £8 to my £2 on the proposition 'It will rain today'; and you will be prepared to wager your £7 to my £3 on 'It will not rain today'. But this is clearly a quite silly pair of bets, since you are guaranteed to lose £5 whatever happens.

It is provable, along the lines of this last example, that people are open to 'Dutch books' if and only if their degrees of belief fail to conform to the probability axioms. Since it seems clearly irrational to have degrees of belief which can lead you to do things which are guaranteed to fail, this shows that everybody's degrees of belief ought rationally to conform to the probability axioms.

Note that the conclusion of this argument is only that a rational person's degrees of belief ought to conform to the probability axioms, not that everybody's degrees of belief will in fact so conform. After all, most people probably have degrees of belief that don't sum to 1 for at least some sets of exclusive and exhaustive propositions. So the most that the subjective interpretation can say is that *rational* degrees of belief are an interpretation of the probability calculus, not that all actual degrees of belief are.

Note also that while the Dutch book argument shows that your degrees of belief ought to conform to the probability calculus, it does not show that you ought to attach any *particular* number to the proposition 'It will rain today'. You can attach 0.7 or 0.1 or 0.435 or whatever number you like to this proposition, provided only that the degree of belief you attach to 'It won't rain today' is 1 minus that number. The Dutch book argument only shows that your degrees of belief must be 'coherent' (that is, must somehow satisfy axioms (1)–(4)); beyond that it is a matter of subjective choice which degrees of belief you have. Different people can attach different 'subjective probabilities' to the same proposition. The requirement is only that for each person the numbers in question satisfy the probability axioms; but these can be quite different numbers for different people.

It is this last point that makes most people think that we need another notion of probability—objective probability—to cover the idea that coins (or dice, or radium atoms) have certain tendencies to land heads (land 'six' up, decay). For these objective tendencies presumably have definite objective values, even if different people have different degrees of belief in the relevant result occurring.

4.6. Objective Probabilities

There are two competing ways of thinking about objective probability, the *frequency* theory and the *propensity* theory. I shall consider them in turn.

4.6.1. *The Frequency Theory*

The traditional way of making sense of objective probabilities is to equate them with the relative frequencies of results. Thus we equate the probability p of result R (heads, 'six', decay) in situation S (coin-toss, throw of a die, radium atom) with:

> the number of Rs/the total number of Ss.

Note that this only allows us to ascribe probabilities to results that happen in repeatable situations where we have some number of Ss, and not to all propositions, as on the subjective theory. But this is not a criticism, since it is arguable that the notion of objective probability only applies to such repeatable situations, and not to once-off propositions like 'Prince Edward will get married this year'.

An obvious problem facing the above definition is to know which 'total number' of trials S we should consider. It cannot normally be the actual trials of kind S, since these will normally be finite in number. The problem here is that we know (since it follows from the probability axioms) that there is always a non-zero probability that the relative frequency after N trials will be different from p, if N is finite. For example, it is entirely possible (indeed highly likely) that 1,000 tosses of coins with a 0.5 objective probability of heads will end up with something other than exactly 500 heads. So there is no guarantee at all that the relative frequency in any finite number of trials will equal the objective probability.

Because of this, the frequency theory standardly defines probabilities, not in terms of frequencies in finite sets of trials, but rather in terms of the proportion of Rs that would occur *if* trial S *were* repeated infinitely many times.

This appeal to infinite sequences of trials raises a technical difficulty. For the notion of a *proportion* of Rs in an infinite sequence of Ss does not make sense. If we toss a fair coin an infinite number of times, then there will be an infinite number of heads and an infinite number of tails. So the proportion of heads in the total number of tosses is infinity divided by infinity, which is nonsense. The way round this difficulty is to equate the probability with the *limit* of the *finite* relative frequency of Rs in the first n Ss, as n gets bigger and bigger. More precisely, we can say that the relative frequency of m Rs in the first n Ss *tends* towards such a *limit* p (and then equate the objective probability with this p) if

> for any e, however small, there is an N, such that, for all $n > N$,
> $-e < m/n - p < +e$.

(This is just the standard mathematical idea of a limit—a number such that, for any tiny region around it, the relative frequency will eventually stay within that region once you have gone far enough along the sequence.)

However, even if the frequency theory can deal with this technical problem raised by infinite sequences, many philosophers still feel unhappy with defining probabilities in terms of hypothetical facts about what *would* happen *if* S happened infinitely often. Since most Ss, like coin-tosses, dice-throws, or indeed atomic-decays, do not actually happen infinitely often, this means that we are trying to define objective probabilities in terms of non-existent, imaginary facts. This has persuaded many philosophers to seek an alternative approach to objective probabilities.

4.6.2. *The Propensity Theory*

The propensity theory of objective probability turns away from the idea of relative frequencies in repeated trials, and argues that we should simply take objective probability as a *primitive* notion which measures the strength of the *propensity* for each particular S to produce R. Propensity theorists normally use the term 'chance' to refer to this quantity. So when they say that the chance is 0.4 that this coin will land heads when I toss it, they simply mean that this particular combination of coin and tosser has a 0.4 tendency to produce heads.

The propensity theory has the disadvantage that it does not define probability, but simply takes it as primitive. On the other hand, it has the advantage that it does not need to appeal to the non-existent infinite sequences of the frequency theory. Which of these two theories you prefer will depend mainly on whether you think the infinite sequences are a price worth paying for an explicit definition.

At first sight it might seem that the propensity theory will find it harder than the frequency theory to explain how we *find out* about objective probabilities. For surely our knowledge of objective probabilities comes from the observation of frequencies. Yet the propensity theory seems to deny any link between objective probabilities and frequencies.

However, propensity theorists can retort that they do recognize a perfectly good link between objective probabilities and frequencies, even if not a link that defines the former in terms of the latter. For they can point out that it is a theorem of the probability calculus that

> in a sequence of n trials each with probability p of result R, the *probability* that the relative frequency of Rs will be close to p can be made as high as you like, by making n big enough.

This does not yield a definition of probability in terms of frequency, since it *uses* the notion of probability in explaining the link between probability and frequency (note the emphasized 'probability' in the statement of the theorem). But it is still a link that entitles us to take frequencies as evidence for probabilities.

True, they don't provide sure-fire evidence, since even for large *n* it is only *probable* that the frequency will be close to the probability, not certain. But this problem ('the problem of statistical inference') is not a problem for the propensity theory alone. After all, even frequency theorists have to find out about objective probabilities on the basis of *finite* frequencies (since we never observe infinite sequences). So they also face the problem that it is at most probable, not certain, that the objective probability (that is, for frequency theories, the frequency in the infinite limit) will be close to the observed frequency.

This problem of statistical inference is just one aspect of the philosophy of probability that we cannot deal with any further here. Our treatment of both objective and subjective probability has done little more than scratch the surface of these topics. But we now have enough to continue our discussion of confirmation theory.

Let me just make one further point before returning to the main line of argument. So far I have said nothing about the connection between subjective and objective probability. There is no question but that these are distinct notions, as I pointed out earlier. But this does not mean that they are unconnected. More specifically, the following principle seems to encapsulate an important connection:

> If you know that the objective probability of *R* at time *t* is *p*, then at *t* your degree of belief in *R* ought to equal *p*.

This idea seems almost too obvious to be worth saying. Of course, if I know that this coin now has an objective probability of 0.5 for heads, I will make my degree of expectation for this outcome equal 50 per cent. However, it is worth observing, before we leave this topic of probability, that none of the theories of objective and subjective probability outlined above offers any obvious explanation of why this principle is true. Once more, there is more to probability than we have been able to deal with here.

4.7. Bayesian Confirmation Theory

I now return to the topic of confirmation theory. In the rest of this section I shall concentrate on Bayesian confirmation theory. Bayesians are philosophers who think that we can use the notion of subjective probability to explicate the relation of confirmation. This is not necessarily the only way of thinking about confirmation. But Bayesianism offers a powerful and uniform way of thinking about issues of confirmation. In particular, as we shall see, it yields natural solutions to the two paradoxes of confirmation described earlier.

The initial assumption made by Bayesian confirmation theory is that our attitudes towards theories are measured by the subjective probabilities we attach to them. So if I fully believe a theory, I give it a subjective probability of 1; whereas if I regard it as a hazardous speculation, I give it a subjective probability close to 0.

Bayesians then say that a piece of evidence *E confirms a theory T* if learning *E* should make people *increase* their probability for *T*. (In the rest of this section I shall tend to omit the qualification 'subjective'; unless I say otherwise, 'probability' will mean 'subjective probability'.)

In order to develop the Bayesian theory further, we need the notion of conditional probability. The *conditional probability of A given B* (written 'Prob(A/B)') is defined as the quotient Prob(*A* and *B*)/Prob(*B*), and can be thought of as the probability-of-*A*-on-the-assumption-that-*B*-is-true. To see why, note that Prob(*B*) is a measure of the likelihood of *B* happening, while Prob(*A* and *B*) is a measure of the likelihood of *A also* happening when *B* happens. So if we divide Prob(*A* and *B*) by Prob(*B*) we get a measure of the likelihood of *A* happening *given* that *B* has happened.

Now consider the case where *E* is some possible evidence and *T* is some theory. Prob(T/E) is then the probability of *T*, on the assumption that *E* is true. Bayesians therefore argue that when you learn *E*, you should increase your probability for *T* to equal this number. So for Bayesians *E will confirm T*, in the sense that discovering *E* will increase the probability we attach to *T*, if and only if *Prob(T/E) is greater than Prob(T)*. (In fact this claim is rather less straightforward than it may seem at first sight. But I shall assume it henceforth. For further discussion see the further reading.)

We can say more about *when E* will confirm *T* if we take note of *Bayes's theorem*, originally discovered by the English clergyman Thomas Bayes in the eighteenth century. This theorem follows quickly from the definition of conditional probability. According to this definition Prob(T/E) = Prob(*T* and *E*)/Prob(*E*), while Prob(E/T) = Prob(*T* and *E*)/Prob(*T*). Putting these two together, we can derive

$$\text{Prob}(T/E) = \text{Prob}(T) \times \text{Prob}(E/T)/\text{Prob}(E).$$

This is Bayes's theorem. Its significance is that it tells us that Prob(T/E) is greater than Prob(T)—that is, *E* confirms *T*—if and only if *Prob(E/T) is bigger than Prob(E)*. This is pre-theoretically just what we would expect. For it says that *E* confirms *T* to the extent that *E* is likely given *T*, but unlikely otherwise. In other words, if *E* is in itself very surprising (like light bending in the vicinity of the sun) but at the same time just what you would expect given your theory *T* (the general theory of relativity) then *E* should make you increase your degree of belief in *T* a great deal. On the other hand, if *E* is no more likely given *T* than it would be on any other theory, then observing *E* provides no extra support for *T*. The movement of the tides, for example, is no great argument for general relativity theory, even though it is predicted by it, since it is also predicted by the alternative Newtonian theory of gravitation.

4.8. The Paradoxes Resolved

Let us now consider how this Bayesian approach to confirmation deals with the paradoxes of confirmation.

4.8.1. *The Raven Paradox*

First, the paradox of the ravens. The assumptions generating this paradox, remember, are (1) that generalizations are confirmed by their instances, and (2) that confirmation bears equally on logically equivalent propositions. The standard Bayesian response to this paradox is to accept both these assumptions, and therewith the apparently absurd conclusion that a white shoe does confirm that all ravens are black. But Bayesians then explain this appearance of absurdity by saying that a white shoe only confirms this hypothesis a *tiny bit*, by comparison with the confirmation it gets from a black raven.

Let me use some simple figures to illustrate the point. Suppose that you initially think that about $\frac{1}{4}$ of physical objects are black, and that about $\frac{1}{10}$ are ravens. (This isn't very realistic, but let's keep the figures simple.) Then, in the absence of any special views about the colours of ravens, your probability for the next object you see being a black raven will be $\frac{1}{40}$, and for its being a non-black non-raven will be $\frac{27}{40}$ (and similarly for its being a non-black raven $\frac{3}{40}$ and a black non-raven $\frac{9}{40}$).

Now consider the conditional probability of a black raven, and a non-black non-raven, on the assumption (T) that all ravens are black. This assumption will tend to increase your probability for both these observations, simply because it decreases the probability that you will see a non-black raven from $\frac{3}{40}$ to zero. Suppose this conditional probability for a black raven is $\frac{2}{40}$, for a non-black non-raven $\frac{28}{40}$ (and for a black non-raven $\frac{10}{40}$).

Now we can apply Bayes's theorem. The initial probability of a black raven is $\frac{1}{40}$, while the conditional probability given T is $\frac{2}{40}$. So, whatever your initial probability for the hypothesis that all ravens are black (equivalently, for all non-blacks are non-ravens), Bayes's theorem tells us that an observation of a black raven will double it. By contrast, where the initial probability of a non-black non-raven is $\frac{27}{40}$, the probability conditional on T is only $\frac{28}{40}$. So the observation of a white shoe will only increase our degree of belief in the hypothesis by $\frac{1}{36}$ths. The point is that the hypothesis that all ravens are black makes the observation of a black raven significantly less surprising than it would otherwise be; whereas the observation of a non-black non-raven, never very surprising to start with, becomes only marginally less so on the hypothesis that all ravens are black. So black ravens confirm the hypothesis a lot; white shoes confirm it scarcely at all.

It is perhaps surprising to learn that white shoes give *any* support to the hypothesis that all ravens are black, even if only a small amount. But we can see that, with any realistic figures, this support would be so minuscule that it would be very odd indeed to say, in an everyday context, that a white shoe gives us any reason to believe that all ravens are black. So this is how Bayesians deal with the raven paradox: they do not deny that white shoes confirm that all ravens are black; they just point out that they confirm it so little as to make no difference in an everyday context.

4.8.2. *The Tacking Paradox*

Now for the 'tacking' paradox. Recall that the assumptions here were (1) that theories are confirmed by the observation of anything they entail, and (2) that any evidence that confirms a theory also confirms its consequences. Many people, as I said earlier, think that there must be something wrong with (1), since it allows that a theory-plus-a-'tacked-on'-part (Newtonian-theory-plus-the-moon-is-green-cheese, say) is confirmed by the predictions of the original theory (the planets move in ellipses), which seems odd.

However, Bayesians are committed to (1). For if some T entails E, then $\text{Prob}(E/T) = 1$. So, as long as E is not itself necessarily true, with an unconditional probability of 1, E must confirm T, by Bayes's theorem.

But Bayesians point out that this is consistent with E only confirming T in the sense that it increases the probability of some part of T, while leaving the probability of the rest of T untouched. So, for example, Bayesians would say the motion of the planets only confirms Newtonian-theory-plus-the-moon-is-green-cheese in the sense that it increases the probability of Newtonian theory itself, while being irrelevant to the green-cheese part of the joint hypothesis.

In line with this, Bayesians will deny (2). For, when some evidence confirms some theory only in the sense that it increases the probability of part of it, while leaving the rest untouched, then we would expect that evidence to confirm the consequences only of that part, and not of the other part. So while the motion of the planets confirms the joint thesis Newtonian-theory-plus-the-moon-is-green-cheese, it does not confirm the consequence that the moon is made of green cheese, or anything that follows from this.

Not only does this Bayesian line offer a natural solution to the tacking paradox, it is also helpful in thinking about the relation between theory and observation generally. Much recent philosophy of science has inferred, from the Duhem–Quine observation that specific theoretical assumptions only generate predictions with the help of auxiliary hypotheses, that the relation between theory and evidence is irredeemably *holistic*, in that it is always the totality of our beliefs about the world that is confirmed or refuted by evidence. But the Bayesian approach shows that even if predictions are generated by a *conjunction* of assumptions, that evidence can support different elements of that conjunction to different degrees.

4.9. Problems for Bayesianism

It should not be forgotten that Bayesianism confirmation theory is derived from the notion of *subjective* probability. As I pointed out earlier, there is nothing in the idea of subjective probability to ensure that different people will attach the same subjective probabilities and conditional probabilities to any set of assumptions, provided they each arrange their own probabilities 'internally' in such a way as to satisfy the probability calculus. Can such a subjective notion really provide a satisfactory basis for the apparently objective notion of how much theories are confirmed by the existing evidence? Surely we do not want to allow that I can be right to hold that the evidence shows relativity theory to have probability 0.8, while you can be equally right to think it has probability 0.2.

Different Bayesians make different responses to this challenge. Some simply say that Bayesianism is only a theory of how to change your probabilities, and not of which probabilities you ought to start or finish with. On this view, Bayes's theorem shows us how to update our subjective probabilities, given that we start with certain initial conditional and unconditional probabilities; but it says nothing about what those initial probabilities ought to be, and therefore nothing about which final probabilities we ought to end up with. There isn't anything wrong with my ending up thinking relativity theory has probability 0.8, while you think it has 0.2, provided we both reached this end-point by updating our initial probabilities in response to the evidence in the way required by Bayes's theorem.

Many Bayesians, however, find the possibility of such divergence worrying, and so offer a more ambitious answer. They say that, whatever your initial degrees of belief, Bayes's theorem will ensure convergence of opinion. The idea is that, given enough evidence, everybody will eventually end up with the same probabilities, even if they have different starting-points. There are a number of theorems of probability theory showing that, within limits, differences in initial probabilities will be 'washed out', in the sense that sufficient evidence and Bayesian updating will lead to effectively identical final degrees of belief. So in the end, argue Bayesians, it does not matter if you start with a high or low degree of belief in relativity theory—for, after a number of observations of light bending, gravitational red-shifts, and so on, you will end up believing it to a degree close to one anyway.

However, interesting as these results are, they do not satisfactorily answer the fundamental philosophical worry. For they do not work for *all* possible initial degrees of belief. Rather they assume that the people in question, while differing among themselves, all draw their initial degrees of belief from a certain range. While this range includes most initial degrees of belief that seem at all plausible, there are nevertheless other possible initial degrees of belief which are consistent with the axioms of probability, but which will not lead to eventual convergence. So, for example, the Bayesians do not in fact explain what is wrong with people who never end up believing relativity theory because they always think it is probable that the course of nature is going to change tomorrow.

This seems to me to show that Bayesianism provides at best a partial account of confirmation. Bayesianism shows us how our initial degrees of belief constrain the way we should respond to new evidence. But it needs to be supplemented by some further account of why some initial degrees of belief are objectively superior to others. Perhaps one way to fill this gap would be to appeal to the kind of 'simplicity' mentioned in Section 3.6 above. But it would take us too far afield to pursue this issue here.

5. EXPLANATION

5.1. The Covering-Law Model

Our main concern so far has been our knowledge of general truths. In this section I shall focus on the use to which this knowledge is put in *explanation*. Both in science and in everyday life the aim of investigation is often to find an explanation for some puzzling phenomenon. But what exactly is an explanation? And how does our knowledge of general truths contribute to our ability to explain?

Most modern discussion of explanation starts with Carl Hempel's 'covering-law model'. Let me first illustrate this model for the case where the item to be explained is some particular event, such as that the ice in your water-pipes froze last Tuesday, or that it rained this morning. According to Hempel, the explanation of any such event conforms to the following schema:

Initial conditions: $I_1, I_2, \ldots, I_n$
Laws: L

Explained event: E.

So, for example, we might explain the fact E that it rained this morning by citing the initial conditions I_1, I_2 that there was a certain level of humidity and that the atmospheric pressure fell to a certain level and the law L that such a fall of pressure in such humidity is always followed by a precipitation of rain.

The law in such an explanation 'covers' the initial conditions and consequent event, in the sense that it shows that the sequence of events behind a particular occurrence is simply an instance of a general pattern. The fact that gets explained, E, is sometimes referred to as the 'explanandum', and the facts that do the explaining, the *I*s and L, as the 'explanans'. Note that while I have represented the law involved in the explanans as a single proposition, L, in most cases we will need a conjunction of simpler laws to see why E follows the relevant Is. For example, we would need both Newton's second law and the law of gravitation to explain why a meteor moves in the way it does.

Note also that, on this model of an explanation, explaining an event is the same thing as *deducing* it from initial conditions and laws. Given the initial conditions and a law which says that in general such initial conditions are followed by an E, then logic alone allows us to infer that the explanandum occurs. Because they involve *deduction* via a *law* in this way, such explanations are often called 'deductive-nomological' explanations, or 'D-N' explanations for short. (There is a variant of the covering-law model which allows probabilistic rather than deterministic laws, and in which this requirement of deducibility is therefore relaxed. So 'covering-law' is strictly a wider term than 'deductive-nomological'. But for the moment let us stick to deductive cases, and leave probabilistic explanations to one side.)

It is worth being clear that the idea of a 'deductive' explanation does not assume that the law L can somehow be 'deduced' from first principles in an a priori way. Such laws still have to be established by *induction* from past observations of results. The idea is simply that, if we have established such a law, then it will deductively imply, together with suitable initial conditions, certain further results.

The covering-law model implies a certain symmetry between explanation and prediction. The structure of explanations, in which we deduce that E had to occur from initial conditions and laws, exactly parallels the structure of predictions, in which we deduce that E is *going* to occur from the same initial conditions and laws. For example, if we can explain its raining this morning by the prior conditions and the relevant law, then we could presumably have predicted its raining beforehand on the basis of the same information. So for the covering-law model the difference between explanation and prediction depends only on whether you know the explanandum before you deduce it from the explanans. If you already know E, then deducing it from initial conditions and laws will serve to explain it. If you do not already know E, then the same deduction will serve to predict it. A prediction tells you what to expect. An explanation shows you that what you already know was only to be expected.

5.2. Theoretical Explanation

In the last section I considered explanations of particular events, such as its raining this morning, or a particular meteor taking a certain path. However, the covering-law model is also designed to accommodate explanations of laws as well as explanations of particular events. For example, suppose you are puzzled about some general law, such as, say, that there is always a rainbow when you look towards rain with the sun at a given angle behind you. I can explain this by

showing that it follows from the laws that (1) sunlight involves a mixture of all wavelengths of light, (2) these different wavelengths refract differently on moving from light to water, and (3) raindrops are of a shape which will lead to internal reflection off the back of the raindrop. Here I have explained one law by reference to other laws. Schematically:

$$\frac{\text{Explanans: } L_1, L_2, \ldots, L_n}{\text{Explanandum: } L.}$$

Because the explanandum here is a general truth, and not a particular event occurring at some specific place and time, it is not necessary that initial conditions be involved in the explanation. But despite this difference, it is still a *deductive* explanation from *laws*, and so still a species of 'deductive-nomological' explanation. Explanations of this kind are often called 'theoretical explanations', to distinguish them from 'particular explanations'.

The possibility of theoretical explanations shows how the covering-law model can respond to a common initial objection. Consider once more the particular explanation of this morning's rain offered in the last section. Somebody might say that it is all very well to attribute this morning's rain to the drop in pressure and the humidity, but object that this is no kind of explanation until you have shown why drops in pressure at high humidity are in general followed by rain.

The covering-law model can respond by insisting that the explanation of a particular rainfall this morning is one thing, and the explanation of the law that falls in pressure at high humidity are followed by rain is another. If you want explanations of both, you can have them. But it does not follow that you have not explained the first, the particular rainfall, until you have also explained the second, the law which accounts for the particular rainfall.

Indeed, it would be obviously self-defeating to require that all explanations should contain explanations of the facts adduced in the explanans. We would be forced into infinite regress. As soon as we explained the laws which originally appeared in the explanans by other laws, we would then have to explain those other laws by further laws, and so on. And, in the case of particular explanations, there would be an additional regress, for we would need to explain the initial conditions mentioned in the explanans (why did the pressure drop? why was the humidity high?), and this would require mention of yet prior initial conditions, which would themselves need to be explained, and so on again.

So it does not make any sense to demand that in an explanation the explanatory facts should always be explained too. This is not because there is anything wrong with asking for further such explanations in specific cases. It is just that we cannot give the answer to an infinite number of questions in a finite time.

Exactly which explanatory questions have to be answered in order to yield explanatory satisfaction is an interesting question. But it probably is not one that admits of any general answer. Whether an explanation is satisfying depends on the practicalities of what is being explained to whom. Any given person has a rough overall picture of the world, in which earlier events lead on to later ones according to familiar patterns. But some phenomena do not fit into these pictures. The role of an explanation is to show how such puzzling phenomena can be fitted in. However, different people can be puzzled by different aspects of a given situation, in that the situation will fail to fit into their respective pictures in different ways. And then different explanations will be needed to satisfy them.

5.3. Do All Explanations Fit the Covering-Law Model, and Vice Versa?

In this section I want to start raising some questions about the adequacy of the covering-law model. The covering-law model was originally proposed by Hempel as an analysis of the intuitive pre-analytic notion of a scientific explanation. So it is possible to ask whether this analysis is adequate. This question has two parts. (A) Is it true that every scientific explanation is an instance of the covering-law pattern? (B) Conversely, does every instance of the covering-law pattern amount to a scientific explanation?

5.3.1. Do All Explanations Fit the Covering-Law Model?

Let me start with (A). Consider this example. Little Katy contracts chicken-pox. You want to know why. You are told that she played with Miranda, who had it. This seems a perfectly cogent explanation. However, it does not seem to conform to the covering-law model. Suppose we think of playing with another child with chicken-pox as an initial condition in a covering-law deduction of Katy getting chicken-pox. Then we need as the law something like 'Whenever a child who has not had chicken-pox plays with another child who has it, the first child gets it too'. But there isn't any such law. There are plenty of cases where children do not come down with chicken-pox after playing with another child with it, even if they have not had it before.

So this is a prima-facie counter-example: an intuitively satisfactory explanation that does not fit the covering-law model.

5.3.2. Are All Instances of the Covering-Law Model really Explanations?

Question (B) raised the converse issue: is every instance of the covering-law pattern really an explanation? Here is one case which is not.

I_1, I_2: The barometer fell this morning; and the humidity was high.
L: Whenever the barometer falls in high humidity, it rains.
―――――――――――――――――――
E: It rained this morning.

This deduction conforms perfectly to the requirements of the covering-law model of a particular explanation. But intuitively it just is not a satisfactory explanation. The barometer falling might account for how you know it is going to rain. But its actually raining is a different fact from your knowing it will rain. And intuitively it seems quite wrong to say that the barometer's fall was responsible for the rain itself.

Here are some similar cases.

I_1, I_2: The shadow of a pole *P* is *n* feet long; and the sun is at angle *a*.
L: Whenever a pole casts a shadow of *n* feet with the sun at angle *a*, the pole itself is *m* feet high.
―――――――――――――――――――
E: Pole *P* is *m* feet high.

I: Star *S* emits red-shifted light.
L: All stars with red-shifted light are rapidly receding.
―――――――――――――――――――
E: Star *S* is rapidly receding.

Both of these seem impeccable cases of covering-law deductions. But, again, it seems quite wrong to say that the pole is *m* feet high because its shadow is *n* feet long, or that the star is rapidly receding because its light is red-shifted.

5.3.3. Explanations that are not Predictions, and Vice Versa

Let me make a general point about the two kinds of counter-example to the covering-law model raised in this section. The covering-law model is committed, as I pointed out earlier, to the view that every explanation is a potential prediction, and vice versa. So if we can find explanations that are not potential predictions, then we will have examples of explanations that do not fit the covering-law model. The example of Katy and the chicken-pox was of this sort. You cannot immediately predict that she will get it just from knowing she played with another infected child. Because it insists that all explanations should be potential predictions, the covering-law model has trouble admitting these prima-facie plausible explanations.

Then we looked at the converse kind of example, predictions that are not in fact explanations, and which therefore get counted as explanations by the covering-law model when they should not be. You can predict the rain from the barometer's fall, or the pole's height from the length of its shadow, or the star's recession from its red-shift. And so, because it accepts that all potential predictions are explanations, the covering-law model has trouble ruling out these prima-facie non-explanations.

5.4. Probabilistic Explanation

Defenders of the covering-law model can make various responses to these counter-examples. Let me first consider counter-examples of type (A), namely, intuitively satisfactory explanations, like that of Katy's chicken-pox, which do not fit the covering-law model.

One possible response here would be to argue that, if Katy's chicken-pox isn't predictable for lack of any law that says she was sure to get it in the circumstances, then, despite first appearances, her playing with Miranda does not explain it. (After all, other children in contact with the infection sometimes do not get chicken-pox. So why suppose that Katy's contact with Miranda is in itself enough to explain her getting the disease?) This line would save the covering-law model of explanation by denying that the apparent counter-example was really a genuine example of an explanation.

However, this move seems unattractive. It would be very odd to deny that Katy got chicken-pox because she played with Miranda. So most covering-law theorists of explanation, from Hempel onwards, have weakened the requirements of the covering-law model to allow that there can be explanations that appeal to *probabilistic* laws, rather than exceptionless ones. After all, in our example it is presumably true that most children in contact with chicken-pox get it themselves, and this means that we can at least anticipate that Katy would get it from Miranda with high probability, if not with certainty. Accordingly Hempel put forward the following model of '*inductive-statistical* explanations' as another species of covering-law explanations alongside 'deductive-nomological explanations'.

Initial conditions: $I_1, I_2, \ldots, I_n$
Probabilistic laws: L, to the effect that *most* $I_1, \ldots, I_n$s are Es
Explained event: E.

Explanations fitting this schema are 'inductive' because the premisses do not deductively imply the conclusion, but only indicate it has a high probability; and they are 'statistical' because they appeal to probabilistic laws rather than exceptionless ones.

Note that Hempel's 'inductive-statistical' model requires that the explanans should give the explanandum a high probability. It is not clear that this is quite the right requirement for probabilistic explanation. Suppose that John Smith gets lung cancer. In explanation we are told that he has smoked fifty cigarettes a day for forty years. Intuitively this seems like a good explanation. But note that

the explanans here does not give the explanandum a *high* probability. Even people who smoke fifty a day for a long period still have a low probability of getting lung cancer in absolute terms. What is true, though, is that they have a much *higher* probability of getting lung cancer than if they did not smoke. Because of this, a number of theorists have suggested that Hempel's requirement of *high* probability be replaced by the different requirement that the initial conditions merely increase the probability of the explanandum, compared to the probability if those initial conditions were absent. (Wesley Salmon has dubbed this the 'statistical-relevance' model, as opposed to the 'inductive-statistical' model, because the requirement is in effect that the initial conditions should be probabilistically relevant to the explanandum.[9])

This issue of probabilistic explanation raises a number of further questions which cannot be resolved here. Most obviously, you might have been wondering whether the probabilistic laws appealed to in such explanations are supposed to be reflections of genuine indeterminism or whether they simply reflect our ignorance of the full set of initial conditions which do determine Katy's chicken-pox, John Smith's cancer, etc. Different answers to this question will lead to different views of probabilistic explanation. But there is little agreement among philosophers on how it should be answered.

5.5. Causation and Explanation

Let me now consider the other kind of counter-example, instances of the covering-law model which are not in fact explanations, such as the deduction of the rain from the barometer's fall, or the pole's height from its shadow's length, or the star's recession from its red-shift.

The obvious reason why these deductions are not in fact explanations is that the initial conditions do not specify the *cause* of the explanandum event. Instead they deduce the explanandum event from a *symptom* (like the barometer's fall) or an *effect* (like the shadow's length, or the red-shift).

The obvious remedy is to add to the covering-law account the further requirement that in explaining particular events the initial conditions should always include the cause of the explanandum event. I think this is the right move. But it calls for a number of comments.

5.5.1. *The Direction of Causation*

In a sense this move simply shifts the original problem into the analysis of causation. The barometer/pole/red-shift counter-examples arose because the

[9] See W. Salmon, *Statistical Explanation and Statistical Relevance* (Pittsburgh, 1971).

original requirements of the covering-law model failed to ensure that the 'causal arrow', so to speak, pointed from the initial conditions to the explanandum. We can remedy the defect by appealing to the existence of such a directed arrow between causally related events, and requiring that genuine explanations proceed in the same direction as this arrow.

But that causation does have such a direction is itself a problematic assumption. Consider Hume's equation of causation with constant conjunction. This in itself does not tell us, given two constantly conjoined events, which lies at the tail of the arrow, and is therefore the cause, and which lies at the point, and is therefore the effect.

So something needs to be added to Hume's constant conjunction analysis to put the direction into causation. How to do this is a matter of active controversy. Hume himself argued that, given two constantly conjoined events, it is always the *earlier* which is the cause and the *later* the effect. But this appeal to temporal precedence is not entirely satisfactory. (After all, the barometer's fall precedes the rain, but still does not cause it. And cannot some causes be simultaneous with their effects?)

I do not propose to pursue this issue any further here. Even if it is unclear how to *account* for causal direction, it is intuitively clear that causation *has* a direction, and that requiring that explanations follow this direction is the way to rule out the barometer/pole/red-shift counter-examples.

5.5.2. Are All Explanations of Particular Events Causal?

It is not clear that it is appropriate to impose the requirement that the explanans should mention a cause on *all* explanations of particular events. Suppose we explain why some frozen substance is water by citing the fact it is H_2O; or suppose we explain why something has a temperature *t* by citing the fact that the mean kinetic energy of its molecules is *k*. These are arguably reasonable explanations. But being made of H_2O does not cause something to be water, so much as constitute its being water. Similarly, having a mean molecular kinetic energy of *k* does not cause the temperature of *t*, but again constitutes it.

Perhaps these are not really explanations in the same sense as most explanations. They do seem slightly peculiar, at least to my ear.

Still, even if we do count them as mainstream explanations, it is of no great importance in the present context. To rule out the barometer/pole/red-shift counter-examples we need to require *some* stronger link between explanans and explanandum than demanded by the original covering-law model. Maybe requiring a specifically causal link is too strong, because we will then rule out the H_2O/mean kinetic energy explanations too. If so, then the solution is simply to say that we need a *metaphysical* link of one sort or the other, in which the explanans either causes *or* constitutes the explanandum.

5.5.3. Teleological Explanations

Perhaps one reason why Hempel and other early proponents of the covering-law model were unwilling to impose this kind of metaphysical link was that there is one important class of explanations in which the explanans neither causes nor constitutes the explanandum. These are the *functional*, or *teleological*, explanations which play such a central role in biology, such as 'Plants contain chlorophyll so that they can photosynthesize' or 'Polar bears are white so that they cannot be seen'. Indeed these explanations are striking precisely because the item that gets explained (chlorophyll, whiteness) is the *cause*, not the effect, of the item that does the explaining (photosynthesis, camouflage).

If we take these explanations at face value, then it is not open to us to require that (non-constitutive) explanations always go from cause to effect. For these explanations seem to go in just the other direction.

Until fairly recently most philosophers of science did take such explanations at face value. Thus Hempel himself regarded teleological explanations as simply another way, alongside normal causal explanations, of exemplifying the covering-law model: the only difference is that in causal explanations the explaining fact (lower temperature) temporally precedes the explained fact (freezing), whereas in functional explanations it is the explained fact (white fur) that comes temporally before the consequence (camouflage) which explains it.

Most contemporary philosophers of science, however, take a different view, and argue that functional explanations, despite appearances, are really a sub-species of causal explanations. On this view, the reference to future effects in functional explanations is merely apparent, and such explanations really refer to past causes. In the biological case, these past causes will be the evolutionary histories which led to the natural selection of the biological trait in question. Thus the functional explanation of the polar bears' colour should be understood as referring us to the fact that their *past* camouflaging led to the natural selection of their whiteness, and not to the fact that they may be camouflaged in the future.

If we take this line on functional explanations, then we can continue to uphold the requirement that all (non-constitutive) explanations should flow from cause to effect, and so deal with the barometer/pole/red-shift difficulty in the way suggested.

BIBLIOGRAPHY

A classic introduction to the problem of induction is section 6 of Bertrand Russell, *The Problems of Philosophy* (Oxford, 1967). For the Popperian response to the problem of induction, see section 1 of Karl Popper, *The Logic of Scientific Discovery*, 1st Eng. edn. (London, 1959). A critical discussion of Popper's views can be found in sections 2 and 3 of Anthony O'Hear, *An Introduction to the Philosophy of Science* (Oxford, 1989). A reliabilist approach to induction is defended in section 4 of David Papineau, *Philosophical Naturalism* (Oxford, 1993). For Goodman's 'new problem of induction', see section 3 of Nelson Goodman, *Fact, Fiction and Forecast* (London, 1954) and the paper by S. Barker and P. Achinstein, 'On the New Riddle of Induction', and the reply by Goodman, 'Positionality and Pictures', in P. H. Nidditch (ed.), *The Philosophy of Science* (Oxford, 1968).

A good general introduction to the problem of distinguishing laws from accidents, as well as a defence of his own non-Humean view, is given by David Armstrong in his *What is a Law of Nature?* (Cambridge, 1983).

Much of the contemporary debate between realists and instrumentalists focuses on the arguments in Bas van Fraassen, *The Scientific Image* (Oxford, 1980). These arguments are discussed further in P. Churchland and C. Hooker (eds.), *Images of Science* (Chicago, 1985).

There are two excellent books on Bayesianism, both of which also provide a general introduction to confirmation theory and concepts of probability: Paul Horwich, *Probability and Evidence* (Cambridge, 1982), and Colin Howson and Peter Urbach, *Scientific Reasoning* (La Salle, Ill., 1989).

The starting-point for all modern discussions of explanation is section 4 of Carl Hempel's *Aspects of Scientific Explanation* (New York, 1965). For more recent debates, see David Ruben, *Explaining Explanation* (London, 1990) and the essays in David Ruben (ed.), *Explanation* (Oxford, 1993).

4

METAPHYSICS

Tim Crane and David Wiggins