14

BAYESIAN VERSUS NON-BAYESIAN APPROACHES TO CONFIRMATION*

Colin Howson and Peter Urbach

The Bayesian notion of confirmation

Information gathered in the course of observation is often considered to have a bearing on the acceptability of a theory or hypothesis (we use the terms interchangeably), either by confirming it or by disconfirming it. Such information may either derive from casual observation or, more commonly, from experiments deliberately contrived in the hope of obtaining relevant evidence. The idea that evidence may count for or against a theory, or be neutral towards it, is a central feature of scientific inference, and the Bayesian account will clearly need to start with a suitable interpretation of these concepts.

Fortunately, there is a suitable and very natural interpretation, for if $P(h)$ measures your belief in a hypothesis when you do not know the evidence e, and $P(h|e)$ is the corresponding measure when you do, e surely confirms h when the latter exceeds the former. So we shall adopt the following as our definitions:

e **confirms or supports** h when $P(h|e) > P(h)$

e **disconfirms or undermines** h when $P(h|e) < P(h)$

e **is neutral with respect to** h when $P(h|e) = P(h)$

One might reasonably take $P(h|e) - P(h)$ as measuring the degree of e's support for h, though other measures have been suggested (e.g., Good, 1950); disagreements on this score will not need to be settled in this book. We shall refer, in the usual way, to $P(h)$ as 'the prior probability of h' and to $P(h|e)$ as h's 'posterior probability' relative to, or in the light of, e. The reasons for this terminology are obvious, but it ought to be noted that the terms have a meaning only in relation to evidence: as Lindley (1970, p. 38) put it, "[t]oday's posterior distribution is tomorrow's prior". It should be remembered too that all the probabilities are evaluated in relation to accepted background knowledge.

The application of Bayes's Theorem

Bayes's Theorem relates the posterior probability of a hypothesis, $P(h|e)$, to the terms $P(h)$, $P(e|h)$, and $P(e)$. Hence, knowing the values of these last three terms, it is possible to

determine whether e confirms h, and, more importantly, to calculate $P(h|e)$. In practice, of course, the various probabilities may only be known rather imprecisely; we shall have more to say about this practical aspect of the question later.

The dependence of the posterior probability on the three terms referred to above is reflected in three striking phenomena of scientific inference. First, other things being equal, the extent to which evidence e confirms a hypothesis h increases with the likelihood of h on e, that is to say, with $P(e|h)$. At one extreme, where e refutes h, $P(e|h) = 0$; hence, disconfirmation is at a maximum. The greatest confirmation is produced, for a given $P(e)$, when $P(e|h) = 1$, which will be met in practice when h logically entails e. Statistical hypotheses are more substantially confirmed the higher the value of $P(e|h)$.

Secondly, the posterior probability of a hypothesis depends on its prior probability, a dependence sometimes discernible in scientific attitudes to ad hoc hypotheses and in frequently expressed preferences for the simpler of two hypotheses. As we shall see, scientists always discriminate, in advance of any experimentation, between theories they regard as more-or-less credible (and, so, worthy of attention) and others.

Thirdly, the power of e to confirm h depends on $P(e)$, that is to say, on the probability of e when it is not assumed that h is true (which, of course, is not the same as assuming h to be false). This dependence is reflected in the scientific intuition that the more surprising the evidence, the greater its confirming power. However, $P(e) = P(e|h)P(h) + P(e|\sim h)P(\sim h)$, so that really, the posterior probability of h depends on the three basic quantities $P(h)$, $P(e|h)$, and $P(e|\sim h)$.

We shall deal in greater detail with each of these facets of inductive reasoning in the course of this chapter.

Falsifying hypotheses

A characteristic pattern of scientific inference is the refutation of a theory, when one of a theory's empirical consequences has been shown to be false in an experiment. This kind of reasoning, with its straightforward and unimpeachable logical structure, exercised such an influence on Popper that he made it the centrepiece of his scientific philosophy.

Although the Bayesian approach was not conceived specifically with this aspect of scientific reasoning in view, it has a ready explanation for it. The explanation relies on the fact that if, relative to background knowledge, a hypothesis h entails a consequence e, then (relative to the same background knowledge) $P(h|\sim e) = 0$. Interpreted in the Bayesian fashion, this means that h is maximally disconfirmed when it is refuted. Moreover, as we should expect, once a theory is refuted, no further evidence can ever confirm it, unless the refuting evidence or some portion of the background assumptions is revoked.

Checking a consequence

A standard method of investigating a deterministic hypothesis is to draw out some of its logical consequences, relative to a stock of background knowledge, and check whether they are true or not. For instance, the General Theory of Relativity was confirmed by establishing that light is deflected when it passes near the sun, as the theory predicts.

It is easy to show, by means of Bayes's Theorem, why and under what circumstances a theory is confirmed by its consequences.

If h entails e, then, as may be simply shown, $P(e|h) = 1$. Hence, from Bayes's Theorem: $P(e|h) = P(h)/P(e)$. Thus, if $0 < P(e) < 1$, and if $P(h) > 0$, then $P(h|e) > P(h)$. It follows that any evidence whose probability is neither of the extreme values must confirm every hypothesis with a non-zero probability of which it is a logical consequence.

Succeeding confirmations must eventually diminish in force, for the theory has an upper limit of probability beyond which no amount of evidence can push it. This too follows from Bayes's Theorem. Suppose e_1, e_2, ..., e_n, ... are consequences of h. Then Bayes's Theorem asserts that

$$p(h|e_1 \& e_2 \& \ldots \& e_n) = \frac{P(h)}{p(h|e_1 \& e_2 \& \ldots \& e_n)}$$

Now

$$P(e_1 \& e_2 \& \ldots \& e_n) = P(e_1)P(e_2 \& \ldots \& e_n|e_1)$$

and

$$P(e_2 \& \ldots \& e_n|e_1) = P(e_2|e_1)P(e_3 \& \ldots \& e_n|e_1 \& e_2).$$

Thus, in general,

$$P(e_1 \& e_2 \& \ldots \& e_n) = P(e_1)P(e_2|e_1) \ldots P(e_n|e_1 \& \ldots \& e_{n-1}).$$

Hence

$$P(h|e_1 \& e_2 \& \ldots \& e_n) = \frac{P(h)}{P(e_1)P(e_2|e_1) \ldots P(e_n|e_1 \& \ldots \& e_{n-1})}$$

Provided $P(h) > 0$, the term $P(e_n|e_1 \& \ldots \& e_{n-1})$ must tend to 1 as n increases. If it did not, the posterior probability of h would at some point exceed 1, which is impossible (Jeffreys, 1961, pp. 43–4). This explains why it is not sensible to test a hypothesis indefinitely, though without more detailed information on the individual's belief-structure, in particular regarding the values of $P(e_n|e_1 \& \ldots \& e_{n-1})$, one could not know the precise point beyond which further predictions of the hypothesis were sufficiently probable not to be worth examining.

Specific categories of a theory's consequences also have a restricted capacity to confirm (Urbach, 1981). Suppose h is the theory under discussion and that h_r is a substantial restriction of that theory. A substantial restriction of Newton's theory might, for example, express the idea that freely falling bodies near the earth descend with a constant acceleration or that the period and length of a pendulum are related by the familiar formula. Since h entails h_r, $P(h) \leq P(h_r)$, and if h_r is much less speculative than its progenitor, it will often be significantly more probable.

Now consider a series of predictions derived from h, but which also follow from h_r. If the predictions are verified, they may confirm both theories, whose posterior

probabilities are given by Bayes's Theorem, thus:

$$p(h|e_1 \& e_2 \& \ldots \& e_n) = \frac{P(h)}{p(e_1 \& e_2 \& \ldots \& e_n)}$$

and

$$p(h_r|e_1 \& e_2 \& \ldots \& e_n) = \frac{P(h_r)}{p(e_1 \& e_2 \& \ldots \& e_n)}.$$

Combining these two equations to eliminate the common denominator, one obtains

$$p(h|e_1 \& e_2 \& \ldots \& e_n) = \frac{P(h)}{P(h_r)} P(h_r|e_1 \& e_2 \& \ldots \& e_n).$$

Since the maximum value of the last probability term in this equation is 1, it follows that however many predictions of h_r have been verified, the main theory, h, can never acquire a posterior probability in excess of $P(h)/P(h_r)$. Hence, the type of evidence characterised by entailment from h_r may well be limited in its capacity to confirm h.

This result explains the familiar phenomenon that repetitions of a particular experiment often confirm a general theory only to a limited extent, for the predictions verified by means of a given kind of experiment (that is, an experiment designed to a specified pattern) do normally follow from and confirm a much-restricted version of the predicting theory. When an experiment's capacity to generate confirming evidence has been exhausted through repetition, further support for h would have to be sought from other experiments, experiments whose outcomes were predicted by different parts of h.

The arguments and explanations in this section rely on the possibility that evidence already accumulated from an experiment may increase the probability of further performances of the experiment producing similar results. Such a possibility is denied by Popperians on the grounds that the probabilities involved are subjective. How then do they explain the fact, attested by every scientist, that by repeating some experiment, one eventually (usually quickly) exhausts its capacity to confirm a given hypothesis? Alan Musgrave (1975) attempted an explanation designed on Popperian lines. He claimed that after a certain, unspecified number of repetitions of an experiment, the scientist would form a generalisation to the effect that whenever the experiment was performed, it would yield a similar result. Musgrave then proposed that the generalisation should be entered into 'background knowledge'. Relative to this newly augmented background knowledge, the experiment is certain to produce a similar result at its next performance. Musgrave then appealed to the principle that evidence confirms a hypothesis in proportion to the difference between its probability relative to the hypothesis together with background knowledge and its probability relative to background knowledge alone. That is, the degree to which e confirms h is proportional to $P(e|h \& b) - P(e|b)$, b being background knowledge. Musgrave then inferred that even if the experiment did produce the expected result when next performed, the hypothesis would receive no new confirmation. Watkins (1984, p. 297) has endorsed this account.

A number of decisive objections may be raised against it, though. First, as we shall show in the next section, although it seems to be a fact and is an essential constituent

of Bayesian reasoning, there is no basis in Popperian methodology for confirmation to depend on the probability of the evidence; Popper simply invoked the principle ad hoc. Secondly, Musgrave's suggestion takes no account of the fact that particular experimental results may be generalised in infinitely many ways. This is a substantial objection, since different generalisations give rise to different expectations about the outcomes of future experiments. Musgrave's account is incomplete without a rule to specify in each case the appropriate generalisation that should be formulated and adopted, and it is hard to imagine how such a rule could be justified within the confines of Popperian philosophy. Finally, the decision to designate the generalisation background knowledge, with the consequent effect on our evaluation of other theories and on our future conduct regarding, for example, whether to repeat certain experiments, is comprehensible only if we have invested some confidence in the theory. But then Musgrave's account tacitly calls on the same kind of inductive considerations as it was designed to circumvent, so its aim is defeated.

The probability of the evidence

The degree to which h is confirmed by e depends, according to Bayesian theory, on the extent to which $P(e|h)$ exceeds $P(e)$. An equivalent way of putting this is to say that confirmation is correlated with the difference between $P(e|h)$ and $P(e|\sim h)$, that is, with how much more probable the evidence is if the hypothesis is true than if it is false. This is obvious from another form of Bayes's Theorem:

$$\frac{P(h|e)}{P(h)} = \frac{1}{P(h) + \dfrac{P(e|\sim h)}{P(e|h)}P(\sim h)}.$$

These facts are reflected in the everyday experience that information that is particularly unexpected or surprising, unless some hypothesis is assumed to be true, supports that hypothesis with particular force. Thus, if a soothsayer predicts that you will meet a dark stranger sometime and you do, your faith in his powers of precognition would not be much enhanced: you would probably continue to think his predictions were just the result of guesswork. However, if the prediction also gave the correct number of hairs on the head of that stranger, your previous scepticism would no doubt be severely shaken.

Cox (1961, p. 92) illustrated this point with an incident in *Macbeth*. The three witches, using their special brand of divination, predicted to Macbeth that he would soon become both Thane of Cawdor and King of Scotland. Macbeth finds both these prognostications almost impossible to believe:

> By Sinel's death, I know I am Thane of Glamis,
> But how of Cawdor?
> The Thane of Cawdor lives, a prosperous gentleman,
> And to be King stands not within the prospect of belief.
> No more than to be Cawdor.

But a short time later he learns that the Thane of Cawdor prospered no longer, was in fact dead, and that he, Macbeth, has succeeded to the title. As a result, Macbeth's

attitude to the witches' powers is entirely altered, and he comes to believe in their other predictions and in their ability to foresee the future.

The following, more scientific, example was used by Jevons (1874, vol. 1, pp. 278–79) to illustrate the dependence of confirmation on the improbability of the evidence. The distinguished scientist Charles Babbage examined numerous logarithmic tables published over two centuries in various parts of the world. He was interested in whether they derived from the same source or had been worked out independently. Babbage (1827) found the same six errors in all but two and drew the "irresistible" conclusion that, apart from these two, all the tables originated in a common source.

Babbage's reasoning was interpreted by Jevons roughly as follows. The theory t_1, which says of some pair of logarithmic tables that they shared a common origin, is moderately likely in view of the immense amount of labour needed to compile such tables *ab initio*, and for a number of other reasons. The alternative, independence theory might take a variety of forms, each attributing different probabilities to the occurrence of errors in various positions in the table. The only one of these which seems at all likely would assign each place an equal probability of exhibiting an error and would, moreover, regard those errors as more-or-less independent. Call this theory t_2 and let e^i be the evidence of i common errors in the tables. The posterior probability of t_1 is inversely proportional to $P(e^i)$, which, under the assumption of only two rival hypotheses, can be expressed as $P(e^i) = P(e^i|t_1)P(t_1) + P(e^i|t_2)P(t_2)$ (This is the theorem of total probability.). Since t_1 entails e^i, $P(e^i) = P(t_1) + P(e^i|t_2)P(t_2)$. The quantity $P(e^i|t_2)$ clearly decreases with increasing i. Hence $P(e^i)$ diminishes and approaches $P(t_1)$, as i increases; and so e^i becomes increasingly powerful evidence for t_1, a result which agrees with scientific intuition.

In fact, scientists seem to regard a few shared mistakes in different mathematical tables as so strongly indicative of a common source that at least one compiler of such tables attempted to protect his copyright by deliberately incorporating three minor errors "as a trap for would-be plagiarists" (L. J. Comrie, quoted by Bowden, 1953, p. 4).

The relationship between how surprising a piece of evidence is on background assumptions and its power to confirm a hypothesis is a natural consequence of Bayesian theory and was not deliberately built in. On the other hand, methodologies that eschew probabilistic assessments of hypotheses seem constitutionally incapable of accounting for the phenomenon. Such approaches would need to be able, first, to discriminate between items of evidence on grounds other than their deductive or probabilistic relation to a hypothesis. And having established such a basis for discriminating, they must show a connection with confirmation. The objectivist school has more-or-less dodged this challenge. An exception is Popper. In tackling the problem, he moved partway towards Bayesianism; however, the concessions he made were insufficient. Thus Popper conceded that, in regard to confirmation, the significant quantities are $P(e|h)$ and $P(e)$, and as we have already reported, he even measured the amount of confirmation (or "corroboration", to use Popper's preferred term) which e confers on h by the difference between these quantities (Popper, 1959, appendix *ix).

But Popper never stated explicitly what he meant by the probability of evidence. On the one hand, he would never have allowed it to have a subjective connotation, for that would have compromised the supposed objectivity of science; on the other hand, he never worked out what objective significance the term could have. His writings suggest that he had in mind some purely logical notion of probability, but there is no adequate account of logical probability. Popper also never explained satisfactorily why

a hypothesis benefits from improbable evidence or, to put the objection another way, he failed to provide a foundation in non-Bayesian terms for the Bayesian confirmation function which he appropriated. (For a discussion and decisive criticism of Popper's account, see Grünbaum, 1976.)

The Bayesian position has recently been misunderstood to imply that if some evidence is known, then it cannot support any hypothesis, on the grounds that known evidence must have unit probability.

The Ravens Paradox

That evidence supports a hypothesis more the greater the ratio $P(e|h)/P(e)$ scotches a famous puzzle first posed by Hempel (1945) and known as the *Paradox of Confirmation* or sometimes as the *Ravens Paradox*. It was called a paradox because its premises were regarded as extremely plausible, despite their counter-intuitive, or in some versions contradictory, implications, and the reference to ravens stems from the paradigm hypothesis ('All ravens are black') which is frequently used to expound the problem. The difficulty arises from three assumptions about confirmation. They are as follows:

1 Hypotheses of the form 'All *R*s are *B*' are confirmed by the evidence of something that is both *R* and *B*. For example, 'All ravens are black' is confirmed by the observation of a black raven. (Hempel called this Nicod's condition, after the philosopher Jean Nicod.)
2 Logically equivalent hypotheses are confirmed by the same evidence. (This is the Equivalence condition.)
3 Evidence of some object not being *R* does not confirm 'All *R*s are *B*'.

We shall describe an object that is both black and a raven with the term *RB*. Similarly, a non-black, non-raven will be denoted $\overline{R}\,\overline{B}$. A contradiction arises for the following reasons: an *RB* confirms 'All *R*s are *B*', on account of the Nicod condition. According to the Equivalence condition, it also confirms 'All non-*B*s are non-*R*s', since the two hypotheses are logically equivalent. But contradicting this, the third condition implies that *RB* does not confirm 'All non-*B*s are non-*R*s'.

The contradiction may be avoided by revoking the third condition, as is sometimes done. (We shall note later another reason for not holding on to it.) However, although the remaining conditions are compatible, they have a consequence which many philosophers have regarded as blatantly false, namely, that by observing a non-black, non-raven (say, a red herring or a white shoe) one confirms the hypothesis that all ravens are black. (The argument is this: 'All non-*B*s are non-*R*' is equivalent to 'All *R*s are *B*'; according to the Nicod condition, the first is confirmed by $\overline{R}\,\overline{B}$; hence, by the Equivalence condition, so is the second.)

If non-black, non-ravens support the raven hypothesis, this seems to imply the paradoxical result that one could investigate that and other generalisations of a similar form just as well by observing white paper and red ink from the comfort of one's writing desk as by studying ravens on the wing. However, this would be a non sequitur. For the fact that *RB* and $\overline{R}\,\overline{B}$ both confirm a hypothesis does not imply that they do so with equal force. Once it is recognised that confirmation is a matter of degree, the conclusion is no longer so counter-intuitive, because it is compatible with $\overline{R}\,\overline{B}$ confirming 'All *R*s are *B*', but to a minuscule and negligible degree.

Indeed, most people do have a strong intuition that an *RB* confirms the ravens hypothesis (h) more than an $\overline{R}\,\overline{B}$. We can appreciate why that might be by consulting Bayes's Theorem as it applies to the two types of datum:

$$\frac{P(h|RB)}{P(h)} = \frac{P(RB|h)}{P(RB)} \;\&\; \frac{p(h|\overline{R}\,\overline{B})}{P(h)} = \frac{P(\overline{R}\,\overline{B}|h)}{P(\overline{R}\,\overline{B})}.$$

These expressions can be simplified. First, $P(RB|h) = P(B|h\&R)P(R|h) = P(R|h) = P(R)$. We arrived at the last equality by assuming that whether some arbitrary object is a raven is independent of the truth of h, which seems plausible to us, at any rate as a good approximation, though Horwich (1982, p. 59) thinks it has no plausibility. By similar reasoning, $P(\overline{R}\;\overline{B}\;|h) = P(\overline{B}|h) = P(\overline{B})$. Also $P(RB) = P(B|R)P(R)$, and $P(B|R) = \Sigma P(B|R\&\theta)P(\theta|R) =$ (assuming independence between θ and R) $\Sigma P(B|R\&\theta)P(\theta)$, where θ represents possible values of the percentage of ravens in the universe that are black (according to h, of course, $\theta = 1$). Finally, $P(B|R\&\theta) = \theta$, for if the percentage of black ravens in the universe is θ, the probability of an arbitrary raven being black is also θ. (This is intuitively correct and is formalised in the so-called Principal Principle.)

Combining all these considerations with the above forms of Bayes's Theorem yields

$$\frac{P(h|RB)}{P(h)} = \frac{1}{\Sigma\theta P(\theta)} \;\&\; \frac{P(h|\overline{R}\,\overline{B})}{P(h)} = \frac{1}{P(\overline{R}|\overline{B})}.$$

Consider first the term $P(\overline{R}|\overline{B})$. Presumably there are vastly more non-black things in the universe than ravens. So even if no ravens are black, the probability of some object about which we know nothing, except that it is not black, being a non-raven must be very high, indeed, practically 1. Hence, $P(h|\overline{R}\,\overline{B}) \approx P(h)$, and, so, the observation that some object is neither a raven nor black provides very little confirmation for h.

According to the equation above, the degree to which *RB* confirms h is inversely proportional to $\Sigma\theta P(\theta)$. This means, for example, that if it is initially very probable that all or virtually all ravens are black, then $\Sigma\theta P(\theta)$ would be large and *RB* would confirm h rather little. While if it is initially relatively probable that most ravens are not black, confirmation could be substantial. Intermediate levels of uncertainty about the proportion of ravens that are black would bring their own levels of confirmation. By contrast, because the class of non-black objects is so much larger than the class of ravens, $\overline{R}\,\overline{B}$ confirms 'All ravens are black' to only a tiny extent, irrespective of $P(\theta)$. Mackie's well-known Bayesian solution to the ravens paradox, is similar and also depends on an assumed large disparity in the number of non-black objects and ravens.

Our Bayesian working of the raven example appears to support the Nicod condition, with the minor limitation that no confirmation is possible, even with positive instances, when the hypothesis has a prior probability of 1. But a Bayesian approach anticipates the violation of Nicod's condition in other circumstances too. And numerous examples have been suggested as plausible instances of such violations. The first of these seems to be due to Good (1961). We shall use an example that is taken, with some modification, from Swinburne (1971). The hypothesis under examination is 'All grasshoppers are located outside the County of Yorkshire'. The observation of a grasshopper just beyond the county border is an instance of this generalisation and, according to Nicod, confirms it. But it might be more reasonably argued that since there are no border controls or

other obstacles restricting the movement of grasshoppers in that area, the observation of one on the edge of the county increases the probability that others have actually entered and hence undermines the hypothesis. In Bayesian terms, this is a case where, relative to background information, the probability of some datum is reduced by a hypothesis—that is, $P(e|h) < P(e)$—which is therefore disconfirmed—in other words, $P(h|e) < P(h)$.

A much more striking example where Nicod's conditions break down was invented by Rosenkrantz (1977, p. 35). Three people leave a party, each with a hat. The hypothesis that none of the three has his own hat is confirmed, according to Nicod, by the observation that person 1 has person 2's hat and by the observation that person 2 has person 1's hat. But since there are only three people, the second observation must *refute* the hypothesis, not confirm it.

Our grasshopper example provides an instance where a datum of the type $\overline{R}\,B$ confirms a generalisation of the form 'All Rs are B'. Imagine that an object which looked for all the world like a grasshopper had been found hopping about just outside Yorkshire and that it turned out to be some other sort of insect. The discovery that the object was not a grasshopper would be relatively unlikely unless the grasshopper hypothesis was true (hence, $P(e) < P(e|h)$); thus it would confirm that hypothesis. If the deceptively grasshopper-like object were within the county boundary, the same conclusion would follow, though the degree of confirmation would be greater. This shows that 'All Rs are B' may also be confirmed by a datum of the $\overline{R}\,\overline{B}$ type. Hence, the impression that non-Rs never confirm such hypotheses may be dispelled.

Horwich (1982) has argued that the raven hypothesis may be differently confirmed, depending on how the black raven was chosen, either by randomly selecting an object from the population of ravens or by making the selection from the population of black objects. (Horwich denotes the evidence that some object is a black raven as either R^*B or RB^*, depending on whether it was discovered by the first selection process or the second.) Prompted by a paper by Kevin Korb (1994), we agree with Horwich that this is so.

But Horwich offers another explanation, which fits poorly with his Bayesian one. For he claims that the datum R^*B is always more powerfully confirming than RB^*, because, he says, only it subjects the raven hypothesis to the risk of falsification. But this surely conflates the process of collecting evidence, which may indeed subject the hypothesis to different risks of refutation, with the evidence itself, which either refutes the hypothesis or does not refute it, and in the case of R^*B and RB^*, it does not.

Our conclusions are, first, that the supposedly paradoxical consequences of Nicod's condition and the Equivalence condition are not problematic, and, secondly, that there are separate reasons for rejecting Nicod's condition, which, moreover, conform to Bayesian principles.

The Duhem problem

The problem

The so-called Duhem (or Duhem-Quine) problem is a problem for theories of science of the type associated with Popper, which emphasise the power of certain evidence to refute a hypothesis. According to Popper's influential views, the characteristic of a theory which makes it 'scientific' is its falsifiability: "Statements or systems of statements,

in order to be ranked as scientific, must be capable of conflicting with possible, or conceivable, observations" (Popper, 1963, p. 39). And, claiming to apply this criterion, Popper (1963, Ch. 1) judged Einstein's gravitational theory to be scientific and Freud's psychology, unscientific. There is a strong flavour of commendation about the term *scientific* which has proved extremely misleading. For a theory that is scientific in Popper's sense is not necessarily true, or even probably true or so much as close to the truth, nor can it be said definitely that it is likely to lead to the truth. In fact, there seems to be no conceptual connection between a theory's capacity to pass Popper's test of scientificness and its having any epistemic or inductive value. There is little alternative, then, so far as we can see, to regarding Popper's demarcation between scientific and unscientific statements as part of a theory about the content and character of what is usually termed science, not as having any normative significance.

Yet as an attempt at understanding the methods of science, Popper's ideas bear little fruit. His central claim was that scientific theories are falsifiable by "possible, or conceivable, observations". This poses a difficulty, for an observation can only falsify a theory (that is, conclusively demonstrate its falsity) if it is itself conclusively certain. But observations cannot be conclusively certain. Popper himself recognised this but seems not to have appreciated its incongruity with his falsificationist thesis. He held every observation report to be fallible; but, reluctant to admit degrees of fallibility or anything of the kind, he concluded that observation reports that are admitted as evidence "are accepted as the result of a decision or agreement; and to that extent they are *conventions*" (Popper, 1959, p. 106; our italics). It is unclear to us to what psychological attitude this sort of acceptance corresponds, but whatever it is, Popper's view of evidence statements seems to pull the rug from under falsificationism: it implies that no theory can really be falsified by evidence. The nearest thing to a refutation would occur when 'conventionally accepted' evidence was inconsistent with a theory, which could then, at best, be described as 'conventionally' rejected. Indeed, Popper conceded this much: "From a logical point of view, the testing of a theory depends upon basic statements whose acceptance or rejection, in its turn, depends upon our *decisions*. Thus it is *decisions* which settle the fate of theories" (Popper, 1959, p. 108).

Watkins is one of those who saw that falsificationism presupposes the existence of some infallibly true observation statements, and he attempted to restore the Popperian position by advancing the claim that such statements do in fact exist. He would agree that statements like 'The hand on this dial is pointing to the numeral 6' are fallible—it is unlikely, but possible, that the person reporting it missaw the position of the hand. But he claimed that introspective perceptual reports, such as 'In my visual field there is now a silvery crescent against a dark blue background', "may rightly be regarded by their authors when they make them as infallibly true" (Watkins, 1984, pp. 79 and 248). But in our view Watkins is wrong, and the statements he regards as infallible are open to exactly the same sceptical doubts as any other observation report. We can illustrate this through Watkins's example: clearly, it is possible, though admittedly not very probable, that the introspector has misremembered and mistaken the shape he usually describes as a crescent or the sensation he usually receives on reporting blue and silvery images. These and other sources of error ensure that introspective reports are not exempt from the rule that non-analytic statements are fallible.

Of course, the kinds of observation statements we have mentioned, if asserted under appropriate circumstances, would never be seriously doubted. That is, although they could be false, they have a force and immediacy that carries conviction; they are

'morally certain', to use the traditional phrase. But if observation statements are merely indubitable, then whether a theory is regarded as refuted by observational data or not must rest ultimately on a subjective feeling of certainty. The fact that such convictions are so strong and uncontroversial may disguise their fallibility, but cannot undo it. Hence, no theory is strictly falsifiable, for none could be conclusively shown to be false by empirical observations. In practice the closest one could get to a refutation would be arriving at the conclusion that a theory that clashes with almost certainly true observations is almost certainly false.

A second objection to Popper's falsifiability criterion, and the one upon which we shall focus for its more general interest, is that it describes as unscientific most of those theories which are usually deemed science's greatest achievements. This is the chief aspect of the well-known criticisms advanced by Polanyi (1962), Kuhn (1970), and Lakatos (1970), amongst others. They have pointed out that, as had already been established by Duhem (1905), many notable theories of science are not falsifiable by what would generally be regarded as observation statements, even if those statements were infallibly true. Predictions drawn from Newton's laws or from the Kinetic Theory of Gases turn out to depend not only on those theories but also on certain auxiliary hypotheses. Hence, if such predictions fail, one is not compelled by logic to infer that the main theory is false, for the fault may lie with one or more of the auxiliary assumptions. The history of science has many occasions when an important theory led to a false prediction and where that theory, nevertheless, was not blamed for the failure. In such cases we find that one or more of the auxiliary assumptions used to derive the prediction was taken to be the culprit. The problem that arose from Duhem's investigations was which of the several distinct theories involved in deriving a false prediction should be regarded as the false element or elements in the assumptions.

The Duhem problem solved by Bayesian means

This problem may be resolved with the help of Bayes's Theorem, as Dorling (1979) has shown, by considering how the individual probabilities of several theories are altered when, as a group, they have been refuted.

Suppose a theory, t, and an auxiliary hypothesis, a, together imply an empirical consequence which is shown to be false by the observation of the outcome e. Let us assume that while the combination of $t \,\&\, a$ is refuted by e, the two components taken individually are not refuted. We wish to consider the separate effects wrought on the probabilities of t and a by the adverse evidence e. The comparisons of interest here are between $P(t|e)$ and $P(t)$, and between $P(a|e)$ and $P(a)$. The conditional probabilities can be expressed using Bayes's Theorem, as follows:

$$P(t|e) = \frac{P(e|t)P(t)}{P(e)} \quad P(a|e) = \frac{P(e|a)P(a)}{P(e)}.$$

In order to evaluate the posterior probabilities of t and of a, one must first determine the values of the various terms on the right-hand sides of these equations. Before attempting this, it is worth noting that the equations convey no expectation that the refutation of $t \,\&\, a$ jointly considered will in general have a symmetrical effect on the separate probabilities of t and of a, nor any reason why the degree of asymmetry may not be considerable in some cases. It is evident that the probability of t changes very little if

$P(e|t) \approx P(e)$, while that of a is reduced substantially just in case $P(e|a)$ is substantially less than $P(e)$. The equations also allow us to discern the factors that determine which hypothesis suffers most in the refutation.

A historical example might best illustrate how a theory that produces a false prediction may still remain very probable; we shall, in fact, use an example that Lakatos (1970, pp. 138–40, and 1968, pp. 174–75) drew heavily on. In 1815, William Prout, a medical practitioner and chemist, advanced the hypothesis that the atomic weights of all the elements are whole number multiples of the atomic weight of hydrogen, the underlying assumption being that all matter is built out of different combinations of some basic element. Prout believed hydrogen to be that fundamental building-block, though the idea was entertained by others that a more primitive element might exist out of which hydrogen itself was composed. Now the atomic weights recorded at the time, though approximately integral when expressed as multiples of the atomic weight of hydrogen, did not match Prout's hypothesis exactly. Those deviations from a perfect fit failed to convince Prout that his hypothesis was wrong however; he instead took the view that there were faults in the methods that had been used to measure the relative weights of atoms. The noted chemist Thomas Thomson drew a similar conclusion. Indeed, both he and Prout went so far as to adjust several reported atomic weights in order to bring them into line with Prout's hypothesis. For instance, instead of accepting 0.829 as the atomic weight (expressed as a proportion of the weight of an atom of oxygen) of the element boron, which was the experimentally reported value, Thomson (1818, p. 340) preferred 0.875 "because it is a multiple of 0.125, which all the atoms seem to be". (Thomson erroneously took 0.125 as the atomic weight of hydrogen, relative to that of oxygen.) Similarly, Prout adjusted the measured atomic weight of chlorine (relative to hydrogen) from 35.83 to 36, the nearest whole number.

Thomson's and Prout's reasoning can be explained as follows: Prout's hypothesis t, together with an appropriate assumption a asserting the accuracy (within specified limits) of the measuring technique, the purity of the chemicals employed, and so forth, implies that the measured atomic weight of chlorine (relative to hydrogen) is a whole number. Suppose, as was the case in 1815, that chlorine's measured atomic weight was 35.83, and call this the evidence e. It seems that chemists of the early nineteenth century, such as Prout and Thomson, were fairly certain about the truth of t, but less so of a, though more sure that a is true than that it is false. Contemporary near-certainty about the truth of Prout's hypothesis is witnessed by the chemist J. S. Stas. He reported (1860, p. 42) that "In England the hypothesis of Dr Prout was almost universally accepted as absolute truth", and he confessed that when he started researching into the matter, he himself had "had an almost absolute confidence in the exactness of Prout's principle" (1860, p. 44). (Stas's confidence eventually faded after many years' experimental study, and by 1860 he had "reached the complete conviction, the entire certainty, as far as certainty can be attained on such a subject, that Prout's law ... is nothing but an illusion", 1860, p. 45.) It is less easy to ascertain how confident Prout and his contemporaries were in the methods by which atomic weights were measured, but it is unlikely that this confidence was very great, in view of the many clear sources of error and the failure of independent measurements generally to produce identical results. On the other hand, chemists of the time must have felt that their methods for determining atomic weights were more likely to be accurate than not, otherwise they would not have used them. For these reasons, we conjecture that $P(a)$ was of the order of 0.6 and that $P(t)$ was around 0.9, and these are

the figures we shall work with. It should be stressed that these numbers and those we shall assign to other probabilities are intended chiefly to illustrate how Bayes's Theorem resolves Duhem's problem; nevertheless, we believe them to be sufficiently accurate to throw light on the progress of Prout's hypothesis. As will become apparent, the results we obtain are not very sensitive to variations in the assumed prior probabilities.

In order to evaluate the posterior probabilities of t and of a, one must fix the values of the terms $P(e|t)$, $P(e|a)$, and $P(e)$. These can be expressed, using the theorem on total probability, as follows:

$$\begin{aligned} P(e) &= P(e|t)P(t) + P(e|\sim t)P(\sim t) \\ P(e|t) &= P(e\,\&\,a|t) + P(e\,\&\sim a|t) \\ &= P(e|t\,\&\,a)P(a|t) + P(e|t\,\&\sim a)P(\sim a|t) \\ &= P(e|t\,\&\,a)P(a) + P(e|t\,\&\sim a)P(\sim a) \end{aligned}$$

Since $t\,\&\,a$, in combination, is refuted by e, the term $P(e|t\,\&\,a)$ is zero. Hence:

$$P(e|t) = P(e|t\,\&\sim a)P(\sim a).$$

It should be noted that in deriving the last equation but one, we have followed Dorling in assuming that t and a are independent, that is, that $P(a|t) = P(a)$ and, hence, $P(\sim a|t) = P(\sim a)$. This seems to accord with many historical cases and is clearly right in the present case. By parallel reasoning to that employed above, we may derive the results:

$$\begin{aligned} P(e|a) &= P(e|\sim t\,\&\,a)P(\sim t) \\ P(e|\sim t) &= P(e|\sim t\,\&\,a)P(a) + P(e|\sim t\,\&\sim a)P(\sim a) \end{aligned}$$

Provided the following terms are fixed, which we have done in a tentative way, to be justified presently, the posterior probabilities of t and of a can be determined:

$$\begin{aligned} P(e|\sim t\,\&\,a) &= 0.01 \\ P(e|\sim t\,\&\sim a) &= 0.01 \\ P(e|t\,\&\sim a) &= 0.02 \end{aligned}$$

The first of these gives the probability of the evidence if Prout's hypothesis is not true but if the method of atomic weight measurement is accurate. Such probabilities were explicitly considered by some nineteenth-century chemists, and they typically took a theory of random assignment of atomic weights as the alternative to Prout's hypothesis (e.g., Mallet, 1880); we shall follow this. Suppose it had been established for certain that the atomic weight of chlorine lay between 35 and 36. (The final results we obtain respecting the posterior probabilities of t and a are, incidentally, not affected by the width of this interval.) The random-allocation theory would assign equal probabilities to the atomic weight of an element lying in any 0.01-wide band. Hence, on the assumption that a is true, but t false, the probability that the atomic weight of chlorine lies in the interval 35.825 to 35.835 is 0.01. We have assigned the same value to $P(e|\sim t\,\&\sim a)$

on the grounds that if a were false because, say, some of the chemicals were impure or the measuring techniques faulty, then, still assuming t to be false, one would not expect atomic weights to be biased towards any particular part of the interval between adjacent integers.

We have set the probability $P(e|t\,\&\sim a)$ rather higher, at 0.02. The reason for this is that although some impurities in the chemicals and some degree of inaccuracy in the method of measurement were moderately likely in the early nineteenth century, chemists would not have considered their techniques entirely haphazard. Thus if Prout's hypothesis were true, but the measuring technique imperfect, the measured atomic weights would have been likely to deviate somewhat from integral values; but the greater the deviation, the less likely, on these assumptions, so the probability of an atomic weight lying in any part of the 35–36 interval would not be distributed uniformly over the interval, but would be more concentrated around the whole numbers.

Let us proceed with the figures we have assumed for the crucial probabilities, noting however that the particular values of the three probability terms are unimportant, only their relative values need be taken into account in the calculation. Thus we would arrive at the same posterior probabilities for a and t with the weaker assumptions that $P(e|\sim t\,\&\,a) = P(e|\sim t\,\&\sim a) = 1/2P(e|t\,\&\sim a)$. We thus obtain

$$\begin{aligned} P(e|\sim t) &= 0.01 \times 0.6 + 0.01 \times 0.4 = 0.01 \\ P(e|t) &= 0.02 \times 0.4 = 0.008 \\ P(e|a) &= 0.01 \times 0.1 = 0.001 \\ P(e) &= 0.008 \times 0.9 + 0.01 \times 0.1 = 0.0082 \end{aligned}$$

Finally, Bayes's Theorem enables us to derive the posterior probabilities in which we were interested:

$$\begin{aligned} P(t|e) &= 0.878 \text{ (Recall that } P(t) = 0.9.) \\ P(a|e) &= 0.073 \text{ (Recall that } P(a) = 0.6.) \end{aligned}$$

These striking results show that evidence of the kind we have-described may exert a sharply asymmetric effect on the probabilities of t and of a. The initial probabilities we assumed seem appropriate for chemists such as Prout and Thomson, and if they are correct, the results deduced from Bayes's Theorem explain why those chemists regarded Prout's hypothesis as being more-or-less undisturbed when certain atomic-weight measurements, diverged from integral values, and why they felt entitled to adjust those measurements to the nearest whole number. Fortunately, these results are relatively insensitive to changes in our assumptions, so the accuracy of those assumptions is not a vital matter as far as our explanation is concerned. For example, if one took the initial probability of Prout's hypothesis (t) to be 0.7, instead of 0.9, keeping the other assignments, we find that $P(e|t) = 0.65$, while $P(a|e) = 0.21$. Hence, as before, after the refutation, Prout's hypothesis is still more likely to be true than false, and the auxiliary assumptions are still much more likely to be false than true. Other substantial variations in the initial probabilities produce similar results, though with so many factors at work, it is difficult to state concisely the conditions upon which these results depend without just pointing to the equations above.

Thus Bayes's Theorem provides a model to account for the kind of scientific reasoning that gave rise to the Duhem problem. And the example of Prout's hypothesis, as well as

others that Dorling (1979 and 1982) has described, show, in our view, that the Bayesian model is essentially correct. By contrast, non-probabilistic theories seem to lack entirely the resources that could deal with Duhem's problem.

A fact that emerges when slightly different values are assumed for the various probabilities in the Prout's hypothesis example is that one or other of the theories may actually become more probable after the conjunction $t \,\&\, a$ has been refuted. For instance, when $P(e|t \,\&\, \sim a)$ equals 0.05, the other probabilities being assigned the same values as before, the posterior probability of t is 0.91, which exceeds its prior probability. This may seem bizarre, but, as Dorling (1982) has argued, it is not so odd when one bears in mind that the refuting evidence normally contains a good deal more information than is required merely to disprove $t \,\&\, a$ and that this extra information may be confirmatory. In general, such confirmation occurs when $P(e) < P(e|t)$, which is easily shown to be equivalent to the condition $P(e|t) > P(e|\sim t)$. In other words, when evidence is easier to explain (in the sense that it receives a higher probability) if a given hypothesis is true than if it is not, then that theory is confirmed by the evidence.

Bad data, and data too good to be true

Bad data. An interesting fact that emerges from the Bayesian analysis is that a successful prediction derived from a combination of two theories, say t and a, does not always redound to the credit of t, even if the prior probability of the evidence is small; indeed, it can even undermine it. We may illustrate this by referring again to the example of Prout's hypothesis.

Suppose the atomic weight of chlorine were 'measured', not in the old-fashioned chemical way, but by concentrating hard on the element in question and picking a number in some random fashion from a given range of numbers. And let us assume that this method assigns a whole-number value to the atomic weight of chlorine. This is just what one would predict on the basis of Prout's hypothesis, if the outlandish measuring technique were reliable. But reliability is obviously most unlikely, and it is equally obvious that, as a result, the measured atomic weight of chlorine adds practically nothing to the probability of Prout's hypothesis, notwithstanding its integral value. This intuition is upheld by Bayes's Theorem, as a simple calculation based on the above formulas shows. (As before, let t be Prout's hypothesis and a the assumption that the measuring technique is accurate. Then set $P(e|t \,\&\, \sim a) = P(e|\sim t \,\&\, \sim a) = P(e|\sim t \,\&\, a) = 0.01$, for reasons similar to those stated earlier, and let $P(a)$ be very small, say 0.0001, for obvious reasons. It then follows that $P(t)$ and $P(t|e)$ are equal to two decimal places.)

This example shows that Leibniz was wrong to declare as a general principle that "It is the greatest commendation of an hypothesis (next to truth) if by its help predictions can be made even about phenomena or experiments not tried". Leibniz and Lakatos, who quoted these words with approval (1970, p. 123), seem to have overlooked the fact that if a prediction can be deduced from a hypothesis only with the assistance of highly questionable auxiliary claims, then that hypothesis may accrue very little credit. This explains why the various sensational predictions which Velikovsky drew from his theory of planetary collisions failed to impress most serious astronomers, even when some of those predictions were to their amazement fulfilled. For instance, Velikovsky's prediction of the existence of large quantities of petroleum on the planet Venus relied not only on his pet theory that various natural disasters in the past had been caused by

collisions between the earth and a comet, but also on a string of unsupported and not very plausible assumptions, such as that the comet in question originally carried hydrogen and carbon, that these had been converted to petroleum by electrical discharges supposedly created in the violent impact with the earth, that the comet had later evolved into the planet Venus, and some others (Velikovsky, 1950, p. 351). (More details of Velikovsky's theory are given in the next section.)

Data too good to be true. Data are sometimes said to be 'too good to be true' when they fit a favoured hypothesis more perfectly than seems reasonable to expect. For instance, suppose all the atomic weights listed in Prout's paper had been whole numbers, exactly. Such a result almost looks as if it was designed to impress, and it is just for this reason that it fails to.

We may analyse this response as follows. Let e be the evidence of, say, 20 atomic-weight measurements, each a perfect whole number. No one could have regarded precise atomic weights measured at the time as absolutely reliable. The most natural view would have been that such measurements are subject to experimental error and, hence, that they would give a certain spread of results about the true value. On this assumption, which we shall label a', it is extremely unlikely that numerous independent atomic-weight measurements would all produce whole numbers, even if Prout's hypothesis were true. So $P(e|t \,\&\, a')$ is extremely small and, clearly, $P(e|\sim t \,\&\, a')$ would be no larger. Now a' has many possible alternatives, one of the more plausible (though initially it might not be very plausible) being that the experiments were consciously or unconsciously rigged in favour of Prout's hypothesis. If this were the only significant alternative (and so, in effect, equivalent to $\sim a'$), $P(e|t \,\&\, \sim a')$ would be very high, as would $P(e|\sim t \,\&\, \sim a')$. It follows from the equations on page 234 above that

$$P(e|t) \approx P(e|t \,\&\, \sim a')P(\sim a') \text{ and}$$
$$P(e|\sim t) \approx P(e|\sim t \,\&\, \sim a')P(\sim a'),$$

and, hence,

$$P(e) \approx P(e|t \,\&\, \sim a')P(\sim a')P(t) + P(e|\sim t \,\&\, \sim a')P(\sim a')P(\sim t).$$

Now, presumably the rigging of the results to produce whole numbers, if it took place, would produce whole numbers equally effectively whether t was true or not; in other words,

$$P(e|t \,\&\, \sim a') = P(e|\sim t \,\&\, \sim a');$$

hence

$$P(e) \approx P(e|t \,\&\, \sim a')P(\sim a').$$

Therefore,

$$P(t|e) = \frac{P(e|t)P(t)}{P(e)} \approx \frac{P(e|t \,\&\, \sim a')P(\sim a')P(t)}{P(e|t \,\&\, \sim a')P(\sim a')} = P(t)$$

Thus e does not confirm t significantly, even though, in a misleading sense, it fits the theory perfectly. This is why it is said to be too good to be true. A similar calculation shows that the probability of a' is diminished and, on the assumptions that we made, this implies that the probability of the experiments having been fabricated is enhanced. (The above analysis is essentially the same as given in Dorling, 1982).

A famous case of data that were allegedly too good to be true is that of Mendel's plant-breeding results. Mendel's genetic theory of inheritance allows one to calculate the probabilities with which certain plants would produce specific kinds of offspring. For instance, under certain circumstances, pea plants of a particular strain may be calculated to yield round and wrinkled seeds with probabilities 0.75 and 0.25, respectively. Mendel obtained seed-frequencies that matched the corresponding probabilities in this and in similar cases remarkably well, suggesting (misleadingly, Fisher contended) substantial support for the genetic theory. Fisher did not believe that Mendel had deliberately falsified his results to appear in better accord with his theory than they really were. To do so, Fisher claimed, would "contravene the weight of the evidence supplied in detail by ... [Mendel's] paper as a whole". But Fisher thought it a "possibility among others that Mendel was deceived by some assistant who knew too well what was expected" (1936, p. 132), an explanation he backed up with some (rather meagre) evidence. Dobzhansky (1967, p. 1589), on the other hand, thought it "at least as plausible" that Mendel had himself discarded results that deviated much from his ideal, in the sincere belief that they were contaminated or that some other accident had befallen them. (For a comprehensive review see Edwards, 1986.)

The argument put forward earlier to show that too-exactly whole-number atomic-weight measurements would not have supported Prout's hypothesis depends on the existence of some sufficiently plausible alternative hypothesis that would explain the data better. We believe that, in general, data are too good to be true relative to one hypothesis only if there are such alternatives. This principle accords with intuition; for if the technique for eliciting atomic weights had long been established as precise and accurate, and if careful precautions had been taken against experimenter bias and deception, all the natural alternatives to Prout's hypothesis could be discounted and the data would no longer seem suspiciously good; they would be straightforwardly good. Fisher, however, did not subscribe to the principle, at least, not explicitly; he believed that Mendel's results told against the genetic theory whatever alternative explanations might suggest themselves. Nevertheless, as just indicated, the consideration of such alternatives played a part in his argument.

Ad hoc hypotheses

As we have seen, an important scientific theory which, in combination with other hypotheses, has made a false prediction may nevertheless emerge relatively unscathed, while one or more of the auxiliary hypotheses are largely discredited. (We are using such expressions in the normal way to describe how hypotheses are received, regarding them as harmless metaphors for obvious and more-or-less precise probabilistic notions. Thus, a hypothesis that is unscathed by negative evidence is one whose posterior and prior probabilities are similar. On the other hand, it is difficult to see what opponents of the Bayesian approach could have in mind when they talk of theories being 'accepted' or 'retained', or 'put forward' or 'saved' or 'vindicated'.) When a set of auxiliary assumptions is discredited in a test, scientists frequently think up new assumptions

which assist the main theory to explain the previously anomalous data. Sometimes these new assumptions give the impression that their role is simply to 'patch up' the theory, and in such cases Francis Bacon called them "frivolous distinctions" (1620, Book I, aphorism xxv). More recently they have been tagged 'ad hoc hypotheses', presumably because they would not have been introduced if the need to bring theory and evidence into line had not arisen.

But although particular ad hoc theories are fairly easy to evaluate intuitively, there is controversy over what general criteria apply. We shall see that the Bayesian approach clarifies the question. First let us consider a few examples of ad hoc theories.

Some examples of ad hoc hypotheses

Velikovsky's theory of collective amnesia. Immanuel Velikovsky, in a daring book called *Worlds in Collision* that attracted a great deal of attention some years ago, put forward the theory that the earth has been subject, at various stages in its history, to cosmic disasters produced by near collisions with massive comets. One of these comets, which went on to make a distinguished career as the planet Venus, is supposed to have passed close by the earth during the Israelites' captivity in Egypt and to have caused many of the various remarkable events of the time, such as the ten plagues and the parting of the Red Sea, related in the Bible. One of the theory's predictions, apparently, is that since no group of people could have missed these tremendous goings-on, if they kept records at all, they would have recorded them. However, many communities failed to note in their writings anything out of the ordinary at that time. But Velikovsky, still convinced by his main theory, put this exceptional behaviour down to what he called a "collective amnesia". He argued that the cataclysms were so terrifying that whole peoples behaved "as if [they had] obliterated impressions that should be unforgettable". There was a need, Velikovsky said, to "uncover the vestiges" of these events, "a task not unlike that of overcoming amnesia in a single person" (1950, p. 288).

Individual amnesia is the issue in the next example.

Dianetics. Dianetics is a theory that purports to analyse the causes of insanity and mental stress, which it sees as the 'misfiling' of information in unsuitable locations in the brain. By refiling these 'engrams', it claims, sanity may be restored, composure enhanced, and, incidentally, the memory vastly improved. Not surprisingly, the therapy is long and expensive, and few people have been through it and borne out the theory's claims. However, one triumphant success, a young student, was announced by the inventor of Dianetics, L. Ron Hubbard, and in 1950 he exhibited this person to a large audience, claiming that she had a "full and perfect recall of every moment of her life". But questions from the floor ("What did you have for breakfast on October 3, 1942?"; "What colour is Mr. Hubbard's tie?", and the like) soon demonstrated that the hapless young woman had a most imperfect memory. Hubbard accounted for this to what remained of the assembly by saying that when the woman first appeared on the stage and was asked to come forward "now", the word "now" had frozen her in "present time" and paralysed her ability to recall the past. (An account of the incident and of the history of Dianetics is given by Miller, 1987.)

An example from psychology. Investigations into the IQs of different groups of people show that average levels of measured intelligence vary. Some environmentalists, so-called, attribute low scores primarily to poor social and educational conditions, an

explanation that ran into trouble when it was discovered that a large group of Eskimos, leading a feckless, poor, and drunken existence, scored very highly on IQ tests. The distinguished biologist Peter Medawar (1974), in an effort to deflect the difficulty away from the environmentalist thesis, tried to explain this unexpected observation by saying that an "upbringing in an igloo gives just the right degree of coziness, security and mutual contact to conduce to a good performance in intelligence tests."

In each of these examples, the theory which was proposed in place of the refuted one seems rather unsatisfactory. It is not likely that they would have been put forward except in response to particular empirical anomalies, hence the label "ad hoc", which suggests that the theory was advanced for the specific purpose of evading a difficulty. However, some theories of this kind cannot be condemned so readily. For instance, an ad hoc alteration which rescued Newtonian theory from a difficulty led directly to the discovery of a new planet and was generally deemed a shining success.

The discovery of the planet Neptune. The planet Uranus was discovered by Sir William Herschel in 1781. Astronomers quickly sought to describe the orbit of the new planet, using Newtonian theory and taking account of the perturbing influence of other known planets, so that predictions could be made concerning its future positions. But discrepancies between predicted and observed positions of Uranus substantially exceeded the admitted limits of experimental error and grew year by year. The possibility that the fault lay with Newton's laws was mooted by a few astronomers, but the prevailing opinion was that there must be some unknown planet providing an extra source of gravitational attraction on Uranus, which ought to be included in the Newtonian calculations. Two astronomers in particular, John Couch Adams and U. J. J. Le Verrier, working independently, were convinced of this, and using all the known sightings of Uranus, they estimated where the hypothetical planet should be. This was a remarkable mathematical achievement, but more importantly, careful telescopic observations and studies of old astronomical charts revealed in 1846 the presence of a planet with the anticipated characteristics. The planet was later called Neptune. Newton's theory was saved, for the time being. (The fascinating story of this episode is told by W. M. Smart, 1947.)

A standard account of adhocness

The common features of the examples we are considering are that a theory t, which we can call the main theory, was combined with an auxiliary hypothesis a, to predict e, when in fact e' occurred, e' being incompatible with e. And in order to retain the main theory in its desired explanatory role, a new auxiliary, a', was proposed which, with t, implies e'. The new theories are ad hoc in the sense that they were advanced "for the sole purpose of saving a hypothesis seriously threatened by adverse evidence" (Hempel, 1966, p. 29). However, many philosophers have distinguished two kinds of ad hoc theory. Theory $t \,\&\, a'$ is of the first kind if it possesses no independent test implications—independent, that is, from the evidence that refuted its predecessor $t \,\&\, a$. It is ad hoc in the second sense if it does have such test implications but none has been verified. Lakatos (1970, p. 175) called the first kind of theory ad hoc_1, the second kind ad hoc_2. Often, the designation *ad hoc* is applied just to the new theory, a', rather than to its conjunction with t.

The term *ad hoc* for hypotheses that do not meet one or other of these conditions seems not to be an old one; its earliest occurrence in English that we know of was

in 1936, in a critical review of a book of psychology. The reviewer, W. J. H. Sprott, commented on some explanations offered in the book of certain aspects of childish behaviour:

> There is a suspicion of '*ad-hoc*-ness' about the 'explanations'. The whole point is that such an account cannot be satisfactory until we can predict the child's movements from a knowledge of the tensions, vectors and valences which are operative, *independent of our knowledge of how the child actually behaved.* So far we seem reduced to inventing valences, vectors and tensions from a knowledge of the child's behaviour.
>
> (Sprott, 1936, p. 249; our emphasis)

Sprott clearly regarded ad hoc theories as unsatisfactory, a view which many philosophers nowadays share. For example, Popper states it as one of his 'requirements' that a theory should not be ad hoc_1:

> We require that the new theory should be *independently testable.* That is to say, apart from explaining all the *explicanda* which the new theory was designed to explain, it must have new and testable consequences (preferably consequences of a *new kind*).
>
> (Popper, 1963, p. 241)

A further requirement laid down by Popper is that the new theory "should pass the independent tests in question", that is, they should not be ad hoc_2. Lakatos (1970) agreed with Popper that a theory is unacceptable if it is ad hoc in either sense; others such as Hempel (1966, p. 29) emphasise only the first sense. Disapproval of ad hoc theories is not new; in the early seventeenth century, Bacon criticized as a "frivolous distinction" the type of hypothesis that is "framed to the measure of those particulars only from which it is derived" (i.e., ad hoc_1 hypotheses). Bacon argued that a hypothesis ought to be "larger and wider" than the observations that gave rise to it and, moreover, that "that largeness and wideness" should be confirmed "by leading us to new particulars" (i.e., the theory should not be ad hoc_2).

The theories advocated by Velikovsky, Medawar, and Hubbard in response to anomalous data are probably ad hoc_1, since they seem to make no independent predictions, though, of course, a closer study of those theories might reverse that judgment. According to the criteria we have discussed, the theories appear therefore to represent unsatisfactory scientific developments, which is intuitively right. The Adams–Le Verrier hypothesis, on the other hand, is not ad hoc in either sense, because it did make new predictions, some of which were verified by telescopic sightings of Neptune. Again, philosophical and intuitive judgment coincides.

Despite this seeming success, we believe the adhocness criterion to be misconceived and unfounded. In setting out our position, we shall show why the criterion must be wrong, and then we shall present the Bayesian view on ad hoc theories.

Why the standard account must be wrong

According to the standard account, all ad hoc hypotheses are unsatisfactory, though ad hoc_1, not surprisingly, is often regarded as worse than ad hoc_2. As we explained,

we do not think any of the attempts to justify the adhocness criterion a priori have been successful. And we shall argue that this is to be expected, since there are positive reasons to reject the criterion, which we shall now set out. Our argument will appeal to counter-examples and to some more general considerations. An attraction of the adhocness criterion, no doubt, is its apparent objectivity and its avoidance of subjective probability, but, as we shall show, the non-Bayesian account has its own subjective aspect, one which, in our view, is very inappropriate.

Consider first some counter-examples to the standard account. Suppose one were examining the hypothesis that a particular urn contains only white counters. Next, imagine that a counter is withdrawn from the urn at random, that after its colour has been noted, it is replaced, and that this operation is repeated 10,000 times. If 4950, say, of the selected counters were red and the rest white, the initial hypothesis and the various necessary auxiliary assumptions, taken together, would be refuted; and it is then natural to conclude that, contrary to the original assumption, the urn contains both red and white counters in approximately equal numbers. This seems a perfectly reasonable inference, the revised hypothesis appears well justified by the evidence, yet there is no independent evidence for it. And if we complicate the example by letting the urn vapourise just after the last counter has been inspected, there will be no possibility of such independent evidence. So the hypothesis about the (late) urn's contents is ad $\text{hoc}_{1\&2}$; but for all that, it seems plausible and satisfactory (Howson, 1984; Urbach, 1991).

Speculating on the contents of an urn is but a humble form of enquiry, which we cite for the simple way it illustrates that a theory can be acceptable even when we have no evidence independent of the observations which caused the theory to be proposed, nor any possibility of such evidence. Hence, the two adhocness criteria are misguided. Examples from the higher sciences confirm this. Take the following case from the science of genetics: suppose it was initially assumed or believed that two characteristics of a certain plant are inherited in accordance with Mendel's principles through the agency of a pair of independently acting genes located on different chromosomes. Imagine now that plant-breeding experiments throw up a surprising number of plants carrying both characteristics, so that the original assumption that the genes act independently is revised in favour of a theory that they are linked on the same chromosome. Again, the revised theory would be strongly confirmed and established as acceptable merely on the evidence that stimulated its formulation and without the necessity of further, independent evidence. (An example of this sort is worked out by Fisher, 1970, Ch. IX.)

The discovery of the planet Neptune illustrates the same point. Adams arrived at what he regarded as the most likely mass and elements of the orbit of the hypothetical planet by the mathematical technique of least squares applied to all the observations that had hitherto been collected on the positions of Uranus. Adams's hypothesis fitted these observations so well that *even before Neptune had been seen through the telescope or detected on astronomical charts*, its existence was contemplated with the greatest confidence by the leading astronomers of the day. For instance, in his retirement address as president of the British Association, Sir John Herschel, after remarking that the previous year had seen the discovery of a new minor planet, went on: "It has done more. It has given us the probable prospect of the discovery of another. We see it as Columbus saw America from the shores of Spain. Its movements have been felt, trembling along the far-reaching line of our analysis, *with a certainty hardly inferior to that of ocular demonstration*" (quoted in Smart, 1947, p. 61; our italics). And the Astronomer Royal,

Sir George Airy, who was initially inclined to believe that the problem with *Uranus* would be resolved by introducing a slight adjustment to the inverse-square law, spoke of "*the extreme probability* of now discovering a new planet in a very short time" (also quoted in Smart, 1947, p. 61; our italics). Neptune was discovered a very short time later.

We turn now to a more general objection to the idea that hypotheses are acceptable only if corroborated by independent evidence. Imagine a scientist who is interested in the conjunction of hypotheses $t \,\&\, a$, whose implication e can be checked in an experiment. The experiment is performed with the result e', incompatible with e, and the scientist advances a new theory, $t \,\&\, a'$, which is consistent with the observations but is ad hoc in one or other of the two senses, that is, there is either no fresh evidence for a' or no possibility of such evidence. The new theory therefore is unacceptable according to the view we are considering.

Suppose, next, that another scientist, working without knowledge of his colleague, also wished to test $t \,\&\, a$ but that he chose a different experiment for this purpose, one with only two possible outcomes: either e or $\sim e$. Of course, he would obtain the latter, and having done so, he would be obliged to revise the refuted theory, to $t \,\&\, a'$, say. This scientist now notices that e' follows from the new theory, and he performs the orthodox experiment to verify e'. The new theory can then count a successful prediction to its credit, and so is not ad hoc. Hence, according to the standard view, it is perfectly acceptable.

This is strange, to say the least, because we have arrived at opposite evaluations of the very same theory, breaching at the same time what we previously called the Equivalence condition and showing that the standard adhocness criterion is inconsistent. Whatever measures might be taken to resolve the inconsistency, it seems to us that one element of the criterion ought to be removed, namely, the significance it attaches to whether the theory concerned was thought up before or after the evidence was known. This introduces into the principles of theory-evaluation considerations concerning the state of the experimenters' minds, which are intuitively irrelevant and incongruous in a methodology with pretentions to objectivity. No such considerations enter the corresponding Bayesian evaluations.

The Bayesian view of ad hoc theories

We have argued, contrary to the standard view, that a theory could be scientific and plausible even if it is ad hoc. An acceptable ad hoc theory is a possibility allowed for by the Bayesian principle that theories should be evaluated according to their probabilities. To illustrate, consider the ad hoc theory a', which we have supposed was put forward in response to some refuting evidence e'. The probability of this theory must be reckoned relative to e' and any other available relevant information, b. The probability calculus places no restrictions on the value of $P(a'|e' \,\&\, b)$; it might, for example, be below 0.5, so that a' would be more likely false than true, or greater than 0.5, when the reverse would be the case. Hence a' does not need the support of new independent predictions in order to be quite plausible and acceptable (Horwich, 1982, pp. 105–8).

Scientists are also interested in whether t in the presence of the newly thought-up a' provides a competent explanation of the previously anomalous e'. It would only do so if $t \,\&\, a'$ was sufficiently credible; since $P(t \,\&\, a'|e' \,\&\, b) \leq P(a'|e' \,\&\, b)$, this would be the case only if a' was itself acceptable, in the sense indicated.

The Bayesian approach, incidentally, explains why people often respond immediately with incredulity, even derision, on first hearing certain ad hoc hypotheses. It is hardly likely that their amusement stems from perceiving, or even thinking that they perceive, that the hypothesis leads to no new predictions. Surely it is more likely that they are reacting to what they see as the utter implausibility of the hypothesis.

The notion of independent evidence

As we have explained, a standard non-Bayesian account of adhocness asserts that a theory consisting of the combination $t \& a$ is only replaced with $t \& a'$ in an acceptable, scientific fashion when a' is successfully tested by evidence independent of that which refuted the first theory. This thesis is often associated with another, rather similar view, namely, that no theory is acceptable unless it is supported by evidence independent of that which prompted its initial proposal, whether that evidence also refuted a predecessor or not. We have shown that these views are neither reasonable nor compatible with scientific practice and, moreover, that they fail to live up to the standards of objectivity to which they aspire. (Howson, 1984, addresses a number of other objections.) One problem with the non-Bayesian criterion of adhocness, which we have not needed to exploit in our criticisms, is that the notion of 'independence' with regard to evidence is left vague and intuitive. Moreover, it seems difficult to give it a satisfactory meaning, except in the context of Bayesian induction.

There is an established notion of probabilistic independence, which, however, is unable to supply a suitable interpretation. For suppose theory h was advanced in response to a refutation by e' and that h both explains the old e' and makes the novel prediction e''. It is the general opinion, certainly shared by Popperians, and a consequence of Bayes's Theorem, that e'' confirms h, provided it is sufficiently improbable relative to background information. As discussed earlier in this chapter (pp. 226–7), such confirmation is available, in particular, when $P(e''|h \& e') > P(e''|e')$. But this inequality is quite compatible with e'' and e' *not* being independent in the probabilistic sense.

Another possible way to interpret the independence notion is in terms of logical independence, so that e' and e'' would be said to be independent just in case neither entails the other. This would mean that if the two bits of evidence were trivially distinct in, say, relating to different times or slightly different places, then they would be independent in the sense employed in the adhocness criterion. But then practically no theory would be ad hoc. Take Medawar's peculiar theory about the Eskimo's cozy style of life, which was propounded in response to some surprising IQ measurements. Presumably, one could infer from the theory that tests applied during the following week to the same group of Eskimos would produce similarly high IQs. But although this prediction is logically independent of the earlier results, its success would not significantly improve the standing of Medawar's theory. Mere logical independence from the old results is clearly insufficient to ensure evidential support.

Intuitively, new evidence supports a theory only when it is substantially different from known results, not just trivially different in the logical sense described, and it is this intuition which, it seems to us, underlies the standard adhocness criterion. The idea that 'different' or 'varied' evidence gives more support to a hypothesis than a similar volume of homogeneous evidence is an old and widely held one. As Hempel put it, "the confirmation of a hypothesis depends not only on the quantity of the favorable evidence available, but also on its variety: the greater the variety, the stronger the

resulting support" (1966, p. 34). So, for example, the report of the rate at which a stone falls to the ground from a given height on a Tuesday is similar to that relating to the stone's fall on a Thursday; it is very different, however, from a report of the trajectory of a planet or of how a given fluid rises in a particular capillary tube. But although the notions of similarity and diversity amongst evidence seem intuitively clear, it is not easy to give them a precise analysis, except, in our view, in probabilistic terms, in the context of Bayesian induction.

The similar instances in the above list have the characteristic that when one of them is known, any other would thereby be anticipated with relatively high probability. This recalls Bacon's characterisation of similarity in the context of inductive evidence. He spoke of observations "with a promiscuous resemblance one to another, insomuch that if you know one you know all" and was probably the first to point out that it would be superfluous to cite more than a small representative sample of such observations in evidence (see Urbach, 1987, pp. 160–64). This idea of similarity between items of evidence is expressed naturally in probabilistic terms by saying that e_1 and e_2 are similar provided $P(e_2|e_1)$ is higher than $P(e_2)$; and one might add that the more the first probability exceeds the second, the greater the similarity. This means that e_2 would provide less support if e_1 had already been cited as evidence than if it was cited by itself.

On the other hand, knowing that one of a pair of dissimilar instances has occurred gives little or no guidance as to whether the other will occur. For example, unless Newton's, or some comparable, theory had already been firmly established, a knowledge of the rate of fall of a given object on some specific occasion would not significantly affect one's confidence that the planet Venus, say, would appear in a particular position in the sky on a designated day. Different pieces of evidence may also have a mutually discrediting effect. An example of this might be the observations of the same constant acceleration of heavy bodies dropped at sea level and the unequal rates of fall of objects released on different mountain tops. Both sets of observations would confirm Newton's laws, but in circumstances where those laws are not already well established, the first set might suggest that all objects falling freely (whether on top of a mountain or not) do so with the same acceleration. In other words, with different instances, say e_3 and e_1, $P(e_3|e_1)$ is either close to or less than $P(e_3)$. Of course, e_3 merely differing from e_1 in this sense does not imply that it supports any hypothesis significantly; whether it does or not depends on its probability. The notion of similarity, as we have characterised it, is reflexive, as it should be; that is, if e_2 is (dis)similar to e_1, then e_1 is (dis)similar to e_2 (this follows directly from Bayes's Theorem).

To summarize, the non-Bayesian appraisal of hypotheses based on the notion of adhocness is ungrounded in epistemology, has highly counterintuitive consequences, and relies on a concept of independence among items of evidence which seems unanalysable except in Bayesian terms. In brief, it is not a success.

Infinitely many theories compatible with the data

The problem

Galileo carried out numerous experiments on freely falling bodies and on bodies rolling down inclined planes in which he examined how long they took to descend various distances. These experiments led him to formulate the well-known law to the effect that $s = ut + 1/2gt^2$, where s is the distance fallen by a freely falling body in time t, u is its

initial downward velocity, and g is a constant. Jeffreys (1961, p. 3) pointed out that Galileo might also have advanced the following as his law:

$$s = ut + \frac{1}{2}gt^2 + f(T)(T - T_1)(T - T_2) \ldots (T - T_n),$$

where T represents the date of the experiment, which could for example be recorded as the number of minutes that have elapsed since the start of the year AD 1600; T_1, T_2, ..., T_n are the specific dates on which Galileo performed his experiments; and f can represent any function of T. Thus Jeffreys's modification stands for an infinite number of alternatives to Galileo's theory. Although all these theoretical alternatives contradict one another and make different predictions about future experiments, the interesting feature of Jeffrey's unorthodox laws of free fall is that they all imply Galileo's experimental data.

This is a particular problem for those non-probabilistic theories of scientific method which hold that the scientific value of a theory is determined just by $P(e|h)$ and, in some versions, by $P(e)$. These philosophical approaches, of which Popper's is one example and maximum-likelihood estimation another, would have to regard the standard law of free fall and those peculiar alternatives described by Jeffreys as equally good scientific theories relative to the evidence available to Galileo, although this is a judgment with which no scientist would agree.

The same point emerges from a well-known example due to Nelson Goodman (1954; for an amusing and illuminating discussion, see Jeffrey, 1983, pp. 187–90). Goodman noted that the evidence of very many and varied green emeralds would normally suggest that all emeralds are green. But he pointed out that that evidence bears the same relation to 'All emeralds are green' as it does to a type of hypothesis he formulated as 'All emeralds are grue'. According to Goodman's definition, something is grue if it was either observed before time t and was green, or was not observed before t and is blue. If t denotes some time after the emeralds described in the evidence were observed, then both the green- and the grue-hypotheses imply that the emeralds observed so far should be green. However, on the assumption that there are unobserved emeralds, the hypotheses are incompatible, differing in their predictions about the colours of emeralds looked at after the critical time. As with Jeffreys's variants of Galileo's theory, the grue-hypothesis represents an infinite number of alternatives to the more natural hypothesis, for t can assume any value, provided it is later than now.

Our examples illustrate a general problem for methodology: that a theory which explains (in the sense of implying or associating a certain probability with) some data is merely one out of an infinite set of rival theories, each of which does exactly the same. The existence of this infinite set of possible explanations, it will be remembered, spelled ruin for any attempt at a positive solution to the problem of induction. The problem with which we are concerned here arises because, in practice, scientists discriminate between possible explanations and typically pick out just one, or at any rate relatively few, as meriting serious attention. An account of scientific method ought to explain how and why they do this.

The Bayesian approach to the problem

This has not proved easy. For the Bayesian, the nature of the problem, at least, is straightforward. Moreover, Bayesian theory does not imply that every hypothesis similarly

related to the data is of equal merit. Suppose one were comparing two theories in the light of the same evidence. Their relative posterior probabilities are given by

$$\frac{P(h_1|e)}{P(h_2|e)} = \frac{P(e|h_1)P(h_1)}{P(e|h_2)P(h_2)}$$

If both theories imply the evidence, then $P(e|h_1) = P(e|h_2) = 1$. And if, in addition, $P(h_1|e)$ exceeds $P(h_2|e)$, then it follows that $P(h_1)$ is larger than $P(h_2)$. More generally, if two theories which explain the data equally well nevertheless have different posterior probabilities, then they must have had different priors too. So theories such as the contrived alternatives to Galileo's law and Goodman's grue-variants must, for some reason, have lower prior probabilities. Indeed, this is clearly reflected in most people finding such hypotheses quite unbelievable. The problem then is to discover the criteria and rationales by which theories assume particular prior probabilities.

Sometimes there is a clear reason why a theory is judged improbable. For instance, suppose the theory concerned a succession of events in the development of a human society; it might, for example, assert that the elasticity of demand for herring remains constant or that the surnames of all future British prime ministers and American presidents will start with the letter Z. These theories, which of course could be true, are however, monstrously improbable. And the reason for this is that the events they describe are influenced by numerous independent processes whose separate outcomes are improbable. The probability that all these processes will turn out to favour the hypotheses in question is therefore the product of many small probabilities, and so itself is very small indeed (Urbach, 1987b). The question of how the probabilities of the causal factors are estimated, of course, remains. This could be answered by reference to other probabilities, in which case the question is just pushed one stage back, or else by some different process that does not depend on probabilistic reasoning. For instance, the simplicity of a hypothesis has been thought to have an influence on its initial probability.

It is worth mentioning here that the equation given above, relating the posterior probabilities of two theories to their prior probabilities, explains an important feature of inductive reasoning. The scientist often prefers a theory which explains the data imperfectly, in that $P(e|h_1) < 1$, to an alternative, h_2, which predicts them with complete accuracy. Thus, even Galileo's data were not in precise conformity with his theory; nevertheless, he did not consider any more-complicated function of u and t to be a better theory of free fall than his own, even though it could have embraced the evidence he possessed more perfectly. According to the above equation, this is because the better explanatory power of the rival hypotheses was offset by their inferior prior probabilities (Jeffreys, 1961, p. 4).

Conclusion

Charles Darwin (1868, vol. 1, p. 8) said that "In scientific investigations it is permitted to invent any hypothesis, and if it explains various large and independent classes of facts it rises to the rank of a well-grounded theory". This is, perhaps, an exaggeration, for not any hypothesis would do; the hypothesis must not be refuted or substantially disconfirmed, nor should it be intrinsically too implausible. With these provisos, Bayesianism, we suggest, is just such a well-grounded hypothesis as Darwin referred to, it arises from

natural and intuitively reasonable attitudes to risk and uncertainty. It is neither refuted nor undermined by any of the phenomena of scientific reasoning. On the contrary, as we have seen, it explains a wide variety of them.

Note

* Originally published as 'Bayesian versus non-Bayesian approaches', Chapter 7 of *Scientific Reasoning: The Bayesian Approach*, 2nd ed., Open Court: Chicago, 1993, pp. 117–64 (omitting sections g, h.2, part of i, j.3, and the exercises). Copyright © 1989, 1993 Open Court Publishing Company. Reprinted with permission.

Bibliography

Babbage, C. (1827), 'Notice Respecting Some Errors Common to Many Tables of Logarithms'. *Memoirs of the Astronomical Society*, 3: 65–67.

Bacon, F. (1620), *Novum Organum*. In *The Works of Francis Bacon*, vol. 4, J. Spedding, R. L. Ellis, and D. D. Heath (eds), London: Longman and Company, 1857–1858.

Bowden, B. V. (1953), 'A Brief History of Computation', in *Faster than Thought*, B. V. Bowden (ed), London: Pitman Publishing.

Cox, R. T. (1961), *The Algebra of Probable Inference*, Baltimore: Johns Hopkins Press.

Darwin, C. (1868), *The Variation of Animals and Plants under Domestication*, 2 vols., London: John Murray.

Dobzhansky, T. (1967), 'Looking Back at Mendel's Discovery', *Science*, 156: 1588–1589.

Dorling, J. (1979), 'Bayesian Personalism, the Methodology of Research Programmes, and Duhem's Problem', *Studies in History and Philosophy of Science*, 10: 177–187.

—— (1982), 'Further illustrations of the Bayesian Solution of Duhem's Problem', unpublished.

Duhem, P. (1905), *The Aim and Structure of Physical Theory*, P. P. Weiner (trans.), Princeton: Princeton University Press, 1954.

Edwards, A. W. F. (1986), 'Are Mendel's Results Really Too Close?', *Biological Reviews of the Cambridge Philosophical Society*, 61: 295–312.

Fisher, R. A. (1936), 'Has Mendel's Work Been Rediscovered?', *Annals of Science*, 1: 115–137.

—— (1970), *Statistical Methods for Research Workers*, 14 ed., Edinburgh: Oliver and Boyd.

Good, I. J. (1950), *Probability and the Weighing of Evidence*, London: Griffin.

—— (1961), 'The Paradox of Confirmation', *British Journal for the Philosophy of Science*, 11: 63–64.

Goodman, N. (1954), *Fact, Fiction, and Forecast*, London: Athlone Press.

Grünbaum, A. (1976), 'Is the Method of Bold Conjectures and Attempted Refutations *Justifiably* the Method of Science?', *British Journal for the Philosophy of Science*, 27: 105–136.

Hempel, C. G. (1945), 'Studies in the Logic of Confirmation', *Mind*, 54: 1–26 and 97–121.

—— (1966), *Philosophy of Natural Science*, Englewood Cliffs, NJ: Prentice-Hall.

Horwich, P. (1982), *Probability and Evidence*, Cambridge: Cambridge University Press.

Howson, C. (1984), 'Bayesianism and Support by Novel Facts', *British Journal for the Philosophy of Science*, 35: 245–251.

Jeffrey, R. (1983), *The Logic of Decision*, 2 ed., Chicago: University of Chicago Press.

Jeffreys, H. (1961), *Theory of Probability*, 3 ed., Oxford: Clarendon Press.

Jevons, W. S. (1874), *The Principles of Science*, London: Macmillan and Co.

Korb, Kevin (1994), 'Infinitely many resolutions of Hempel's paradox', in *Theoretical aspects of reasoning about knowledge*, R. Fagin (ed.), Asilomar, California: V. Morgan Kaufmann, pp. 138–149.

Kuhn, T. S. (1970), *The Structure of Scientific Revolutions*, 2 ed., Chicago: University of Chicago Press.

Lakatos, I. (1970), 'Falsificationism and the Methodology of Scientific Research Programmes', in *Criticism and the Growth of Knowledge*, I. Lakatos and A. Musgrave (eds.), Cambridge: Cambridge University Press.

Lindley, D. V. (1970), 'Bayesian Analysis in Regression Problems', in *Bayesian Statistics*, D. L. Meyer and R. O. Collier (eds), Itasca, IL: F. E. Peacock.

Mallet, J. W. (1880), 'Revision of the Atomic Weight of Aluminium', *Philosophical Transactions*, 171: 1003–1035.

Medawar, P. (1974) 'More Unequal Than Others', *New Statesman*, 87: 50–51.

Miller, R. (1987) *Bare-faced Messiah*, London: Michael Joseph.

Musgrave, A. (1975), 'Popper and "Diminishing Returns From Repeated Tests" ', *Australasian Journal of Philosophy*, 53: 248–253.

Polanyi, M. (1962), *Personal Knowledge*, 2 ed., London: Routledge and Kegan Paul.

Popper, K. R. (1959), *The Logic of Scientific Discovery*, London: Hutchinson.

—— (1963), *Conjectures and Refutations*, London: Routledge and Kegan Paul.

Rosenkrantz, R. D. (1977), *Inference, Method, and Decision: Towards a Bayesian Philosophy of Science*, Dordrecht: Reidel.

Smart, W. M. (1947), 'John Couch Adams and the Discovery of Neptune', *Occasional Notes of the Royal Astronomical Society*, number 11.

Sproutt, W. J. H. (1936), 'Review of K. Lewin's *A Dynamical Theory of Personality*', *Mind*, 45: 246–251.

Stas, J. S. (1860), 'Researches on the Mutual Relations of Atomic Weights', *Bulletin de l'Academie Royale de Belgigue*, 208–336.

Swinburne, R. G. (1971), 'The Paradoxes of Confirmation—A Survey', *American Philosophical Quarterly*, 8: 318–329.

Thomson, T. (1818), 'Some Additional Observations of the Weights of the Atoms of Chemical Bodies', *Annals of Philosophy*, 12: 338–350.

Urbach, P. (1981), 'On the Utility of Repeating the "Same" Experiment', *Australasian Journal of Philosophy*, 59: 151–162.

—— (1987), *Francis Bacon's Philosophy of Science*, La Salle, IL: Open Court.

—— (1987b), 'The Scientific Standing of Evolutionary Theories of Society' *The LSE Quarterly*, 1: 23–42.

—— (1991), 'Bayesian Methodology: Some Criticisms Answered', *Ratio (new series)*, 4: 170–184.

Velikovsky, I. (1950) *Worlds in Collision*, London: Victor Gollancz. (Page references are to 1972 reprint, Sphere Books, Ltd.)

Watkins, J. W. N. (1984), *Science and Scepticism*, London: Hutchinson and Princeton: Princeton University Press.