SCIENTIFIC REASONING: THE BAYESIAN APPROACH

COLIN HOWSON

AND

PETER URBACH

Second Edition

. . . if this [probability] calculus be condemned, then the whole of the sciences must also be condemned.
—Henri Poincaré

Our assent ought to be regulated by the grounds of probability.
—John Locke

Open Court
Chicago and La Salle, Illinois

CHAPTER 15

The Objections to the Subjective Bayesian Theory

a INTRODUCTION

In the preceding chapters we have developed the theory of subjective or personalistic Bayesianism as a theory of inductive inference. We have shown that it offers a highly satisfactory explanation of standard methodological lore in the domains of both statistical and deterministic science; and we have also argued at length that all the alternative accounts of inductive inference—like Popper's or Fisher's—achieve their explanatory goals, where they achieve them at all, only at the cost of quite arbitrary stipulations. However, the subjective Bayesian theory itself has been the object of much critical attention, to such an extent that it is still regarded in some influential quarters as vitiated by hopeless difficulties. These difficulties, in our view, stem from misunderstanding, and in this final chapter we shall do our best to dispel them.

Of the standard criticisms some—due largely to Popper and his followers—are answered relatively simply and quickly, and we shall deal with these first.

b THE BAYESIAN THEORY IS PREJUDICED IN FAVOUR OF WEAK HYPOTHESES

Discussing theories of inductive inference which assess the empirical support of hypotheses by changes in their probabilities on receipt of the relevant new data, Watkins (1987, p. 71) asserts that such theories are "prejudiced" in favour of logically weaker hypotheses. This is a favourite charge of the Popperian school and is frequently made by its eponymous founder; for example, Popper (1959, p. 363; his italics) writes that "[scientists] have to choose between high probability and high informative content, since *for logical reasons they cannot have both*".

Such a charge is quite baseless. There is *nothing* in logic or the probability calculus which precludes the assignment of even probability 1 to any statement, however strong, as long as it is not a contradiction, of course. The only other way in which probabilities depend on logic is in their decreasing monotonically from entailed to entailing statements. But this again does not preclude anybody from assigning any consistent statement as large a probability as they wish. Popper's thesis that a necessary concomitant of logical strength is low probability is simply incorrect.

Glymour attempts to argue a variant of Popper's objection, but this, too, is easily rebutted. Glymour claims that since the observable consequences of scientific theories are at least as probable as the theories themselves, then in a Bayesian account one is unable to account for our entertaining theories at all:

> On the probabilist view, it seems, they are a gratuitous risk. The natural answer is that theories have some special function that their collection of observable consequences cannot serve; the function most frequently suggested is explanation. . . . [But] whatever explanatory power may be, we should certainly expect that goodness of explanation will go hand in hand with warrant for belief, yet if theories explain and their observational consequences do not, the Bayesian must deny the linkage. (Glymour, 1980, pp. 84–85)

The Bayesian certainly does want to justify the quest for theories in terms of a desire for explanation that a congeries of observational laws cannot by itself provide; but he would also, for very good reason, deny the linkage Glymour alleges between explanatory power and warrant for belief. Indeed, counter-examples to the claim that any such linkage exists are only too easy to find: a tautology, to take an obvious one, has maximal warrant for belief and minimal explanatory power. This does not, of course, imply that what we take to be good explanations do not tend to have correspondingly high probabilities on the available evidence. They do. But Glymour's premiss makes the additional claim that an increase in "warrant for belief" should imply an increase in explanatory power. That premiss is clearly false, and Glymour's objection collapses.

It is strange that Glymour and the Popperians should converge, from quite different starting points, in charging Bayesians with an implicit denial of the value of deep explanatory theories. Glymour thinks that good explanatory theories by that token justify a correspondingly large claim to belief, whereas the Popperians assert that such theories merit the lowest possible degree of belief. Whatever their starting points, however, the charge of Glymour, Popper, et al., that Bayesians must in principle undervalue theories is patently false. Perhaps a simple analogy will dispel any lingering doubts that may remain. A jury has always at least two mutually inconsistent hypotheses to consider: that the accused is guilty is one, and that the accused is not guilty and there is some alternative explanation of the known facts is the other. The jury has to determine which of these is the more probable hypothesis, given the evidence. Imagine their surprise at being informed that they are thereby committed, on their return to the court, to announcing that their favoured conclusion is the restatement of the evidence! (*see also* Horwich, 1982, p. 132). Scientists, like the court, want information of a specific sort combined with the assurance that it is credible information; and these demands *can,* despite Popper's claim to the contrary, simultaneously be met. The Bayesian theory tells us how.

c THE PRIOR PROBABILITY OF UNIVERSAL HYPOTHESES MUST BE ZERO

Popper, we noted in the previous section, asserts that it is impossible for a hypothesis to possess both high informative content and high probability. In particular, he asserts that the probability of a universal hypothesis must, for logical reasons, be zero (1959a, appendices *vii and *viii). He occasionally remarks (for example, 1959a, p. 381) that the constraints imposed by the probability calculus alone require that the only consistent assignment of a probability to such a hypothesis is zero.

Were Popper correct, then that would be the end of our enterprise in this book, for the truth of Popper's thesis would imply that we could never regard unrestricted universal laws

as confirmed by observational or experimental data, since if $P(h) = 0$, then $P(h \mid e) = 0$ also, whatever finite sample data e may consist of.

Popper's thesis is quite untrue, however. Even in Carnap's so-called continuum of inductive methods (Carnap, 1952; *see* our discussion in Chapter 4), one of those methods (corresponding to $\lambda = 0$), assigns, in an obviously consistent way, positive probabilities to a class of strictly universal hypotheses over an infinite domain. And Hintikka's systems of inductive logic almost invariably assign positive prior probabilities to consistent universal sentences, as we also noted in Chapter 4, whether the domain of individuals is finite or infinite. It is even possible to assign positive probabilities to *all* the non-contradictory sentences in a language powerful enough to include all of science and mathematics (Horn and Tarski, 1948, theorem 2.5).

Popper's arguments for his zero-probability claim are really designed to show something considerably less ambitious than the false thesis that there can be no consistent assignment of a non-zero probability to a universal hypothesis. What they aim at showing is that the assignment of positive probabilities to universal hypotheses involves a quite unacceptable degree of arbitrariness. He has three main arguments. We shall review them briefly and conclude that none succeeds (the discussion follows Howson, 1973).

1. Popper points out that only one among the 2^n state descriptions in $L(A,n)$ (*see* Chapter 4, section **c.1**) satisfies the universal sentence $\forall x\, A(x)$. Hence, the proportion of 'possible worlds' satisfying that sentence is zero in the limit as n tends to infinity. This well-known property of $m^\dagger$ is maintained, moreover, even if we expand the language L to incorporate any number of predicates. But Popper fails to provide any reason why we should adopt that particular a priori distribution. As we argued in Chapter 4, no equiprobability distribution over any partition of logical space qualifies as representing genuine epistemic neutrality. $m^\dagger$, for example, assigns the statement 'There are universal laws' the a priori probability 0, and the statement 'There are no universal laws' probability 1, *an assignment which is irrevocable however strong the evidence might be that phenomena exemplify lawlike behaviour.*

To be fair to Popper, he is not entirely wholehearted in his endorsement of $m^\dagger$, but he claims that $m^\dagger$ *is* correct in making all the sentences $A(a_i)$ probabilistically independent with constant probability; and this is clearly enough to imply that in the limit as n tends to infinity $P[\forall x\, A(x)]$ is equal to 0. The $A(a_i)$ should be probabilistically independent, according to Popper, because "every other assumption would amount to postulating *ad hoc* a kind of after-effect" (1959, p. 367). Popper nowhere argues for the constant probability assumption, and indeed, such an assumption appears to be quite arbitrary. Nor does he notice that independence, by an argument exactly parallel to his own, should amount to postulating *ad hoc* a *lack* of after-effect. If one postulate is unaceptable, so ought to be the other. But every distribution P will either make the $A(a_i)$ independent or it will not, so either way P will be unacceptable.

2. Let $e^n = e_1 \& \ldots \& e_n$. If $h \vdash e_i$ for each i, $1 \leq i \leq n$, then $P(h \mid e^n) = \dfrac{P(h)}{P(e^n)}$. $P(e^n) \leq P(e^{n-1})$, so if $P(h) > 0$, then $P(e^n)$ must tend to a non-zero limit as n tends to infinity. Since $P(e^n) = P(e_n \mid e^{n-1}) \ldots P(e_2 \mid e_1)P(e_1)$, it follows that $P(e_n \mid e^{n-1})$ tends to 1. Popper now invites us to consider all the 'grue' variants h_k *of* h, where h_k entails $e_1, e_2, \ldots, e_k$, and also $\sim e_{k+1}, \sim e_{k+2}, \ldots$ If each $P(h_k) > 0$ also, and it seems quite unjustified to deny this and concede it for h, then $P(c_{kn} \mid c_k^{n-1})$ also tends to 1, where $c_{ki} = e_i$ if $i \leq k$, and $c_{k_i} = \sim e_i$ if $i > k$. Hence $P(c_{kn} \mid c_k^{n-1})$ tends to 1. In that case, claims Popper, there must be an m such that both $P(e_m \mid e^{m-1}) > \frac{1}{2}$ and $P(\sim e_m \mid e^{m-1}) > \frac{1}{2}$ (1959a, p. 371), which is a contradiction. But this does not follow at all; it presupposes (*see* exercise 3 at the end of this chapter) that the sequence of functions $f_n(k) = P(c_{k\,n} \mid c_k^{n-1})$ converges *uniformly* to 1 over the set $k \geq 1$, which there is no reason whatever to believe true.

3. Consider the sequence of polynomial theories

$$h_n\colon y = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0,$$

where x can take any value and the a_i are undetermined or adjustable *non-zero* real parameters, to be evaluated from the data. Popper claims (1959a, pp. 381–82) that we must accept that $P(h_{i+1}) \geq P(h_i)$ for all i, since the smaller the index i, the

more opportunities there are for falsifying h_i (in general, it will require n observations to determine the n parameters in h_n). Given that the a_i are non-zero, the h_i are mutually inconsistent, and so $\Sigma P(h_i) \leq 1$, from which it follows that $P(h_i) = 0$ for all i. This argument too fails, because Popper's premiss equating testability with improbability is false. That fewer independent observations are required to test h_i than h_{i+1} does not imply that h_{i+1} should be regarded as less likely to be false than h_i. As we have already argued in Chapter 7, section **j.3**, in supposing otherwise Popper is confusing pragmatic with epistemological considerations.

So none of Popper's arguments for $P(h) = 0$, where h is a universal generalisation, succeeds. Moreover, Popper's recommendation actually turns out to be impossible to implement consistently. For as we saw above, the assignment of probability 0 to universal sentences in $L(A,n)$ entails that the meta-level *universal* sentence 'There are no laws' must be assigned probability 1. So Popper's position is worse than being arbitrary; it is incoherent.

Nevertheless, Popper has called our attention to a fact which deserves some comment, namely, that the history of science is the history of great explanatory theories eventually being refuted. In view of this, ought we not rationally to expect all theories to be eventually refuted? It is far from clear that such bleak pessimism really is the lesson taught by the history of science. The mere fact that succeeding extensions of the observational base of science have caused the demise of many an explanatory theory does not demonstrate the appropriateness of total scepticism, nor does it even make it plausible. If up till now I have failed to find the thimble, I do not conclude, and certainly ought not to conclude, that further quest is hopeless. Of course, science is not hunt the thimble, but this does not destroy the point of the analogy, which is that a number of past failures to discover the truth does not by itself imply that one will not one day be successful.

Pessimism on that particular score is certainly not something to which many practitioners of science subscribe. There is a great deal of biographical and anecdotal evidence which suggests that, on the contrary, some very illustrious scientists are, if anything, overoptimistic. Einstein's confidence sometimes bordered on the hubristic: when, after the reports of the

1919 eclipse expedition, someone asked him what he would have felt had the result not confirmed General Relativity, he is said to have replied, "Then I would have to pity the dear Lord. The theory is still correct". Even where there is doubt about a theory, that doubt is often accompanied by a belief that some substantial part of the theory is true or approximately true. In such cases, it *is believed* that a suitably modified form of that theory will turn out to be true. To sum up, there is no evidence that scientists regard general theories as invariably false and no evidence that they ought to.

To sum up: (a) Popper's logical arguments for the probability of laws being zero are all invalid; (b) it is both irrationally dogmatic and incoherent to make these probabilities uniformly zero; and (c) the history of science offers no evidence or encouragement for the view that they should be zero.

d PROBABILISTIC INDUCTION IS IMPOSSIBLE

This dramatic claim is made by Popper and Miller (1983), who also supply what purports to be a rigorous proof of it. This proceeds as follows. According to Bayesian theories of support or confirmation, whether they are subjectively based or not, evidence e supports hypothesis h if and only if $P(h \mid e) > P(h)$. Suppose that h entails e, modulo background information including initial conditions and so forth. It follows from theorem 17, Chapter 2, that e supports h if and only if $P(h) > 0$ and $P(e) < 1$. Suppose that these latter conditions are satisfied also, so that h is (it seems) supported by e. Popper and Miller demonstrate that if, in addition, $P(h \mid e) < 1$, then $\sim e \vee h$ is *counter-supported* by e, in the sense that its posterior probability relative to e is *less* than its prior probability (the proof is very straightforward, and we leave it as exercise 4 at the end of this chapter).

This simple theorem of the probability calculus is given a dramatic significance by Popper and Miller, for they claim that $\sim e \vee h$, for reasons which we shall give later, represents *the excess or inductive content of* h *over* e, or that part of h's content going beyond e, and they interpret their result as stating that this excess content is always counter-supported by e. All support, conclude Popper and Miller, is really *deductive* sup-

port (since e entails e); the support given by e to the genuinely inductive content of h is always negative.

While various elements in Popper's and Miller's argument have been challenged, nobody (apart from Rivadulla, 1991) seems to have noticed that, if correct, their argument proves that *counter-induction is valid:* it is surely just as bad, from their anti-inductivist point of view, to show that $\sim e \vee h$ is *counter-supported* by e as if they had proved that it was supported by e. *Both* are species of inductive inference, since in both cases e gives information about what is claimed to transcend it. In one, e tells us that the allegedly inductive content of h is more likely to be true than it was before, and in the other, e tells us that that content is less likely to be true.

Popper and Miller obtain their odd result only because they adopt an eccentric interpretation of the idea of the excess content of one set of sentences with respect to another. Suppose, with Popper and Miller, we define the content of h to be its class $Cn(h)$ of (non-tautologous) consequences. We want to define the excess content of h over e. How do we do it? There is a fairly standard answer for defining the excess of A over B where A and B are any two sets: the excess of A over B is the largest subset of A which includes nothing in B, and this is, of course, just the set-theoretic difference $A - B$. But Popper and Miller define the content of h excess to that of e to be the largest subset of $Cn(h)$ which contains nothing in $Cn(e)$ *and which is also deductively closed.* And this is equal to $Cn(\sim e \vee h)$.

But *why* stipulate that the difference between $Cn(h)$ and $Cn(e)$ is deductively closed? Popper and Miller have, as they admit, chosen to represent the difference of $Cn(h)$ with respect to $Cn(e)$ in the algebraic context of Tarski's well-known 'calculus of deductive systems', which requires that the results of performing operations of sum, difference, and product on deductively closed sets, like $Cn(h)$ and $Cn(e)$, *must themselves be deductively closed sets.* Indeed, the Tarski difference between $Cn(h)$ and $Cn(e)$ is $Cn(\sim e \vee h)$. But we repeat, why choose this method of defining the difference between $Cn(h)$ and $Cn(e)$, rather than that represented by the set-theoretical difference, of which $Cn(\sim e \vee h)$ is, it is not difficult to see, a (small) proper subset? (In fact, $Cn(\sim e \vee h)$ contains only those consequences of h which are also consequences of $\sim e$.) It is like looking through a distorting lens which makes a good deal of

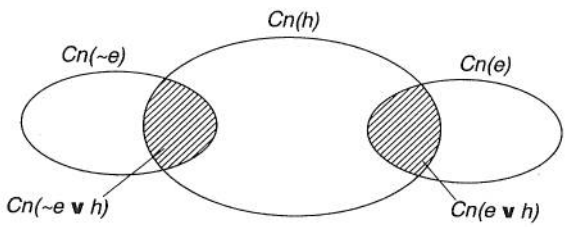

FIGURE 1

what is actually present—the consequences of h which are consequences neither of e nor of $\sim e$—disappear from view (figure 1; e and h are represented as logically independent).

The same distorting effect is achieved with Popper's well-known *numerical measure* of content $ct(a) = P(\sim a)$. It is easy to verify, and we shall leave it as an exercise, that $ct(e) + ct(\sim e \vee h) = ct(h)$ when h entails e. The reason for this is that $P(\sim a)$ determines a measure $P[M(\sim a)]$ on the algebra of extensions $M(a)$ of sentences of a (in the terminology of Chapter 2, section **f**). It is straightforward to show that this algebra, under the mapping which associates the consequence-class $Cn(a)$ of a with the set of models of $\sim a$, is isomorphic to the Tarski algebra of consequence classes. So we can regard Popper's content measure as a measure on the Tarski algebra itself, inheriting the distorting features of that algebra which we have noted (*see also* Howson, 1993).

Howson and Franklin (1986) investigate numerical content-measures which reflect more faithfully the structure of the underlying consequence classes than does ct and show that in this respect more adequate measures are afforded by $\text{Inf}(a) = -\log P(a)$, the so-called information measure (its expected value is Shannon's entropy), and $[1 - P(a)]:P(a)$, or the odds against a. It turns out that any adequate measure based on a probability function P must be a decreasing function $f(P)$ which, unlike $1 - P$, is strictly convex in some

subinterval of the unit interval; the reader is referred to Howson and Franklin's paper for further details.

Popper and Miller, in one of their publications (1987), defend their adoption of $Cn(\sim e \vee h)$ as the excess content of h over e by pointing out, correctly, that in any larger class there will be statements which will share non-tautologous consequences with e, and which consequently, in their terminology, will be to some extent 'deductively dependent' on e. For example, if a and b are logically independent, then setting $h = a \,\&\, b$ and $e = a$, $C_n(\sim e \vee h)$ will not even contain b; however, this is just as it should be, if Popper and Miller are correct, because b will share non-tautologous consequences with a.

But this is to impose the condition that the difference between $Cn(h)$ and $Cn(e)$ should be a *hereditary* one, in the sense that not only should the difference itself be a set disjoint from $Cn(e)$, but so should the set of consequences of every member of that set. Such a stipulation is quite arbitrary, and we recommend that it be ignored. If it is, Popper and Miller's strange version of induction ceases to be provable: which is as it should be.

■ e THE PRINCIPAL PRINCIPLE IS INCONSISTENT (MILLER'S PARADOX)

Miller (1966) produces an interesting argument which purports to demonstrate that the principle which, following Lewis, we have called the Principal Principle, is inconsistent. According to that principle, on which the Bayesian analysis of statistical inference rests, the (subjective) probability that an event described by the sentence a will occur at a trial of type T is equal to r, if our data are confined to the information that the physical probability of a, relative to the conditions T, is r. We can write this concisely as the equation

(1) $P[a \mid P^*(a) = r] = r,$

where P is a degree-of-belief probability function and P^* the physical probability function.

Miller's argument is as follows. Let r be $\frac{1}{2}$. Then by (1)

$$P[a \mid P^*(a) = \tfrac{1}{2}] = \tfrac{1}{2}.$$

But clearly, $P^*(a) = \frac{1}{2}$ if and only if $P^*(a) = P^*(\sim a)$; and the probability calculus tells us that we can substitute equivalent statements, whence we obtain

(2) $P[a \mid P^*(a) = P^*(\sim a)] = \tfrac{1}{2}.$

However, we can also instantiate (1) thus:

(3) $P[a \mid P^*(a) = P^*(\sim a)] = P^*(\sim a),$

and combining (2) and (3) we infer that $P^*(\sim a) = \frac{1}{2}$, which is odd, since no factual premiss of any kind has been employed in the derivation. While this result may not have the form of an outright contradiction, it very quickly leads to one. For we can repeat the reasoning above with the two substitution instances $P^*(a) = \frac{2}{3}$, so that $P^*(a) = 2P^*(\sim a)$, whence we would infer that $P^*(a) = \frac{2}{3}$, in explicit contradiction to $P^*(a) = \frac{1}{2}$.

Were Miller's derivation formally sound, the consequences for the Bayesian theory of statistical inference would be little short of disastrous, for the characteristic and often striking properties of the posterior probabilities of statistical hypotheses are due to the behaviour of the *likelihood function* $g(i) = P(e \mid h_i)$, where $\{h_i\}$ is a family of alternative hypotheses about the value of some physical probability distribution. But, as we also noted in Chapter 13, odds different from those based on the Principal Principle are demonstrably unfair, and this tells us that *something* must be wrong with Miller's argument. The question is, what? Miller's error is difficult to spot because it is concealed by the notation which, precisely in consequence of its being well adapted to smooth exposition and development of the theory, does not make explicit all the distinctions which are nevertheless implicit.

The erroneous step in Miller's derivation is to take (3) to be a substitution instance of (1). (3) is *not* a substitution instance of (1); it makes a quite different *type* of assertion from (1), and it will help the reader see why to turn back to and re-read

Chapter 2, section **d,** where random-variable statements are introduced and their meaning discussed, for (1), though it may not look like it, is an equation involving random variables. This fact is obscured by our tendency to regard $P^*(a)$ as a number, or scalar. But in the context of a discussion in which '$P^*(a) = r$' is itself a statement assigned a probability value (by the function P), $P^*(a)$ is *not* a number: it is something which takes a range of possible values—those possible values being, of course, all the real numbers in the closed unit interval. And a quantity which takes different values in different possible states of the world is a function, and when it is a function over whose values there is a probability distribution, it is a *random variable.*

Let us accordingly write $P^*(a)$ in (1) explicitly as a random variable X. So (1) says

(4) $P(a \mid X = r) = r,$

for all r, $0 \leq r \leq 1$. Now recall, from Chapter 2, section **f,** that we can replace the sentences in (4) by the sets of possible worlds making them true. So we can rewrite (4) as

(5) $P[M(a) \mid M(X = r)] = r.$

Looking back at Chapter 2 again, we see that $M(X = r) = \{w : X(w) = r\}$, where the w's are the members of the outcome space of the stochastic experiment relative to which the distribution P^* is defined. So (5) can be written

(6) $P[M(a) \mid \{w : X(w) = r\}] = r.$

But (3) has the form

(7) $P[M(a) \mid \{w : X(w) = Y(w)\}] = Y(w),$

where Y is another random variable equal to $1 - X$; Y is of course $P^*(\sim a)$ and $P^*(\sim a) = 1 - P^*(a)$. But it is now obvious that (7) is not a legitimate substitution instance of (5); it contravenes the logical rule that terms involving so-called free variables, like w, must not be substituted into contexts in which those free variables become bound, as the operator $\{w: \ldots\}$ binds $Y(w)$. (*See* Mendelson, 1964, p. 48, for the statement of this rule for logical quantifier operators.) It follows that (3) is not a legitimate substitution instance of (1) and the derivation of Miller's paradoxical conclusion cannot proceed. The Principal Principle is consistent.

■ f THE PARADOX OF IDEAL EVIDENCE

Suppose you are contemplating a long sequence of tosses—say, 1000—of a coin, about which you initially know nothing except that it looks like an ordinary coin. You are asked on two separate occasions what your degree of belief in the outcome of the 1000th toss is (i) now, before the tosses commence, and (ii) after the outcomes of 999 have been recorded and the 1000th toss is about to take place. In view of the sparseness of your initial data, your degree of belief *now* that it will land heads at the 1000th toss is, let us suppose, one half, more or less. Now suppose that the 999 tosses record an observed number of heads around 500—say, 503. You compute, using standard Bayesian calculations, the conditional probability on this data of a head at the 1000th toss and find that it is still the value, approximately one half, specified in your prior distribution over $\{H,T\}$. Your degree of belief in the coin's landing heads at the 1000th toss, obtained by conditionalisation from this conditional probability after observing the outcomes of the 999 tosses, *is therefore unchanged by the very extensive statistical data now available;* indeed, it is probabilistically independent of it.

Popper, and many others, regard this as highly damaging to a theory like the Bayesian theory which, since it registers states of belief by a *single* number, apparently cannot distinguish your posterior degree of belief, based as it is on very weighty evidence, from your prior belief, based on virtually none at all (we shall hear this same criticism echoed by Shafer later in this chapter, section **k**). According to Popper, the example above, which he terms 'the paradox of ideal evidence' (1959a, p. 407), points to the need for at least a *two-dimensional* representation of your belief state, where one of the dimensions is the ordinary probability of h in the light of the evidence, and the other represents in some way the *weight* of the evidence on which that probability is based. Your initial degree of belief in a head was based on sparse and imprecise

data, and therefore its weight index would be very small. Your later degree of belief, though numerically the same as your first, was based on very weighty data, by contrast. The latter subjective probability, therefore, according to the proponents of this weight of evidence theory, is more *epistemically reliable* (Gärdenfors and Sahlin, 1989, p. 321).

But it has been pointed out many times that the standard Bayesian account makes just these sorts of discriminations anyway, without needing to introduce a two-dimensional representation of belief states. The objective probability of heads in Popper's coin-tossing experiment is a random variable X in the Bayesian theory, taking as values all the real numbers x, $0 \leq x \leq 1$. Let $P_0(x)$ and $P_{999}(x)$ be the prior and posterior subjective-probability density distributions, respectively, over these values, i.e., before any tosses have taken place and after 999 have taken place. That we tend not to regard the initial data as providing grounds for discriminating very much between the values of X in a fairly broad band around the value $x=\frac{1}{2}$ will be reflected in the prior density $P_0(x)$ being rather diffuse. The data specifying the outcomes of the 999 tosses will, on the other hand (as we saw in Chapter 13, section **e**), cause the likelihood function $x^{503}(1-x)^{496}$ to be sharply peaked in the region $x=\frac{1}{2}$. Hence the posterior density $P_{999}(x)$ will concentrate virtually all the posterior probability in that region. So we see that the difference between the two 'weights of evidence' is reflected in the large variance of the prior density of X and the very small variance of the posterior density, and we could, if we so wished, define the respective weights as inversely proportional to these variances.

This is not all. Let h be the sentence 'A head will occur on the 1000th toss'. We shall leave it as an exercise for the reader to show that the prior probability $P_0(h)$ and the posterior probability $P_{999}(h)$ of h are respectively equal to the expectations $P_0(h) = E_0(X)$, $P_{999}(h) = E_{999}(X)$, where $E_0(X)$ is the expected value of X relative to the prior density $P_0(x)$ and $E_{999}(X)$ is the expected value of X relative to the posterior density $P_{999}(x)$. There is a clear sense in which the posterior probability $P_{999}(h)$ is epistemically more reliable than $P_0(h)$, even though they are (approximately) numerically equal. For the mean of X relative to increasing evidence will be increasingly insensitive to change: the relative frequency of heads will eventually converge to some value, to which your successive subjective probabilities of heads conditional on the observed outcomes will also converge. Therefore those subjective probabilities, *even if they remain numerically much the same,* are increasingly unlikely to alter as a result of the acquisition of more data. They are, in short, more reliable. Even where it is not assumed that the sequence of relative frequencies converges (if it is not generated by a collective), it follows from Dawid's well-known theorem on calibration (Dawid, 1982) that still you must, on pain of inconsistency, believe (in fact, *be certain*) that your conditional probabilities on accumulating data will converge on the average to the observed relative frequency, and hence be increasingly reliable in the sense of being unlikely to be disturbed significantly (we shall discuss Dawid's theorem at greater length in section **m** below).

The 'paradox' of ideal evidence is like many of the other objections to the Bayesian approach, in that on closer examination it reveals the strength, rather than the weakness, of that approach. In this particular case, we can see that the Bayesian theory, without any further assumptions or modifications being necessary, makes just those discriminations which its critics charge it with lacking.

g HYPOTHESES CANNOT BE SUPPORTED BY EVIDENCE ALREADY KNOWN

g.1 $P(h \mid e) = P(h)$ if e Is Known When h Was Proposed

In a much-discussed essay, severely critical of the Bayesian approach, Clark Glymour observed that

> Newton argued for universal gravitation using Kepler's second and third laws, established before the *Principia* was published. The argument that Einstein gave in 1915 for his gravitational field equations was that they explained the anomalous advance of the perihelion of Mercury, established more than half a century earlier. . . . Old evidence can in fact confirm new theory, but according to Bayesian kinematics it cannot. (1980, p. 86)

By "Bayesian kinematics" Glymour here means simply the principle of Bayesian conditionalisation. Glymour's objection is grounded on the fact that relative to a stock of background information including e, $P(e)$ is 1, whence $P(e \mid h)$ is 1 also, so that it follows immediately from Bayes's Theorem that $P(h \mid e) = P(h)$. Thus e does not raise the prior probability of h and hence, according to the Bayesian, does not confirm it.

Glymour's argument has the rather strange consequence that *no* data, whether obtained before or after the hypothesis is proposed, can, within a Bayesian theory of confirmation, confirm *any* hypothesis, for even if the hypothesis h is proposed before evidence e is collected, then by the time someone comes to do the Bayes's Theorem calculation, the terms $P(e)$, $P(e \mid h)$ must again be set equal to 1, since by that time e will of course be known and hence in background information. But few would infer in such a case that according to the Bayesian theory e did not support h. And not even the most committed opponent of that theory claims it to be damaged by this demonstration, for it is clear that the theory has been incorrectly used. It is equally clear where the mistake lies, namely, in relativising all the probabilities to the *totality* of current knowledge. They should, of course, have been relativised to current knowledge *minus* e. The reason for the restriction is, of course, that *your current assessment of the support of* h *by* e *measures the extent to which, in your opinion, the addition of* e *to your current stock of knowledge would cause a change in your degree of belief in* h.

It might be objected that the relativisation of the probabilities in Bayes's Theorem calculations to what is strictly a fictitious state of background information is simply ad hoc: it is a device which avoids the otherwise embarrassing necessity of setting $P(e)$ and $P(e \mid h)$ equal to 1, but it does so at the cost of being in conflict with core Bayesian principles. This charge is untrue. Core Bayesian principles simply state the conditions—obedience to the probability calculus—for a set of degrees of belief, relative to a stock of background information, to determine a corresponding set of odds which are not demonstrably unfair. There is absolutely nothing in this which asserts that in computing levels of support, one's subjective probabilities must define degrees of belief relative to the totality of one's current knowledge. On the contrary; the support of h by e is gauged according to the effect which one believes a knowledge

of e *would now have* on one's degree of belief in h, on the (counter-factual) supposition that one does not yet know e.

A more serious objection to this sort of relativisation of the probabilities in calculations of support is that in general there simply is no uniquely determined set $K - e$ which is the result of deleting the information represented by a sentence e from a database K—except, of course, in the case where e is not in K, so that then $K - e$ is just K. Even if e is separately represented in K, matters are not straightforward, for the result of simply removing e from K will depend on how K is axiomatised. For example, $K_1 = \{a,e\}$ and $K_2 = \{e \rightarrow a,e\}$ represent exactly the same stock of information, since they have the same consequences, but subtracting $\{e\}$ set-theoretically from each will leave $\{a\}$ in the first case and $\{e \rightarrow a\}$ in the second.

Much has been written on this problem (the contemporary state of the discussion is nicely summarised in Gärdenfors, 1989), and it seems that there is no *uniformly* satisfactory way to characterise the operation of deleting e from K. One briefly canvassed idea was to define $K - e$ as the so-called *full meet contraction function* on K,e, namely, the intersection of all those subsets of K which do not entail e. In the case where K is the deductive closure $Cn(k)$ of some sentence k (which might, of course, be some arbitrarily large conjunction), this results in the familiar set $Cn(\sim e \vee k)$. $Cn(\sim e \vee k)$, recall, is Popper and Miller's characterisation of the excess content of k over e. We found good reason to reject $Cn(\sim e \vee k)$ for that role, and $Cn(\sim e \vee k)$ has also been uniformly rejected as supplying a meaning for $K - e$, on the ground that it is too small, containing as it does consequences only of $\sim e$ and k (Gärdenfors, 1989, p. 79; Makinson, 1985, p. 359).

It is not the case, however, that in the sorts of situations of interest to us $K - e$ cannot be given a definite and satisfactory meaning. Indeed it can, for we have the paradigm example of how it can be done in Bayes's Theorem calculations using newly recorded evidence. Nobody, and this includes Glymour himself, denies that a perfectly proper way of deleting e from the current K in this case is simply to subtract it set-theoretically from the set of sentences represented by K, in which e will be an isolated point, or to put it more grandly, an independent 'axiom'. That K could be axiomatised differently, as the set $\{a,e\}$ could be reaxiomatised $\{e \rightarrow a,e\}$, raises no

difficulty of principle or practice, for it is simply a fact that the agent represents his or her own knowledge in a particular way. Given this, and given that e is usually (though not, of course, always) an independent item in this representation, the state of the individual's knowledge-base on the counter-factual assumption that e is not part of it is uniquely determined. Where e is not an independent item in K, the situation is less easy to deal with, and we may just have to conclude that the notion of what the probability of e, and of e given h, would be, were one counter-factually assumed not to know e, is not defined. But that such counter-factually based probabilities do not always exist should not blind us to the fact that in many cases they do, and consequently that in many cases the notion of support relative to known evidence is also perfectly meaningful.

Glymour considers this type of response quite sympathetically, but contends that nevertheless in

> actual historical cases . . . there is no single counterfactual degree of belief in the evidence ready to hand, for belief in the evidence sentence may have grown gradually—in some cases it may have even waxed, waned, and waxed again. (1980, p. 88)

He cites as an example the data on the perihelion of Mercury; there were different values obtained for this, over a period of several decades, by different methods, and employing mathematical techniques sometimes without rigorous justification. Glymour contrasts this situation with the results of tossing a coin a specified number of times, where he thinks it does make sense to talk of the probability of that outcome, as if it had not yet occurred. But in the case of Mercury's estimated perihelion advance, "there is no single event, like the coin-flipping, that makes the perihelion anomaly virtually certain" (Glymour, 1980, p 88).

But whether there is as much epistemic warrant for the data in 1915 about the magnitude of Mercury's perihelion advance as there is about the number of heads we have just observed in a sample of a hundred tosses of a coin is beside the point. We may be more tentative about some data, and about other data, less. The Bayesian theory we are proposing is a theory of inference from data; we say nothing about whether it

is correct to accept the data or even whether your commitment to the data is absolute. It may not be, and you may be foolish to repose in it the confidence you actually do. The Bayesian theory of support is a theory of how the acceptance as true of some evidential statement affects your belief in some hypothesis. How you came to accept the truth of the evidence, and whether you are correct in accepting it as true, are matters which, from the point of view of the theory, are simply irrelevant. Glymour's disquisition on the frailty of much scientific data is therefore, however valuable in its own right, beside the point of evaluating the adequacy of the Bayesian theory of inference.

In the course of his discussion, Glymour proposes an altogether different explanation of why it is that old evidence appears to confirm a new theory: it may not be the "old result that confirms a new theory, but rather the new discovery that the new theory entails (and thus explains) the old [result]" (Glymour, 1980, p. 92). This observation prompts Glymour to propose a new criterion for old evidence e to be taken as confirming a new theory h, namely that

(8) $P[h \mid e \,\&\, (h \vdash e)] > P(h)$,

where the probability calculus is weakened appropriately, by replacing the conditions on axioms 2 and 3 by 'if it is known that t is a tautology, then . . .', and 'if it is known that $a \vdash \sim b$, then . . .' (we have modified Glymour's notation in (8) slightly and omitted explicit reference to background information). This emendation of the classical theory is sympathetically endorsed by Niiniluoto (1983), and is further examined by Garber (1983), though the same idea seems first to have been proposed and developed by I. J. Good (starting with Good, 1977), who calls the resulting notion of probability "dynamic" or "evolving" probability.

Despite the considerable interest provoked by Glymour's novel explanation of how old evidence confirms new theories, we believe that it is refuted by the observation that scientists often build their theories around some particular item of data, which is thereafter cited as being among the evidence in support of those theories. Einstein, and many others, believed the invariance of the speed of light provided powerful evidence for the Theory of Special Relativity, despite the fact that the

theory was constructed to explain the phenomenon. Newton, and many others, believed that the explanation of Kepler's Laws by his theory of gravitation was one of the most important pieces of evidence for that theory, though the entailment of an approximate form of Kepler's Laws was the principal constraint on the form of the theory. In such cases, therefore, the ground for believing the theory supported by the evidence cannot be the discovery that the theory entails it, for that is exactly what the theory was designed to do.

g.2 Evidence Doesn't Confirm Theories Constructed to Explain It

To fully sustain our objection to Glymour, however, we have to answer those who believe such evidence inadmissible, and there is a body of opinion which does. We have already (in Chapter 7, section **j.4**) encountered—and showed to be untenable—the doctrine that theories, to be acceptable, must be supported by evidence independent of that which they were constructed to explain. Here, however, we face a more radical doctrine: *such evidence fails to support at all.* This view, which has attracted considerable following among philosophers (though not among scientists), is invariably defended by the following argument, recently stated by Giere, but which goes back to Peirce if not earlier:

> If the known facts were used in constructing the model and were thus built into the resulting hypothesis . . . then the fit between these facts and the hypothesis provides no evidence that the hypothesis is true [since] these facts had no chance of refuting the hypothesis. (Giere, 1984, p. 161)

Despite beguiling many, the argument is radically unsound. Whatever plausibility it might seem to possess is quickly removed by noting that no *fact* has a "chance" of refuting anything. If e is a factual statement and h a hypothesis, then e either refutes h or it does not, and it does or does not do so whether h was designed to explain or embody e or not. Giere, and all the many other people (like Worrall, 1989) who have been swayed by this argument, are confusing an experimental set-up, E, with one of its possible outcomes, e. This is similar to identifying a random variable with one of its values.

Once the distinction is made, the argument collapses, for even if h was constructed from one particular outcome of E, E could, if it was a well-designed experiment, have produced another inconsistent with h.

Of course, had it done so, it would not have been this h that you considered, but another, h', designed to accommodate that other outcome. Can one not reformulate the objection, then, in something like the following way: there is no chance—if you are careful—of the h you produce being inconsistent with the e you want explained by it. This is undeniably, indeed trivially, true. *But it still does not follow that* h *is necessarily unsupported by* e. Recall that it is necessary for h to be supported by e that, in Bayesian terms, $P(e \mid h)$ is larger than $P(e \mid \sim h)$, and there seems to be no reason to suppose that, just because you contrived h to explain e, $P(e \mid h)$ is not larger than $P(e \mid \sim h)$.

Redhead, however, has recently taken up the argument for this 'null support thesis' by claiming that in the circumstances envisaged $P(e \mid \sim h)$ will not only not be smaller than $P(e \mid h)$, *but it will necessarily be equal to 1* (1986). Redhead's argument is that by "filtering" out all the hypotheses which do not explain e, you are in effect assigning them all probability 0. We leave it as an exercise to show that such an assignment entails that $P(e \mid \sim h)$ is equal to 1. But there is a fatal flaw in this argument. For it follows from $P(e \mid \sim h) = 1$, together with $P(e \mid h) = 1$, that $P(e)$ itself must be equal to 1 (since $P(e) = P(e \mid h)P(h) + P(e \mid \sim h)P(\sim h) = P(h) + P(\sim h) = 1$). This, as we know, would imply that no hypothesis at all, built to explain e or not, could ever be supported by it, and that, we know, is false. And accepting that, as we must, means rejecting the assignment $P(e) = 1$, and, by implication, $P(e \mid \sim h) = 1$.

There are also simple counter-examples to Redhead's thesis. Suppose we are removing tickets from an urn. We remove all n and discover that r of the tickets are red. Let e record this observation. We now formulate the hypothesis h: 'There were r red tickets in the urn'. Clearly, $P(e \mid \sim h)$ is so small as to be practically zero, yet h was constructed to be the most plausible hypothesis which explains e.

g.3 The Principle of Explanatory Surplus

We shall now consider a slightly more-sophisticated version of the thesis that evidence on which a theory is constructed

cannot be used as evidence for it. Suppose that h is constructed with the help of a body e of data. Suppose also that we could decompose e into two parts, that which was actually used in the construction of the theory (e') and the remainder, e''. It remains a widely held view that while e may well support h, it does so only because the 'explanatory surplus' e'' supports h: e' never does (the term 'explanatory surplus' is due to Gillies, 1989; both he and Worrall, 1978, are philosophers who have recently held this position). It might be, for example, that e' reports the minimum necessary number of observations to fix a set of adjustable parameters in h. There is no guarantee that the surplus data e'' will satisfy h; e'' might even falsify it, and that is why e'' but not e' can be regarded as potentially supporting evidence for h.

It is clear that we are back with the argument that only data which has a 'chance' of conflicting with h can support it. We declared the argument unsound, because the experiment E may well have possible outcomes inconsistent with h. But if h is guaranteed, *whatever* outcome E delivers, to be consistent with that outcome—because h's parameters are determined by that outcome, for example—does that not show that the argument is valid for such cases? No. The appearance to the contrary arises because we are confusing two things: (i) the hypothesis h before its parameters are determined from the non-surplus data e'; and (ii) the hypothesis with its parameters evaluated, call it h'. It is true that h is not in general supported by e', because e' will in general give no information about the truth-value of h. *But it is h', not h, which was constructed with the help of e'.* And as regards h' we repeat what we said earlier: the data source E may well have possible outcomes inconsistent with h', and therefore be in principle capable of falsifying h'. And e' might well support h'.

Indeed, it is easy to think of cases where it will. Consider the following simple example, where h is a pure linear hypothesis $y = ax+b$, with the two parameters a and b undetermined. Let e' be two independent joint observations of y for specified values of x, determining a and b. h is not supported at all by e' if e' is totally independent of h; e' could be any pair of joint values. Suppose this is the case, so that $P(h \mid e') = P(h)$. The x-values are part of the experimental specification, so that relative to this we have $h' \Leftrightarrow h \,\&\, e'$. It is now easy to show that

while h is not supported by e', h' certainly is. First, we have that $P(h' \mid e') = P(h \,\&\, e' \mid e') = P(h \mid e) = P(h)$. We can assume that the specific values of y revealed by e' have probability less than 1, in which case it follows immediately that $P(h' \mid e') > P(h')$. Q.E.D.

We can also see from this example why the practice of merely saving the phenomena is so universally disparaged. If h is constructed with enough free parameters to be able to accommodate whatever data are to be explained, but there is no independent reason to believe h true, that is just another way of saying that $P(h)$ is very small. But as we see from the reasoning above, $P(h' \mid e') = P(h)$, i.e., $P(h' \mid e')$ is exactly as large as the prior probability of h, which is reckoned to be small. So although $P(h' \mid e')$ will in general exceed $P(h')$, the support e' gives to h' as measured by the difference between $P(h' \mid e')$ and $P(h')$ can never be considerable.

■ h PREDICTION SCORES HIGHER THAN ACCOMMODATION

It is not true that a hypothesis constructed to accommodate some evidence e is never supported by e. Nor is it even true that a hypothesis is unsupported by just the part of the data actually used in its construction. But an influential tradition, including Leibniz and later Whewell, has claimed that independent prediction of data nevertheless confers *more* support, for given data e and hypothesis h, than if h had merely accommodated e. Furthermore, it is often claimed, the Bayesian theory lacks the means even to discriminate between evidence incorporated as a deliberate constraint in the construction of a hypothesis h and evidence independently by h. Hence, since independent prediction is—it is alleged—such an important methodological criterion by which theories are evaluated as to their empirical adequacy, the Bayesian theory must be seriously inadequate.

But it is false that the Bayesian theory cannot make the discrimination between evidence accommodated and evidence independently predicted. It does so in a variety of ways, one of which turns on prior probabilities. Consider, for example, the case of a hypothesis h with some or all of its parameters left

undetermined. Suppose there is some data e which a rival hypothesis h'' predicts directly, but which h can only account for once its parameters have been fixed by e, and e is just sufficient to do this. Let h' be the result of fixing those parameters. If h and h'' start off with equal prior probabilities, and e is independent of h, it is easy to show that the adjusted form h' of h will in general obtain much less support from e than will h''. For the difference between the posterior and prior probabilities of h'' is $P(h'' \mid e) - P(h'') = \frac{P(h'')[1 - P(e)]}{P(e)}$, since h'' entails e. Similarly, the difference between the posterior and prior probabilities of h' is $\frac{P(h')[1 - P(e)]}{P(e)}$. But if the prior probability of e is less than 1, as it standardly will be, we shall certainly have $P(h) < P(h')$, for by the reasoning of the previous section $P(h') = P(h \,\&\, e) = P(h)P(e)$. But by assumption, $P(h'') = P(h)$, so the incremental support of h'' exceeds that of h', the adjusted hypothesis.

The disparity in supports in this example between the predicting and the accommodating hypotheses reflects the disparity in prior probabilities of h' and h''. But this is not always the reason. In the following example, discussed by Maher (1988), the enhanced support for the independently predicting hypothesis depends crucially on incorporating into background information the fact that the prediction was independent of knowledge of the evidence. In this example, an individual predicts the outcomes of a coin tossed 100 times. Compare the following two scenarios: (i) The subject predicts all the outcomes in advance, and it is discovered that after 99 tosses none are wrong. Let h be the prediction of the whole 100 tosses, and e the description of the outcomes of the first 99. (ii) the subject again 'predicts' h, but only *after* learning e. We should be inclined to repose more trust in the prediction h in (i) than in (ii), because we believe there is evidence that the subject in (i) has some privileged knowledge of the apparatus. By contrast, there is no evidence that the subject in (ii) has such knowledge.

Let us see how this reasoning can be expressed formally. Let H be the hypothesis that the subject in each case has privileged knowledge of the apparatus. Consider (i), where the fact that the subject predicts the statement h goes into background information (i.e., into what we knew before learning e). Assuming that H has a non-zero prior probability, however small, it will have that probability raised to somewhere close to 1 by e. For given that background information includes the information that e was predicted, $P(e \mid \sim H)$ can be interpreted as the probability that the subject got e right by chance, and for the whole 99 outcomes this probability will be exceedingly small. By Bayes's Theorem $P(H \mid e)$ is therefore close to 1. Now H entails h which entails e, since background information includes the information that the subject, who is right according to H, predicted h which implies e. It follows that

$$P(h \mid e) = P(H \mid e) + P(h \mid \sim H \,\&\, e)P(\sim H \mid e),$$

and so $P(h \mid e)$ will also be close to 1.

In (ii), the background information relative to which P is computed is now that the subject knows the outcomes of the first 99 tosses and predicts that the next toss will be a head (say). Relative to this information, H no longer implies h or e, though $H \,\&\, e$ entails h. The probability of e conditional on H and on $\sim H$ is the same, we can suppose, as its probability conditional on the hypothesis of chance. This being so, it is straightforward to show by Bayes's Theorem that $P(H \mid e) = P(H)$. Hence

$$P(h \mid e) = P(H) + P(h \mid \sim H \,\&\, e)P(\sim H),$$

which is approximately equal to $P(h \mid \textit{chance} \,\&\, e)$ if $P(H)$ is small (an extended discussion is in Howson and Franklin, 1991).

■ i THE PROBLEM OF SUBJECTIVISM

i.1 Entropy, Symmetry, and Objectivity

Possibly the most popular of all the objections to the subjective Bayesian theory is that it is too subjective. Fisher, in his remark which we quoted in Chapter 4, section **c.2,** that results concerning the measurement of belief "are useless for scientific purposes", summed up what many thought and still think to be a crucial objection. Science is objective to the extent that the procedures of inference in science are. But if those

procedures reflect purely personal beliefs to a greater or lesser extent, as they appear to do if they are constrained only to follow Bayes's Theorem, with no condition other than mere consistency being imposed on the forms of the priors, then the inductive conclusions so generated will also reflect those purely personal opinions. Echoing Fisher, E. T. Jaynes claims that

> the most elementary requirement of consistency demands that two persons with the same relevant prior information should assign the same prior probabilities. Personalistic doctrine makes no attempt to meet this requirement . . . the notion of personalistic probability belongs to the field of psychology and has no place in applied statistics. Or, to state this more constructively, objectivity requires that a statistical analysis should make use, not of anybody's personal opinions, but rather the specific factual data on which those opinions are based. (Jaynes, 1968, p. 228)

Jaynes developed his ideas well beyond the programmatic stage. One of his most influential suggestions is embodied in the *Principle of Maximum Entropy:* the distribution allegedly containing no information beyond that contained in 'the specific factual data' is the distribution, where it exists, which has maximum *entropy* subject to the constraints imposed by those data. Formally, the entropy of a discrete distribution is $-\Sigma p_i \log p_i$ ($0 \log 0$ is equal to 0). Intuitively, the entropy of a distribution is its degree of diffuseness, so that the more concentrated it is, the smaller is its entropy. The uniform distribution, $p_i = n^{-1}$ is the maximum entropy distribution for the constraint that X takes n values.

The principle of maximum entropy turns out to be far from unproblematic in practice (and also in principle, but we shall come to that shortly). For many sets of constraints, maximum entropy distributions do not exist, and where they do, they can lead to strongly counter-intuitive results (Shimony, 1985). Also, there is a technical problem in extending the principle to the continuous case: if X is continuously distributed, then the limit of the sequence $-\Sigma p_i \log p_i$, obtained by chopping up the range of X into smaller and smaller subintervals such that p_i is the probability that X lies in the ith subinterval, is infinity. Shannon, who introduced entropy as a measure of uncertainty, defined continuous entropy as the integral from minus infinity

to plus infinity of $-P(x)\log P(x)$, where $P(x)$ is the probability density of X. This quantity is not the limiting case of the discrete entropy, however, nor (as is well-known) is it invariant under change of variable (in fact, it is the expected value of the logarithm of a probability *density,* or probability per unit of X).

Jaynes was well aware of the technical problems posed by extending the entropy functional to continuous distributions, and he proposed a version of the cross-entropy we encountered in Chapter 6 (Jaynes, 1983, p. 59), though, containing as it does a strange point-density function, it can scarcely be considered a more successful solution than Shannon's. Jaynes has, however, advanced a quite distinct prior-determining criterion for continuous distributions. This follows an earlier idea of Jeffreys's, which was to choose that prior for a given problem which satisfies suitable invariance conditions. The invariance conditions suggested by Jaynes are those allegedly implicit in the problem at hand. For example, if nothing is specified about the position of some parameter other than that it lies somewhere on the real line, then according to Jaynes this means that the prior distribution over its possible locations should be translation-invariant. The resulting distribution turns out to be uniquely specified by this condition, and it is the uniform density distribution over the line.

That density does not, however, integrate to one; it diverges. So does the density x^{-1} of the distribution, claimed to be obtained by assuming scale-invariance, for a variable X which takes non-negative values only. Such distributions are called *improper,* and though they infringe the axiom of the probability calculus which demands that the probability of certainty is one, they are nevertheless considered legitimate by many, usually for the reason that they can be regarded as handy approximations of 'proper' distributions. It is also often cited in their defence that in certain cases improper priors can be combined in Bayes's Theorem with ordinary likelihoods to generate perfectly proper posterior distributions. This is true, but it doesn't justify their use. Improper priors are strictly inconsistent with the probability axioms, and there can be no guarantee, therefore, that they will not lead to contradictions elsewhere, as the so-called marginalisation paradoxes discovered by Dawid, Stone, and Zidek (1973) bear witness.

There are other difficulties with using the symmetries implicit in the problem at hand to determine the appropriate prior distribution. Determining exactly what these symmetries are is by no means as objective a matter as it is made out to be. Even having determined what you think they are, the resulting constraints may fail to determine a distribution uniquely (underdetermination) or at all (overdetermination: the constraints are inconsistent). Even the apparently straightforward requirement of scale-invariance does not unambiguously determine the density x^{-1} (Milne, 1983).

Despite the technical and other problems to which they frequently lead, Jaynes's ideas have been very influential, and the ideal that he enunciates, of conjuring up a prior distribution which, in the context of a given problem, expresses only the 'specific factual data', continues to inspire. Thus Rosenkrantz, an enthusiastic supporter of Jaynes, defends the uniform prior distributions that often arise within Jaynes's theory by pointing out an analogy with current cosmological practice:

> Steady-date cosmologists, to take one of myriad instances, start off by assuming the laws of physics are the same in temporally and spatially remote regions of the universe. This, they urge, is surely the simplest assumption. But it is more than that. To assume that different laws obtained a billion years ago would be entirely arbitrary; it would be to import knowledge we do not in fact possess. (Rosenkrantz, 1977, p. 54)

But we don't know that the laws were the same either. Rosenkrantz has failed to see that *any* assumption 'imports knowledge', the assumption that things were essentially the same just as much as the assumption that they were not. So it is with a uniform or any other prior distribution, as we argued at length in Chapter 3. And this objection strikes at the heart of Jaynes's programme. *No prior probability or probability-density distribution expresses merely the available factual data; it inevitably expresses some sort of opinion about the possibilities consistent with the data.* Jaynes's and Rosenkrantz's 'objective' priors may well not embody the opinions of any one person; but they are as far removed from the data as if they did.

Nor, it seems to us, is it a tenable claim that the distribution which maximises entropy is "the one which is maximally

noncommittal with regard to missing information" (Jaynes, 1957, p. 623). Any distribution, in our opinion, is as informative as any other insofar as it supplies a definite probability to every Borel set. In particular, the flat distributions determined by the Principle of Maximum Entropy, where there are no constraints other than those supplied by the probability axioms, are not less committed than any other; they merely have a different shape.

i.2 Simplicity

Another frequently canvassed 'objective' criterion for determining priors, at any rate up to a rank ordering, is that which states that hypotheses should be ordered in prior probability according to how *simple* they are. There is no claim to informationlessness for the priors admitted by this criterion, however. Jeffreys, a working scientist when not writing works on general methodology, famously incorporated such a Simplicity Postulate into his Bayesian theory. But he did so, not because he felt that he was thereby not trespassing beyond the data, but because he believed that simplicity was, as a matter of empirical fact, the dominant factor in people's prior evaluation of theories.

That this is so is doubtful, however, if only because it is notoriously difficult to know exactly what people refer to under the name of simplicity, and there seems no good reason to suppose that they are all referring to the same thing. Some people, for example, maintain that simplicity resides in an organic unity exemplified by the fundamental principles of the simple theory. Others say that it resides in the fewness of the adjustable parameters which the theory introduces (this was Jeffreys's view). Yet others say it resides in the ease with which computations can be done within the theory. Some, like Einstein on occasion, when they cite simplicity as a principal virtue, seem to appeal to some inarticulable personal aesthetic. But all these notions, even where any clear sense can be made of them, are to a great extent independent of each other.

And it is certainly not always easy to make clear sense of them, even of the apparently perspicuous (and popular) idea that a linear hypothesis is simpler than a nonlinear one, for whether an equation is linear or not depends on the choice of coordinates: the equation of the unit circle is the quadratic

$x^2 + y^2 = 1$ in Cartesian coordinates, but it takes the linear form $r = 1$ in polar coordinates, to choose a well-known example. Even Jeffreys's own coordinate-independent characterisation of simplicity as paucity of independent parameters is, on inspection, far from ambiguous. Newton's theory, for example, might be thought to possess very few undetermined parameters—some people claim that it contains only one, the gravitational constant. But as applied, say, in the kinetic theory of gases, it contains of the order of 10^{23} undetermined parameters, and when further degrees of freedom are added, the number rises correspondingly. (On the other hand the charge, made by Popper [1959a] and recently repeated by Watkins, that Jeffreys's Simplicity Postulate is inconsistent is not true and arises from misunderstanding what the postulate says [Howson, 1988]).

Even if there were a univocal notion of simplicity, and even if people were invariably to favour simpler hypotheses, this would still not, we believe, warrant the adoption of a Simplicity Postulate as an axiomatic component of our Bayesian theory. For reasons we have already been at pains to emphasise, we believe that the addition of *any* criterion for determining prior distributions is unwarranted in a theory which purports to be a theory of consistent degrees of belief, and nothing more. How you should determine your prior distributions is simply not something that comes within the scope of such a theory; nor, as we shall argue shortly, is it even desirable that it should.

The alternative to including rules for determining priors is *not,* however, as Jaynes, Fisher, and others (including, regrettably, even some Bayesian personalists) believe, a Bayesian theory condemned to be no more than a record of the whims of individual psychology. That quite fallacious inference has been possibly more damaging to rational methodology than any other. The charge of excessive subjectivism rests largely, as we have seen, on the fact that certain quantities, specifically the prior distributions, are not determined within the theory. But this no more means that the theory is subjective than it means that deductive logic, which does not determine the truth-values of the premisses of a deduction, is unduly 'subjective'.

The analogy with deductive logic is entirely appropriate here. Deductive logic is both a theory of (truth-value) consistency, and thereby also a theory of deductively valid inferences from premisses whose truth-values are exogenously given. Inductive logic—which is how we regard the subjective Bayesian theory—is a theory of (degree of belief) consistency, and thereby also a theory of inference from some exogenously given data and prior distribution of belief to a posterior distribution. And as far as the canons of correct inference are concerned, neither logic allows freedom to individual discretion: both are quite *impersonal* and objective.

Nor is the subjective Bayesian theory empty of definite and valuable information. Quite the contrary: the emphasis of the preceding chapters has been very much on just how great is its power to explain and justify methodological practice. In addition it does, to as great an extent as is desirable, incorporate Jaynes's requirement that "two persons with the same relevant prior information" assign the same prior probabilities. It does so asymptotically, as their data garnered from experience grow without bound. Even then, as we point out in Chapter 14, it characteristically does not take all that much sample data to diminish the different distributions to the point where they are practically identical. Experience is allowed to dominate prior beliefs, in other words, though in a controlled way; disagreement is not eradicated at once, but its effect usually falls off quickly.

All this is as it should be. People do have diverse opinions, and this diversity is a source of strength, not weakness, just as—to employ a metaphor Popperians will appreciate—diversity in a gene pool is a source of strength, a token of its ability to adapt quickly to a changing environment. That such diversity is explicitly allowed for in the personalist Bayesian theory is therefore a point strongly in its favour, not against it. The prescription of the same 'objective' prior probability for everybody in the same knowledge state is a prescription for stagnation and eventual catastrophe, as is the suppression of dissent quite generally. A tolerance of diversity enables a society of theories as much as of people to anticipate unexpected developments: some 'crank' or other, like Einstein, will have foreseen them.

■ j PEOPLE ARE NOT BAYESIANS

In their summary of an influential piece of empirical work, Kahneman and Tversky deliver themselves of the following judgment:

> The view has been expressed . . . that man, by and large, follows the correct Bayesian rule, but fails to appreciate the full impact of evidence [they cite W. Edwards, 1968], and is therefore conservative.
>
> The usefulness of the normative Bayesian approach to the analysis and the modeling of subjective probability depends primarily not on the accuracy of the subjective estimates, but rather on whether the model captures the essential determinant of the judgment process. The research discussed in this paper suggests that it does not. . . . In his evaluation of evidence, man is apparently not a conservative Bayesian: he is not Bayesian at all. (Kahneman and Tversky, 1972, p. 46)

It has been the burden of the foregoing chapters that a Bayesian theory is capable of explaining standard modes of scientific inference where other theories are not. Yet the empirical studies Kahneman and Tversky refer to are taken by these authors to indicate very strongly that people do not use Bayesian reasoning where the Bayesian theory appears to say that they should.

Other studies seem to reinforce Kahneman and Tversky's conclusions. Let us briefly consider a recent one (Cosmides and Tooby, 1992). A questionnaire informs respondents that disease D has a prevalence of 0.1% in the population. Test T to detect D has a false positive rate of 5%. What is the chance of someone having D, given that they have tested positive?

A majority of people gave the answer 95%. It turned out that among these were some who did not understand what 'false positive' meant. The false positive rate is the percentage of those who do not have the disease who test positive. Also omitted from the original questionnaire was the information about the true positive rate (100%). With this information supplied, the allegedly correct answer for the chance that someone does have D who tests positive is slightly under 2%, and it is obtained by transforming the data into probabilities in the following way. Let e be 'a person tests positive', and d be 'they have D'; if $P(e \mid d) = 1$, $P(e \mid \sim d) = 5\%$, and $P(d) = .1\%$, then $P(d \mid e)$ is easily calculated by Bayes's Theorem to be approximately $\frac{1}{51}$.

Even when the additional information about the meaning of 'false positive', and the value of the true positive rate, had been supplied, a majority of respondents still gave much too high a figure for their answer. This seems to be symptomatic of a widespread phenomenon, first identified by Kahneman and Tversky, of so-called 'neglect of the base rate'—i.e. of the tendency to ignore the incidence in the population of the disease D. However, readers of Chapter 13 will be aware that the identification of 'single-case' probabilities (which is presumably what is meant by asking for the chance that someone has D) with frequency probabilities (which is presumably what the percentages are intended to represent) has long been and still is a subject of controversy. Add in the continued puzzlement about the relation between conditional statements involving probabilities and conditional probabilities (Lewis 1976), and the fact that informed opinion on the correct probability model(s) to apply to experience is still far from unanimous, and it is hardly surprising that untrained people have difficulty in obtaining the 'correct' answers to problems like the one above (for a similarly sceptical discussion see also Levi, 1985).

Cosmides and Tooby themselves claim that results like the ones they record do not at all show that people are not Bayesian reasoners (though by 'Bayesian' they mean only 'capable of using Bayes's Theorem'; they certainly do not mean 'Bayesian' in the sense which it is given in this book). Pursuing a line of inquiry stemming from Gigerenzer (1991), they vary the test format to show that if the questions asked, and the information supplied, are put in terms of frequencies in populations of a specified size—e.g. "out of 1000 people who are perfectly healthy, 50 of them test positive for the disease" (p. 59)—then the vast majority of respondents do obtain the 'correct' answer 1 in 51. But if caution is advised over Kahneman and Tversky's interpretation of their results, then it is positively urged for Cosmides and Tooby's interpretation of theirs, which is no less than that people do possess a faculty for probabilistic, even Bayesian, reasoning, which is manifested however for a purely frequentist concept of probability.

A more accurate conclusion is that their respondents are competent at whole number arithmetic, which is anyway hardly surprising in view of the fact that they are often university students. But with *probabilistic* reasoning, and especially with reasoning about frequency probabilities, Cosmides and Tooby's results have very little to do at all, despite their dramatic claims. As Chapter 13 also demonstrates, the relation between sample frequencies and frequency probabilities is far from direct, and can only be forged successfully within a theory of non-frequency probabilities. To claim that people are statistical reasoners because they can get the right answer to the disease problem when it is posed in terms of numbers in an actual finite reference population is fallacious, and the data tells us little except, as we have said, about the subjects' ability to do arithmetic.

It does seem, however, that people do not always follow Bayesian canons of reasoning. In fact, we should be extremely surprised if subjects were invariably to employ impeccable Bayesian reasoning, even where the calculations were mathematically tractable. It is instructive to compare Kahneman and Tversky's, and others', apparently negative results with a rather striking and very uniform result (one of the present authors has tested it himself on a group of American freshman and sophomore students) of an experiment, devised by P. C. Wason (1966) to test subjects' performance of a simple deductive task. Four cards are placed flat on a table. Each card has an integer printed on one face and a letter on the other. The uppermost faces of the cards are

and the subjects are asked to name those cards, and only those cards, which need to be turned over in order to determine whether the statement, 'if a card has a vowel on one side, then it has an even number on the other', is false or not. Wason discovered that the vast majority of his subjects indicated either the pair of cards E and 4, or only the card 4. The correct answer is, of course, the pair E and 7.

This empirical result has proved to be remarkably persistent:

> Time after time our subjects fall into error. Even some professional logicians have been known to err in an embarrassing fashion, and only the rare individual takes us by surprise and gets it right. It is impossible to predict who he will be. This is all very puzzling. . . . (Wason and Johnson-Laird, 1972, p. 173)

Puzzled Wason and Johnson-Laird may be, but about one thing they are clear: these subjects did get the answer wrong. Moreover, even the subjects themselves eventually agreed on that. Now this observation has an obvious relevance to Kahneman's and Tversky's dramatic claim, made in the light of evidence analogous to Wason's, that we are not Bayesians. Wason has shown, by this and other empirical studies, that we are not consistently deductive logicians in practice—but he has not shown, nor did he claim to have shown, that we are not deductive logicians in some other important sense. For we ourselves nevertheless constructed those deductive standards and consciously attempt to meet them, even though we sometimes fail, and in some cases nearly always fail. By the same token, it is not prejudicial to the conjecture that *what we ourselves take to be correct inductive reasoning* is Bayesian in character that there should be observable and sometimes systematic deviations from Bayesian precepts.

■ k THE DEMPSTER-SHAFER THEORY

k.1 Belief Functions

The personalist Bayesian theory is a mathematical theory of uncertain reasoning. It is not the only one, however, and the last twenty years have seen the emergence of several others. With the rapid development of artificial intelligence, and in particular of rule-based expert systems, it has become a matter of some practical urgency to find an acceptable mathematical model of uncertainty, and the supply of candidates has been at least commensurate with the demand.

One of the most influential rivals to the personalist Bayesian theory is that of *belief functions* put forward by Glenn Shafer (1976), and it is usually referred to as the Dempster-Shafer theory, because one of its central principles is Demp-

ster's rule for combining evidence. Because the Dempster-Shafer theory derives much of its current support from Shafer's influential objections to the personalist Bayesian theory as a theory of evidence, we shall outline the Dempster-Shafer theory, argue that it itself is inadequate, and then proceed to consider—and rebut—Shafer's criticisms of the Bayesian theory.

The fundamental idea of the Dempster-Shafer (DS) theory is the equation of degree of belief with evidential support. Formally, it takes expression in a class of functions, called *basic probability assignments,* defined on the subsets of what is called a *frame of discernment,* denoted by Θ. The latter corresponds to a class of exclusive and exhaustive hypotheses. The basic probability assignment to a subset A of Θ is constrained to lie in the closed unit interval, and is intended to be that amount of belief you commit to A (which represents the *disjunction* of all its constituent hypotheses; the subsets correspond to the $M(a)$ in Chapter 2, section **f**) but not to any proper subset of A, i.e., not to any stronger proposition. The empty set, representing a contradiction, is assigned the basic probability number 0 by any assignment, and the sum of all the basic probability numbers assigned to the subsets of Θ is 1.

Basic probability assignments, despite the name, are not formally probability functions, and only in the exceptional case that the singletons of Θ and nothing else are assigned non-zero basic probability numbers will the associated belief function Bel be. Bel(A) is supposed to represent the total belief assigned directly and indirectly to A (by every B such that B implies A) and is the sum of the basic probability numbers assigned to A itself and the proper subsets of A. It follows that Bel(A) also lies between 0 and 1 inclusive, that Bel$(\emptyset) = 0$ and Bel$(\Theta) = 1$ and that the basic probability assignment is uniquely determined by Bel. Those subsets of Θ assigned positive basic probability numbers are called the *focal elements* of Bel.

In this theory, the body of data causing your belief to be as it is, is not represented explicitly but only implicitly in the Bel function: Bel(A) is allegedly the total support given by the evidence to A. Since Bel is not required to obey the probability axioms, there is no systematic relation between Bel(A) and Bel(A^*), where A^* is the complement of A with respect to Θ (note that it is possible for A and its complement both to be assigned belief 0). But there is nevertheless a relation between A and A^* in the Dempster-Shafer theory, expressed in the function Pl$(A) = 1 -$ Bel(A^*). Pl(A) is the *plausibility* of A, and it is intended to represent the degree to which the evidence *fails* to support A^*. Since Bel determines and is determined by the basic probability assignment, it follows that so also is Pl, and expressing Pl in terms of that basic probability assignment, it is easy to see that Bel$(A) \leq$ Pl(A).

The pair [Bel(A),Pl(A)] is called a *belief interval* for A, and when (and only when) Bel = Pl, then it is easy to show that Bel is a probability function. The analogy with lower and upper probabilities seems very powerful, and indeed formally Bel and Pl are lower and upper probabilities determined by a closed convex set Π of probability functions; i.e., Bel(A) is the minimum of the values of $P(A)$ for every P in Π, while Pl(A) is the maximum of $P(A)$ for every P in Π (Kyburg, 1987). The converse is not true, however: it is not the case that every such family of probability functions determines a belief function.

The analogy breaks down more seriously when it comes to the method for updating on additional evidence employed in the respective theories. In the DS theory, this is achieved by Dempster's rule for combining the belief functions representing the distinct bodies of evidence; this may give a different overall belief function than that obtained by updating Bel by conditionalisation in a frame in which some subset B (representing the conditioning statement) becomes established with certainty. Dempster's rule states that if m_1 and m_2 are the basic probability assignments associated with two belief functions Bel_1 and Bel_2 over the same frame and generated by independent bodies of evidence, then the combined $m(A)$ is the orthogonal sum $\Sigma m_1(C)m_2(D)$ of products of basic probability numbers $m_1(C),m_2(D)$ of all pairs C,D of subsets of Θ whose intersection is equal to A. However, if there are two sets X and Y such that $X \cap Y = \emptyset$ and $m_1(X),m_2(Y) > 0$, then m would not be well-defined by this scheme, since $m(\emptyset)$ would be non-zero. These pairs must be omitted from the sum, which is then normalised by dividing by $1 - \Sigma\, m_1(X)m_2(Y)$ for every pair X,Y whose intersection is empty. The combined belief function $\text{Bel}_1 \oplus \text{Bel}_2$ is defined in the usual way from m. Note that it will not be defined when there is some A such that $\text{Bel}_1(A) = 1$ and $\text{Bel}_2(A^*) = 1$; this would mean that Bel_1 contradicts Bel_2, and

intuitively two such functions cannot be consistently welded into a new one.

$\text{Bel}_1 \oplus \text{Bel}_2$ is associative and commutative. When for some B, $m_2(B) = 1$, so that B and every proposition implied by it is rendered certain by the new evidence, then Bel_1 and Bel_2 are combinable if and only if $\text{Bel}_1(B^*) < 1$, and the combined function $\text{Bel}(A \mid B) = \text{Bel}_1 \oplus \text{Bel}_2(A)$ is such that

$$\textbf{(1)}\ \text{Bel}(A \mid B) = \frac{\text{Bel}_1(A \cup B^*) - \text{Bel}_1(B^*)}{1 - \text{Bel}_1(B^*)}$$

(1) is called *Dempster's Rule of Conditioning*. It easily follows that $\text{Pl}(A \mid B) = \frac{\text{Pl}_1(A \cap B)}{\text{Pl}_1(B).}$ Also, if Bel_1 is a probability function, then so is $\text{Bel}(. \mid B)$ and (1) reduces to ordinary Bayesian conditionalisation: $\text{Bel}(A \mid B) = \frac{\text{Bel}_1(A \cap B)}{\text{Bel}_1(B).}$ The contrast with lower and upper probabilities becomes apparent in the inequality

$$\min_{P \in \Pi} \text{P(A} \mid \text{B)} \leq \text{Bel}(A \mid B) \leq \text{Pl}(A \mid B) \leq \max_{P \in \Pi} \text{P(A} \mid \text{B)},$$

where Π is the family of probability functions determined by $\text{Bel}(. \mid B)$. Kyburg (1987) constructs a simple example where the outer inequalities are strict and $\text{Bel}(A \mid B) = \text{Pl}(A \mid B)$, though $\text{Bel}_1(A) \neq \text{Pl}_1(A)$.

k.2 What Are Belief Functions?

That is enough by way of exposition. Bel and Pl cannot, as the results of the previous paragraph demonstrate, easily be interpreted as lower and upper probabilities. How, then, is the formal apparatus of the DS theory to be interpreted? Shafer has, of course, already provided an answer of sorts to this question: $\text{Bel}(A)$ is a measure of the degree to which a contextually implied body of data is evidence in favour of A, and Dempster's Rule of Combination then allegedly tells us how evidence from distinct sources combines to support each proposition in the frame of discernment.

The trouble with this answer is that it does not, as it stands, provide us with any means of assessing how adequately Bel and Dempster's rule fulfil these roles. Why should an adequate measure of evidence, and the belief it evokes, satisfy

the conditions imposed by Shafer? In logicians' terminology, we have a syntax but as yet very little in the way of a semantics for it. The contrast with the Bayesian theory in this respect could not be more marked, for in the latter, support is measured explicitly in terms of changes in consistent degrees of belief induced by clearly specified evidence, and it is *demonstrable* that consistent measures of belief have the formal structure of probabilities. You may not want support to be defined in this way, but at least you know why, if it is, it must have the formal structure it has; for that structure is then unambiguously determined.

Shafer has, however, attempted to provide a more definite key to understanding his theory in terms of what he calls a set of *canonical examples*. These are randomly coded messages. You receive a coded message, and you know that the codes are chosen at random from a specified set, where the chance of the ith code being chosen is p_i (Shafer, 1981a). The decoded messages are all of the form 'The true hypothesis is in A' for some A, where A is a subset of some frame of discernment, so with probability p_i the true message is 'The truth is in A_i'. Let $m(A)$ be the total chance that the message was 'The truth is in A', i.e., $m(A)$ is the sum of all p_i such that $A_i = A$. $\text{Bel}(A)$ is therefore the chance that the message *implies* that the truth is in A. We are also asked to assume that the true message is itself true and that the coded message is our only evidence, so that $\text{Bel}(A)$ measures our degree of belief that the truth is in A. According to Shafer, "Our task, when we assess evidence using belief functions, is to choose values of $m(A)$ that make the canonical 'coded-message' example most like that evidence" (1981a, p. 6).

Now suppose that two messages are transmitted in such a way that the code selected for the first is probabilistically independent of that selected for the second. If $\text{Bel}_1(C)$ and $\text{Bel}_2(D)$ correspond to the chances that the first message implies that the truth is in C and that the second implies that the truth is in D, then $\text{Bel}_1 \oplus \text{Bel}_2(A)$ turns out to give the chance that the two messages jointly imply that the truth is in A (the multiplication of the m-values in the orthogonal sum results from the codes being chosen independently).

It is no doubt a picturesque idea that bodies of observational data are randomly encrypted messages sent by God or

Nature, but in the late twentieth century it hardly provides a convincing justification for the syntax of belief functions. If this is all that can be mustered to support its adoption, then it is hardly enough to warrant the considerable degree of acceptance that the theory has received by workers in artificial intelligence. In a variant semantics recently advanced by Pearl (1990), the randomly encoded messages are replaced by randomly selected theories, and Bel(*A*) is equated with the total chance that *A* is a consequence of the theory selected. No direct significance is given to the basic probability assignments. But this account is really no advance in plausibility on Shafer's own (which theories are in the population of theories? who put them there? who or what arranged the selection mechanism?).

These scenarios, we submit, are not to be taken seriously. Nor do we believe they are by those who nevertheless take the DS syntax seriously. They do so because they accept at their face value the objections brought by Shafer against the only worked out and plausible alternative, the Bayesian theory, and which his own theory is expressly designed to avoid. We shall show that these objections are either based on desiderata for a theory of evidence that in fact no adequate theory should embody, or else they arise from a misunderstanding of what the Bayesian theory actually says.

Let us deal with the second point first. Just as he adduces 'canonical examples' designed to make his own theory intelligible, Shafer adduces them also for the Bayesian theory, and then points out that they are inappropriate for a theory of belief based on evidence. The 'canonical examples' for the Bayesian theory, according to Shafer, are states of affairs chosen by a chance mechanism (1981). It is these chance mechanisms that allegedly support the probability distributions over the hypothesis-evidence spaces. Now while it is true that any probability distribution *can* be modelled by a suitable arrangement of random drawings from urns (for example), or limiting cases based on these, it does not follow that they must be, and in the personalist Bayesian theory they certainly are not. At the risk of stating the truth too often, we repeat that the probability distributions over hypotheses are degree-of-belief distributions and nothing else.

Shafer does nevertheless recognise this fact but maintains that it represents a retreat "from the idea that the prior

Bayesian belief function is based on any particular evidence" (1976, pp. 26–27). It is not entirely clear what this is intended to mean. If it means that Bayesian priors may reflect very diverse bodies of accumulated evidence, then it is true also of Shafer's own belief functions. The only retreat that we can think of which may be what Shafer is referring to, is the abandonment of any attempt to represent current epistemic probability distributions as posterior distributions obtained by conditioning successively on each piece of evidence thrown up during the agent's lifetime, relative to an original prior distribution purporting to represent total ignorance.

That such a programme existed is true: it was Carnap's, as we know. It was abandoned, and rightly, because the choice of any distribution prior to all experience would of necessity be quite arbitrary, as also would be the choice of language within which to represent all the diverse 'deliverances of experience'. But the personalist theory has never professed such grandiose and unattainable ambitions. It is merely a theory of inference from prior to posterior distributions, as we keep stressing, and it is unfair, to put it mildly, to charge it with retreating from a position it never occupied and never wanted to occupy. What Shafer fails to point out is how close his own theory actually is in this respect to the personalist theory. Both incorporate prior distributions which each theory represents simply as given, and both incorporate the effect of new evidence *given* those priors. They just do it in different ways.

k.3 Representing Ignorance

The mention of *ignorance* brings us to the most influential of Shafer's objections, and it is an objection to *any* probabilistic attempt to represent uncertain reasoning. The objection is that no such theory can adequately represent ignorance between alternatives. This charge appears to carry weight, in view of the well-known difficulties with the Principle of Indifference, difficulties on which Shafer himself dwells. He points out that the natural way to represent total ignorance is to assign null belief to each of the alternatives stated, and that in his theory, this is possible because there are belief functions that assign the value 0 not only to each atom of the frame but also even to *every* proper subset. In a probabilistic theory ignorance can, it seems, be expressed only by a uniform

distribution of probabilities over atoms. Shafer sees two problems with this. (i) It might be that an atom in one frame is composite in another; e.g., a counter may be classified as blue or non-blue, or it may be classified as blue, green, red, etc. Yet a uniform distribution over one classification leads to a skewed distribution over the other, as we know. (ii) You are not allowed to be ignorant about $\sim h$ whatever your state of belief about h; for the probability of $\sim h$ is fixed once you have determined that of h.

Our answer to these objections is very simple: it is that the notion of pure ignorance *is not well-defined,* and it is a virtue, not a vice, of a probabilistic theory that it brings this fact out very clearly. Such a theory shows you quite explicitly that you cannot be *uniformly unopinionated* between all possible sets of alternatives, as we saw in Chapter 4, and this is surely intuitively correct. If you have equal degrees of belief in each of the numbers from 0 to 10 being called, then obviously you cannot, or at any rate certainly should not, have equal degrees of belief in the propositions '0 will be called' and 'A non-zero number will be called'. But in Shafer's theory you can, and that is why Shafer's theory is *false* as a theory of reasonable belief.

That an increase in belief for h is automatically a decrease in belief for its negation is also, we believe, a very basic intuition; it is, of course, an immediate consequence of the probability axioms. That Shafer's denial of it, and the embodiment of that denial in his theory, should be taken as a reason for rejecting the Bayesian theory rather than his own we find perplexing. In our opinion, the difficulties in giving any coherent interpretation of the syntax of the Dempster-Shafer theory merely reflect the fact that no coherent interpretation exists. The theory is radically at odds with sound reasoning, and all of Shafer's attempts to convict the Bayesian theory of that offence merely have the unintended consequence of convicting his own.

■ I EVALUATING PROBABILITIES WITH IMPRECISE INFORMATION

It is often said that one of the factors standing in the way of a widespread use of the Bayesian theory is that the component

probabilities—the likelihoods and especially the priors—of a Bayes's Theorem calculation are often not readily computable, because the data are too vague, or too numerous and diverse, or all these things together. This is a common complaint of workers in artificial intelligence, among others, who want a neatly packaged set of algorithms to apply, and who consequently tend to favour classical techniques (which are now frequently produced in the form of easily usable software packages), or even to invent their own (like the MYCIN and EMYCIN calculi of certainty factors).

Much in fact has been done to render Bayesian techniques more algorithmic. Bayesian software is now produced, there are Bayesian expert systems, and the recent and very promising representation of probabilistic relationships in so-called Bayesian networks is a specifically computer-oriented development of the Bayesian formalism (Pearl, 1988, contains a comprehensive account). But citing this work should not divert attention away from the fact that the Bayesian theory is not set up primarily as a source of algorithms but as a general logic of consistent belief. It is no criticism of this function that there is a vast array of inferential problems for which no algorithmic solution exists and where fallible personal judgment will consequently play a major role.

Indeed, most problems of uncertain inference—whether a witness's testimony is reliable, what the weather will do tomorrow, how the disintegration of the Soviet Union will affect world politics in the next ten years, and so on—are of a characteristically diffuse kind. Even those which social scientists frequently have to deal with often involve no precisely characterised statistical or other mathematical model. Shafer tells us that applied statisticians who have tried to apply Bayesian method to problems in which no such model can be assumed have even found more difficulty in evaluating likelihoods than the notorious priors (1981b). In the circumstances stated, this is not at all surprising, but we do not infer from this that the Bayesian methodology is inapplicable. *On the contrary: the Bayesian methodology, being a general theory of uncertainty, is always applicable.* Indeed, it is indispensable.

It is surprising how readily a complaint about the world—the data it supplies are often complex and difficult to interpret—is turned into a complaint against a theory which acknowl-

edges its vagaries and into praise for one which does not. For example, that classical techniques appear to offer easily computable solutions where the Bayesian theory does not (because of its dependence on priors) is frequently cited as not merely a practical but a *theoretical* disadvantage of the latter. The fact is, however, as we have often emphasised, that classical procedures can do this only because they systematically ignore relevant information that the Bayesian conscientiously attempts to represent.

And the way such information is represented will be a matter of personal judgment (though expert systems have been developed whose prior probabilities are based on the pooled opinions of many specialists). To want some universal 'inference engine' into which you can feed data and which will duly output probabilities is a natural enough desire. Carnap, among others, tried to build one. His attempt failed, and if in this uncertain world there is one thing we can be certain of—and we can be certain precisely because the world is uncertain—it is that all attempts must fail. The complaint that the Bayesian leaves too much to individual judgment and not enough to formula is therefore fundamentally misconceived.

■ m ARE WE CALIBRATED?

We have already mentioned more than once the concept of a calibrated probability measure. The basic idea is simple. Suppose that S is an indefinitely continued sequence of propositions each describing the occurrence of a single event. For example, S might be a sequence of weather predictions for successive days. Let P be a subjective probability measure defined on a domain including the members of S. Select the subsequence S_r of those members of S assigned probability r conditional on all information gathered by the owner of P up to that point. P is *calibrated for* S if, for all r, the relative frequency of truths in S_r converges to r.

The calibration criterion is generalised in Dawid's classic paper (1982) to the condition that relative frequencies in any rule-selected subsequence of S converge to the average conditional probability of the propositions selected. S is now explicitly an infinite sequence, and Dawid, using results from the theory of martingales (a martingale is a mathematical representation of a fair game), derives the remarkable and surprising theorem that with P-probability one, P is calibrated for all sequences S.

Calibration is frequently held up as a criterion of how well your probability function matches or reflects empirical reality, and a deep-seated worry of many subjectivists is that the subjectivist theory clearly provides no guarantee of any such matching. Scoring-rules, of the type we discussed in Chapter 5, section **c.2,** are often suggested as a method of inducing better calibration by imposing penalties for miscalibration; indeed, they have been used with considerable success.

But Dawid's theorem tells us that we should be certain *a priori* that we are calibrated—when we know perfectly well that we may be very badly miscalibrated indeed. In other words, it seems to be true both (i) that we should recognise the very real possibility, if not likelihood, of miscalibration, and (ii) that we should nevertheless be certain that we shall be calibrated. Here there seems to be at best a paradox, at worst an outright contradiction, as Dawid himself pointed out in his paper. Can it be resolved?

The answer is—fortunately—that it can. As Joseph B. Kadane pointed out in the discussion of Dawid's paper (1982, p. 610), Dawid's theorem is a theorem about the properties of a conditional probability function *in the limit where complete information exists* (this is a condition of the theorem's validity). Thus it is really hardly surprising to learn that you ought to believe that your probabilities, conditional on the evidence to hand, will *eventually* converge to the empirical relative frequencies—although it remains of course logically possible that they will not, a fact allowed for in the convergence only occurring 'with P-probability one'.

Where the members of S are of the form '$X_i = 1$', where the X_i are a sequence of independent, identically distributed random variables, we already know that, given suitable regularity conditions, with probability one the relative frequencies will converge to the posterior distribution; i.e., the probability function is calibrated for such sequences. This is just a

consequence of the Strong Law of Large Numbers, and Dawid's theorem can therefore be regarded as an interesting extension of that result.

■ n RELIABLE INDUCTIVE METHODS

Hume has a famous argument (1739, Book I, Part III) which appears to show that inductive arguments will inevitably at some point presuppose what they set out to establish; they will be circular, in other words. As we saw in Chapter 4, section **c,** Carnap's λ calculus affords a striking vindication of Hume's thesis.

Hume's thesis has now achieved the status of conventional wisdom, though attempts are periodically made to circumvent his arguments. One of the more recent, due to Reichenbach (1949), exploits the possibility that an inductive method may generate hypotheses which demonstrably converge to the truth in the limit as data accumulate, *whatever those data might be.* The notion of an inductive method, or 'truth detecting paradigm', as it has been called, which robustly identifies the truth in the limit as the evidence accumulates, was taken up by Putnam, and more recently developed by Osherson, Glymour, Weinstein, Kelly, and others, and made the foundation of a new discipline, called *Formal Learning Theory.* One of the conclusions they draw from their researches is that the Bayesian theory can be formally demonstrated to be suboptimal in the class of all inductive methods. We shall examine this claim, but to do so means that we shall have to examine briefly first the basic ideas and theses of Formal Learning Theory itself.

The precise characterisation of an inductive method within formal learning theory varies from author to author, but the differences seem to be inessential. We can define (roughly following Kelly, 1990) such a method to be any function F which associates with a given hypothesis and a finite data sequence a truth-value. F is reliable, or successful, for a specified class K (usually required to be countable) of structures and hypothesis h if, for each w in K, F stabilises on the truth-value of h in w when given enough data about w.

Clearly, for any given K and h one of the first questions to

be asked is whether such a function exists. The question can be extended by considering additional constraints. A natural one is that the hypothesis conjectured true at any stage be consistent with the data. Another it might seem natural to impose, and which we shall consider later at some length, is that the method be effective, i.e. performable by a suitably idealised computer equivalent in power to a Turing machine.

Among the many interesting and deep results which have been obtained (see Glymour's survey article, 1991, for some of them), are those which appear to suggest that the Bayesian theory determines an inductive method which is demonstrably suboptimal. Whether this is so or not turns on two issues: (i) whether the Bayesian theory determines an inductive method in this sense at all; and (ii) whether, even if it could be made to do so in some acceptable manner, the condition of its being effective is a desirable one.

As to (i), Bayesianism does not, on the face of it, prescribe a procedure for outputting the truth-values of hypotheses in response to data inputs. It is a great advance, as we see it, that the Bayesian theory does not have to resort to these crude and misleading qualitative categories. However, formally speaking we could take the range of the function F to be the set $\{0,1\}$ rather than the set {true, false}, and say that F successfully identifies the truth just in case any hypothesis to which F eventually assigns only the value 1 is in fact true. The Bayesian theory can then be taken to determine an inductive method by setting $F(h) = 1$ if a suitable probabilistic condition is satisfied; e.g. if $P(h) > \frac{1}{2}$, or even if h is just the mode of the posterior distribution.

Granted some such definition, it can be shown that if any method at all reliably identifies the truth over all possible states of affairs left open by some specified background information, then there is a Bayesian method which does (Juhl, 1993, Theorem 2.1). But this happy situation appears to change when it is required that the Bayesian method be computable. Putnam had earlier shown by a diagonalisation argument that, given that the probability function P was Turing-computable, then it is possible to construct a hypothesis h such that *if h is true* then its so-called instance confirmation (the probability that the next observation will yield an

instance of h, conditional on the evidence up to that point) will never rise above one half (Putnam, 1975). So it looks as if the answer to the question whether there is always a reliable *computable* Bayesian learner is likely to be negative. Osherson, Stob, and Weinstein (1988), and, using a simpler argument Juhl (op. cit.), argue that it is, and the latter shows that there is always a reliable computable non-Bayesian learner.

How damaging to the Bayesian theory are results like these? In our opinion, not at all. In the first place, it is not clear exactly how a computable Bayesian should be characterised. Juhl's, and Osherson, Stob, and Weinstein's, negative results require that a computable Bayesian be one which has a computable procedure for selecting the mode of the posterior distribution. However, it is difficult to see how this is possible unless at the very least the probability function itself be computable. Yet this is an exceedingly, and unrealistically, strong condition to impose. We can easily see why by noting that no computable probability function can be strictly positive, in the sense of assigning the unconditional probability 0 only to contradictions. For a computable, strictly positive function would immediately provide a decision procedure for logical truth (truth in all interpretations of the extralogical items in the sentence), and a celebrated theorem of Church (1936) shows this to be impossible.

Computable probability functions are a very artificial and restricted class (In Gaifman and Snir's definability ordering for probability functions on a first order language (1980, §3), the computable functions are, in terms of what they ascribe probability 0 to, the most dogmatic). To restrict the Bayesians to these is rather like restricting mathematicians to arithmetical methods only in their proofs. Nor is this the only reason for disputing the inference to the suboptimality of Bayesian procedures from claims about what 'computable Bayesians' can't do. For it turns out that computable non-Bayesian procedures succeed where 'computable Bayesian' ones do not because the former allow hypotheses to be proposed at certain stages of data acquisition which are actually inconsistent with those data. This sort of liberalism is hardly a feature on which to base a claim to superiority.

Nothing in the preceding discussion, however, should be taken to depreciate the new and powerful discipline of formal learning theory, or the intrinsic interest and depth of its results. One in particular, due to Kelly, provides additional information about a well-known Bayesian convergence-of-opinion theorem. This theorem (Halmos, 1950, p. 215, Theorem B) says that if the hypothesis h is that the infinite sequence w, whose initial segment $s(w)$ is being observed, is in a set H of sequences, then $P(h|s(w))$ tends to 1 if w is in H and 0 if not, except on at most a class C of sequences of a priori probability 0. Kelly's result (1990, Theorem 2.2) implies that H occurs at a certain rather low level in the Borel hierarchy relativised to the complement of C (the level of a set in this hierarchy indicates the degree of quantificational complexity involved in defining it in terms of a basis consisting of those subsets of C determined by specifying only finitely many sequence coordinates).

■ o FINALE

One of the reasons why one expects deductive reasoning to exercise a more-or-less widely felt and obeyed constraint on the way people reason is because it is truth-preserving. Probabilistic reasoning also possesses a characteristic which authorises it to exercise no less a regulatory function: its rules, as we observed in Chapter 5, are broken on pain of committing inconsistency. It is, we suggest, for this reason that divergence from the norm set by the probability calculus is so widely regarded as deviant. Certainly, ever since people chose to express their uncertainty in terms of the odds they thought fair, they have felt themselves explicitly constrained by the axioms of the probability calculus, and while it was not until this century that it was explicitly proved that obedience to the calculus is a necessary condition for fairness, there can be little doubt that that result was taken for granted.

The discovery of the probability calculus, together with the usual formula connecting (fair) odds and probabilities, can now be seen to be part of the great scientific renaissance of the seventeenth century. The probability calculus became the foundation of a mathematical theory of uncertainty, of enormous potential scope and power, which simultaneously generated a quantitative logic of inductive inference and bound

together the new mathematical concept of probability with another developed at about the same time, utility, to produce a theory of rational action. The mathematicians of the eighteenth century, and to a lesser extent the nineteenth, divided their time between developing the new physics and extending the probability calculus and the theory of inductive inference and rational decision based on it: among these pioneers, Huyghens, James and Daniel Bernoulli, Laplace, and Poisson stand out as pre-eminent.

On the way, however, paradoxes began to appear in the programme, mostly connected with the Principle of Indifference but also—as a criterion of rational action—with the principle of expected utility. These problems, especially those within the theory of probability itself, seemed at one time, in the early years of this century, so intractable that many people, like Fisher and Popper (as we have seen), wrote off the account of probability on which the programme was based. But they were wrong: in the middle years of this century, shortly after Fisher and Popper penned their obituaries, secure foundations were finally laid. Ramsey first, and then von Neumann and Morgenstern, put utility theory on a consistent basis, and Ramsey and de Finetti realised that an adequate theory of epistemic probability can dispense with pseudo-objective principles like that of Indifference without giving up its claim to impose quite objective standards of consistency in reasoning involving such probabilities. The probabilities might be personal, but the constraints imposed on them by the condition of consistency are certainly not—a distinction still not widely grasped even today, and whose failure to be appreciated continues to vitiate so much contemporary discussion.

We have written this book in an attempt to convince believers in 'objective' standards in science that there is nothing subjective in the Bayesian theory *as a theory of inference:* its canons of inductive reasoning are quite impartial and objective. We want this simple truth to be more widely appreciated. We maintain also that this is the *only* theory which is adequate to the task of placing inductive inference on a sound foundation, and we believe that we have demonstrated this fact in this and the previous chapters.

Why, though, we have been asked, is there nothing in this book about *rational decisions?* If we now have in the personalist Bayesian theory a satisfactory foundation of epistemic probability, as we claim, and if the theory of utility is also now set on a satisfactory basis, why not include an exposition of what, admittedly, is a highly successful and fertile discipline, that of Bayesian decision theory—especially, it is often argued, since any adequate account of theory-choice must include reference to some theory or other of rational decision-making?

We choose to stop at this point because we are far from convinced that decision theory does have any useful, let alone crucial, role to play in a theory of *inference,* which is what we have been expounding in this book. In fact, we believe that it does not. The intellectual tools one needs to assess such things as empirical support, weight of evidence, and probability itself, have nothing to do with making decisions, in any but a purely trivial sense. In our view, decision theory presupposes these tools, not they it. And while Bayesian decision theory is a worthwhile and rewarding study, it is not one which we are concerned with here. If we have presented an acceptable account of scientific inference, we are satisfied enough.

■ EXERCISES

1. Show that $P(e^n) = P(e_n \mid e^{n-1})\mathrm{x} \ldots xP(e_2 \mid e_1)xP(e_1)$.
2. Show that if $h \vdash e_i$ for every i and $P(h) > 0$, then $P(e_n \mid e^{n-1})$ tends to 1.
3. A sequence of functions $f_n(x)$ is said to converge uniformly to $g(x)$ over a subset X of its domain if for any positive real number r, however small, there is a number m such that for all x in X and all $n > m$ the absolute value of the difference between $f_n(x)$ and $g(x)$ is less than r. Define c_{ki} as in section **c** and suppose h entails e_i for all i. Show that if $f_n(k) = P(c_{kn} \mid c_{k(n-1)})$ converges uniformly to the constant function $g(k) = 1$ over the set $\{k : k \geq 1\}$, then there is a number m such that $P(e_m \mid e_m - 1) > \frac{1}{2}$ and $P(\sim e_m \mid e_{m-1}) > \frac{1}{2}$.
4. Show that $P(\sim e \vee h \mid e) - P(\sim e \vee h) < 0$ if $P(h \mid e) < 1$.

5. Show that all the consequences of $\sim e \vee h$ are consequences of $\sim e$.

6. Suppose $h \vdash e$. Show

   (i) that $\sim e \vee h$ shares no non-tautologous consequences with e
   (ii) that $\sim e \vee h$ is the strongest statement which conjoined with e is equivalent to h.

7. Let $Ct(a) = 1 - P(a)$. Show that $ct(e \vee h) + ct(\sim e \vee h) = ct(h)$. Hence show that when h entails e, $ct(e) + ct(\sim e \vee h) = ct(h)$.

8. Define $S(h,e) = P(h \mid e) - P(h)$. Show

   (i) that $S(h,e) = S(\sim e \vee h,e) + S(e \vee h,e)$
   (ii) that $S(h,e) = S(\sim e \,\&\, h,e) + S(e \,\&\, h,e)$.

9. Let $Cn(a)$ be the set of consequences of a, and define $Cn(a).Cn(b) = Cn(a \vee b)$ and $Cn(a) + Cn(b) = Cn(a \,\&\, b)$ and $Cn(a)' = Cn(\sim a)$. Define a mapping $Cn(a)^* = M(a)^c$, where $M(a)^c$ is the set-theoretic complement of $M(a)$ with respect to $M(t)$, t a tautology. Show that * is an isomorphism, i.e., that $[Cn(a)']^* = Cn(a)^{*c}$, $[Cn(a).Cn(b)]^* = Cn(a)^* \cap Cn(b)^*$, and $[Cn(a) + Cn(b)]^* = Cn(a)^* \cup Cn(b)^*$.

10. Where h, X, and $P_m(h)$ and $E_m(X)$ are defined as in section **f**, show that $P_m(h) = E_m(X)$, $m \leq 1000$.

11. Show from the stated properties of basic probability assignments that $0 \leq \text{Bel} \leq 1$.

12. A *simple support function* f assigns the value s to every set in a frame of discernment Θ which includes some specified set A, 1 to Θ, and 0 to every other set in the frame, where $0 \leq s \leq 1$. Show that f is a belief function, with basic probability numbers $m(A) = s$, $m(\Theta) = 1 - s$, and $m(B) = 0$ for every other subset B of Θ.

13. Show how Bel determines a unique basic probability assignment.

14. Where Pl(A) is defined as in section **k.1**, show that $\text{Pl}(A|B) = \dfrac{\text{Pl}(A \cap B)}{\text{Pl}(B)}$.

15. Show that every probability function defined on the set of all subsets of Θ is a belief function. Exhibit a frame Θ and a basic probability assignment to the subsets of Θ whose associated belief function is not a probability function.

16. Show that a necessary and sufficient condition for a belief function to be a probability function is that $\text{Bel}(A) = 1 - \text{Bel}(A^*)$ for every A.