Assigned: 26 Jan '07 Topic: What is science? (part II)

Again, remember that our ultimate question is whether computer science is a science.

Required

- *Papineau, David (1996), "Philosophy of Science" [PDF], in Nicholas Bunnin & E.P. Tsui-James (eds.), The Blackwell Companion to Philosophy (Oxford: Blackwell): 290-324.*
  * **pp.290-295, 298-310; skim the rest.**

Strongly Recommended

- Read the rest of Papineau 1996

- *Popper, Karl R. (1953), "Science: Conjectures and Refutations", from his Conjectures and Refutations: The Growth of Scientific Knowledge (New York: Harper & Row, 1962).*

9
Philosophy of Science

David Papineau

The philosophy of science can usefully be divided into two broad areas. On the one hand is the epistemology of science, which deals with issues relating to the justification of claims to scientific knowledge. Philosophers working in this area investigate such questions as whether science ever uncovers permanent truths, whether objective decisions between competing theories are possible and whether the results of experiment are clouded by prior theoretical expectations. On the other hand are topics in the metaphysics of science, topics relating to philosophically puzzling features of the natural world described by science. Here philosophers ask such questions as whether all events are determined by prior causes, whether everything can be reduced to physics and whether there are purposes in nature. You can think of the difference between the epistemologists and the metaphysicians of science in this way. The epistemologists wonder whether we should believe what the scientists tell us. The metaphysicians worry about what the world is like, if the scientists are right. Readers will wish to consult chapters on EPISTEMOLOGY (chapter 1), METAPHYSICS (chapter 2), PHILOSOPHY of MATHEMATICS (chapter 10), PHILOSOPHY OF SOCIAL SCIENCE (chapter 11) and PRAGMATISM (chapter 26).

1 The Epistemology of Science

1.1 The Problem of Induction

Much recent work in the epistemology of science is a response to the problem of induction. Induction is the process whereby scientists decide, on the basis of various observations or experiments, that some theory is true. At its simplest, chemists may note, say, that on a number of occasions samples of sodium heated on a Bunsen burner have glowed bright orange, and on this basis conclude that in general *all* heated sodium will glow bright orange. In more complicated cases, scientists may move from the results of a series of complex experiments to the conclusion that some fundamental physical principle is true. What all such inductive inferences have in common, however, is that they start with particular premises about a *finite* number of past observations, yet end up with a general conclusion about how nature will *always* behave. And this is where the problem lies. 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 may change, and that the patterns we have observed so far turn out to be a poor guide to the future? Even if all heated sodium has glowed orange up till now, who is to say it will not start glowing blue sometime in the next century?

In this respect induction contrasts with deduction. In deductive inferences the premises 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 premises leaves no room for the conclusion to be anything but true. But in an inductive inference this does not hold. To take the simplest case, if you are told, for properties *A* and *B*, 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.

The problem of induction seems to pose a threat to all scientific knowledge. All scientific discoveries worth their 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 the authority of all these laws in question. For if our evidence is simply that these laws have worked so far, then how can we be sure that they will not be disproved by future occurrences?

1.2 Popper's falsificationism

One influential response to the problem of induction is due to Sir Karl Popper (1902–94). In Popper's view (1959a, 1963, 1972), science does not rest on induction in the first place. Popper 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 premise, 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 *A*s are *B*s. There is no room here for the premise to be true and the conclusion false. If we discover that some body falls with *increasing* acceleration (say because it falls from a great height, and so is

subject to a greater gravitational force as it nears the earth), then we know for sure that all bodies do not fall with constant acceleration. 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 and refutations. Scientific theories are put forward as hypotheses, and they are replaced by new hypotheses when they are falsified. However, if scientific theories are always conjectural in this way, 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 (1959a, ch. 2). 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 astrologists 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 the 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 that 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. Similary, 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.

1.3 *The failings of falsificationism*

At first sight Popper seems to offer an extremely attractive account of science. He explains its superiority over other forms of belief, while at the same time apparently freeing it from any problematic dependence on induction. Certainly his writings have struck a chord within the scientific community. Popper is one of the few philosophers ever to have become a Fellow of the Royal Society, an honour usually reserved for eminent scientists.

In the philosophical world, however, Popper's views are more controversial. This is because many philosophers feel that his account of science signally fails to solve the problem with which he begins, namely, the problem of induction (for example, see Ayer, 1956, pp. 71–5; Worrall, 1989). The central objection to his position 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 makes science 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 *do not*. Useful scientific knowledge comes in the form 'All As are Bs', not 'It's false that all As are Bs'. Since Popper only accounts for the latter kind of knowledge, he seems leave out what is most interesting and important about science.

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 of whether technologists should *believe* those conjectures, and *rely* on their predictions when, say, they administer 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. After all, Popper claims to have solved the problem of induction. But 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: (1) when I jump from this tenth floor window I shall crash painfully into the ground; (2) when I jump from the window I will float like a feather to a gentle landing. Intuitively, it is more rational to believe (1), which assumes that the future will be like the past, than (2), 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 that believing (2) is no less rational than believing (1).

Something has gone wrong. *Of course* believing (1) is more rational than believing (2). In saying this, I do not want to deny that there is a *problem* of induction. Indeed it is precisely *because* believing (1) is more rational than believing (2) that induction is problematic. Everybody, Popper aside, can see that believing (1) is more rational than believing (2). The problem is then to explain *why* believing (1) is more rational than believing (2), in the face of the apparent invalidity of induction. So Popper's denial of the rational superiority of (1) over (2) is not so much a *solution* to the problem of induction, but simply a refusal to recognize the problem in the first place.

Even if it fails to deal with induction, 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 suggested, theories are often mere conjectures when they are first put forward, and they may remain conjectures as the intitial 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.

Such examples 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. It is only those philosophers who have been bemused by the problem of induction who view these theories as being no better than initial hypotheses. Everybody else who is acquainted with the evidence has no doubt that these theories are proven truths.

1.4 Bayesianism

If we insist, against Popper, that we are fully entitled to believe at least some scientific theories on the basis of past evidence, then we are committed to finding some solution to the problem of induction. One currently popular account of the legitimacy of induction is found within Bayesianism, named after Thomas Bayes (c.1701–61) (Horwich, 1982, Howson and Urbach, 1989).

Bayesians are philosophers who hold that our beliefs, including our beliefs in scientific theories, come in degrees. Thus, for example, I can believe to degree 0.5 that it will rain today, in the sense that I think there is a 50 per cent likelihood of rain today. Similarly, I might attach a 0.1 degree of belief to the theory that the strong nuclear and electro-weak forces are the same force – I think it unlikely, but allow that there is a one-in-ten possibility it may turn out true.

As these examples indicate, Bayesians think of degrees of belief as the extent to which you subjectively take something to be PROBABLE (pp. 161–2). Accordingly, they argue that your degrees of belief ought to satisfy the axioms of the probability calculus. (See the box below for the 'Dutch Book Argument' for this thesis.) It is important to realize, however, that while Bayesians think of degrees of belief as probabilities in this mathematical sense, they still think of them as *subjective* probabilities. In particular, they allow that it can be perfectly rational for different people to attach *different* subjective probabilities to the same proposition – you can believe that it will rain today to degree 0.2, while I believe this to degree 0.5. What rationality does require, according to the Bayesians, is only that if you have a subjective probability of 0.2 for rain, then you must have one of 0.8 for its not raining, while if I have 0.5 for rain, then I must have 0.5 for its not raining. That is, both of us must accord, in our different ways, with the theorem of the probability calculus that Prob(p) = 1 – Prob(not-p).

At first sight, this element of subjectivity might seem to disqualify Bayesianism as a possible basis for scientific rationality. If we are all free to attach whatever degrees of belief we like to scientific theories, provided only that we are faithful to the structure of the probability calculus, then what is to stop each of us from supporting different theories, depending only on individual fads or prejudices? But Bayesians have an answer, namely, that it does not matter what prejudices you start with, as long as you *revise* your degrees of belief in a rational way.

Bayesians derive their account of how to revise degrees of belief, as well as their name, from Bayes' theorem, originally proved by Thomas Bayes in a paper published in 1763. Bayes' theorem states:

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

The simple proof of this theorem is given in the box. But the philosophical significance of the theorem is that it suggests a certain procedure for revising your degrees of belief in response to new evidence. Suppose that H is some hypothesis, and E is some newly discovered evidence. Then Bayesians argue

that, when you discover E, you should adjust your degree of belief in H in line with the right-hand side of the above equation: that is, you should increase it to the extent that you think E is likely given H, 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 H (the general theory of relativity), then E should make you increase your degree of belief in H a great deal. On the other hand, if E is no more likely given H than it would be on any other theory, then observing E provides no extra support for H. The movement of the tides, for example, is no great argument for general relativity, even though it is predicted by it, since it is also predicted by the alternative Newtonian theory of gravitation.

Note in particular that this strategy for updating degrees of belief in response to evidence can be applied to inductive inferences. Consider the special case where H is some universal generalization – all bodies fall with constant acceleration, say – and the evidence E is that some particular falling body has been observed to accelerate constantly. If this observation was something you did not expect at all, then Bayesianism tells you that you should increase your degree of belief in Galileo's law significantly, for it is just what Galileo's law predicts. Of course, once you have seen a number of such observations, and become reasonably convinced of Galileo's law, then you will cease to find new instances surprising, and to that extent will cease to increase your degree of belief in the law. But that is as it should be. Once you are reasonably convinced of a law, then there is indeed little point in gathering further supporting instances, and so it is to the credit of Bayesianism that it explains this.

The Bayesian account of how to revise degrees of belief seems to make good sense. In addition, it promises a solution to the problem of induction, since it implies that positive instances give us reason to believe scientific generalizations.

There are, however, problems facing this account. For a start, a number of philosophers have queried whether Bayes' theorem, which after all is little more than an arithmetical truth, can constrain what degrees of belief we adopt in the future (see the box below). And even if we put this relatively technical issue to one side, it is unclear how far the Bayesian account really answers the worry raised above, that the subjectivity of degrees of belief will allow different scientists to commit themselves arbitrarily to different theories. The Bayesian answer to this worry was that Bayes' theorem will at least constrain these different scientists to revise their degrees of belief in response to the evidence in similar ways. But, even so, it still seems possible that the scientists will remain on different tracks, if they start at different places. If two scientists are free to attach different prior degrees of belief in Galileo's law, and both update those degrees of belief according to Bayes' theorem when they learn the evidence, will they not still end up with different *posterior* degrees of beliefs?

The standard answer to the objection is to appeal to *convergence* of opinion. The idea is that, given enough evidence, everybody will *eventually* end up in the same place, 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 Galileo's law – for after 1,000 observations of constantly falling bodies 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 questions about inductive reasoning. For they do not work for all possible initial degrees of belief. Rather, they assume that the scientists at issue, while differing among themselves, all draw their initial degrees of belief from a certain range. While this range includes all the initial degrees of belief that seem at all intuitively plausible, there are nevertheless other possible initial degrees of belief that 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 Galileo's law because they are always convinced that the course of nature is going to change tomorrow. Of course Bayesians are right to regard such people are irrational. But they do not explain why they are irrational. So they fail to show why all thinkers must end up with the same attitude to scientific theories. And in particular they fail to solve the problem of induction, since they do not show why all rational thinkers must expect the future to be like the past.

Bayesianism

The Dutch Book Argument

The axioms of probability require that

(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.

Bayesians appeal to the 'Dutch book argument' to show why subjective degrees of belief should conform to these axioms. Imagine that your degrees of belief did not so conform. You believe proposition P to degree y, say, and yet do *not* believe not-P to degree 1 – y. (You thus violate the conjunction of axioms (2) and (4), because P or not-P is a necessary truth.) Then it will be possible for somebody to induce you to make bets on P and not-P in such a way that you will lose whatever happens. A set of bets that guarantee that you will lose whatever happens is called a 'Dutch book'. The undesirability of such a set of bets thus provides an argument that any rational person's subjective degrees of belief should satisfy the axioms of the probability calculus.

Bayes' Theorem

The conditional probability of P given Q – Prob(P/Q) – is defined as Prob(P and Q)/Prob(Q). Intuitively, Prob(P/Q) signifies the probability of P on the assumption that Q is true. It immediately follows from this definition that

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

This is Bayes' theorem. As you can see, it says that the conditional probability Prob(H/E) of some hypothesis H given evidence E is greater than Prob(H) to the extent that E is improbable in itself, but probable given H.

Bayesian Updating

Bayesians recommend that if you observe some evidence E, then you should *revise* your degree of belief in H, and set your new $\text{Prob}_{t'}(H)$ equal to your *previous* conditional degree of belief in H given E, $\text{Prob}_t(H/E)$, where t is the time before you learn E, and t' after. Bayes' theorem, applied to your subjective probabilities at t, then indicates that this will *increase* your degree of belief in H to the extent that you previously thought E to be subjectively improbable in itself, but subjectively probable given H.

This Bayesian *recommendation*, that you revise your degree of belief in H by setting it equal to your old conditional degree of belief in H given E, should be distinguished from Bayes' *theorem*. Bayes' theorem is a trivial consequence of the definition of conditional probability, and constrains your degrees of belief at *a given time*. The Bayesian recommendation, by contrast, specifies how your degrees of belief should change *over time*. Bayes' theorem is uncontentious, but it is a matter of active controversy whether there is any satisfactory way of defending the Bayesian recommendation (Hacking, 1967, Teller, 1973).

1.5 Instrumentalism versus realism

At this stage let us leave the problem of induction for a while and turn to a different difficulty facing scientific knowledge. 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, does it not 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 general lines of response to 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 experienced. But 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 SKEPTICAL (pp. 46–58) attitude to such postulates, saying that we have no entitlement to *believe* them (as opposed to using them as an instrument for calculations).

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. One aspect of scientific practice invoked in this connection has been the *unification* of different kinds of theories in pursuit of a single 'theory of everything' (Friedman, 1984); other features of science appealed to by realists have included the use of theories to *explain* observable phenomena (Boyd, 1980), and the reliance on theories to make novel *predictions* (Smart, 1963). For, so the realist argues, these aspects of scientific practice only make 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 no parallel reason why a bunch of instruments should be unifiable into one big 'instrument of everything'. And similarly, the realist will argue, there seems no reason to expect a mere calculating instrument, as opposed to a true description of an underlying reality, to yield a genuine explanation of some past occurrence, or a reliable prediction of a future one.

However, this form of argument tends to be inconclusive. 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. As an example of the first response, they could argue that the unification of science is motivated, not by the pursuit of one underlying truth, but simply by the desirability of having a single all-purpose calculating instrument rather than a rag-bag of different instruments for different problems. The second kind of response would be to deny that unification is essential to science to start with. Thus Nancy Cartwright argues 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 which have

shown themselves to deliver the right answer to the kind of problem at hand (Cartwright, 1983).

Similar responses can be made by instrumentalists to the arguments from explanation and prediction. Instrumentalists can either retort that there is no reason why the status of theories as calculating instruments should preclude them from giving rise to predictions and explanations; or they can query whether scientific theories really do add to our ability to predict and explain to start with. Not all these lines of response are equally convincing. But between them they give instrumentalism plenty of room to counter the initial realist challenge.

1.6 *Theory, observation and incommensurability*

A different line of argument against instrumentalism focuses on the distinction between what is observable and what is not. This distinction is crucial to instrumentalism, in that instrumentalists argue that claims about observable phenomena are unproblematic, but claims about unobservables are not. However, a number of writers have queried this distinction, arguing that observation reports are not essentially different from claims about unobservables, since they too depend on theoretical assumptions about the underlying structure of reality. Norwood Hanson (1958) has argued, for example, that scientists before and after Copernicus *saw* different things when they looked at the Sun: whereas pre-Copernicans regarded the Earth as stationary and so saw the Sun revolving round it, post-Copernican scientists saw the Sun as stationary and the Earth as rotating. Similarly, Hanson (1963) argues that the photographic plate which looks like a squiggly mess to a lay observer is *seen* as displaying a well-defined electron–positron pair by an experienced particle physicist. Examples like these undermine the distinction between what is observable and what is not, since they show that even judgements made in immediate response to sensory stimulation are influenced by fallible theories about reality.

Nor is the point restricted to recherché observations of astronomical bodies or subatomic particles. Even immediate perceptual judgements about the colour, shape and size of medium-sized physical objects can be shown to depend on theoretical assumptions implicit in our visual systems. Perhaps the best-known illustration is the Müller–Lyer illusion (see figure 9.1 opposite), which shows how our visual system uses complex assumptions about the normal causes of certain kinds of retinal patterns to draw conclusions about the geometry of physical figures. And analogous illusions can be used to demonstrate the presence of other theoretical presuppositions in our visual and other sensory systems.

As I said, in the first instance the unclarity of the observable–unobservable distinction counts against instrumentalism rather than realism. After all, it is instrumentalism, not realism, which *needs* the distinction, since instrumentalism says that we should be sceptical about unobservable claims, but not observable ones, whereas realism is happy to regard both kinds of claims as belief-worthy, so does not mind if they cannot be sharply distinguished.

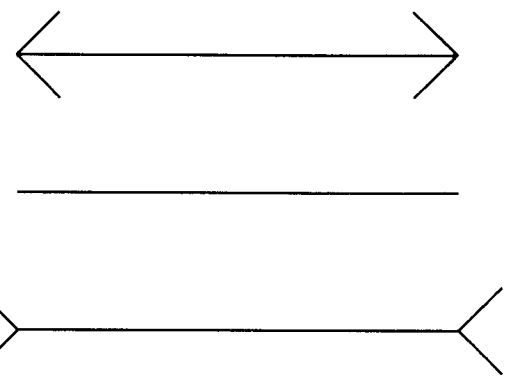

Figure 9.1 The Müller–Lyer illusion. Although all three lines are the same length, the top line, with inward-pointing arrowheads, appears to be shorter, and the bottom line, with outward-pointing arrowheads, appears to be longer, than the 'neutral' middle line.

However, there is another way of responding to doubts about the theory–observation distinction. For note that the arguments against the observable–unobservable distinction do not in fact vindicate the realist belief-worthiness of claims about *unobservables*; rather, they attack realism from the bottom up, and undermine the belief-worthiness of claims about *observables*, by showing that even observational claims depend on fallible theoretical assumptions. Obviously, if there is no observable–unobservable distinction, then all scientific claims are in the same boat. But on reflection it seems that the boat they all end up in is the instrumentalist boat of sceptical disbelief, not the realist one of general faith in science.

A number of influential recent philosophers of science, most prominently T. S. Kuhn (1962) and Paul Feyerabend (1976), have embraced this conclusion wholeheartedly, and maintained that no judgements made within science, not even observational judgements, can claim the authority of established truth. Rather, they argue, once scientists have embraced a theory about the essential nature of their subject matter, such as geocentrism, or Newtonian dynamics, or the wave theory of light, they will interpret all observational judgements in the light of that theory, and so will never be forced to recognize the kind of negative observational evidence that might show them that their theory is mistaken. Kuhn and Feyerabend independently lit on the term *incommensurable* to express the view that there is no common yardstick, in the form of theory-independent observation judgements, which can be used to decide objectively on the worth of scientific theories. Instead, they argue, decisions on scientific theories are never due to objective observational evidence, but are always relative to the presuppositions, interests and social milieu of the scientists involved.

Kuhn's and Feyerabend's blanket relativism has provoked much discussion among philosophers of science, but won few whole-hearted converts. Much of the discussion has focused on the status of observations. Most philosophers of science are prepared to accept that all observational judgements in some sense presuppose some element of theory. But many balk at the conclusion that observations therefore never have any independent authority to decide scientific questions. After all, they point out, most simple observations, such as that a pointer is adjacent to a mark on a dial, presuppose at most a minimal amount of theory, about rigid bodies, say, and about basic local geometry. Since such minimal theories are themselves rarely at issue in serious scientific debates, this minimal amount of theory-dependence provides no reason why observations of pointer readings should not be used to settle scientific disputes. If a scientific theory about the behaviour of gases, say, predicts that a pointer will be at a certain place on a dial, and it is observed not to be, then this decides against the theory about gases. It is not to the point to respond that, in taking the pointer reading at face value, we are making assumptions about rigid bodies and local geometry. For nothing in the debate about gases provides any reason to doubt these assumptions. And this of course is why scientists take such pains to work out what their theories imply about things like pointer readings – since observations of pointer readings do not depend on anything *contentious*, they will weigh with all sides in the scientific debate.

So, despite the arguments of Kuhn and Feyerabend, nearly all philosophers are realists about pointer readings and similar observable phenomena. But this still leaves us with the original disagreement between realism and instrumentalism about less directly observable entities. For, even if claims about pointer readings are uncontroversially belief-worthy, instrumentalists can still argue that theories about viruses, atoms and gravitational waves are nothing more than useful fictions for making calculations.

The realist response, as I said, is that the observable facts provide good indirect evidence for these theoretical entities, even if we cannot observe them directly. However, there are two strong lines of argument that instrumentalists can use to cast doubt on this suggestion. In the next two sections I shall discuss 'the underdetermination of theory by data' and 'the pessimistic meta-induction from past falsity'.

1.7 The underdetermination of theory by evidence

The argument from underdetermination asserts that, given any theory about unobservables that fits the observable facts, there will be other incompatible theories that 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 that fits any set of observable facts? One popular argument for this conclusion stems from the 'Duhem–Quine thesis'. According to this thesis, any particular scientific theory can always be defended in the face of contrary observations by adjusting auxiliary hypotheses. For example, when the Newtonian theory of gravitation was threatened by observations of anomalous movements by the planet Mercury, it could always be defended by postulating a hitherto unobserved planet, say, or an inhomogeneous mass distribution in the Sun. This general strategy for defending theories against contrary evidence 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.

Another argument for underdetermination starts, not with competing theories, but with some given theory. Suppose that all the predictions of some particular theory are accurate. We can construct a 'de-Ockhamized' version of this theory (reversing William of Ockham's 'razor' which prescribes that 'entities are not to be multiplied beyond necessity'), by postulating some unnecessarily complicated unobservable mechanism which nevertheless yields a new theory with precisely the same observational consequences as the original one.

Both of these lines of reasoning can be used to argue that more than one theory about unobservables will always fit any given set of observational data. Does this make realism about unobservables untenable? Many philosophers conclude that it does. But this is too quick. 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 that we need take their flat-earthism seriously. Similarly, even though Newtonian gravitational theory can in principle be made consistent with all the contrary evidence, this is no reason not to believe general relativity. Nor is our ability to 'cook up' a de-Ockhamized version of general relativity a reason to stop believing the standard version unencumbered with unnecessary entities.

Nevertheless, as I said, many contemporary philosophers of science do move directly from the premise that different theories are consistent with the observational evidence to the conclusion that none of them can be regarded as the truth. This is because many of them address this issue from an essentially Popperian perspective. For if you follow Popper in rejecting induction, then you will not believe that evidence ever provides positive support for any theory, except in the back-handed sense that the evidence can fail to falsify it. Accordingly you will think that all theories that have not been falsified are on a par, and in particular that any two theories that are both consistent with the evidence are equally well-supported by it.

So the arguments for underdetermination do present a problem to Popperians, since Popperians have no obvious basis for discriminating among different theories consistent with the data. But, as I pointed out above, these need not worry those of us who diverge from Popper in thinking unfalsified

theories can be better or worse supported by evidence, for we can simply respond to the underdetermination arguments by observing that some underdetermined theories are better supported by the evidence than others.

Now that we have returned to Popper, it is worth noting that the Duhem–Quine argument also raises a more specific problem for Popperians. Recall that Popper's overall philosophy raised the 'problem of demarcation', the problem of how to distinguish science from other kinds of conjecture. Popper's answer was that science, unlike astrology, or Marxism and psychoanalytic theory, is falsifiable. But the Duhem–Quine argument shows that even such eminently scientific theories as Newtonian physics are not falsifiable in any straightforward sense, since they can always save themselves in the face of failed predictions by adjusting auxiliary hypotheses.

Not only does this cast doubt on Popper's dismissal of Marxism and psychoanalysis as unscientific, but it seems to undermine his whole solution to the demarcation problem. If such paradigmatic scientific theories as Newtonian physics are not falsifiable, then it can scarcely be falsifiability that distinguishes science from non-science. Still, this is Popper's problem, not ours (see Harding, 1975). If we do not reject induction, then we do not have a problem of demarcation. For we can simply say that what distinguishes successful scientific theories from non-science is that the observational evidence gives us inductive reason to regard scientific theories as true.

The arguments in the latter part of this section have presupposed that a certain kind of inductive argument is legitimate. The kind of inductive argument relevant to underdetermination is not simple 'enumerative' induction, from observed As being Bs to 'All As are Bs', but rather inferences from any collection of observational data to the most plausible theory about unobservables that is consistent with that data. But these are species of the same genus; indeed, enumerative inductions can themselves be interpreted as treating 'All As are Bs' as the most plausible extrapolation consistent with the observed As being Bs. My attitude to this more general category of inductive inferences remains the same as my attitude to enumerative induction, which I outlined earlier. We do not yet have an explanation of *why* inductive inferences are legitimate, and to that extent we still face a problem of induction. But it is silly to try to solve that problem by denying that inductive inferences are ever legitimate. And that is why the underdetermination of theory by data does not constitute a good argument for instrumentalism. For to assume that we are never entitled to believe a theory, if there are others consistent with the same data, is simply to assume the illegitimacy of induction.

The Underdetermination of Theory by Observational Data (UTD)

There are two arguments for the UTD. The first is based on the 'Duhem–Quine thesis', originally formulated by the French philosopher and historian Pierre Duhem (1861–1916) and later revived by the American logician W. V. O. Quine (b. 1908). Duhem (1951) and Quine (1951) point out that a scientific theory T does not normally imply predictions P on its own, but only in conjunction with auxiliary hypotheses H.

$$T \,\&\, H \Rightarrow P$$

So when P is falsified by observation, this does not refute T, but only the conjunction of T & H.

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

So T can be retained, and indeed still explain 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 UTD follows quickly. 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 argument, first put forward by physicists like Henri Poincaré (1854–1912) and Ernst Mach (1838–1916) at the turn of the century, has a different starting point. Imagine that T_1 is the complete truth about physical reality, and that it implies observational facts O. Then we can always construct some 'de-Ockhamized' T_2 which postulates more complicated unobservable mechanisms but makes just the same observational predictions O. (Glymour 1980, ch. 5.)

For example, suppose we start with standard assumptions about the location of bodies in space–time and about the forces acting on them. A de-Ockhamized theory might then postulate that all bodies, including all measuring instruments, are accelerating by 1ft/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 UTD, 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's and T_2's auxiliary hypotheses may need to be revised; the de-Ockhamization 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-Ockhamization argument assumes that all future observations are as T_1 predicts.

1.8 The pessimistic meta-induction

I turn now to the other argument against realism. This argument takes as its premise the fact 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. (This is called a 'meta-induction' because its subject matter is not the natural world, but scientific theories about the natural world.)

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 does it not 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? (See Laudan, 1981.)

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 (chickenpox is due to a herpes virus), or the nature of everyday physical phenomena (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. Paleoanthropologists want to know how many hominid species were present on earth 3 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 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 did not 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 sub-class of theories which are not supported by adequate evidence.

1.9 *Naturalized epistemology of science*

In the last decade a number of philosophers of science have turned to a *naturalized* approach to scientific knowledge (Kitcher 1992). In place of traditional attempts to establish criteria for scientific theory-choice by *a priori* philosophical investigation, the naturalized approach regards science itself as a subject for *a posteriori* empirical investigation. Accordingly, naturalized epistemologists look to the history, sociology, and psychology of science, rather than to first principles, to identify criteria for the acceptability of scientific theories.

One apparent difficulty facing this kind of naturalized epistemology of science is that it is unclear how empirical investigation can ever yield anything more than *descriptive* information about how scientists actually operate. Yet any epistemology of science worth its name ought also to have a *normative* content – it ought to *prescribe* how scientists should reason, as well as describe how they do reason. David HUME (chapter 21) first pointed out that there is a logical gap between 'is' and 'ought'. A naturalized epistemology based on the empirical study of science seems fated to remain on the wrong side of this gap.

However, there is room for naturalized epistemologists to reply to this charge. They can agree that the empirical study of science cannot by itself yield prescriptions about how science ought to be done. But empirical study can still be *relevant* to such prescriptions. Suppose it is agreed that *technological fertility*, in the sense of generating technological advances, is a virtue in a scientific theory. Then the history, sociology and psychology of science might be able to show us that certain kinds of research strategies are effective at developing technologically fertile theories. More generally, given any agreed theoretical end Y, empirical study can show that research strategy X is an effective *means* to that end. The empirical study of science can thus yield the *hypothetical prescription* that, if you want Y, then you *ought* to adopt means X. It is this kind of hypothetical prescription that naturalized philosophers of science seek to establish: they look to the history, sociology and psychology of science to show us that scientists who *choose* theories on grounds X will in general *achieve* theories with characteristic Y.

Can the naturalized study of science tell us which research strategies are an effective means to theoretical *truth*? Different naturalized philosophers of science give different answers to this question. Many are suspicious of the idea of theoretical truth, and instead prefer to stick to the study of how to achieve more practical ends like technological fertility, simplicity, and predictive

accuracy. However, there seems no good reason for this restriction. There is nothing obviously incoherent in the idea of looking to the empirical study of science to tell us which research strategies have proved a good way of developing true theories. Indeed, the discussion of the 'pessimistic meta-induction' in the previous section amounted to the sketch of just such an investigation, in that it appealed to the history of science to decide whether or not the standard procedures of scientific theory-choice succeed in identifying true theories. It is not difficult to imagine more detailed and specific studies of this kind of issue.

Let me now return briefly to the issue with which I began, namely, the problem of induction. It is possible that the naturalized study of how to get at the scientific truth will enable us to make headway with this problem. For an empirical investigation into science might be able to show us that a certain kind of *inductive* inference is in general a reliable guide to scientific truth. And this would then provide a kind of vindication of that inductive method (see Papineau, 1993, ch. 5).

It is true that this kind of defence of induction will inevitably involve an element of circularity. For when we infer that certain kinds of induction are *in general* a reliable guide to truth, on the basis of evidence from the history of science, this will itself be an inductive inference. It is a matter of some delicacy, however, whether this circularity is vicious.

Defenders of this naturalized defence of induction will point out that, from their point of view, a legitimate criterion of theory-choice need not be an *a priori* guide to truth, but only an empirically certifiable one. Given this, the original argument against induction, that it is not logically valid, will not worry naturalized philosophers of science. Induction may not provide any *a priori* guarantee for its conclusions; but from the naturalized point of view, this does not show that induction is in any way illegitimate, since it leaves it open that induction may be an empirically reliable guide to the truth. And if there is nothing to show that induction is illegitimate, naturalized philosophers of science can then argue, why should we not use it to investigate the worth of inductive inferences? Maybe this is less satisfying a defence of induction than we might originally have hoped for. But perhaps it is defence enough.

2 The Metaphysics of Science

2.1 Causation

Many issues in the metaphysics of science hinge on the notion of *causation*. This notion is as important in science as it is in everyday thinking, and much scientific theorizing is concerned specifically to identify the *causes* of various phenomena. However, there is little philosophical agreement on what it means to say that one event is the cause of another.

Modern discussion of causation starts with David Hume, who argued that causation is simply a matter of CONSTANT CONJUNCTION (p. 583). According to Hume (1978), one event causes another if and only if events of the type to which the first event belongs regularly occur in conjunction with events of the

type to which the second event belongs. This formulation, however, leaves a number of questions open. Firstly, there is the problem of distinguishing genuine *causal laws* from *accidental regularities*. Not all regularities are sufficiently lawlike to underpin causal relationships. Being a screw in my desk could well be constantly conjoined with being made of copper, without its being true that these screws are made of copper because they are in my desk. Secondly, the idea of constant conjunction does not give a *direction* to causation. Causes need to be distinguished from effects. But knowing that A-type events are constantly conjoined with B-type events does not tell us which of A and B is the cause and which the effect, since constant conjunction is itself a symmetric relation. Thirdly, there is a problem about *probabilistic causation*. When we say that causes and effects are constantly conjoined, do we mean that the effects are always found with the causes, or is it enough that the causes make the effects probable?

Many philosophers of science this century have preferred to talk about *explanation* rather than causation. According to the covering-law model of explanation, something is explained if it can be deduced from premises which include one or more laws. As applied to the explanation of particular events, this implies that one particular event can be explained if it is linked by a law to some other particular event. However, while they are often treated as separate theories, the covering-law account of explanation is at bottom little more than a variant of Hume's constant conjunction account of causation. This affinity shows up in the fact that the covering-law account faces essentially the same difficulties as Hume: (1) in appealing to deductions from 'laws', it needs to explain the difference between genuine laws and accidentally true regularities; (2) it omits the requisite directionality, in that it does not tell us why we should not 'explain' causes by effects, as well as effects by causes; after all, it is as easy to deduce the height of a flagpole from the length of its shadow and the laws of optics, as to deduce the length of the shadow from the height of the pole and the same laws; (3) are the laws invoked in explanation required to be exceptionless and deterministic, or is it acceptable, say, to appeal to the merely probabilistic fact that smoking makes cancer more likely, in explaining why some particular person developed cancer?

In what follows I shall discuss these three problems in order (treating them as problems that arise equally both for the analysis of causation and the analysis of explanation). After that I shall consider some further issues in the metaphysics of science.

The Covering-Law Model of Explanation

According to this model (originally proposed by Hempel and Oppenheim, 1948, and further elaborated in Hempel, 1965) one statement (the *explanandum*) is explained by other statements (the *explanans*) if and only if the explanans contains one or more *laws*, and the explanandum can be *deduced* from the explanans. In the simplest case, where the explanandum is some particular statement to the effect that some individual *a* has property E, we might therefore have:

a has C
For all *x*, if *x* has C, then *x* has E
---
a has E

For example, we might deduce that a piece of litmus paper turned red, from the law that all litmus paper placed in acid turns red, together with prior condition that this piece of litmus paper was in fact placed in acid. The model can accommodate more complicated explanations of particular events, and can also allow explanations of laws themselves, as when we deduce Kepler's law that all planets move in ellipses, say, from Newton's law of universal gravitation and his laws of motion.

As applied to the explanation of particular events, the covering-law model implies a symmetry between *explanation* and *prediction*. For the information that, according to the model, suffices for the explanation of some known event should also enable us to predict that event if we did not yet know of it. Many critics have fastened on this implication of the model, however, and pointed out that we can often predict when we do not have enough information to explain (as when we predict the height of the flagpole from its shadow) and can often explain when we could not have predicted (as when we explain X's cancer on the basis of X's smoking).

These examples suggest that genuine explanations of particular events need to cite genuine causes, and that the reason the covering-law model runs into counter-examples is that it adds nothing to the inadequate constant conjunction analysis of causation, except that it substitutes the term 'law' for 'constant conjunction'. To get a satisfactory account of explanation we need, firstly, to recognize that explanations of particular events must mention causes, and, secondly, to improve on the constant conjunction analysis of causation.

There is a variant of the covering-law model which allows non-deterministic explanation as well as deterministic ones. This is termed the 'inductive–statistical (I–S)' model, by contrast with the original 'deductive–nomological (D–N)' model. An example would be:

a drinks 10 units of alcohol per diem
For *p* per cent of *x*s, if *x* drinks 10 units of alcohol per diem, *x* has a damaged liver
---
a has a damaged liver

Here the explanandum cannot be *deduced* from the explanans, but only follows with an *inductive* probability of *p*; and the inference appeals to a *statistical* regularity, rather than an exceptionless *nomological* generalization. In Hempel's original version of this model, it was required that the probability of the explanandum be *high*. A better requirement, however, as explained in the section on probabilistic causation below, is that the particular facts in the explanans need only make the probability of the explanandum *higher* than it would otherwise have been.

2.2 *Laws and accidents*

There are two general strategies for distinguishing laws from accidentally true generalizations. The first stands by Hume's idea that causal connections are mere constant conjunctions, and then seeks to explain why some constant conjunctions are better than others. That is, this first strategy accepts the principle that causation involves nothing more than certain events always

happening together with certain others, and then seeks to explain why some such patterns – the 'laws' – matter more than others – the 'accidents'. The second strategy, by contrast, rejects the Humean presupposition that causation involves nothing more than happenstantial co-occurrence, and instead postulates a relationship of 'necessitation', a kind of 'cement', which links events that are connected by law, but not those events (like being a screw in my desk and being made of copper) that are only accidentally conjoined.

There are a number of versions of the first Humean strategy. The most successful, originally proposed by F. P. Ramsey (1903–30), and later revived by David Lewis (1973), holds that laws are those true generalizations that can be fitted into an ideal system of knowledge. The thought here is that the laws are those patterns that are somehow explicable in terms of basic science, either as fundamental principles themselves, or as consequences of those principles, while accidents, although true, have no such explanation. Thus, 'All water at standard pressure boils at 100° C' is a consequence of the laws governing molecular bonding; but the fact that 'All the screws in my desk are copper' is not part of the deductive structure of any satisfactory science. Ramsey neatly encapsulated this idea by saying that laws are 'consequences of those propositions which we should take as axioms if we knew everything and organized it as simply as possible in a deductive system' (1978, p. 130).

Advocates of the alternative non-Humean strategy object that the difference between laws and accidents is not a *linguistic* matter of deductive systematization, but rather a *metaphysical* contrast between the kind of links they report. They argue that there is a link in nature between *being at 100° C* and *boiling*, but not between *being in my desk* and *being made of copper*, and that this is nothing to do with how the description of this link may fit into theories. According to D. M. Armstrong (1983), the most prominent defender of this view, the real difference between laws and accidents is simply that laws report relationships of natural *necessitation*, while accidents only report that two types of events *happen* to occur together.

Armstrong's view may seem intuitively plausible, but it is arguable that the notion of necessitation simply re-states the problem, rather than solving it. Armstrong says 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, Armstrong and other defenders of this view say very little about what this extra strength amounts to, except that it distinguishes laws from accidents. Armstrong's critics argue that a satisfactory account of laws ought to cast more light than this on the nature of laws.

2.3 *The direction of causation*

Hume said that the *earlier* of two causally related events is always the cause, and the *later* the effect. However, there are a number of objections to using the earlier–later 'arrow of time' to analyse the directional 'arrow of causation'. For

a start, it seems in principle possible that some causes and effects could be simultaneous. More seriously, the idea that time is directed from 'earlier' to 'later' itself stands in need of philosophical explanation – and one of the most popular explanations is that the idea of 'movement' from earlier to later depends on the fact that cause–effect pairs always have a given orientation in time. However, if we adopt such a 'causal theory of the arrow of time', and explain 'earlier' as the direction in which causes lie, and 'later' as the direction of effects, then we will clearly need to find some account of the direction of causation which does not itself assume the direction of time.

A number of such accounts have been proposed. David Lewis (1979) has argued that the asymmetry of causation derives from an 'asymmetry of overdetermination'. The overdetermination of present events by past events – consider a person who dies after simultaneously being shot and struck by lightning – is a very rare occurrence. By contrast, the multiple 'overdetermination' of present events by future events is absolutely normal. This is because the future, unlike the past, will always contain multiple traces of any present event. To use Lewis' example, when the President presses the red button in the White House, the future effects do not only include the dispatch of nuclear missiles, but also his fingerprint on the button, his trembling, the further depletion of his gin bottle, the recording of the button's click on tape, the emission of light waves bearing the image of his action through the window, the warming of the wire from the passage of the signal current, and so on, and on, and on.

Lewis relates this asymmetry of overdetermination to the asymmetry of causation as follows. If we suppose the cause of a given effect to have been absent, then this implies the effect would have been absent too, since (apart from freaks like the lightning–shooting case) there will not be any other causes left to 'fix' the effect. By contrast, if we suppose a given effect of some cause to have been absent, this does not imply the cause would have been absent, for there are still all the other traces left to 'fix' the cause. Lewis argues that these counterfactual considerations suffice to show why causes are different from effects.

Other philosophers appeal to a probabilistic variant of Lewis' asymmetry. Following Reichenbach (1956), they note that the different causes of any given type of effect are normally probabilistically independent of each other; by contrast, the different effects of any given type of cause are normally probabilistically correlated. For example, both obesity and high excitement can cause heart attacks, but this does not imply that fat people are more likely to get excited than thin ones; on the other hand, the fact that both lung cancer and nicotine-stained fingers can result from smoking *does* imply that lung cancer is more likely among people with nicotine-stained fingers. So this account distinguishes effects from causes by the fact that the former, but not the latter, are probabilistically dependent on each other.

2.4 *Probabilistic causation*

The just-mentioned probabilistic account of the *direction* of causation is normally formulated as part of a more general *theory of probabilistic causation*. Until relatively recently philosophers assumed that the world fundamentally conforms to deterministic laws, and that probabilistic dependencies between types of events, such as that between smoking and lung cancer, merely reflected our ignorance of the full causes. The rise of quantum mechanics, however, has persuaded most philosophers that determinism is false, and that some events, like the decay of a radium atom, happen purely as a matter of chance. A particular radium atom may decay, but on another occasion an identical atom in identical circumstances might well not decay.

Accordingly, a number of philosophers of science have put forward models of causation which require only that causes probabilify, rather than determine, their effects. The earliest such model was the 'inductive–statistical' version of the covering-law model of explanation (Hempel, 1965). Unlike deterministic 'deductive–nomological' explanations, such inductive–statistical explanations required only that prior conditions and laws imply a *high* probability for the event to be explained, not that this event will certainly happen. However, even this seems too strong a requirement for probabilistic causation. After all, smoking unequivocally causes lung cancer, but even heavy smokers do not have a *high* probability of lung cancer, in the sense of a probability close to one. Rather, their smoking increases their probability of lung cancer, not to a high figure, but merely from a low to a less low figure, but still well below 50 per cent. So more recent models of probabilistic causation simply require that causes should *increase* the probability of their effects, not that they should give them a high probability (Salmon, 1971).

This kind of model needs to guard against the possibility that the probabilistic association between putative cause and putative effect may be *spurious*, like the probabilistic association between barometers falling and subsequent rain. Such associations are not due to a causal connection between barometer movements and rain, but rather to both of these being joint effects of a *common cause*, namely, in our example, falls in atmospheric pressure. The obvious response to this difficulty is to say that we have a cause–effect relationship between A and B if and only if A increases the probability of B, and this association is *not* due to some common cause C. However, this is obviously incomplete as an analysis of causation, since it uses the notion of (common) *cause* in explaining causation.

It would solve this problem if we could analyse the notion of common cause in probabilistic terms. It seems to be a mark of common causes that they probabilistically '*screen off*' the associations between their joint effects, in the sense that, if we consider cases where the common cause is present and where it is absent separately, then the probabilistic association between the joint effects will disappear. For example, if it is given that the atmospheric pressure *has* fallen, then a falling barometer does not make it any more likely that it will rain; and similarly, if the atmospheric pressure *has not* fallen, a faulty falling reading on a barometer is no probable indicator of impending rain. (Numerically, if C is a common cause, and A and B its joint effects, we will find that A and B are associated – Prob (B/A) > Prob(B) – but that C and its absence render A irrelevant to B – Prob(B/A & C) = Prob(B/C) and Prob(B/A & not-C) = Prob(B/not-C).) It remains a matter of some debate, however, whether this

characteristic probabilistic structure of common causes is enough to allow a complete explanation of causation in probabilistic terms, or whether further non-probabilistic considerations need to be introduced.

2.5 *Probability*

Philosophical interest in probabilistic causation has led to a resurgence of interest in the philosophy of probability itself. Probability raises philosophical puzzles in its own right, quite apart from its connection with causation. What exactly is the 'probability' of a given event? The only part of the answer that is uncontroversial is that probabilities are quantities that satisfy the axioms of the probability calculus I specified earlier when discussing Bayesianism. But this leaves plenty of room for alternative philosophical views, for there are a number of different ways of interpreting these axioms.

One interpretation is the *subjective* theory of probability, which equates probabilities with subjective degrees of belief. This is the interpretation assumed by Bayesian confirmation theory. Most philosophers are happy to agree that subjective degrees of belief exist, and that the Dutch Book Argument (see the above box on Bayesianism) shows why they ought to conform to the axioms of probability. But many, if not all, philosophers argue that we need a theory of *objective* probability in addition to this subjective account.

One possible objective interpretation is the *frequency* theory, originally put forward by Richard von Mises (1957). According to this theory, the probability of a given kind of result is the number of times this result occurs, divided by the total number of occasions on which it might have occurred. So, for example, the probability of heads on a coin toss is the proportion of heads in some wider class of coin tosses.

This theory, however, faces a number of difficulties. For a start, it has problems in dealing with 'single-case probabilities'. Consider a particular coin toss. We can consider it as a member of the class of all coin tosses, or of all tosses of coins with that particular shape, or of all tosses made in just that way, or so on. However, these different 'reference classes' may well display different frequencies of heads. Yet intuitively it seems that there ought to be a unique value for the probability of heads on a particular toss of a particular coin. Perhaps this difficulty can be dealt with by specifying that the single-case probability should equal the relative frequency in the reference class of all tosses that are *similar in all relevant respects* to the particular toss in question. But there remain difficulties about which respects should count as 'relevant' in this sense.

In addition, there is the problem that many of these more specific reference classes will only be finite in extent. Coins with a certain distinctive shape may only be tossed in some given way ten times in the whole history of the universe. Yet the probability of heads on these tosses is unlikely to be equal to the relative frequency in the ten tosses, for luck may well yield a disproportionately high, or low, number of heads in ten tosses. Because of this, frequency theorists standardly appeal, not to actual reference classes, but to hypothetical infinite sequences, and equate the probability with the *limit* that the relative frequency *would* tend to *if* the relevant kind of trial were repeated an infinite number of times. Critics of the frequency theory object that this reliance on hypothetical infinite reference sequences makes probabilities inadmissibly abstract.

Because of these difficulties, many contemporary philosophers of probability have adopted the '*propensity*' theory of probability in place of the frequency theory. The earliest version of this theory, proposed by Popper (1959b), simply modified the frequency theory by specifying that only those relative frequencies generated by repeated trails on a given '*experimental set-up*' should count as genuine probabilities. This arguably deals with the problem of single-case probabilities, but it still leaves us with hypothetical reference classes. To avoid this, later versions of the propensity theory do not define probabilities in terms of frequencies at all, but simply take probabilities to be primitive propensities of particular situations to produce given results.

This kind of propensity theory does not seek to define objective probabilities in terms of frequencies, but in effect simply takes single-case probabilities as primitive (see Mellor 1971). But it can still recognize a connection between probabilities and frequencies. For, as long as propensities are assumed to obey the axioms of the probability calculus (though this assumption itself merits some debate), it will follow that, in a sufficiently long sequence of independent trials in each of which the propensity to produce B is p, the overall *propensity* for the observed frequency of B to differ by more than a given amount from p can be made arbitrarily small, in accord with the Law of Large Numbers. (Note how the italicized second use of 'propensity' in this claim prevents it serving as a definition of propensities in terms of frequencies.)

Both the frequency theory and the propensity theory have their strengths. The frequency theory has the virtue of offering an explicit definition of probability, where the propensity theory takes probabilities as primitive. On the other hand, the propensity theory has no need to assume hypothetical reference sequences, whereas these are essential to the frequency theory.

It may seem that the frequency theory, because it offers an explicit definition, is better able than the propensity theory to explain how we find out about probabilities. But this is an illusion. The trouble is that the frequency theory's explicit definition is in terms of frequencies in INFINITE SEQUENCES. But our evidence is always in the form of frequencies in *finite* samples. So the problem of explaining how we can move from frequencies in finite samples to knowledge of probabilities is as much a problem for the frequency theory as for the propensity theory. (There are various suggestions about how to solve this 'problem of statistical inference', none of them universally agreed. My present point is merely that this problem of statistical inference arises in just the same way for both frequency and propensity theorists.)

In the face of continued debate about the interpretation of objective probability, some philosophers have turned to physics, and in particular to the notion of probability used in quantum mechanics, to resolve the issue. Unfortunately quantum mechanics is no less philosophically controversial

than the notion of probability. There are different philosophical interpretations of the formal theory of quantum mechanics, each of which involves different understandings of probability. Because of this it seems likely that philosophical disputes about probability will continue until there is an agreed interpretation of quantum mechanics.

The Interpretation of Quantum Mechanics

Modern quantum mechanics says that the state of any given system of microscopic particles is fully characterized by its 'wave function'. However, instead of specifying the exact positions and velocities of the particles, as is done in classical mechanics, this 'wave function' only specifies *probabilities* of the particles displaying certain values of position, and velocity, *if* appropriate measurements are made. *Schrödinger's equation* then specifies how this wave function evolves smoothly and deterministically over time, analogously to the way that Newton's laws of motion specify how the positions and velocities of macroscopic objects evolve over time – except that Schrödinger's equation again only decribes changes in probabilities, not exact values.

On the orthodox interpretation of quantum mechanics, quantum probabilities change into actualities only when '*measurements*' are made. If you measure the position of a particle, say, *then* its position assumes a definite value, even though nothing before the measurement determined exactly what this value would be.

There is something puzzling about this, however, since any overall system of measured particle and measuring instrument is itself just another system of microscopic particles, which might therefore be expected to evolve smoothly according to Schrödinger's equation, rather than to jump suddenly to some definite value for position. To account for this, the orthodox interpretation says that in addition to the normal Schrödinger evolution, there is a special kind of change which occurs in 'measurements', when the wave function suddenly '*collapses*' to yield a definite value for the measured quantity.

The '*measurement problem*' is the problem of explaining exactly when, and why, these collapses occur. The story of 'Schrödinger's cat' makes the difficulty graphic. Suppose that some unfortunate cat is sitting next to a poison dispenser which is wired up to emit cyanide gas if an electron emitted from some source turns up on the right half of some position-registering plate, but not if the electron turns up on the left half. The basic quantum mechanical description of this situation says that it is both possible that the electron will turn up on the right half of the plate and that it will turn up on the left half, and therefore both possible that the poison is emitted and that it is not, and therefore both possible that the cat is alive and that it is dead. One of these possibilities only becomes actual when the wave function of the whole system collapses. But when does that happen? When the electron is emitted? When it reaches the plate? When the cat dies or not? Or only when a human being looks at the cat to see how it is faring?

There seems no principled way to decide between these answers. Because of this, many philosophers reject the orthodox view that physical systems are completely characterized by their wave functions, and conjecture that, in addition to the variables quantum mechanics recognizes, there are various 'hidden variables' which always specify exact positions and velocities for all physical particles. It is difficult, however, for such *hidden variable theories* to reproduce the surprising phenomena predicted by quantum mechanics, without postulating mysterious mechanisms that seem inconsistent with other parts of physics.

A more radical response to the measurement problem is to deny that the wave function ever does collapse, and somehow to make sense of the idea that reality contains both a live cat and a dead cat. This 'many-worlds' interpretation of quantum mechanics flies in the face of common sense, but its theoretical attractions are leading an increasing number of philosophers to take it seriously.

2.6 Teleology

We normally explain some particular fact by citing its *cause*: for example, we explain why some water freezes by noting that its temperature fell below 0° C. There are cases, however, where we seem to explain items by citing their *effects* instead. In particular, this kind of explanation is common in biology. We often explain some biological trait by showing how it is useful to the organism in question: for instance, the explanation of the polar bear's white fur is that it camouflages it; the explanation of human sweating is that it lowers body temperature, and so on. Similar explanations are also sometimes offered in anthropology and sociology.

Until fairly recently most philosophers of science took such functional or teleological explanations at face value, as an alternative to causal explanation, in which items are explained, not by their causal antecedents, but by showing how they contribute to the well-being of some larger system. Carl Hempel's covering-law model of explanation embodied an influential version of this attitude. According to Hempel, causal explanations and functional explanations are simply two different ways 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 *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 all explanations of particular facts are really causal, and that functional explanations, despite appearances, are really a *subspecies* of causal explanations. On this view, the reference to future facts 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 that led to the natural selection of the biological trait in question. Thus the functional explanation of the polar bear's 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*. Similarly, any acceptable functional explanations in anthropology or sociology should be understood as referring us *back* in time to the conscious intentions or unconscious selection processes which caused the facts to be explained (see Wright, 1973, Neander, 1991). (There remains the terminological matter of whether functional explanations understood in this way ought still to be called 'teleological'. Traditional usage reserves the term 'teleology' for distinctively non-causal explanations in terms of future results.

But most contemporary philosophers are happy to describe disguised causal explanations that make implicit reference to selection mechanisms as 'teleological'.)

2.7 The logic of natural selection

The Darwinian theory of evolution by natural selection is not just important to functional explanation, but to biological thinking in general (see Sober 1993). This raises a number of further philosophical issues. An initial question is whether Darwin's theory has any predictive content. Darwin's theory explains the evolution of species in terms of the differential survival of the *'fittest'* animals. But if 'fittest' simply means 'those animals that in fact survive', then the theory will collapse into a definitional truth, and will be unable to predict how species will evolve in the future. However, we do not have to understand 'fitness' in this entirely empty way. At a minimum, we can read 'fit' as signifying those kinds of characteristics that have proved helpful to survival in the *past*; it is then no longer a matter of definition, but predictively informative, to say that animals with these characteristics will outcompete others in the future.

A related charge is that Darwinian theory is guilty of 'adaptationism'. According to this criticism, Darwinian biologists mistakenly presuppose that all biological traits must serve some purpose, and in consequence invent 'just so stories', like Rudyard Kipling's children's tales, to provide fictional histories in which each biological trait has provided some benefit to its owners. In reality, however, the anti-Darwinians continue, many biological features serve no purposes at all, but are simply side-effects of other traits, or results of random genetic drift. So, for example, it would be a mistake to suppose that any useful purpose is served by the small size of insects; rather the smallness of insects is simply an automatic upshot of their lack of internal skeletons.

In response to this criticism, most Darwinians admit that *some* biological traits serve no function, but point out that it certainly does not follow that *all* traits serve no function. What is more, they argue, there is plenty of genuine evidence to show that specific biological traits have been historically favoured because of specific functions, and that this process has been important to the evolution of species.

Another question raised by Darwin's theory is what *kinds of entities* are selected in the competition to survive? Is it biological *species*, or *groups* of individuals within species, or single *individuals*, or the *genes* inside individuals? In recent years there has been a heated controversy about which of these entities are the central *'units of selection'*. Biologists influenced by mathematical population genetics have argued that we should view genes as the units of selection, since evolution always involves changes in the frequencies of alternative genes within a larger population. Other biologists, however, have argued that this is just a matter of 'bookkeeping', and that the causal processes responsible for evolution always involve the survival of individuals and even groups. Some progress in unravelling this knotty issue has been made by distinguishing 'replicators', namely, the genes that carry the

instructions for building bodies from generation to generation, from 'vehicles', such as individuals and groups, whose survival is normally, but not always, the prerequisite for gene survival.

Work on the logic of natural selection has been associated with the development of *sociobiology*, which seeks to identify the evolutionary purpose of the characteristic *social behaviour* of animals, including humans. An obvious and often-made objection to this programme is that much social behaviour, especially in higher mammals, is due to individual learning and culture, and so not a product of evolutionarily determined genes. More extreme defenders of sociobiology, especially in its popularized versions, respond that most human behaviour really does depend on nature rather than nurture, despite widespread liberal opinion to the contrary, including even such apparently cultural behaviour as style of dress, or choice of marriage partner. A better defence of sociobiology, however, is to admit that environment is often as important for behaviour as genes, but to point out that it is still important to understand the evolutionary advantages provided by the genes that do influence behaviour.

2.8 Theoretical reduction

Another philosophical question about biology is whether it can be *reduced* to such lower level (in the sense of ontologically more basic) sciences as chemistry and physics. Obviously, this is an issue that arises not just for biology, but also for such other 'special' natural sciences as geology and meteorology, and also for such human sciences as psychology and sociology.

One science is said to 'reduce' to another if its categories can be defined in terms of the categories of the latter, and its laws explained by the laws of the latter. *Reductionists* argue that all sciences form a hierarchy in which the higher can always be reduced to the lower. Thus, for example, biology might be reduced to physiology, physiology to chemistry, and eventually chemistry to physics.

Reductionism can be viewed either historically or metaphysically. The historical question is whether science characteristically progresses by later theories reducing earlier ones. The metaphysical question is whether the different areas of science describe different realities, or just the one physical reality described at different levels of detail. Though often run together, these are different questions.

Taken as a general thesis, historical reductionism is false. Recall the earlier discussion of the 'pessimistic meta-induction from past falsity'. This involved the claim that new theories characteristically show their predecessors to be false. To the extent that this claim is true, historical reductionism is false: for a new theory can scarcely explain why an earlier theory was true, if it shows it is false.

In the earlier discussion I argued that there are some areas of science, like molecular biology and medical science, to which the pessimistic meta-induction does not apply. If this is right, then we can expect that in these areas

new theories will indeed normally reduce old ones. But I did not dispute that there are other areas of science, like cosmology and fundamental particle physics, in which the normal fate of old theories is to be thrown out. It follows that we must reject historical reductionism, understood as the thesis that *all* science proceeds by new theories reducing old ones.

This does not mean, however, that metaphysical reductionism is false. Even if science proceeds towards the overall truth by fits and starts, there may be general reasons for expecting that this overall truth, when eventually reached, will reduce to physical truth.

One possible such argument stems from the *causal interaction* between the phenomena discussed in the special sciences and physical phenomena. Biological, geological and meteorological events all unquestionably have physical effects. It is difficult to see how they could do this unless they are made of physical components.

It is doubtful, however, whether this suffices to establish full-scale reductionism, as opposed to the weaker thesis (sometimes called 'token-identity') according to which each *particular* higher-level event is identical with some *particular* physical event. Thus, for example, it might be true that one animal's aggressive behaviour can be equated with a given sequence of physical movements, and another animal's aggressive behaviour can be equated with another sequence of physical movements, without there being any uniform way of defining 'aggressive behaviour', for all animals, in terms of physical movements. The case-by-case token-identity will explain how each instance of aggressive behaviour can have physical effects, like causing intruding animals to move away. But without any uniform definition of 'aggressive behaviour' in terms of physical movements there is no question of reducing ethology (the science of behaviour) to physics, and so no question of explaining ethological laws by physical laws. Instead, the laws of ethology and other special sciences will be *sui generis*, identifying patterns whose instances vary in their physical make-up, and which therefore cannot possibly be explained in terms of physical laws alone (see Fodor 1974).

Acknowledgements

I would like to thank Stathis Psillos for helping me with this chapter.

Further Reading

For an introduction to the problem of induction, and Popper's solution, see Popper, *The Logic of Scientific Discovery* (1959a) especially ch. 1. The problem and Popper's solution are further discussed in O'Hear, *An Introduction to the Philosophy of Science* (1989). There are two excellent introductions to

Bayesian philosophy of science: Horwich, *Probability and Evidence* (1982) and Howson and Urbach, *Scientific Reasoning* (1989).

The best modern defence of instrumentalism is Van Fraassen, *The Scientific Image* (1980). Churchland and Hooker (eds), *Images of Science* (1985) offers a good collection of essays on the realism–instrumentalism debate. Kitcher, *The Advancement of Science* (1993) contains a strong defence of realism.

The classic works on the theory-dependence of observation and the incommensurability of theories are Hanson, *Patterns of Discovery* (1958), Kuhn, *The Structure of Scientific Revolutions* (1962) and Feyerabend, *Against Method* (1976). A good collection of essays on these issues is Hacking (ed.), *Scientific Revolutions* (1981). The best sources for the underdetemination of theories by evidence and the pessimistic meta-induction are respectively Quine, 'Two Dogmas of Empiricism' (1951) and Laudan, 'A Confutation of Convergent Realism' (1981). For a survey of recent work on naturalized epistemology, see Kitcher's 'The Naturalists Return' (1992).

Most modern discussions of explanation begin with the title essay in Hempel, *Aspects of Scientific Explanation* (1965). Explanation and its relation to causation are further explored by the essays in Ruben (ed.), *Explanation* (1993). Armstrong, *What is a Law of Nature?* (1983) provides an excellent account of the general problem of distinguishing laws from accidents, as well as his own solution. Chapter 7 of O'Hear, *An Introduction to the Philosophy of Science* (1989) contains a good introduction to both probability and probabilistic causation. The best contemporary non-specialist discussion of the problems of quantum mechanics is to be found in chapters 11–13 of Lockwood, *Mind, Brain and the Quantum* (1989).

Sober, *Philosophy of Biology* (1993) explains the ways in which Darwinian thinking is important to the philosophy of biology. The view that events discussed in the special sciences are token-identical, but not reducible, to physical events is defended in Fodor's 'Special Sciences' (1974).

References

Armstrong, D. 1983: *What is a Law of Nature?* Cambridge: Cambridge University Press.

Ayer, A. J. 1956: *The Problem of Knowledge.* London: Macmillan.

Boyd, R. 1980: Scientific Realism and Naturalistic Epistemology. In P. Asquith and R. Giere (eds), *PSA 1980* vol. 2, East Lansing, Mi: Philosophy of Science Association, 613–62.

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

Churchland, P. and Hooker, C. (eds), 1985: *Images of Science.* Chicago: University of Chicago Press.

Duhem, P. 1951 [1906]: *The Aim and Structure of Physical Theory* (translated by P. Wiener). Princeton, NJ: Princeton University Press.

Feyerabend, P. 1976: *Against Method.* London: New Left Books.

Fodor, J. 1974: Special Sciences. *Synthèse*, 28, 97–115.

Friedman, M. 1984: *Foundations of Spacetime Theories*. Princeton, NJ: Princeton University Press.
Glymour, C. 1980: *Theory and Evidence*. Princeton, NJ: Princeton University Press.
Hacking, I. 1967: Slightly More Realistic Personal Probability. *Philosophy of Science*, 34, 311–25.
—— (ed.) 1981: *Scientific Revolutions*. Oxford: Oxford University Press.
Hanson, N. R. 1958: *Patterns of Discovery*. Cambridge: Cambridge University Press.
—— 1963: *The Concept of the Positron*. Cambridge: Cambridge University Press.
Harding, S. (ed.), 1975: *Can Theories be Refuted?* Dordrecht: Reidel.
Hempel, C. 1965: *Aspects of Scientific Explanation*. New York: Free Press.
Hempel, C. and Oppenheim, P. 1948: Studies in the Logic of Explanation. *Philosophy of Science*, 15, 135–75.
Horwich, P. 1982: *Probability and Evidence*. Cambridge: Cambridge University Press.
Howson, C. and Urbach, P. 1989: *Scientific Reasoning*. La Salle: Open Court.
Hume, D. 1978 [1739]: *A Treatise of Human Nature* (edited by P. H. Nidditch). Oxford: Clarendon Press.
Kitcher, P. 1992: The Naturalists Return. *Philosophical Review*, 101, 53–114.
—— 1993: *The Advancement of Science*. New York: Oxford University Press.
Kuhn, T. S. 1962: *The Structure of Scientific Revolutions*. Chicago: Chicago University Press.
Laudan, L. 1981: A Confutation of Convergent Realism. *Philosophy of Science*, 48.
Lewis, D. 1973: *Counterfactuals*. Oxford: Blackwell.
—— 1979: Counterfactual Dependence and Time's Arrow. *Noûs*, 13, 455–76.
Lockwood, M. 1989: *Mind, Brain and the Quantum*. Oxford: Blackwell.
Mellor, D. 1971: *The Matter of Chance*. Cambridge: Cambridge University Press.
Mises, R. von 1957: *Probability, Statistics and Truth*. London: Allen and Unwin.
Neander, K. 1991: The Teleological Notion of Function. *Australasian Journal of Philosophy*, 69, 454–68.
O'Hear, A. 1989: *An Introduction to the Philosophy of Science*. Oxford: Clarendon Press.
Papineau, D. 1993: *Philosophical Naturalism*. Oxford: Blackwell.
Popper, K. 1959a: *The Logic of Scientific Discovery*. London: Hutchison.
—— 1959b: The Propensity Interpretation of Probability. *British Journal for the Philosophy of Science*, 10, 25–42.
—— 1963: *Conjectures and Refutations*. London: Routledge and Kegan Paul.
—— 1972: *Objective Knowledge*. Oxford: Clarendon Press.
Quine, W. V. O. 1951: Two Dogmas of Empiricism. In his *From a Logical Point of View*, New York: Harper, 20–46.
Reichenbach, H. 1956: *The Direction of Time*. Berkeley, CA: University of California Press.
Ramsey, F. 1978 [1929]: General Propositions and Causality. In his *Foundations*, D. H. Mellor (ed.), London: Routledge and Kegan Paul.
Ruben, D.-H. (ed.) 1993: *Explanation*. Oxford: Oxford University Press.
Salmon, W. 1971: *Statistical Explanation and Statistical Relevance*. Pittsburgh, PA: University of Pittsburgh Press.
Smart, J. 1963: *Philosophy and Scientific Realism*. London: Routledge and Kegan Paul.
Sober, E. 1993: *Philosophy of Biology*. Oxford: Oxford University Press.
Teller, P. 1973: Conditionalization and Observation. *Synthèse*, 28, 218–58.
Van Fraassen, B. 1980: *The Scientific Image*. Oxford: Clarendon Press.
Worrall, J. 1989: Why Both Popper and Watkins Fail to Solve the Problem of Induction. In F. D'Agostino and I. Jarvie (eds), *Freedom and Rationality*, Dordrecht: Kluwer, 257–96.
Wright, L. 1973: Functions. *Philosophical Review*, 82, 139–68.

Discussion Questions

1 Is it any less rational to accept induction than it is to accept deduction?

2 Is it always a mistake to save a theory when it has been falsified?

3 How can we show that it is more rational to believe some hypotheses than to believe others?

4 How do scientific theories move beyond the stage of conjecture?

5 Is belief a matter of degree?

6 Does Bayes' theorem help us to deal rationally with evidence?

7 Do scientific claims about unobservable entities differ in status from scientific claims about observable entities?

8 Is instrumentalism mistaken if it cannot account for some features of scientific practice? How can we determine what features are really a part of scientific practice?

9 Are general theories or piecemeal procedures more important to our basic characterization of science?

10 Does all observation depend on theoretical assumptions? What are the implications of your answer for an account of unobservables?

11 'Given any theory about unobservables which fits the observed facts, there will always be other incompatible theories which fit the same facts.' Does this make realism about unobservables untenable?

12 Can we be realists in some areas of science and instrumentalists in others?

13 Should we look to the history, sociology, and psychology of science, rather than to first principles, to identify criteria for the acceptability of scientific theories?

14 Can a naturalized study of science tell us which research strategies are an effective means to theoretical *truth*?

15 Is there any good reason for a philosopher to prefer to talk about explanation rather than causation?

16 How can we distinguish laws from accidentally true generalizations?

17 Must we posit an ideal system of knowledge in order to understand the notion of 'law'? What if there cannot be such a system?

18 Do scientific laws involve necessity? In what sense?

19 How can we explain the direction of causation?

20 Can we accept a model of causation according to which causes probabilify, rather than determine, their effects?

21 What problem poses greater difficulties for frequency theories of probability: 'single-case probabilities' or a reliance on hypothetical infinite reference sequences?

22 If 'propensities' are primitive, can a propensity theory give us any insight into the nature of probability?

23 How are interpretations of quantum mechanics relevant to philosophical disputes about probability?

24 Can we give up common sense in favour of a 'many-worlds' reality containing both Schrödinger's cat alive and Schrödinger's cat dead?

25 Are teleological explanations an alternative to causal explanations or a kind of causal explanation?

26 Does 'survival of the fittest' help to explain which species survive?

27 How do we determine which biological traits serve a function and which do not? What explains those traits which do not serve a function?

28 How can we settle whether species, groups, individuals or genes are the appropriate units of selection in a theory of natural selection?

29 How can we determine whether all sciences form a hierarchy in which the higher can always be reduced to the lower?

30 Do different areas of science describe different realities, or just one physical reality at different levels of detail?

10
Philosophy of Mathematics

Mary Tiles

Since the time of ancient Greece, mathematics has been intimately tied to philosophy, both as a model of knowledge and as an object of philosophical reflection. Are numbers real? What is a proof? Is mathematics more certain than other knowledge? Can finite minds have knowledge of infinity? How can mathematics apply to the world? In this chapter, changing philosophical conceptions of mathematics, changes in the historical context of mathematical thought and changes in mathematics itself are explored in relation to a basic question: How can theoretical reasoning about non-concrete mathematical objects be both secure and so useful? Platonic, Aristotelian and Kantian approaches to mathematics are examined in their own right and to place in context the problems and proposed solutions of the major modern schools of logicism, formalism and intuitionism. Recent developments which do not seek foundations for mathematics are also considered. Many of the discussions of great historical figures in this volume, especially Frege and Russell (see chapter 27), are relevant to the present chapter. Readers will also wish to consult chapters on EPISTEMOLOGY (chapter 1), METAPHYSICS (chapter 2), PHILOSOPHY OF LANGUAGE (chapter 3), PHILOSOPHY OF LOGIC (chapter 4), and PHILOSOPHY OF SCIENCE (chapter 9).

Introductions to the philosophy of mathematics often begin where Körner's influential introduction (Körner, 1960) began, outlining three positions: logicism, formalism and intuitionism. These were the three contending schools to emerge from nineteenth-century mathematical moves to provide rigorous foundations for mathematical analysis (including infinitesimal calculus). The problem for the philosophy of mathematics has been that (1) these seemed to represent all the reasonable positions available and (2) in the light of Gödel's incompleteness theorems and other results proved by Turing, Church, Skolem and Tarski in the 1930s, neither logicism nor formalism seemed a philosophically viable position. Intuitionism, whilst having its philosophical credentials intact, was unacceptable to most mathematicians because it involved discarding parts of classically accepted mathematics. This apparent impasse partly explains the decline of interest in the philosophy of mathematics since the first part of this century, when for a while, with the work of Russell and Whitehead, and the Vienna Circle logical positivists, it seemed to occupy centre stage.

Hao Wang, in his perceptive retrospective analysis of the philosophy of mathematics of this period (Wang, 1988), explains this trajectory in terms of

Blackwell Companions to Philosophy

This benchmark student reference series offers a comprehensive survey of philosophy as a whole. Written by many of today's leading figures, each volume provides lucid and engaging coverage of the key figures, terms, and movements of the main subdisciplines of philosophy. Each essay is fully cross-referenced and supported by a selected bibliography. Taken together, it provides the ideal basis for course use and an invaluable work of reference.

Already published:

1 A Companion to Ethics
Edited by Peter Singer

2 A Companion to Aesthetics
Edited by David Cooper

3 A Companion to Epistemology
Edited by Jonathan Dancy and Ernest Sosa

4 A Companion to Contemporary Political Philosophy
Edited by Robert E. Goodin and Philip Pettit

5 A Companion to the Philosophy of Mind
Edited by Samuel Guttenplan

6 A Companion to Metaphysics
Edited by Jaegwon Kim and Ernest Sosa

Forthcoming:

7 A Companion to the Philosophy of Law and Legal Theory
Edited by Dennis Patterson

8 A Companion to the Philosophy of Religion
Edited by Philip Quinn and Charles Taliaferro

9 A Companion to the Philosophy of Language
Edited by Crispin Wright and Bob Hale

The Blackwell Companion to Philosophy

edited by Nicholas Bunnin and E. P. Tsui-James

David Papineau,
"Philosophy of Science"

BLACKWELL Reference

There could be no fairer destiny for any . . . theory than that it should point the way to a more comprehensive theory in which it lives on, as a limiting case.

ALBERT EINSTEIN

1

SCIENCE: CONJECTURES AND REFUTATIONS

Mr. Turnbull had predicted evil consequences, . . . and was now doing the best in his power to bring about the verification of his own prophecies.

ANTHONY TROLLOPE

I

WHEN I received the list of participants in this course and realized that I had been asked to speak to philosophical colleagues I thought, after some hesitation and consultation, that you would probably prefer me to speak about those problems which interest me most, and about those developments with which I am most intimately acquainted. I therefore decided to do what I have never done before: to give you a report on my own work in the philosophy of science, since the autumn of 1919 when I first began to grapple with the problem, '*When should a theory be ranked as scientific?*' or '*Is there a criterion for the scientific character or status of a theory?*'

The problem which troubled me at the time was neither, 'When is a theory true?' nor, 'When is a theory acceptable?' My problem was different. I *wished to distinguish between science and pseudo-science;* knowing very well that science often errs, and that pseudo-science may happen to stumble on the truth.

I knew, of course, the most widely accepted answer to my problem: that science is distinguished from pseudo-science—or from 'metaphysics'—by its *empirical method*, which is essentially *inductive*, proceeding from observation or experiment. But this did not satisfy me. On the contrary, I often formulated my problem as one of distinguishing between a genuinely empirical method and a non-empirical or even a pseudo-empirical method—that is to say, a method which, although it appeals to observation and experiment, nevertheless

A lecture given at Peterhouse, Cambridge, in Summer 1953, as part of a course on developments and trends in contemporary British philosophy, organized by the British Council; originally published under the title 'Philosophy of Science: a Personal Report' in British Philosophy in Mid-Century, *ed. C. A. Mace*, 1957.

does not come up to scientific standards. The latter method may be exemplified by astrology, with its stupendous mass of empirical evidence based on observation—on horoscopes and on biographies.

But as it was not the example of astrology which led me to my problem I should perhaps briefly describe the atmosphere in which my problem arose and the examples by which it was stimulated. After the collapse of the Austrian Empire there had been a revolution in Austria: the air was full of revolutionary slogans and ideas, and new and often wild theories. Among the theories which interested me Einstein's theory of relativity was no doubt by far the most important. Three others were Marx's theory of history, Freud's psycho-analysis, and Alfred Adler's so-called 'individual psychology'.

There was a lot of popular nonsense talked about these theories, and especially about relativity (as still happens even today), but I was fortunate in those who introduced me to the study of this theory. We all—the small circle of students to which I belonged—were thrilled with the result of Eddington's eclipse observations which in 1919 brought the first important confirmation of Einstein's theory of gravitation. It was a great experience for us, and one which had a lasting influence on my intellectual development.

The three other theories I have mentioned were also widely discussed among students at that time. I myself happened to come into personal contact with Alfred Adler, and even to co-operate with him in his social work among the children and young people in the working-class districts of Vienna where he had established social guidance clinics.

It was during the summer of 1919 that I began to feel more and more dissatisfied with these three theories—the Marxist theory of history, psycho-analysis, and individual psychology; and I began to feel dubious about their claims to scientific status. My problem perhaps first took the simple form, 'What is wrong with Marxism, psycho-analysis, and individual psychology? Why are they so different from physical theories, from Newton's theory, and especially from the theory of relativity?'

To make this contrast clear I should explain that few of us at the time would have said that we believed in the *truth* of Einstein's theory of gravitation. This shows that it was not my doubting the *truth* of those other three theories which bothered me, but something else. Yet neither was it that I merely felt mathematical physics to be more *exact* than the sociological or psychological type of theory. Thus what worried me was neither the problem of truth, at that stage at least, nor the problem of exactness or measurability. It was rather that I felt that these other three theories, though posing as sciences, had in fact more in common with primitive myths than with science; that they resembled astrology rather than astronomy.

I found that those of my friends who were admirers of Marx, Freud, and Adler, were impressed by a number of points common to these theories, and especially by their apparent *explanatory power*. These theories appeared to be able to explain practically everything that happened within the fields to which they referred. The study of any of them seemed to have the effect of an

intellectual conversion or revelation, opening your eyes to a new truth hidden from those not yet initiated. Once your eyes were thus opened you saw confirming instances everywhere: the world was full of *verifications* of the theory. Whatever happened always confirmed it. Thus its truth appeared manifest; and unbelievers were clearly people who did not want to see the manifest truth; who refused to see it, either because it was against their class interest, or because of their repressions which were still 'un-analysed' and crying aloud for treatment.

The most characteristic element in this situation seemed to me the incessant stream of confirmations, of observations which 'verified' the theories in question; and this point was constantly emphasized by their adherents. A Marxist could not open a newspaper without finding on every page confirming evidence for his interpretation of history; not only in the news, but also in its presentation—which revealed the class bias of the paper—and especially of course in what the paper did *not* say. The Freudian analysts emphasized that their theories were constantly verified by their 'clinical observations'. As for Adler, I was much impressed by a personal experience. Once, in 1919, I reported to him a case which to me did not seem particularly Adlerian, but which he found no difficulty in analysing in terms of his theory of inferiority feelings, although he had not even seen the child. Slightly shocked, I asked him how he could be so sure. 'Because of my thousandfold experience,' he replied; whereupon I could not help saying: 'And with this new case, I suppose, your experience has become thousand-and-one-fold.'

What I had in mind was that his previous observations may not have been much sounder than this new one; that each in its turn had been interpreted in the light of 'previous experience', and at the same time counted as additional confirmation. What, I asked myself, did it confirm? No more than that a case could be interpreted in the light of the theory. But this meant very little, I reflected, since every conceivable case could be interpreted in the light of Adler's theory, or equally of Freud's. I may illustrate this by two very different examples of human behaviour: that of a man who pushes a child into the water with the intention of drowning it; and that of a man who sacrifices his life in an attempt to save the child. Each of these two cases can be explained with equal ease in Freudian and in Adlerian terms. According to Freud the first man suffered from repression (say, of some component of his Oedipus complex), while the second man had achieved sublimation. According to Adler the first man suffered from feelings of inferiority (producing perhaps the need to prove to himself that he dared to commit some crime), and so did the second man (whose need was to prove to himself that he dared to rescue the child). I could not think of any human behaviour which could not be interpreted in terms of either theory. It was precisely this fact—that they always fitted, that they were always confirmed—which in the eyes of their admirers constituted the strongest argument in favour of these theories. It began to dawn on me that this apparent strength was in fact their weakness.

With Einstein's theory the situation was strikingly different. Take one

typical instance—Einstein's prediction, just then confirmed by the findings of Eddington's expedition. Einstein's gravitational theory had led to the result that light must be attracted by heavy bodies (such as the sun), precisely as material bodies were attracted. As a consequence it could be calculated that light from a distant fixed star whose apparent position was close to the sun would reach the earth from such a direction that the star would seem to be slightly shifted away from the sun; or, in other words, that stars close to the sun would look as if they had moved a little away from the sun, and from one another. This is a thing which cannot normally be observed since such stars are rendered invisible in daytime by the sun's overwhelming brightness; but during an eclipse it is possible to take photographs of them. If the same constellation is photographed at night one can measure the distances on the two photographs, and check the predicted effect.

Now the impressive thing about this case is the *risk* involved in a prediction of this kind. If observation shows that the predicted effect is definitely absent, then the theory is simply refuted. The theory is *incompatible with certain possible results of observation*—in fact with results which everybody before Einstein would have expected.[1] This is quite different from the situation I have previously described, when it turned out that the theories in question were compatible with the most divergent human behaviour, so that it was practically impossible to describe any human behaviour that might not be claimed to be a verification of these theories.

These considerations led me in the winter of 1919–20 to conclusions which I may now reformulate as follows.

(1) It is easy to obtain confirmations, or verifications, for nearly every theory—if we look for confirmations.

(2) Confirmations should count only if they are the result of *risky predictions*; that is to say, if, unenlightened by the theory in question, we should have expected an event which was incompatible with the theory—an event which would have refuted the theory.

(3) Every 'good' scientific theory is a prohibition: it forbids certain things to happen. The more a theory forbids, the better it is.

(4) A theory which is not refutable by any conceivable event is non-scientific. Irrefutability is not a virtue of a theory (as people often think) but a vice.

(5) Every genuine *test* of a theory is an attempt to falsify it, or to refute it. Testability is falsifiability; but there are degrees of testability: some theories are more testable, more exposed to refutation, than others; they take, as it were, greater risks.

(6) Confirming evidence should not count *except when it is the result of a genuine test of the theory*; and this means that it can be presented as a serious but unsuccessful attempt to falsify the theory. (I now speak in such cases of 'corroborating evidence'.)

[1] This is a slight oversimplification, for about half of the Einstein effect may be derived from the classical theory, provided we assume a ballistic theory of light.

(7) Some genuinely testable theories, when found to be false, are still upheld by their admirers—for example by introducing *ad hoc* some auxiliary assumption, or by re-interpreting the theory *ad hoc* in such a way that it escapes refutation. Such a procedure is always possible, but it rescues the theory from refutation only at the price of destroying, or at least lowering, its scientific status. (I later described such a rescuing operation as a '*conventionalist twist*' or a '*conventionalist stratagem*'.)

One can sum up all this by saying that *the criterion of the scientific status of a theory is its falsifiability, or refutability, or testability.*

II

I may perhaps exemplify this with the help of the various theories so far mentioned. Einstein's theory of gravitation clearly satisfied the criterion of falsifiability. Even if our measuring instruments at the time did not allow us to pronounce on the results of the tests with complete assurance, there was clearly a possibility of refuting the theory.

Astrology did not pass the test. Astrologers were greatly impressed, and misled, by what they believed to be confirming evidence—so much so that they were quite unimpressed by any unfavourable evidence. Moreover, by making their interpretations and prophecies sufficiently vague they were able to explain away anything that might have been a refutation of the theory had the theory and the prophecies been more precise. In order to escape falsification they destroyed the testability of their theory. It is a typical soothsayer's trick to predict things so vaguely that the predictions can hardly fail: that they become irrefutable.

The Marxist theory of history, in spite of the serious efforts of some of its founders and followers, ultimately adopted this soothsaying practice. In some of its earlier formulations (for example in Marx's analysis of the character of the 'coming social revolution') their predictions were testable, and in fact falsified.[2] Yet instead of accepting the refutations the followers of Marx re-interpreted both the theory and the evidence in order to make them agree. In this way they rescued the theory from refutation; but they did so at the price of adopting a device which made it irrefutable. They thus gave a 'conventionalist twist' to the theory; and by this stratagem they destroyed its much advertised claim to scientific status.

The two psycho-analytic theories were in a different class. They were simply non-testable, irrefutable. There was no conceivable human behaviour which could contradict them. This does not mean that Freud and Adler were not seeing certain things correctly: I personally do not doubt that much of what they say is of considerable importance, and may well play its part one day in a psychological science which is testable. But it does mean that those 'clinical observations' which analysts naïvely believe confirm their theory cannot do this any more than the daily confirmations which astrologers find

[2] See, for example, my *Open Society and Its Enemies*, ch. 15, section iii, and notes 13–14.

in their practice.[3] And as for Freud's epic of the Ego, the Super-ego, and the Id, no substantially stronger claim to scientific status can be made for it than for Homer's collected stories from Olympus. These theories describe some facts, but in the manner of myths. They contain most interesting psychological suggestions, but not in a testable form.

At the same time I realized that such myths may be developed, and become testable; that historically speaking all—or very nearly all—scientific theories originate from myths, and that a myth may contain important anticipations of scientific theories. Examples are Empedocles' theory of evolution by trial and error, or Parmenides' myth of the unchanging block universe in which nothing ever happens and which, if we add another dimension, becomes Einstein's block universe (in which, too, nothing ever happens, since everything is, four-dimensionally speaking, determined and laid down from the beginning). I thus felt that if a theory is found to be non-scientific, or 'metaphysical' (as we might say), it is not thereby found to be unimportant, or insignificant, or 'meaningless', or 'nonsensical'.[4] But it cannot claim to be backed by empirical evidence in the scientific sense—although it may easily be, in some genetic sense, the 'result of observation'.

(There were a great many other theories of this pre-scientific or pseudo-

[3] 'Clinical observations', like all other observations, are *interpretations in the light of theories* (see below, sections iv ff.); and for this reason alone they are apt to seem to support those theories in the light of which they were interpreted. But real support can be obtained only from observations undertaken as tests (by 'attempted refutations'); and for this purpose *criteria of refutation* have to be laid down beforehand: it must be agreed which observable situations, if actually observed, mean that the theory is refuted. But what kind of clinical responses would refute to the satisfaction of the analyst not merely a particular analytic diagnosis but psycho-analysis itself? And have such criteria ever been discussed or agreed upon by analysts? Is there not, on the contrary, a whole family of analytic concepts, such as 'ambivalence' (I do not suggest that there is no such thing as ambivalence), which would make it difficult, if not impossible, to agree upon such criteria? Moreover, how much headway has been made in investigating the question of the extent to which the (conscious or unconscious) expectations and theories held by the analyst influence the 'clinical responses' of the patient? (To say nothing about the conscious attempts to influence the patient by proposing interpretations to him, etc.) Years ago I introduced the term '*Oedipus effect*' to describe the influence of a theory or expectation or prediction *upon the event which it predicts* or describes: it will be remembered that the causal chain leading to Oedipus' parricide was started by the oracle's prediction of this event. This is a characteristic and recurrent theme of such myths, but one which seems to have failed to attract the interest of the analysts, perhaps not accidentally. (The problem of confirmatory dreams suggested by the analyst is discussed by Freud, for example in *Gesammelte Schriften*, III, 1925, where he says on p. 314: 'If anybody asserts that most of the dreams which can be utilized in an analysis . . . owe their origin to [the analyst's] suggestion, then no objection can be made from the point of view of analytic theory. Yet there is nothing in this fact', he surprisingly adds, 'which would detract from the reliability of our results.')

[4] The case of astrology, nowadays a typical pseudo-science, may illustrate this point. It was attacked, by Aristotelians and other rationalists, down to Newton's day, for the wrong reason—for its now accepted assertion that the planets had an 'influence' upon terrestrial ('sublunar') events. In fact Newton's theory of gravity, and especially the lunar theory of the tides, was historically speaking an offspring of astrological lore. Newton, it seems, was most reluctant to adopt a theory which came from the same stable as for example the theory that 'influenza' epidemics are due to an astral 'influence'. And Galileo, no doubt for the same reason, actually rejected the lunar theory of the tides; and his misgivings about Kepler may easily be explained by his misgivings about astrology.

scientific character, some of them, unfortunately, as influential as the Marxist interpretation of history; for example, the racialist interpretation of history—another of those impressive and all-explanatory theories which act upon weak minds like revelations.)

Thus the problem which I tried to solve by proposing the criterion of falsifiability was neither a problem of meaningfulness or significance, nor a problem of truth or acceptability. It was the problem of drawing a line (as well as this can be done) between the statements, or systems of statements, of the empirical sciences, and all other statements—whether they are of a religious or of a metaphysical character, or simply pseudo-scientific. Years later—it must have been in 1928 or 1929—I called this first problem of mine the '*problem of demarcation*'. The criterion of falsifiability is a solution to this problem of demarcation, for it says that statements or systems of statements, in order to be ranked as scientific, must be capable of conflicting with possible, or conceivable, observations.

III

Today I know, of course, that this *criterion of demarcation*—the criterion of testability, or falsifiability, or refutability—is far from obvious; for even now its significance is seldom realized. At that time, in 1920, it seemed to me almost trivial, although it solved for me an intellectual problem which had worried me deeply, and one which also had obvious practical consequences (for example, political ones). But I did not yet realize its full implications, or its philosophical significance. When I explained it to a fellow student of the Mathematics Department (now a distinguished mathematician in Great Britain), he suggested that I should publish it. At the time I thought this absurd; for I was convinced that my problem, since it was so important for me, must have agitated many scientists and philosophers who would surely have reached my rather obvious solution. That this was not the case I learnt from Wittgenstein's work, and from its reception; and so I published my results thirteen years later in the form of a criticism of Wittgenstein's *criterion of meaningfulness*.

Wittgenstein, as you all know, tried to show in the *Tractatus* (see for example his propositions 6.53; 6.54; and 5) that all so-called philosophical or metaphysical propositions were actually non-propositions or pseudo-propositions: that they were senseless or meaningless. All genuine (or meaningful) propositions were truth functions of the elementary or atomic propositions which described 'atomic facts', i.e.—facts which can in principle be ascertained by observation. In other words, meaningful propositions were fully reducible to elementary or atomic propositions which were simple statements describing possible states of affairs, and which could in principle be established or rejected by observation. If we call a statement an 'observation statement' not only if it states an actual observation but also if it states anything that *may* be observed, we shall have to say (according to the *Tractatus*, 5 and 4.52) that every genuine proposition must be a truth-function of, and

therefore deducible from, observation statements. All other apparent propositions will be meaningless pseudo-propositions; in fact they will be nothing but nonsensical gibberish.

This idea was used by Wittgenstein for a characterization of science, as opposed to philosophy. We read (for example in 4.11, where natural science is taken to stand in opposition to philosophy): 'The totality of true propositions is the total natural science (or the totality of the natural sciences).' This means that the propositions which belong to science are those deducible from *true* observation statements; they are those propositions which can be *verified* by true observation statements. Could we know all true observation statements, we should also know all that may be asserted by natural science.

This amounts to a crude verifiability criterion of demarcation. To make it slightly less crude, it could be amended thus: 'The statements which may possibly fall within the province of science are those which may possibly be verified by observation statements; and these statements, again, coincide with the class of *all* genuine or meaningful statements.' For this approach, then, *verifiability, meaningfulness, and scientific character all coincide.*

I personally was never interested in the so-called problem of meaning; on the contrary, it appeared to me a verbal problem, a typical pseudo-problem. I was interested only in the problem of demarcation, i.e. in finding a criterion of the scientific character of theories. It was just this interest which made me see at once that Wittgenstein's verifiability criterion of meaning was intended to play the part of a criterion of demarcation as well; and which made me see that, as such, it was totally inadequate, even if all misgivings about the dubious concept of meaning were set aside. For Wittgenstein's criterion of demarcation—to use my own terminology in this context—is verifiability, or deducibility from observation statements. But this criterion is too narrow (*and* too wide): it excludes from science practically everything that is, in fact, characteristic of it (while failing in effect to exclude astrology). No scientific theory can ever be deduced from observation statements, or be described as a truth-function of observation statements.

All this I pointed out on various occasions to Wittgensteinians and members of the Vienna Circle. In 1931–2 I summarized my ideas in a largish book (read by several members of the Circle but never published; although part of it was incorporated in my *Logic of Scientific Discovery*); and in 1933 I published a letter to the Editor of *Erkenntnis* in which I tried to compress into two pages my ideas on the problems of demarcation and induction.[5] In this letter

[5] My *Logic of Scientific Discovery* (1959, 1960, 1961), here usually referred to as *L.Sc.D.*, is the translation of *Logik der Forschung* (1934), with a number of additional notes and appendices, including (on pp. 312–14) the letter to the Editor of *Erkenntnis* mentioned here in the text which was first published in *Erkenntnis*, 3, 1933, pp. 426 f.

Concerning my never published book mentioned here in the text, see R. Carnap's paper '*Ueber Protokollstäze*' (On Protocol-Sentences), *Erkenntnis*, 3, 1932, pp. 215–28 where he gives an outline of my theory on pp. 223–8, and accepts it. He calls my theory 'procedure B', and says (p. 224, top): 'Starting from a point of view different from Neurath's' (who developed what Carnap calls on p. 223 'procedure A'), 'Popper developed procedure B as

and elsewhere I described the problem of meaning as a pseudo-problem, in contrast to the problem of demarcation. But my contribution was classified by members of the Circle as a proposal to replace the verifiability criterion of *meaning* by a falsifiability criterion of *meaning*—which effectively made nonsense of my views.[6] My protests that I was trying to solve, not their pseudo-problem of meaning, but the problem of demarcation, were of no avail.

My attacks upon verification had some effect, however. They soon led to complete confusion in the camp of the verificationist philosophers of sense and nonsense. The original proposal of verifiability as the criterion of meaning was at least clear, simple, and forceful. The modifications and shifts which were now introduced were the very opposite.[7] This, I should say, is now seen even by the participants. But since I am usually quoted as one of them I wish to repeat that although I created this confusion I never participated in it. Neither falsifiability nor testability were proposed by me as criteria of meaning; and although I may plead guilty to having introduced both terms into the discussion, it was not I who introduced them into the theory of meaning.

Criticism of my alleged views was widespread and highly successful. I have yet to meet a criticism of my views.[8] Meanwhile, testability is being widely accepted as a criterion of demarcation.

---

part of his system.' And after describing in detail my theory of tests, Carnap sums up his views as follows (p. 228): 'After weighing the various arguments here discussed, it appears to me that the second language form with procedure B—that is in the form here described—is the most adequate among the forms of scientific language at present advocated . . . in the . . . theory of knowledge.' This paper of Carnap's contained the first published report of my theory of critical testing. (See also my critical remarks in *L.Sc.D.*, note 1 to section 29, p. 104, where the date '1933' should read '1932'; and ch. 11, below, text to note 39.)

[6] Wittgenstein's example of a nonsensical pseudo-proposition is: 'Socrates is identical'. Obviously, 'Socrates is not identical' must also be nonsense. Thus the negation of any nonsense will be nonsense, and that of a meaningful statement will be meaningful. ***But the negation of a testable (or falsifiable) statement need not be testable***, as was pointed out, first in my *L.Sc.D.*, (e.g. pp. 38 f.) and later by my critics. The confusion caused by taking testability as a criterion of *meaning* rather than of *demarcation* can easily be imagined.

[7] The most recent example of the way in which the history of this problem is misunderstood is A. R. White's 'Note on Meaning and Verification', *Mind*, 63, 1954, pp. 66 ff. J. L. Evans's article, *Mind*, 62, 1953, pp. 1 ff., which Mr. White criticizes, is excellent in my opinion, and unusually perceptive. Understandably enough, neither of the authors can quite reconstruct the story. (Some hints may be found in my *Open Society*, notes 46, 51 and 52 to ch. 11; and a fuller analysis in ch. 11 of the present volume.)

[8] In *L.Sc.D.* I discussed, and replied to, some likely objections which afterwards were indeed raised, without reference to my replies. One of them is the contention that the falsification of a natural law is just as impossible as its verification. The answer is that this objection mixes two entirely different levels of analysis (like the objection that mathematical demonstrations are impossible since checking, no matter how often repeated, can never make it quite certain that we have not overlooked a mistake). On the first level, there is a logical asymmetry: one singular statement—say about the perihelion of Mercury—can formally falsify Kepler's laws; but these cannot be formally verified by any number of singular statements. The attempt to minimize this asymmetry can only lead to confusion. On another level, we may hesitate to accept any statement, even the simplest observation statement; and we may point out that every statement involves *interpretation in the light of theories*, and that it is therefore uncertain. This does not affect the fundamental asymmetry, but it is important: most dissectors of the heart before Harvey observed the wrong things—those, which they expected to see. There can never be anything like a completely safe observation,

IV

I have discussed the problem of demarcation in some detail because I believe that its solution is the key to most of the fundamental problems of the philosophy of science. I am going to give you later a list of some of these other problems, but only one of them—the *problem of induction*—can be discussed here at any length.

I had become interested in the problem of induction in 1923. Although this problem is very closely connected with the problem of demarcation, I did not fully appreciate the connection for about five years.

I approached the problem of induction through Hume. Hume, I felt, was perfectly right in pointing out that induction cannot be logically justified. He held that there can be no valid logical[9] arguments allowing us to establish '*that those instances, of which we have had no experience, resemble those, of which we have had experience*'. Consequently '*even after the observation of the frequent or constant conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience*'. For 'shou'd it be said that we have experience'[10]—experience teaching us that objects constantly conjoined with certain other objects continue to be so conjoined—then, Hume says, 'I wou'd renew my question, *why from this experience we form any conclusion beyond those past instances, of which we have had experience*'. In other words, an attempt to justify the practice of induction by an appeal to experience must lead to an *infinite regress*. As a result we can say that theories can never be inferred from observation statements, or rationally justified by them.

I found Hume's refutation of inductive inference clear and conclusive. But I felt completely dissatisfied with his psychological explanation of induction in terms of custom or habit.

It has often been noticed that this explanation of Hume's is philosophically not very satisfactory. It is, however, without doubt intended as a *psychological* rather than a philosophical theory; for it tries to give a causal explanation of a psychological fact—*the fact that we believe in laws*, in statements asserting regularities or constantly conjoined kinds of events—by asserting that this fact is due to (i.e. constantly conjoined with) custom or habit. But even this reformulation of Hume's theory is still unsatisfactory; for what I have just called a 'psychological fact' may itself be described as a custom or habit—

free from the dangers of misinterpretation. (This is one of the reasons why the theory of induction does not work.) The 'empirical basis' consists largely of a mixture of *theories* of lower degree of universality (of 'reproducible effects'). But the fact remains that, relative to whatever basis the investigator may accept (at his peril), he can test his theory only by trying to refute it.

[9] Hume does not say 'logical' but 'demonstrative', a terminology which, I think, is a little misleading. The following two quotations are from the *Treatise of Human Nature*, Book I, Part III, sections vi and xii. (The italics are all Hume's.)

[10] This and the next quotation are from *loc. cit.*, section vi. See also Hume's *Enquiry Concerning Human Understanding*, section IV, Part II, and his *Abstract*, edited 1938 by J. M. Keynes and P. Sraffa, p. 15, and quoted in *L.Sc.D.*, new appendix *VII, text to note 6.

the custom or habit of believing in laws or regularities; and it is neither very surprising nor very enlightening to hear that such a custom or habit must be explained as due to, or conjoined with, a custom or habit (even though a different one). Only when we remember that the words 'custom' and 'habit' are used by Hume, as they are in ordinary language, not merely to *describe* regular behaviour, but rather to *theorize about its origin* (ascribed to frequent repetition), can we reformulate his psychological theory in a more satisfactory way. We can then say that, like other habits, *our habit of believing in laws is the product of frequent repetition*—of the repeated observation that things of a certain kind are constantly conjoined with things of another kind.

This genetico-psychological theory is, as indicated, incorporated in ordinary language, and it is therefore hardly as revolutionary as Hume thought. It is no doubt an extremely popular psychological theory—part of 'common sense', one might say. But in spite of my love of both common sense and Hume, I felt convinced that this psychological theory was mistaken; and that it was in fact refutable on purely logical grounds.

Hume's psychology, which is the popular psychology, was mistaken, I felt, about at least three different things: (*a*) the typical result of repetition; (*b*) the genesis of habits; and especially (*c*) the character of those experiences or modes of behaviour which may be described as 'believing in a law' or 'expecting a law-like succession of events'.

(*a*) The typical result of repetition—say, of repeating a difficult passage on the piano—is that movements which at first needed attention are in the end executed without attention. We might say that the process becomes radically abbreviated, and ceases to be conscious: it becomes 'physiological'. Such a process, far from creating a conscious expectation of law-like succession, or a belief in a law, may on the contrary begin with a conscious belief and destroy it by making it superfluous. In learning to ride a bicycle we may start with the belief that we can avoid falling if we steer in the direction in which we threaten to fall, and this belief may be useful for guiding our movements. After sufficient practice we may forget the rule; in any case, we do not need it any longer. On the other hand, even if it is true that repetition may create unconscious expectations, these become conscious only if something goes wrong (we may not have heard the clock tick, but we may hear that it has stopped).

(*b*) Habits or customs do not, as a rule, *originate* in repetition. Even the habit of walking, or of speaking, or of feeding at certain hours, *begins* before repetition can play any part whatever. We may say, if we like, that they deserve to be called 'habits' or 'customs' only after repetition has played its typical part; but we must not say that the practices in question originated as the result of many repetitions.

(*c*) Belief in a law is not quite the same thing as behaviour which betrays an expectation of a law-like succession of events; but these two are sufficiently closely connected to be treated together. They may, perhaps, in exceptional cases, result from a mere repetition of sense impressions (as in the case of the

stopping clock). I was prepared to concede this, but I contended that normally, and in most cases of any interest, they cannot be so explained. As Hume admits, even a single striking observation may be sufficient to create a belief or an expectation—a fact which he tries to explain as due to an inductive habit, formed as the result of a vast number of long repetitive sequences which had been experienced at an earlier period of life.[11] But this, I contended, was merely his attempt to explain away unfavourable facts which threatened his theory; an unsuccessful attempt, since these unfavourable facts could be observed in very young animals and babies—as early, indeed, as we like. 'A lighted cigarette was held near the noses of the young puppies', reports F. Bäge. 'They sniffed at it once, turned tail, and nothing would induce them to come back to the source of the smell and to sniff again. A few days later, they reacted to the mere sight of a cigarette or even of a rolled piece of white paper, by bounding away, and sneezing.' [12] If we try to explain cases like this by postulating a vast number of long repetitive sequences at a still earlier age we are not only romancing, but forgetting that in the clever puppies' short lives there must be room not only for repetition but also for a great deal of novelty, and consequently of non-repetition.

But it is not only that certain empirical facts do not support Hume; there are decisive arguments of a *purely logical* nature against his psychological theory.

The central idea of Hume's theory is that of *repetition, based upon similarity* (or 'resemblance'). This idea is used in a very uncritical way. We are led to think of the water-drop that hollows the stone: of sequences of unquestionably like events slowly forcing themselves upon us, as does the tick of the clock. But we ought to realize that in a psychological theory such as Hume's, only repetition-for-us, based upon similarity-for-us, can be allowed to have any effect upon us. We must respond to situations as if they were equivalent; *take* them as similar; *interpret* them as repetitions. The clever puppies, we may assume, showed by their response, their way of acting or of reacting, that they recognized or interpreted the second situation as a repetition of the first: that they expected its main element, the objectionable smell, to be present. The situation was a repetition-for-them because they responded to it by *anticipating* its similarity to the previous one.

This apparently psychological criticism has a purely logical basis which may be summed up in the following simple argument. (It happens to be the one from which I originally started my criticism.) The kind of repetition envisaged by Hume can never be perfect; the cases he has in mind cannot be cases of perfect sameness; they can only be cases of similarity. Thus *they are repetitions only from a certain point of view*. (What has the effect upon me of a repetition may not have this effect upon a spider.) But this means that, for logical reasons, there must always be a point of view—such as a system of

[11] *Treatise*, section xiii; section xv, rule 4.

[12] F. Bäge, 'Zur Entwicklung, etc.', *Zeitschrift f. Hundeforschung*, 1933; cp. D. Katz, *Animals and Men*, ch. VI, footnote.

expectations, anticipations, assumptions, or interests—*before* there can be any repetition; which point of view, consequently, cannot be merely the result of repetition. (See now also appendix *x, (1), to my *L.Sc.D.*)

We must thus replace, for the purposes of a psychological theory of the origin of our beliefs, the naïve idea of events which *are* similar by the idea of events to which we react by *interpreting* them as being similar. But if this is so (and I can see no escape from it) then Hume's psychological theory of induction leads to an infinite regress, precisely analogous to that other infinite regress which was discovered by Hume himself, and used by him to explode the logical theory of induction. For what do we wish to explain? In the example of the puppies we wish to explain behaviour which may be described as *recognizing or interpreting* a situation as a repetition of another. Clearly, we cannot hope to explain this by an appeal to earlier repetitions, once we realize that the earlier repetitions must also have been repetitions-for-them, so that precisely the same problem arises again: that of *recognizing or interpreting* a situation as a repetition of another.

To put it more concisely, similarity-for-us is the product of a response involving interpretations (which may be inadequate) and anticipations or expectations (which may never be fulfilled). It is therefore impossible to explain anticipations, or expectations, as resulting from many repetitions, as suggested by Hume. For even the first repetition-for-us must be based upon similarity-for-us, and therefore upon expectations—precisely the kind of thing we wished to explain.

This shows that there is an infinite regress involved in Hume's psychological theory.

Hume, I felt, had never accepted the full force of his own logical analysis. Having refuted the logical idea of induction he was faced with the following problem: how do we actually obtain our knowledge, as a matter of psychological fact, if induction is a procedure which is logically invalid and rationally unjustifiable? There are two possible answers: (1) We obtain our knowledge by a non-inductive procedure. This answer would have allowed Hume to retain a form of rationalism. (2) We obtain our knowledge by repetition and induction, and therefore by a logically invalid and rationally unjustifiable procedure, so that all apparent knowledge is merely a kind of belief—belief based on habit. This answer would imply that even scientific knowledge is irrational, so that rationalism is absurd, and must be given up. (I shall not discuss here the age-old attempts, now again fashionable, to get out of the difficulty by asserting that though induction is of course logically invalid if we mean by 'logic' the same as 'deductive logic', it is not irrational by its own standards, as may be seen from the fact that every reasonable man applies it *as a matter of fact*: it was Hume's great achievement to break this uncritical identification of the question of fact—*quid facti?*—and the question of justification or validity—*quid juris?*. (See below, point (13) of the appendix to the present chapter.)

It seems that Hume never seriously considered the first alternative. Having

cast out the logical theory of induction by repetition he struck a bargain with common sense, meekly allowing the re-entry of induction by repetition, in the guise of a psychological theory. I proposed to turn the tables upon this theory of Hume's. Instead of explaining our propensity to expect regularities as the result of repetition, I proposed to explain repetition-for-us as the result of our propensity to expect regularities and to search for them.

Thus I was led by purely logical considerations to replace the psychological theory of induction by the following view. Without waiting, passively, for repetitions to impress or impose regularities upon us, we actively try to impose regularities upon the world. We try to discover similarities in it, and to interpret it in terms of laws invented by us. Without waiting for premises we jump to conclusions. These may have to be discarded later, should observation show that they are wrong.

This was a theory of trial and error—of *conjectures and refutations*. It made it possible to understand why our attempts to force interpretations upon the world were logically prior to the observation of similarities. Since there were logical reasons behind this procedure, I thought that it would apply in the field of science also; that scientific theories were not the digest of observations, but that they were inventions—conjectures boldly put forward for trial, to be eliminated if they clashed with observations; with observations which were rarely accidental but as a rule undertaken with the definite intention of testing a theory by obtaining, if possible, a decisive refutation.

v

The belief that science proceeds from observation to theory is still so widely and so firmly held that my denial of it is often met with incredulity. I have even been suspected of being insincere—of denying what nobody in his senses can doubt.

But in fact the belief that we can start with pure observations alone, without anything in the nature of a theory, is absurd; as may be illustrated by the story of the man who dedicated his life to natural science, wrote down everything he could observe, and bequeathed his priceless collection of observations to the Royal Society to be used as inductive evidence. This story should show us that though beetles may profitably be collected, observations may not.

Twenty-five years ago I tried to bring home the same point to a group of physics students in Vienna by beginning a lecture with the following instructions: 'Take pencil and paper; carefully observe, and write down what you have observed!' They asked, of course, *what* I wanted them to observe. Clearly the instruction, 'Observe!' is absurd.[13] (It is not even idiomatic, unless the object of the transitive verb can be taken as understood.) Observation is always selective. It needs a chosen object, a definite task, an interest, a point of view, a problem. And its description presupposes a descriptive language, with property words; it presupposes similarity and classification, which in its turn presupposes interests, points of view, and problems. 'A hungry animal',

[13] See section 30 of *L.Sc.D.*

writes Katz,[14] 'divides the environment into edible and inedible things. An animal in flight sees roads to escape and hiding places. . . . Generally speaking, objects change . . . according to the needs of the animal.' We may add that objects can be classified, and can become similar or dissimilar, *only* in this way—by being related to needs and interests. This rule applies not only to animals but also to scientists. For the animal a point of view is provided by its needs, the task of the moment, and its expectations; for the scientist by his theoretical interests, the special problem under investigation, his conjectures and anticipations, and the theories which he accepts as a kind of background: his frame of reference, his 'horizon of expectations'.

The problem 'Which comes first, the hypothesis (*H*) or the observation (*O*),' is soluble; as is the problem, 'Which comes first, the hen (*H*) or the egg (*O*)'. The reply to the latter is, 'An earlier kind of egg'; to the former, 'An earlier kind of hypothesis'. It is quite true that any particular hypothesis we choose will have been preceded by observations—the observations, for example, which it is designed to explain. But these observations, in their turn, presupposed the adoption of a frame of reference: a frame of expectations: a frame of theories. If they were significant, if they created a need for explanation and thus gave rise to the invention of a hypothesis, it was because they could not be explained within the old theoretical framework, the old horizon of expectations. There is no danger here of an infinite regress. Going back to more and more primitive theories and myths we shall in the end find unconscious, *inborn* expectations.

The theory of inborn *ideas* is absurd, I think; but every organism has inborn *reactions* or *responses*; and among them, responses adapted to impending events. These responses we may describe as 'expectations' without implying that these 'expectations' are conscious. The new-born baby 'expects', in this sense, to be fed (and, one could even argue, to be protected and loved). In view of the close relation between expectation and knowledge we may even speak in quite a reasonable sense of 'inborn knowledge'. This 'knowledge' is not, however, *valid a priori*; an inborn expectation, no matter how strong and specific, may be mistaken. (The newborn child may be abandoned, and starve.)

Thus we are born with expectations; with 'knowledge' which, although not *valid a priori*, is *psychologically or genetically a priori*, i.e. prior to all observational experience. One of the most important of these expectations is the expectation of finding a regularity. It is connected with an inborn propensity to look out for regularities, or with a *need* to *find* regularities, as we may see from the pleasure of the child who satisfies this need.

This 'instinctive' expectation of finding regularities, which is psychologically *a priori*, corresponds very closely to the 'law of causality' which Kant believed to be part of our mental outfit and to be *a priori* valid. One might thus be inclined to say that Kant failed to distinguish between psychologically *a priori* ways of thinking or responding and *a priori* valid beliefs. But I do

[14] Katz, *loc. cit.*

not think that his mistake was quite as crude as that. For the expectation of finding regularities is not only psychologically *a priori*, but also logically *a priori*: it is logically prior to all observational experience, for it is prior to any recognition of similarities, as we have seen; and all observation involves the recognition of similarities (or dissimilarities). But in spite of being logically *a priori* in this sense the expectation is not valid *a priori*. For it may fail: we can easily construct an environment (it would be a lethal one) which, compared with our ordinary environment, is so chaotic that we completely fail to find regularities. (All natural laws could remain valid: environments of this kind have been used in the animal experiments mentioned in the next section.)

Thus Kant's reply to Hume came near to being right; for the distinction between an *a priori* valid expectation and one which is both genetically *and* logically prior to observation, but not *a priori* valid, is really somewhat subtle. But Kant proved too much. In trying to show how knowledge is possible, he proposed a theory which had the unavoidable consequence that our quest for knowledge must necessarily succeed, which is clearly mistaken. When Kant said, 'Our intellect does not draw its laws from nature but imposes its laws upon nature', he was right. But in thinking that these laws are necessarily true, or that we necessarily succeed in imposing them upon nature, he was wrong.[15] Nature very often resists quite successfully, forcing us to discard our laws as refuted; but if we live we may try again.

To sum up this logical criticism of Hume's psychology of induction we may consider the idea of building an induction machine. Placed in a simplified 'world' (for example, one of sequences of coloured counters) such a machine may through repetition 'learn', or even 'formulate', laws of succession which hold in its 'world'. If such a machine can be constructed (and I have no doubt that it can) then, it might be argued, my theory must be wrong; for if a machine is capable of performing inductions on the basis of repetition, there can be no logical reasons preventing us from doing the same.

The argument sounds convincing, but it is mistaken. In constructing an induction machine we, the architects of the machine, must decide *a priori* what constitutes its 'world'; what things are to be taken as similar or equal; and what *kind* of 'laws' we wish the machine to be able to 'discover' in its 'world'. In other words we must build into the machine a framework determining what is relevant or interesting in its world: the machine will have its 'inborn' selection principles. The problems of similarity will have been solved for it by its makers who thus have interpreted the 'world' for the machine.

[15] Kant believed that Newton's dynamics was *a priori* valid. (See his *Metaphysical Foundations of Natural Science*, published between the first and the second editions of the *Critique of Pure Reason*.) But if, as he thought, we can explain the validity of Newton's theory by the fact that our intellect imposes its laws upon nature, it follows, I think, that our intellect *must succeed* in this; which makes it hard to understand why *a priori* knowledge such as Newton's should be so hard to come by. A somewhat fuller statement of this criticism can be found in ch. 2, especially section ix, and chs. 7 and 8 of the present volume.

VI

Our propensity to look out for regularities, and to impose laws upon nature, leads to the psychological phenomenon of *dogmatic thinking* or, more generally, dogmatic behaviour: we expect regularities everywhere and attempt to find them even where there are none; events which do not yield to these attempts we are inclined to treat as a kind of 'background noise'; and we stick to our expectations even when they are inadequate and we ought to accept defeat. This dogmatism is to some extent necessary. It is demanded by a situation which can only be dealt with by forcing our conjectures upon the world. Moreover, this dogmatism allows us to approach a good theory in stages, by way of approximations: if we accept defeat too easily, we may prevent ourselves from finding that we were very nearly right.

It is clear that this *dogmatic attitude*, which makes us stick to our first impressions, is indicative of a strong belief; while a *critical attitude*, which is ready to modify its tenets, which admits doubt and demands tests, is indicative of a weaker belief. Now according to Hume's theory, and to the popular theory, the strength of a belief should be a product of repetition; thus it should always grow with experience, and always be greater in less primitive persons. But dogmatic thinking, an uncontrolled wish to impose regularities, a manifest pleasure in rites and in repetition as such, are characteristic of primitives and children; and increasing experience and maturity sometimes create an attitude of caution and criticism rather than of dogmatism.

I may perhaps mention here a point of agreement with psycho-analysis. Psycho-analysts assert that neurotics and others interpret the world in accordance with a personal set pattern which is not easily given up, and which can often be traced back to early childhood. A pattern or scheme which was adopted very early in life is maintained throughout, and every new experience is interpreted in terms of it; verifying it, as it were, and contributing to its rigidity. This is a description of what I have called the dogmatic attitude, as distinct from the critical attitude, which shares with the dogmatic attitude the quick adoption of a schema of expectations—a myth, perhaps, or a conjecture or hypothesis—but which is ready to modify it, to correct it, and even to give it up. I am inclined to suggest that most neuroses may be due to a partially arrested development of the critical attitude; to an arrested rather than a natural dogmatism; to resistance to demands for the modification and adjustment of certain schematic interpretations and responses. This resistance in its turn may perhaps be explained, in some cases, as due to an injury or shock, resulting in fear and in an increased need for assurance or certainty, analogous to the way in which an injury to a limb makes us afraid to move it, so that it becomes stiff. (It might even be argued that the case of the limb is not merely analogous to the dogmatic response, but an instance of it.) The explanation of any concrete case will have to take into account the weight of the difficulties involved in making the necessary adjustments—difficulties which may be considerable, especially in a complex

and changing world: we know from experiments on animals that varying degrees of neurotic behaviour may be produced at will by correspondingly varying difficulties.

I found many other links between the psychology of knowledge and psychological fields which are often considered remote from it—for example the psychology of art and music; in fact, my ideas about induction originated in a conjecture about the evolution of Western polyphony. But you will be spared this story.

VII

My logical criticism of Hume's psychological theory, and the considerations connected with it (most of which I elaborated in 1926–7, in a thesis entitled 'On Habit and Belief in Laws' [16]) may seem a little removed from the field of the philosophy of science. But the distinction between dogmatic and critical thinking, or the dogmatic and the critical attitude, brings us right back to our central problem. For the dogmatic attitude is clearly related to the tendency to *verify* our laws and schemata by seeking to apply them and to confirm them, even to the point of neglecting refutations, whereas the critical attitude is one of readiness to change them—to test them; to refute them; to *falsify* them, if possible. This suggests that we may identify the critical attitude with the scientific attitude, and the dogmatic attitude with the one which we have described as pseudo-scientific.

It further suggests that genetically speaking the pseudo-scientific attitude is more primitive than, and prior to, the scientific attitude: that it is a pre-scientific attitude. And this primitivity or priority also has its logical aspect. For the critical attitude is not so much opposed to the dogmatic attitude as super-imposed upon it: criticism must be directed against existing and influential beliefs in need of critical revision—in other words, dogmatic beliefs. A critical attitude needs for its raw material, as it were, theories or beliefs which are held more or less dogmatically.

Thus science must begin with myths, and with the criticism of myths; neither with the collection of observations, nor with the invention of experiments, but with the critical discussion of myths, and of magical techniques and practices. The scientific tradition is distinguished from the pre-scientific tradition in having two layers. Like the latter, it passes on its theories; but it also passes on a critical attitude towards them. The theories are passed on, not as dogmas, but rather with the challenge to discuss them and improve upon them. This tradition is Hellenic: it may be traced back to Thales, founder of the first *school* (I do not mean 'of the first *philosophical* school', but simply 'of the first school') which was not mainly concerned with the preservation of a dogma.[17]

The critical attitude, the tradition of free discussion of theories with the

[16] A thesis submitted under the title '*Gewohnheit und Gesetzerlebnis*' to the Institute of Education of the City of Vienna in 1927. (Unpublished.)

[17] Further comments on these developments may be found in chs. 4 and 5, below.

aim of discovering their weak spots so that they may be improved upon, is the attitude of reasonableness, of rationality. It makes far-reaching use of both verbal argument and observation—of observation in the interest of argument, however. The Greeks' discovery of the critical method gave rise at first to the mistaken hope that it would lead to the solution of all the great old problems; that it would establish certainty; that it would help to *prove* our theories, to *justify* them. But this hope was a residue of the dogmatic way of thinking; in fact nothing can be justified or proved (outside of mathematics and logic). The demand for rational proofs in science indicates a failure to keep distinct the broad realm of rationality and the narrow realm of rational certainty: it is an untenable, an unreasonable demand.

Nevertheless, the role of logical argument, of deductive logical reasoning, remains all-important for the critical approach; not because it allows us to prove our theories, or to infer them from observation statements, but because only by purely deductive reasoning is it possible for us to discover what our theories imply, and thus to criticize them effectively. Criticism, I said, is an attempt to find the weak spots in a theory, and these, as a rule, can be found only in the more remote logical consequences which can be derived from it. It is here that purely logical reasoning plays an important part in science.

Hume was right in stressing that our theories cannot be validly inferred from what we can know to be true—neither from observations nor from anything else. He concluded from this that our belief in them was irrational. If 'belief' means here our inability to doubt our natural laws, and the constancy of natural regularities, then Hume is again right: this kind of dogmatic belief has, one might say, a physiological rather than a rational basis. If, however, the term 'belief' is taken to cover our critical acceptance of scientific theories —a *tentative* acceptance combined with an eagerness to revise the theory if we succeed in designing a test which it cannot pass—then Hume was wrong. In such an acceptance of theories there is nothing irrational. There is not even anything irrational in relying for practical purposes upon well-tested theories, for no more rational course of action is open to us.

Assume that we have deliberately made it our task to live in this unknown world of ours; to adjust ourselves to it as well as we can; to take advantage of the opportunities we can find in it; and to explain it, *if* possible (we need not assume that it is), and as far as possible, with the help of laws and explanatory theories. *If we have made this our task, then there is no more rational procedure than the method of trial and error—of conjecture and refutation:* of boldly proposing theories; of trying our best to show that these are erroneous; and of accepting them tentatively if our critical efforts are unsuccessful.

From the point of view here developed all laws, all theories, remain essentially tentative, or conjectural, or hypothetical, even when we feel unable to doubt them any longer. Before a theory has been refuted we can never know in what way it may have to be modified. That the sun will always rise and set within twenty-four hours is still proverbial as a law 'established by induction beyond reasonable doubt'. It is odd that this example is still in use, though it

may have served well enough in the days of Aristotle and Pytheas of Massalia —the great traveller who for centuries was called a liar because of his tales of Thule, the land of the frozen sea and the *midnight sun*.

The method of trial and error is not, of course, simply identical with the scientific or critical approach—with the method of conjecture and refutation. The method of trial and error is applied not only by Einstein but, in a more dogmatic fashion, by the amoeba also. The difference lies not so much in the trials as in a critical and constructive attitude towards errors; errors which the scientist consciously and cautiously tries to uncover in order to refute his theories with searching arguments, including appeals to the most severe experimental tests which his theories and his ingenuity permit him to design.

The critical attitude may be described as the conscious attempt to make our theories, our conjectures, suffer in our stead in the struggle for the survival of the fittest. It gives us a chance to survive the elimination of an inadequate hypothesis—when a more dogmatic attitude would eliminate it by eliminating us. (There is a touching story of an Indian community which disappeared because of its belief in the holiness of life, including that of tigers.) We thus obtain the fittest theory within our reach by the elimination of those which are less fit. (By 'fitness' I do not mean merely 'usefulness' but truth; see chapters 3 and 10, below.) I do not think that this procedure is irrational or in need of any further rational justification.

VIII

Let us now turn from our logical criticism of the *psychology of experience* to our real problem—the problem of *the logic of science*. Although some of the things I have said may help us here, in so far as they may have eliminated certain psychological prejudices in favour of induction, my treatment of the *logical problem of induction* is completely independent of this criticism, and of all psychological considerations. Provided you do not dogmatically believe in the alleged psychological fact that we make inductions, you may now forget my whole story with the exception of two logical points: my logical remarks on testability or falsifiability as the criterion of demarcation; and Hume's logical criticism of induction.

From what I have said it is obvious that there was a close link between the two problems which interested me at that time: demarcation, and induction or scientific method. It was easy to see that the method of science is criticism, i.e. attempted falsifications. Yet it took me a few years to notice that the two problems—of demarcation and of induction—were in a sense one.

Why, I asked, do so many scientists believe in induction? I found they did so because they believed natural science to be characterized by the inductive method—by a method starting from, and relying upon, long sequences of observations and experiments. They believed that the difference between genuine science and metaphysical or pseudo-scientific speculation depended solely upon whether or not the inductive method was employed. They

believed (to put it in my own terminology) that only the inductive method could provide a satisfactory *criterion of demarcation.*

I recently came across an interesting formulation of this belief in a remarkable philosophical book by a great physicist—Max Born's *Natural Philosophy of Cause and Chance.*[18] He writes: 'Induction allows us to generalize a number of observations into a general rule: that night follows day and day follows night . . . But while everyday life has no definite criterion for the validity of an induction, . . . science has worked out a code, or rule of craft, for its application.' Born nowhere reveals the contents of this inductive code (which, as his wording shows, contains a 'definite criterion for the validity of an induction'); but he stresses that 'there is no logical argument' for its acceptance: 'it is a question of faith'; and he is therefore 'willing to call induction a metaphysical principle'. But why does he believe that such a code of valid inductive rules must exist? This becomes clear when he speaks of the 'vast communities of people ignorant of, or rejecting, the rule of science, among them the members of anti-vaccination societies and believers in astrology. It is useless to argue with them; I cannot compel them to accept the same criteria of valid induction in which I believe: the code of scientific rules.' This makes it quite clear that *'valid induction' was here meant to serve as a criterion of demarcation between science and pseudo-science.*

But it is obvious that this rule or craft of 'valid induction' is not even metaphysical: it simply does not exist. No rule can ever guarantee that a generalization inferred from true observations, however often repeated, is true. (Born himself does not believe in the truth of Newtonian physics, in spite of its success, although he believes that it is based on induction.) And the success of science is not based upon rules of induction, but depends upon luck, ingenuity, and the purely deductive rules of critical argument.

I may summarize some of my conclusions as follows:

(1) Induction, i.e. inference based on many observations, is a myth. It is neither a psychological fact, nor a fact of ordinary life, nor one of scientific procedure.

(2) The actual procedure of science is to operate with conjectures: to jump to conclusions—often after one single observation (as noticed for example by Hume and Born).

(3) Repeated observations and experiments function in science as *tests* of our conjectures or hypotheses, i.e. as attempted refutations.

(4) The mistaken belief in induction is fortified by the need for a criterion of demarcation which, it is traditionally but wrongly believed, only the inductive method can provide.

(5) The conception of such an inductive method, like the criterion of verifiability, implies a faulty demarcation.

(6) None of this is altered in the least if we say that induction makes theories only probable rather than certain. (See especially chapter 10, below.)

[18] Max Born, *Natural Philosophy of Cause and Chance*, Oxford, 1949, p. 7.

IX

If, as I have suggested, the problem of induction is only an instance or facet of the problem of demarcation, then the solution to the problem of demarcation must provide us with a solution to the problem of induction. This is indeed the case, I believe, although it is perhaps not immediately obvious.

For a brief formulation of the problem of induction we can turn again to Born, who writes: '. . . no observation or experiment, however extended, can give more than a finite number of repetitions'; therefore, 'the statement of a law—B depends on A—always transcends experience. Yet this kind of statement is made everywhere and all the time, and sometimes from scanty material.'[19]

In other words, the logical problem of induction arises from (*a*) Hume's discovery (so well expressed by Born) that it is impossible to justify a law by observation or experiment, since it 'transcends experience'; (*b*) the fact that science proposes and uses laws 'everywhere and all the time'. (Like Hume, Born is struck by the 'scanty material', i.e. the few observed instances upon which the law may be based.) To this we have to add (*c*) *the principle of empiricism* which asserts that in science, only observation and experiment may decide upon the *acceptance or rejection* of scientific statements, including laws and theories.

These three principles, (*a*), (*b*), and (*c*), appear at first sight to clash; and this apparent clash constitutes the *logical problem of induction.*

Faced with this clash, Born gives up (*c*), the principle of empiricism (as Kant and many others, including Bertrand Russell, have done before him), in favour of what he calls a 'metaphysical principle'; a metaphysical principle which he does not even attempt to formulate; which he vaguely describes as a 'code or rule of craft'; and of which I have never seen any formulation which even looked promising and was not clearly untenable.

But in fact the principles (*a*) to (*c*) do not clash. We can see this the moment we realize that the acceptance by science of a law or of a theory is *tentative only*; which is to say that all laws and theories are conjectures, or tentative *hypotheses* (a position which I have sometimes called 'hypotheticism'); and that we may reject a law or theory on the basis of new evidence, without necessarily discarding the old evidence which originally led us to accept it.[20]

The principle of empiricism (*c*) can be fully preserved, since the fate of a theory, its acceptance or rejection, is decided by observation and experiment —by the result of tests. So long as a theory stands up to the severest tests we can design, it is accepted; if it does not, it is rejected. But it is never inferred, in any sense, from the empirical evidence. There is neither a psychological nor

[19] *Natural Philosophy of Cause and Chance*, p. 6.

[20] I do not doubt that Born and many others would agree that theories are accepted only tentatively. But the widespread belief in induction shows that the far-reaching implications of this view are rarely seen.

a logical induction. *Only the falsity of the theory can be inferred from empirical evidence, and this inference is a purely deductive one.*

Hume showed that it is not possible to infer a theory from observation statements; but this does not affect the possibility of refuting a theory by observation statements. The full appreciation of this possibility makes the relation between theories and observations perfectly clear.

This solves the problem of the alleged clash between the principles (*a*), (*b*), and (*c*), and with it Hume's problem of induction.

X

Thus the problem of induction is solved. But nothing seems less wanted than a simple solution to an age-old philosophical problem. Wittgenstein and his school hold that genuine philosophical problems do not exist;[21] from which it clearly follows that they cannot be solved. Others among my contemporaries do believe that there are philosophical problems, and respect them; but they seem to respect them too much; they seem to believe that they are insoluble, if not taboo; and they are shocked and horrified by the claim that there is a simple, neat, and lucid, solution to any of them. If there is a solution it must be deep, they feel, or at least complicated.

However this may be, I am still waiting for a simple, neat and lucid criticism of the solution which I published first in 1933 in my letter to the Editor of *Erkenntnis*,[22] and later in *The Logic of Scientific Discovery.*

Of course, one can invent new problems of induction, different from the one I have formulated and solved. (Its formulation was half its solution.) But I have yet to see any reformulation of the problem whose solution cannot be easily obtained from my old solution. I am now going to discuss some of these re-formulations.

One question which may be asked is this: how do we really jump from an observation statement to a theory?

Although this question appears to be psychological rather than philosophical, one can say something positive about it without invoking psychology. One can say first that the jump is not from an observation statement, but from a problem-situation, and that the theory must allow us *to explain* the observations which created the problem (that is, *to deduce* them from the theory strengthened by other accepted theories and by other observation statements, the so-called initial conditions). This leaves, of course, an immense number of possible theories, good and bad; and it thus appears that our question has not been answered.

But this makes it fairly clear that when we asked our question we had more in mind than, 'How do we jump from an observation statement to a theory?' The question we had in mind was, it now appears, 'How do we jump from an observation statement to a *good* theory?' But to this the answer is: by jumping first to *any* theory and then testing it, to find whether it is good or not; i.e.

[21] Wittgenstein still held this belief in 1946; see note 8 to ch. 2, below.

[22] See note 5 above.

by repeatedly applying the critical method, eliminating many bad theories, and inventing many new ones. Not everybody is able to do this; but there is no other way.

Other questions have sometimes been asked. The original problem of induction, it was said, is the problem of *justifying* induction, i.e. of justifying inductive inference. If you answer this problem by saying that what is called an 'inductive inference' is always invalid and therefore clearly not justifiable, the following new problem must arise: how do you justify your method of trial and error? Reply: the method of trial and error is a *method of eliminating false theories* by observation statements; and the justification for this is the purely logical relationship of deducibility which allows us to assert the falsity of universal statements if we accept the truth of singular ones.

Another question sometimes asked is this: why is it reasonable to prefer non-falsified statements to falsified ones? To this question some involved answers have been produced, for example pragmatic answers. But from a pragmatic point of view the question does not arise, since false theories often serve well enough: most formulae used in engineering or navigation are known to be false, although they may be excellent approximations and easy to handle; and they are used with confidence by people who know them to be false.

The only correct answer is the straightforward one: because we search for truth (even though we can never be sure we have found it), and because the falsified theories are known or believed to be false, while the non-falsified theories may still be true. Besides, we do not prefer *every* non-falsified theory —only one which, in the light of criticism, appears to be better than its competitors: which solves our problems, which is well tested, and of which we think, or rather conjecture or hope (considering other provisionally accepted theories), that it will stand up to further tests.

It has also been said that the problem of induction is, 'Why is it *reasonable* to believe that the future will be like the past?', and that a satisfactory answer to this question should make it plain that such a belief is, in fact, reasonable. My reply is that it is reasonable to believe that the future will be very different from the past in many vitally important respects. Admittedly it is perfectly reasonable to *act* on the assumption that it will, in many respects, be like the past, and that well-tested laws will continue to hold (since we can have no better assumption to act upon); but it is also reasonable to believe that such a course of action will lead us at times into severe trouble, since some of the laws upon which we now heavily rely may easily prove unreliable. (Remember the midnight sun!) One might even say that to judge from past experience, and from our general scientific knowledge, the future will *not* be like the past, in perhaps most of the ways which those have in mind who say that it will. Water will sometimes not quench thirst, and air will choke those who breathe it. An apparent way out is to say that the future will be like the past *in the sense that the laws of nature will not change*, but this is begging the question. We speak of a 'law of nature' only if we think that we have before us a regularity which does not change; and if we find that it changes then we shall not

continue to call it a 'law of nature'. Of course our search for natural laws indicates that we hope to find them, and that we believe that there are natural laws; but our belief in any particular natural law cannot have a safer basis than our unsuccessful critical attempts to refute it.

I think that those who put the problem of induction in terms of the *reasonableness* of our beliefs are perfectly right if they are dissatisfied with a Humean, or post-Humean, sceptical despair of reason. We must indeed reject the view that a belief in science is as irrational as a belief in primitive magical practices —that both are a matter of accepting a 'total ideology', a convention or a tradition based on faith. But we must be cautious if we formulate our problem, with Hume, as one of the reasonableness of our *beliefs*. We should split this problem into three—our old problem of demarcation, or of how to *distinguish* between science and primitive magic; the problem of the rationality of the scientific or critical *procedure*, and of the role of observation within it; and lastly the problem of the rationality of our *acceptance* of theories for scientific and for practical purposes. To all these three problems solutions have been offered here.

One should also be careful not to confuse the problem of the reasonableness of the scientific procedure and the (tentative) acceptance of the results of this procedure—i.e. the scientific theories—with the problem of the rationality or otherwise *of the belief that this procedure will succeed*. In practice, in practical scientific research, this belief is no doubt unavoidable and reasonable, there being no better alternative. But the belief is certainly unjustifiable in a theoretical sense, as I have argued (in section v). Moreover, if we could show, on general logical grounds, that the scientific quest is likely to succeed, one could not understand why anything like success has been so rare in the long history of human endeavours to know more about our world.

Yet another way of putting the problem of induction is in terms of probability. Let t be the theory and e the evidence: we can ask for $P(t,e)$, that is to say, the probability of t, given e. The problem of induction, it is often believed, can then be put thus: construct a *calculus of probability* which allows us to work out for any theory t what its probability is, relative to any given empirical evidence e; and show that $P(t,e)$ increases with the accumulation of supporting evidence, and reaches high values—at any rate values greater than $\frac{1}{2}$.

In *The Logic of Scientific Discovery* I explained why I think that this approach to the problem is fundamentally mistaken.[23] To make this clear, I introduced there the distinction between *probability* and *degree of corroboration or confirmation*. (The term 'confirmation' has lately been so much used and misused that I have decided to surrender it to the verificationists and to use for my own purposes 'corroboration' only. The term 'probability' is best

[23] *L.Sc.D.* (see note 5 above), ch. x, especially sections 80 to 83, also section 34 ff. See also my note 'A Set of Independent Axioms for Probability', *Mind*, N.S. 47, 1938, p. 275. (This note has since been reprinted, with corrections, in the new appendix *ii of *L.Sc.D.* See also the next note but one to the present chapter.)

used in some of the many senses which satisfy the well-known calculus of probability, axiomatized, for example, by Keynes, Jeffreys, and myself; but nothing of course depends on the choice of words, as long as we do not *assume*, uncritically, that degree of corroboration must also be a probability —that is to say, that it must satisfy the calculus of probability.)

I explained in my book why we are interested in theories with a *high degree of corroboration*. And I explained why it is a mistake to conclude from this that we are interested in *highly probable* theories. I pointed out that the probability of a statement (or set of statements) is always the greater the less the statement says: it is inverse to the content or the deductive power of the statement, and thus to its explanatory power. Accordingly every interesting and powerful statement must have a low probability; and *vice versa*: a statement with a high probability will be scientifically uninteresting, because it says little and has no explanatory power. Although we seek theories with a high degree of corroboration, *as scientists we do not seek highly probable theories but explanations; that is to say, powerful and improbable theories.*[24] The opposite view—that science aims at high probability—is a characteristic development of verificationism: if you find that you cannot verify a theory, or make it certain by induction, you may turn to probability as a kind of '*Ersatz*' for certainty, in the hope that induction may yield at least that much.

I have discussed the two problems of demarcation and induction at some length. Yet since I set out to give you in this lecture a kind of report on the work I have done in this field I shall have to add, in the form of an Appendix, a few words about some other problems on which I have been working, between 1934 and 1953. I was led to most of these problems by trying to think out the consequences of the solutions to the two problems of demarcation and induction. But time does not allow me to continue my narrative, and to tell you how my new problems arose out of my old ones. Since I cannot even start a discussion of these further problems now, I shall have to confine my-

[24] A definition, in terms of probabilities (see the next note), of $C(t,e)$, i.e. of the degree of corroboration (of a theory t relative to the evidence e) satisfying the demands indicated in my *L.Sc.D.*, sections 82 to 83, is the following:

$$C(t, e) = E(t,e)\,(1 + P(t)P(t,e)),$$

where $E(t,e) = (P(e,t) - P(e))/(P(e,t) + P(e))$ is a (non-additive) measure of the explanatory power of t with respect to e. Note that $C(t,e)$ is not a probability: it may have values between -1 (refutation of t by e) and $C(t, t) < +1$. Statements t which are lawlike and thus non-verifiable cannot even reach $C(t,e) = C(t,t)$ upon empirical evidence e. $C(t,t)$ is the *degree of corroborability* of t, and is equal to the *degree of testability* of t, or to the *content* of t. Because of the demands implied in point (6) at the end of section I above, I do not think, however, that it is possible to give a complete formalization of the idea of corroboration (or, as I previously used to say, of confirmation).

(Added 1955 to the first proofs of this paper:)

See also my note 'Degree of Confirmation', *British Journal for the Philosophy of Science*, **5**, 1954, pp. 143 ff. (See also **5**, pp. 334.) I have since simplified this definition as follows (*B.J.P.S.*, 1955, **5**, p. 359:)

$$C(t,e) = (P(e,t) - P(e))/(P(e,t) - P(et) + P(e))$$

For a further improvement, see *B.J.P.S.* **6**, 1955, p. 56.

self to giving you a bare list of them, with a few explanatory words here and there. But even a bare list may be useful, I think. It may serve to give an idea of the fertility of the approach. It may help to illustrate what our problems look like; and it may show how many there are, and so convince you that there is no need whatever to worry over the question whether philosophical problems exist, or what philosophy is really about. So this list contains, by implication, an apology for my unwillingness to break with the old tradition of trying to solve problems with the help of rational argument, and thus for my unwillingness to participate wholeheartedly in the developments, trends, and drifts, of contemporary philosophy.

APPENDIX: SOME PROBLEMS IN THE PHILOSOPHY OF SCIENCE

My first three items in this list of additional problems are connected with the calculus of probabilities.

(1) The frequency theory of probability. In *The Logic of Scientific Discovery* I was interested in developing a consistent theory of probability as it is used in science; which means, a statistical or frequency theory of probability. But I also operated there with another concept which I called 'logical probability'. I therefore felt the need for a generalization—for a formal theory of probability which allows different *interpretations*: (*a*) as a theory of the logical probability of a statement relative to any given evidence; including a theory of absolute logical probability, i.e. of the measure of the probability of a statement relative to zero evidence; (*b*) as a theory of the probability of an event relative to any given *ensemble* (or 'collective') of events. In solving this problem I obtained a simple theory which allows a number of further interpretations: it may be interpreted as a calculus of contents, or of deductive systems, or as a class calculus (Boolean algebra) or as propositional calculus; and also as a calculus of *propensities*.[25]

[25] See my note in *Mind*, *loc. cit.* The axiom system given there for elementary (i.e. non-continuous) probability can be simplified as follows ('$\bar{x}$' denotes the complement of x; 'xy' the intersection or conjunction of x and y):

(A1)	$P(xy) > P(yx)$	(Commutation)
(A2)	$P(x(yz)) > P((xy)z)$	(Association)
(A3)	$P(xx) > P(x)$	(Tautology)
(B1)	$P(x) > P(xy)$	(Monotony)
(B2)	$P(xy) + P(x\bar{y}) = P(x)$	(Addition)
(B3)	$(x)\,(Ey)\,(P(y) \neq O$ *and* $P(xy) = P(x)P(y))$	(Multiplication)
(C1)	*If* $P(y) \neq O$, *then* $P(x,y) = P(xy)/P(y)$	(Definition of relative
(C2)	*If* $P(y) = O$, *then* $P(x,y) = P(x,x) = P(y,y)$	Probability)

Axiom (C2) holds, in this form, for the finitist theory only; it may be omitted if we are prepared to put up with a condition such as $P(y) \neq O$ in most of the theorems on relative probability. For relative probability, (A1) — (B2) and (C1) — (C2), is sufficient; (B3) is not needed. For absolute probability, (A1) — (B3) is necessary and sufficient: without (B3)

(2) This problem of a *propensity interpretation of probability* arose out of my interest in Quantum Theory. It is usually believed that Quantum Theory has to be interpreted statistically, and no doubt statistics is essential for its empirical tests. But this is a point where, I believe, the dangers of the testability theory of meaning become clear. Although the tests of the theory are statistical, and although the theory (say, Schrödinger's equation) may imply statistical consequences, it need not have a statistical meaning: and one can give examples of objective propensities (which are something like generalized forces) and of fields of propensities, which can be measured by statistical methods without being themselves statistical. (See also the last paragraph of chapter 3, below, with note 35.)

(3) The use of statistics in such cases is, in the main, to provide *empirical tests* of theories which need not be purely statistical; and this raises the question of the *refutability of statistical statements*—a problem treated, but not to my full satisfaction, in the 1934 edition of my *The Logic of Scientific Discovery*. I later found, however, that all the elements for constructing a satisfactory solution lay ready for use in that book; certain examples I had given allow a mathematical characterization of a class of infinite chance-like

---

we cannot, for example, derive the definition of absolute in terms of relative probability,

$$P(x) = P(x,\overline{x\bar{x}})$$

nor its weakened corollary

$$(x)(Ey)\,(P(y) \neq O \text{ and } P(x) = P(x,y))$$

from which (B3) results immediately (by substituting for '$P(x,y)$' its definiens). Thus (B3), like all other axioms with the possible exception of (C2), expresses part of the intended meaning of the concepts involved, and we must not look upon $1 > P(x)$ or $1 > P(x,y)$, which are derivable from (B1), with (B3) or with (C1) and (C2), as 'inessential conventions' (as Carnap and others have suggested).

(Added 1955 to the first proofs of this paper; see also note 31, below.)

I have since developed an axiom system for *relative probability* which holds for finite *and* infinite systems (and in which absolute probability can be defined as in the penultimate formula above). Its axioms are:

(B1) $P(x,z) > P(xy,z)$
(B2) If $P(y,y) \neq P(u,y)$ then $P(x,y) + P(\bar{x},y) = P(y,y)$
(B3) $P(xy,z) = P(x,yz)P(y,z)$
(C1) $P(x,x) = P(y,y)$
(D1) If $((u)P(x,u) = P(y,u))$ then $P(w,x) = P(w,y)$
(E1) $(Ex)\,(Ey)\,(Eu)\,(Ew)\,P(x,y) \neq P(u,w)$

This is a slight improvement on a system which I published in *B.J.P.S.*, 6, 1955, pp. 56 f.; 'Postulate 3' is here called 'D1'. (See also *vol. cit.*, bottom of p. 176. Moreover, in line 3 of the last paragraph on p. 57, the words 'and that the limit exists' should be inserted, between brackets, before the word 'all'.)

(Added 1961 to the proofs of the present volume.)

A fairly full treatment of all these questions will now be found in the new addenda to *L.Sc.D.*

I have left this note as in the first publication because I have referred to it in various places. The problems dealt with in this and the preceding note have since been more fully treated in the new appendices to *L.Sc.D.* (To its 1961 American Edition I have added a system of only three axioms; see also section 2 of the *Addenda* to the present volume.)

sequences which are, in a certain sense, the *shortest sequences* of their kind.[26] A statistical statement may now be said to be testable by comparison with these 'shortest sequences'; it is refuted if the statistical properties of the tested *ensembles* differ from the statistical properties of the initial sections of these 'shortest sequences'.

(4) There are a number of further problems connected with the interpretation of the formalism of a quantum theory. In a chapter of *The Logic of Scientific Discovery* I criticized the 'official' interpretation, and I still think that my criticism is valid in all points but one: one example which I used (in section 77) is mistaken. But since I wrote that section, Einstein, Podolski, and Rosen have published a thought-experiment which can be substituted for my example, although their tendency (which is deterministic) is quite different from mine. Einstein's belief in determinism (which I had occasion to discuss with him) is, I believe, unfounded, and also unfortunate: it robs his criticism of much of its force, and it must be emphasized that much of his criticism is quite independent of his determinism.

(5) As to the problem of determinism itself, I have tried to show that even classical physics, which is deterministic in a certain *prima facie* sense, is misinterpreted if used to support a deterministic view of the physical world in Laplace's sense.

(6) In this connection, I may also mention the *problem of simplicity*—of the simplicity of a theory, which I have been able to connect with the content of a theory. It can be shown that what is usually called the simplicity of a theory is associated with its logical improbability, and not with its probability, as has often been supposed. This, indeed, allows us to deduce, from the theory of science outlined above, why it is always advantageous to try the simplest theories first. They are those which offer us the best chance to submit them to severe tests: the simpler theory has always a higher degree of testability than the more complicated one.[27] (Yet I do not think that this settles all problems about simplicity. See also chapter 10, section xviii, below.)

(7) Closely related to this problem is the problem of the *ad hoc* character of a hypothesis, and of degrees of this *ad hoc* character (of '*ad hocness*', if I may so call it). One can show that the methodology of science (and the history of science also) becomes understandable in its details if we assume that the aim of science is to get explanatory theories which are as little *ad hoc* as possible: a 'good' theory is not *ad hoc*, while a 'bad' theory is. On the other hand one can show that the probability theories of induction imply, inadvertently but necessarily, the unacceptable rule: always use the theory which is the most *ad hoc*, i.e. which transcends the available evidence as little as possible. (See also my paper 'The Aim of Science', mentioned in note 28 below.)

(8) An important problem is the problem of the *layers of explanatory hypotheses* which we find in the more developed theoretical sciences, and of

[26] See *L.Sc.D.*, p. 163 (section 55); see especially the new appendix *xvi.
[27] Ibid., sections 41 to 46. But see now also ch. 10, section xviii.

the relations between these layers. It is often asserted that Newton's theory can be induced or even deduced from Kepler's and Galileo's laws. But it can be shown that Newton's theory (including his theory of absolute space) strictly speaking contradicts Kepler's (even if we confine ourselves to the two-body problem[28] and neglect the mutual attraction between the planets) and also Galileo's; although approximations to these two theories can, of course, be deduced from Newton's. But it is clear that neither a deductive nor an inductive inference can lead, from consistent premises, to a conclusion which contradicts them. These considerations allow us to analyse the logical relations between 'layers' of theories, and also the idea of an *approximation*, in the two senses of (*a*) The theory x is an approximation to the theory y; and (*b*) The theory x is 'a good approximation to the facts'. (See also chapter 10, below.)

(9) A host of interesting problems is raised by *operationalism*, the doctrine that theoretical concepts have to be defined in terms of measuring operations. Against this view, it can be shown that *measurements presuppose theories*. There is no measurement without a theory and no operation which can be satisfactorily described in non-theoretical terms. The attempts to do so are always circular; for example, the description of the measurement of length needs a (rudimentary) theory of heat and temperature-measurement; but these, in turn, involve measurements of length.

The analysis of operationalism shows the need for a *general theory of measurement*; a theory which does not, naïvely, take the practice of measuring as 'given', but explains it by analysing its function in the testing of scientific hypotheses. This can be done with the help of the doctrine of degrees of testability.

Connected with, and closely parallel to, operationalism is the doctrine of *behaviourism*, i.e. the doctrine that, since all test-statements describe behaviour, our theories too must be stated in terms of possible behaviour. But the inference is as invalid as the phenomenalist doctrine which asserts that since all test-statements are observational, theories too must be stated in terms of possible observations. All these doctrines are forms of the verifiability theory of meaning; that is to say, of inductivism.

Closely related to operationalism is *instrumentalism*, i.e. the interpretation of scientific theories as practical instruments or tools for such purposes as the

[28] The contradictions mentioned in this sentence of the text were pointed out, for the case of the many-body problem, by P. Duhem, *The Aim and Structure of Physical Theory* (1905; trans. by P. P. Wiener, 1954). In the case of the two-body problem, the contradictions arise in connection with Kepler's third law, which may be reformulated for the two-body problem as follows. 'Let S be any *set of pairs* of bodies such that *one* body of each pair is of the mass of our sun; then $a^3/T^2 = constant$, for any set S.' Clearly this contradicts Newton's theory, which yields for appropriately chosen units $a^3/T^2 = m_0 + m_1$ (where m_0 = mass of the sun = *constant*, and m_1 = mass of the second body, which varies with this body). But '$a^3/T^2 = constant$' is, of course, an excellent approximation, *provided* the varying masses of the second bodies are all negligible compared with that of our sun. (See also my paper 'The Aim of Science', *Ratio*, 1, 1957, pp. 24 ff., and section 15 of the *Postscript* to my *Logic of Scientific Discovery*.)

prediction of impending events. That theories may be used in this way cannot be doubted; but instrumentalism asserts that they can be best understood as instruments; and that this is mistaken, I have tried to show by a comparison of the *different functions* of the formulae of applied and pure science. In this context the problem of the *theoretical* (i.e. non-practical) function of predictions can also be solved. (See chapter 3, section 5, below.)

It is interesting to analyse from the same point of view the function of language—as an instrument. One immediate finding of this analysis is that we use descriptive language in order to talk *about the world*. This provides new arguments in favour of *realism*.

Operationalism and instrumentalism must, I believe, be replaced by 'theoreticism', if I may call it so: by the recognition of the fact that we are always operating within a complex framework of theories, and that we do not aim simply at correlations, but at explanations.

(10) The problem of *explanation* itself. It has often been said that scientific explanation is reduction of the unknown to the known. If pure science is meant, nothing could be further from the truth. It can be said without paradox that scientific explanation is, on the contrary, the reduction of the known to the unknown. In pure science, as opposed to an applied science which takes pure science as 'given' or 'known', explanation is always the logical reduction of hypotheses to others which are of a higher level of universality; of 'known' facts and 'known' theories to assumptions of which we know very little as yet, and which have still to be tested. The analysis of degrees of explanatory power, and of the relationship between genuine and sham explanation and between explanation and prediction, are examples of problems which are of great interest in this context.

(11) This brings me to the problem of the relationship between explanation in the natural sciences and historical explanation (which, strangely enough, is logically somewhat analogous to the problem of explanation in the pure and applied sciences); and to the vast field of problems in the methodology of the social sciences, especially the problems of *historical prediction*; *historicism* and *historical determinism*; and *historical relativism*. These problems are linked, again, with the more general problems of determinism and relativism, including the problems of linguistic relativism.[29]

(12) A further problem of interest is the analysis of what is called 'scientific objectivity'. I have treated this problem in several places, especially in connection with a criticism of the so-called 'sociology of knowledge'.[30]

(13) One type of solution of the problem of induction should be mentioned here again (see section iv, above), in order to warn against it. (Solutions of this kind are, as a rule, put forth without a clear formulation of the problem which they are supposed to solve.) The view I have in mind may be described

[29] See my *Poverty of Historicism*, 1957, sections 28 and note 30 to 32; also the Addendum to vol. ii of my *Open Society* (added to the 4th edition 1962).

[30] *Poverty of Historicism*, section 32; *L.Sc.D.*, section 8; *Open Society*, ch. 23 and Addendum to vol. ii (Fourth Edition). The passages are complementary.

as follows. It is first taken for granted that nobody seriously doubts that we do, *in fact*, make inductions, and successful ones. (My suggestion that this is a myth, and that the apparent cases of induction turn out, if analysed more carefully, to be cases of the method of trial and error, is treated with the contempt which an utterly unreasonable suggestion of this kind deserves.) It is then said that the task of a theory of induction is to describe and classify our inductive policies or procedures, and perhaps to point out which of them are the most successful and reliable ones and which are less successful or reliable; and that any further question of justification is misplaced. Thus the view I have in mind is characterized by the contention that the distinction between the factual problem of describing how we argue inductively (*quid facti?*), and the problem of the justification of our inductive arguments (*quid juris?*) is a misplaced distinction. It is also said that the justification required is unreasonable, since we cannot expect inductive arguments to be 'valid' in the same sense in which deductive ones may be 'valid': induction simply is not deduction, and it is unreasonable to demand from it that it should conform to the standards of logical—that is, deductive—validity. We must therefore judge it by its own standards—by inductive standards—of reasonableness.

I think that this defence of induction is mistaken. It not only takes a myth for a fact, and the alleged fact for a standard of rationality, with the result that a myth becomes a standard of rationality; but it also propagates, in this way, a principle which may be used to defend *any* dogma against *any* criticism. Moreover, it mistakes the status of formal or 'deductive' logic. (It mistakes it just as much as those who saw it as the systematization of our factual, that is, psychological, 'laws of thought'.) For deduction, I contend, is not valid because we choose or decide to adopt its rules as a standard, or decree that they shall be accepted; rather, it is valid because it adopts, and incorporates, the rules by which truth is transmitted from (logically stronger) premises to (logically weaker) conclusions, and by which falsity is re-transmitted from conclusions to premises. (This re-transmission of falsity makes formal logic the *Organon of rational criticism*—that is, of refutation.)

One point that may be conceded to those who hold the view I am criticizing here is this. In arguing from premises to the conclusion (or in what may be called the 'deductive direction'), we argue from the truth or the certainty or the probability of the premises to the corresponding property of the conclusion; while if we argue from the conclusion to the premises (and thus in what we have called the 'inductive direction'), we argue from the falsity or the uncertainty or the impossibility or the improbability of the conclusion to the corresponding property of the premises; accordingly, we must indeed concede that standards such as, more especially, *certainty*, which apply to arguments in the deductive direction, do not also apply to arguments in the inductive direction. Yet even this concession of mine turns in the end against those who hold the view which I am criticizing here; for they assume, wrongly, that we may argue in the inductive direction, though not to the certainty, yet to the *probability* of our 'generalizations'. But this assumption

is mistaken, for all the intuitive ideas of probability which have ever been suggested.

This is a list of just a few of the problems of the philosophy of science to which I was led in my pursuit of the two fertile and fundamental problems whose story I have tried to tell you.[31]

[31] (13) was added in 1961. Since 1953, when this lecture was delivered, and 1955, when I read the proofs, the list given in this appendix has grown considerably, and some more recent contributions which deal with problems not listed here will be found in this volume (see especially ch. 10, below) and in my other books (see especially the new appendices to my *L.Sc.D.*, and the new *Addendum* to vol. ii of my *Open Society* which I have added to the fourth edition, 1962). See especially also my paper 'Probability Magic, or Knowledge out of Ignorance', *Dialectica*, **11**, 1957, pp. 354–374.

CONJECTURES AND REFUTATIONS

The Growth of Scientific Knowledge

by

KARL R. POPPER

HARPER TORCHBOOKS
Harper & Row, Publishers
New York and Evanston