# Bayes's Theorem

Richard Swinburne

[Swinburne, Richard, 2002, “Bayes's Theorem”, British Academy.]

## Preface

The introduction to this volume and the four papers of Elliott Sober, Colin Howson, Philip Dawid, and John Earman are developed versions of papers given at a British Academy symposium on Bayes’s Theorem on Saturday March 10th 2001.). We have also added to the volume a short paper by David Miller dealing with the applicability of Bayes’s Theorem to physical probability, an issue not explored in the symposium papers but important for a full treatment of the applicability of Bayes’s Theorem. We have also added to the original paper by Thomas Bayes which introduced his theorem to the world, including the introductory and concluding sections of that paper written by Richard Price when he presented the paper to the Royal Society on 23rd December 1763. The volume is completed by a short historical introduction to the original paper, written by G. A. Barnard, originally published, together with the version of that paper printed here, in Biometrika 45(1958), 293-315.

## Contents

## Introduction*

Richard Swinburne

Kinds of Probability

Bayes's Theorem is concerned with probability. When from the seventeenth century onward people began to talk about things being probable in somewhat like modern senses and reflected on what they meant, sometimes they supposed that there was only one kind of probability and sometimes they supposed that there were two kinds of probability - one a feature of the physical world, and the other the probability on evidence that something was the case in the physical world. Among modern philosophers the distinction was made most sharply and influentially by Rudolf Carnap who called the former 'probability$_2$' and the latter 'probability$_1$'.[1] In my view Carnap seriously underestimated the number of kinds of probability, that is, the number of things which have usefully been called 'probability' in either ordinary language or technical literature, and which are susceptible to philosophical analysis and mathematical articulation.

Carnap's 'probability$_2$' is statistical probability. A statistical probability is simply a proportion in an actual class or in a hypothetical class, that is a class generatable by a repeatable process. Ordinary language expressions in which the probability is said to concern 'a' member, i.e. any member, of some class are naturally so interpreted. 'The probability of an inhabitant of New Hampshire in the year 2000 voting for the Republican presidential candidate', or 'the probability of a toss of this coin being heads' are naturally interpreted as assertions about actual proportions or proportions in a class generatable by some procedure, for example tossing the coin for some time. Where the proportion concerns an infinite class, there must normally (if the assertion is to have clear content) be an understanding of the order in which members of the class are taken; and the assertion is then to be read as claiming that when more and more members of the class are taken in that order (eventually) the proportion diverges less and less from the stated value.[2] Where the proportion concerns a hypothetical class of concrete objects (whether finite or infinite), there must be an understanding of what must remain constant and what is allowed to vary (and how) when new members of the class are generated. What is the probability of a toss of an evenly balanced coin with a head and a tail landing heads? Suppose the world is deterministic, that is which event occurs at a given time is fully determined by each previous state of the universe (within some interval of time). Then if we specify fully enough

the conditions in which each toss is to be made - the exact angle of toss, distance above the table of the toss, momentum imparted to the coin, distribution of air molecules in the room etc., the probability will be 1 (or almost 1) for many sets of conditions, and 0 (or almost 0) for many other sets of conditions. Which of these sets of conditions are we talking about? Normally, in ordinary talk, neither. We suppose that these conditions may vary. But how? In the proportions in which when people normally toss coins, they do actually vary. For example, if most tosses of recent years are made approximately equally often from within areas of equal angular width within a range of from $90^0$ to the floor to $30^0$, we suppose a similar variation in tosses when talking of the proportion of heads resulting from an infinite series of tosses. And so on for all the other circumstances which affect the result. Given this ordinary context, then the probability of heads is ½. I shall represent the statistical probability of a B being A by Pr (A□B). That there are statistical probabilities (at any rate for finite classes) is uncontentious. Detailed modern explications of statistical probability begin with John Venn's Logic of Chance (1866)[3] and continue through Richard von Mises's Probability, Statistics and Truth (1928)[4] and Hans Reichenbach's The Theory of Probability (1949).[5]

There is, however, another feature of the physical world which most of us believe to exist, and which is often called 'probability' but which Carnap did not discuss in his main book on probability. Most of us think that the occurrence of events is predetermined by prior causes, either totally or to some limited extent; and talk of "probability" is used to measure the extent to which some outcome is predetermined to happen. I shall call this measure a measure of physical (or natural) probability. It is the kind of probability analysed by a 'propensity' theory of probability, such as that of Karl Popper.[6] An event having a probability of 1, is it being predetermined to happen - that is, physically necessary; an event having a probability of 0, is it being predetermined not to happen - that is, physically impossible.[7] Intermediate values measure the extent of the bias in nature towards the event happening or not happening. To say that the probability now of this atom of $C_{14}$ decaying within 5,600 years is ½, is to say that given the whole state of the world now, it is not predetermined whether or not the atom will decay within 5,600 years, but that nature has an equal propensity to cause decay and to cause no decay within that time. Physical probability is relative to time - as the time at which or by which the event is predicted to happen or not to happen, draws near, so (if that probability is not 1 or 0) the probability of its occurrence may change. Most of us think that the physical probability of almost all macroscopic events is very close to 1 or 0, but many of us think that because of the indeterminism of Quantum Theory many microscopic events have intermediate degrees of physical

probability. (There are also, I must add, a few philsophers[8] who think, as did Hume, that there are statistical patterns in nature, but not underlying propensities in individual causes which produce them. But Humeanism is very much a minority view). Given the existence of physical probability, there are a few logical entailments between it and statistical probability. For example, if the physical probability of every A of being B at some time is 1, then the statistical probability of an A being B at that time will also be 1. There are only limited such entailments. For example, it does not follow deductively from the physical probability of each A being B having a certain particular value other than 1 or 0, that the statistical probability of an A being B will have any particular value at all. Even if the physical probability is ½, the very improbable may happen (even in the infinite sequence) and every A be B. However, plausibly, it will be very improbable (in the sense of logical probability to be delineated below) that this will happen, and we need an account of the criteria of logical probability which yields this result.

Carnap's 'probability$_1$' is a measure of the extent to which some proposition *e* (which may state some evidence) makes another proposition *h* (which may state some hypothesis) likely to be true. It has a value 1 when *e* makes *h* certain (that is, when given *e*, *h* is certainly true), 0 when *e* makes not -*h* (the negation of *h*) certain, and intermediate values as *e* gives intermediate degrees of support to *h*. The evidence may include evidence of statistical or physical probability; and the hypothesis may also concern statistical or physical probability. It is the probability of h on the total evidence available at some time, e, by which it is rational to be guided in our actions at that time. Most of us in our unphilosophical and unmathematical moments think that there are objective truths about whether such-and-such evidence makes such-and-such a hypothesis very probable, or only fairly probable or very improbable; though we doubt whether very precise values can often be given to this degree of probability. When this evidential (or epistemic) probability is understood as measuring the objectively correct degree of evidential support I shall call it logical (or inductive) probability. Detailed modern explications of evidential probability began with J. M. Keynes's A Treatise on Probability (1921).[9] Keynes supposed that (at any rate often and approximately) there are true values for the extent to which one proposition makes another one probable, and so his account is an account of logical probability.[10] Keynes's work was developed by Rudolf Carnap who sought in his Logical Foundations of Probability (1950) to give an explication of what was involved in what he regarded as an objective concept of 'evidential probability', his 'probability$_1$.'

Strangely (to my mind), many philosophers and statiticians in their philosophical and mathematical moments deny that there is such a thing as logical probability. Those of us who think otherwise allow that very precise values cannot often be given to the probabilities involved, that

the probabilities take very rough values and that often all we can say truly is that this hypothesis is more probable than that one on such-and-such evidence, or that this hypothesis is more probable on this evidence than on that evidence. Those of us who believe that there is such a thing as logical probability, usually think also that humans have roughly the same criteria as each other for assessing evidential support. We must, however, acknowledge that humans differ greatly in their ability to apply these criteria. Thus while the logical probability of *h* on *e* is 1 if *e* entails *h*, only someone who can recognize the entailment can see that the probability has that value. Many writers on probability, however, consider that there are no objectively correct standards for assessing the probability on evidence of particular hypotheses. Each person has their own standards; and so we cannot talk of logical probability but only of each person's subjective probability which is a measure of the extent to which that person treats some evidence as supporting some hypothesis. Detailed explications of subjective probability, and the logical constraints to which rational measures of it are subject began with the work of F. P. Ramsey (1926) and Bruno de Finetti (1930).[11]

Probability Axioms

While there are these various kinds of probability, almost all writers consider that the same axioms (re-expressed as relating different kinds of entities - classes or propositions) govern (or in case of subjective probability, ought to govern) both statistical and evidential probability (that is, both logical and subjective probability). These axioms were classically codified by Kolmogorov[12], but were taken for granted or stated in more or less the same form for two or three centuries previously. As axioms of statistical probability, they can be stated as follows: for all classes A, B, and C:

1. Pr (A∩B) ≥ 0
2. If B⊆ A, Pr (A∩B) = 1
3. If A∩B∩C = 0, Pr (A∪B/C) = Pr (A∩C) + Pr (B∩C)
4. Pr (A∩B∩C) = Pr (A∩B∩C) ÷ Pr (B∩C)

(3 and 4 do not apply if the class C has no members).

For finite classes, these are simple arithmetic truths; and they are naturally extended to the infinite domain.

Re-expressed as expressing relations between propositions, and P( | ) is an operator governing propositions, they become: for all propositions q, r and s:

1. P (q∨r) ≥ 0
2. If N (r -> q), P (q∨r) = 1

3. If N ¬ (q & r & s) (that is, not all three can be true together), P (qvr⏐s) = P (q⏐s) + P (r⏐s)

4. P (q & r⏐s) = P (q⏐r & s) P (r⏐s)

(3 and 4 do not apply if N ¬ s, that is if s is impossible.)[13]

Classes are individuated extensionally by their members, and so any two classes which have the same members are the same class. Propositions, however, are individuated by their intensions (what they mean) and not by the worlds in which they would be true. Thus 'the number of pebbles in my box is 3' and 'the number of pebbles in my box is √27' are different propositions, although they mutually entail each other and so are true in the same worlds. It follows from axioms 2 and 4 that if p and q are necessarily equivalent, for any r, P (q|s) = P (r|s), but (in order to have a kind of probability which takes account of all necessary equivalences) we need to add

5. If N (q ¬ r), P (s⏐q) = P (s⏐r).

In order to interpret these axioms as axioms of physical probability, that is, propensity to cause, we need to confine the atomic propositions[14] to ones denoting time-indexed total world states (events), and read the 'N' (necessarily) as 'of physical neccesity'. It is, however, disputed whether, even so, the resulting axiom system is satisfactory in view of the asymmetry of causation (that if Q is (part of) the cause of R, R is not (part of) the cause of Q), which I shall for simplicity's sake assume to hold in the form that effects must be later than their causes. For then if Q (denoted by 'q') is earlier than R (denoted by 'r'), P(q|r) will always equal 0 (for R can have no propensity to cause Q) and so therefore - it follows from the axioms will P(r|q) (unless P(q) = 0 in which case P(r|q) is undefined). So there cannot be any propensities at all! Hence, either we need a different axiom system for physical probability[15] or we need a more carefully expressed semantics for physical probability, in which, for example, all probabilities are time-relative. P(q|r) is then interpreted as the propensity at a particular time, say the present moment, for the world to develop in such a way that Q occurs if R occurs. That propensity will exist whether Q is earlier or later than R. This way of interpreting the axioms is expounded by David Miller in section 6 of this book.[16] There are different solutions of these kinds in the literature. There is, however, very widespread agreement[17] that the traditional axioms provide a satisfactory set for evidential probability (subjective or logical). For this purpose, the propositions may report anything whatsoever, and the 'N' is to be read as 'of logical necessity'. P(h|e) is the probability of h given e. If we use the notation to designate subjective probability, we can, if required, add a subscript to indicate whose subjective probability is being considered. (Some of the other contributors to this volume have represented evidential probability as a relation between the states of affairs designated by propositions rather

than as a relation between propositions, and so have expressed it in ways such as P(Q|R) rather than P(q|r).

The normal technique for justifying the axioms, treated as axioms of subjective probability, is to show that in some way your desires are inevitably not going to be satisfied unless you assess the probabilities that actions will achieve your desired goals in ways that are consistent with the axioms of the calculus. The simplest case of this arises with betting. If you judge that the probability of an event E is p and are prepared to make any bet for or against *p* at the corresponding odds, e.g. so that if you bet £x that E will occur, you win £($\frac{1-p}{p}$) x if E occurs

and lose £x if it fails to occur; and if you bet £x that E will not occur you win £($\frac{p}{1-p}$) x if E does not occur but lose £x if E does occur, someone can always make a “Dutch book” against you, i.e. bet with you in such a way that you will inevitably lose money - unless your probability judgments are consistent with the axioms of the calculus. So, if contrary to the axioms of the calculus you judge that in a two-horse race (with no ties allowed), both horses have a probability of _ of winning and so you bet £2 on each at 1-2, then you will inevitably lost £1 whatever the result of the race. This kind of argument has been extended so as to apply where people do not bet in a literal sense, but where the gains and losses resulting from their actions are measured in terms of their value to the agent rather than in monetary terms, so as to show that being guided by probabilities which do not conform to the constraints of the calculus will lead in various circumstances to inevitable loss. But, even if these extensions are successful, at most they seem to provide a constraint on the judgments of probability which a given agent should make at a given time. They seem to provide no reason why judgments of probability by a given agent at different times taken together should conform to the calculus. Subjective probability theorists have always had difficulty in justifying ‘conditionalization’ - moving from the conditional probability of h given e being p, when it is not known whether or not e, to the probability of h which ought to guide action, when e becomes known and forms the total available evidence, being p.

In any case, the only constraint to which subjective probability is subject is a coherency constraint of the kind described. Some of the craziest judgments of what makes what probable are consistent with the axioms of the calculus. It is perfectly compatible with the calculus, when having observed the sun rise innumerable times in the past (e) and given any other knowledge about the past (k), to judge that there is a 0.99 probability that it will never rise again (h) - to claim that P (h□e & k) = 0.99. So long as you are careful not to ascribe a value other than 0.01 to the probability that the sun will rise again, and so values summing to no more than 0.01 to the probabilities that it will rise tomorrow, it will rise the next day, and so on, your attributions are fully coherent. That leads

many of us to think that there are further constraints on attributions of value to probabilities other than conformity to the calculus, criteria for attributing correct values to some of the probabilities which will enable correct values to be attributed to other probabilities by means of the calculus. Such rules will yield the logical probabilities of propositions on various pieces of evidence as opposed to various people's differing subjective probabilities.

From where are we to get these criteria? From a study of many different examples of when we would regard a certain hypothesis as rendered very probable by certain evidence, less probable by other evidence, very improbable on yet other evidence, we can gradually distil criteria to which correct judgments should conform. We study the actual practice of scientists and historians and detectives, in those cases where we think that these experts made the right judgment. We then add innumerable further imaginary cases, and reflect on what would be right to say about which evidence renders which hypothesis how probable in each of them. Sometimes it will be clear that not merely certain comparative judgments but certain numerical judgments are correct. On the sole evidence that 500 out of 1000 tosses of this coin are heads, surely the probability that the next toss will be heads is 0.5. Not every investigator will agree in their judgments about all cases, but there is - I suggest - a very large measure of rough agreement about what makes what how probable. From all these examples we can extract general criteria for determining logical probability. These criteria will include conformity to the calculus - for we assess probability judgments which do not conform to the calculus as false, for reasons additional to those considered earlier.

## Bayes's Theorem

In 1763 Richard Price presented to the Royal Society an edited version of the paper of Thomas Bayes, entitled 'An Essay Towards solving a Problem in the Doctrine of Chances'.[18] Its principal result was interpreted by Price as providing 'a sure foundation for all our reasonings concerning past facts', that is, as a claim about evidential probability. It is on that form of it that subsequent theoretical interest has been focused and to that form which the contributors to this volume devote their attention. (However, as far as the text of Bayes's essay is concerned, he might as well or instead be thinking of his theorem as concerned with physical probability). Taking for granted the axioms which I have called 1, 2 and 5, Bayes states explicitly the axioms which I have called 3 and 4, and thence derives his 'Proposition 5', which is the theorem in his text the closest (in my view)[19] to what has subsequently been called Bayes's Theorem. As stated by Bayes it is concerned with the probability of a first event having happened given that a second event

has happened. Denoting the first one by r and the second one by s, and our prior or background evidence by k, Bayes's Theorem then reads:

$$P(r \mid s \,\&\, k) = \frac{P(r \,\&\, s \mid k)}{P(s \mid k)}$$

There is, however, given the axioms, no need for 'r' to denote a first event and 's' a second one. Allowing them to be any propositions at all, and naming them 'h' and 'e' after their most usual application ('h' for hypothesis and 'e' for new evidence) and using axiom 4 for the value of the probabiliy of a conjunction, we get: (given that $P(e \mid k) \neq 0$)

$$P(h \mid e \,\&\, k) = \frac{P(e \mid h \,\&\, k)\, P(h \mid k)}{P(e \mid k)}$$

(where $P(e \mid k) = P(e \mid h \,\&\, k)\, P(h \mid k) + P(e \mid \sim h \,\&\, k)\, P(\sim h \mid k)$.)

It is this which I shall henceforward call Bayes's Theorem. If we interpret it as concerned with evidence (e) and hypothesis (h), it says that the posterior probability of h (its probability, given e and k) equals its predictive power (P(e|h & k), often - rather misleadingly - called the 'likelihood of h')[20] multiplied by its prior probability (its probability given k alone) divided by the prior probability of e. The formula in brackets states that the prior probability of e is the sum of the probabilities of its occurrence in the different possible world states of h being true and ~h (not-h) being true. Most thinkers have no problem in regarding Bayes's Theorem as an acceptable theorem of subjective probability. The controversial issue is its range of application. Is there a 'logical probability' to which it applies; are there criteria for ascribing correct values to its terms? The view that there is such a logical probability which governs the traditional axioms of the calculus and so Bayes's Theorem, is sometimes called objective Bayesisanism. So are there such criteria? In all areas of inquiry, or only in some?

Bayesianism - Subjective and Objective

Any division of evidence between new evidence of observation (e) and background evidence (k) is arbitrary. But it is sometimes useful to make such a division in order to assess the probability of some hypothesis when we are taking certain aspects of the wider situation for granted. In considering a hypothesis (h) about how neon behaves at low temperatures, we may have both experimental evidence with respect to neon (e) and evidence about how other gases behave at low temperatures (k); and it may be helpful to call the latter 'background evidence' and see it as supporting the hypothesis about how neon

behaves. But the crucial problems about logical probability are most evident if we suppose that we do not have any contingent evidence from a wider field about how things behave in some narrow field. In that case *k* is some mere tautology, and P(h|k) and P(e|k) may be called the intrinsic probabilities of *h* and *e* respectively.

Even in this situation there are some extremely obvious criteria for ascribing values which would be accepted by everyone. These typically concern the value of P (e□h & k), when h is some hypothesis purporting to explain e. For example, the logical probability of a particular event on the sole evidence of a statistical probability, P (Bx□Ax & Pr (B□A) = p), is p. (The logical probability that this coin will land heads, on the sole evidence that it is tossed and that 60% of tosses of this coin so far have landed heads is 0.6). Also, on the sole evidence that there is a physical probability of *p* at *t* that *x* will be B at (t+1), the logical probability on that evidence alone that *x* wil be B at (t+1), is also p.[21] But many doubt whether there are any objective rules for ascribing intrinsic probabilities. In the main text of Bayes's paper, when he discusses the example of a ball thrown on to a table or plane he makes an explicit assumption of prior probability - that 'there shall be the same probability that [a ball] rests on any one equal part of the plane as another, and that it must necesarily rest somewhere upon it'. He provides at that place no criteria for reaching judgments of prior probability. But in the Scholium he tells us to calculate prior probability on the basis of a principle of indifference - which we may express, more precisely, that we should assume (in the absence of experiment) for any variable about which we know nothing except that it lies between certain possible values, that it is equally probable that it will lie between any one interval of a given length as between any other interval of that length. But, as various paradoxes show, different ways of describing a set-up suggest different equiprobable ranges. Take the von Kries paradox, as stated by Keynes.[22] Suppose the specific volume of a liquid to lie between 1 and 3, and so its specific density to lie between 1 and _. A principle of indifference applied to the former would suggest that the specific volume is equally likely to lie between 1 and 2 as between 2 and 3; and so - equivalently - that the specific density is equally likely to lie between 1 and ½ as between ½ and _. Whereas, directly applied to the specific density, a principle of indifference would suggest that the specific density is equally likely to lie between 1 and _, as between _ and _; which entails that it is not equally likely to lie between 1 and ½ as between ½ and _. So which is the right way of allocating prior probabilities? And when the space of alternative hypotheses are not just hypotheses differing in respect of which intervals of the same variables they declare to be equiprobable (and so with a different distribution for the probability density), but much more substantially - in terms of the kinds of entity and property they postulate and the mathematical relations the values of their property

have to each other, how on earth is any comparision of prior probability possible? Surely one can make no judgments of prior probability in advance of any evidence?

Prior Probability and Simplicity

The objectivist Bayesian answer is that if one could not make judgments of relative probability before evidence, one could not make them afterwards either. For compatible with any finite collection of data - both new observational evidence and background evidence about what has happened in the past in a wider field - there are always an infinite number of incompatible hypotheses which yield those data with probability 1 (or any lesser probability you care to choose). (There are an infinite number of curves which pass through a finite number of points, and otherwise diverge wildly. And an infinite number of those curves will pass through any finite number of new points you choose). So, if we want to say, as normally we do, that despite yielding the data with probability 1, some of these hypotheses are more probable than others, there must be a priori factors which are determining this. These are the factors which determine the prior probability of a hypothesis on zero relevant evidence which I call its intrinsic probability. In my view a full and careful analysis of the procedures of investigation, along the lines described earlier, will reveal that there are two kinds of factor - scope and simplicity. The greater scope a hypothesis has - the more entities about which a hypothesis makes claims and the more and more precise claims it makes about them - the lower (for a given degree of simplicity) is its prior probability. In my view the simplicity of a hypothesis is a matter of it postulating few entities of few kinds attributing to them few properties of few kinds, concerning properties more readily observable, fewness of laws, individual laws relating few variables, fewness of terms in equations stating a law and the mathematical primitiveness of the mathematical objects and relations utilized in laws.[23] To the extent to which a hypothesis is simpler on the balance of these various facets of simplicity it is, in my view, simpler overall and so (among hypotheses of equal scope) has greater intrinsic probability. By one property Q being more readily observable than another one R, I mean that the predicate ‘R’ rigidly designating R is introduced by a definition in terms (at least in part) of the predicate ‘Q’ designating Q but not vice versa; thus if ‘grue’ is introduced by a definiton such as that an object is ‘grue’ at a time t iff it is green and C is before 2050, or blue and C is after 2050 (and ‘grue’ is introduced by means of paradigm examples of objects which are green), then green is a property more readily observable than grue. If, on the other hand, ‘grue’ is introduced by paradigm examples of objects which are grue (and green is introduced in the same way), then since before 2050 ‘grue’ and ‘green’ will have the same paradigm examples, these words will mean the same. The general force of this requirement is to lead us, other

things being equal, to ascribe greater probability to hypotheses concerning properties such as ‘green’ and ‘pointing to 7 on the dial’, than to hypotheses concerning the physicist’s properties of enthalpy?, isospin? or hypercharge. But other things are often not equal, and hypotheses concerning the latter properties become very probable because of their great predictive power in comparison with hypotheses concerning the former properties. When the simplest hypothesis fails to predict well, the next simple hypothesis often acquires the greater posterior probability.

There is an interesting and superficially much more precise and unified account of the simplicity of a hypothesis than the rough account which I have just given, and which has had some considerable influence among physicists, first put forward by T. J. Solomonoff.[24] He proposed giving hypotheses intrinsic Bayesian probabilities in terms of the reciprocal of the minimum nbumber of conputational symbols (in ‘bits’, 0’s or 1’s that is) needed to express that hypothesis, called its ‘string length’. To give an example of string length - a random strong of $10^8$ bits would have a string lengthof about, $10^8$ (because you cannot summarize the string by a short equation), whereas a string of $10^8$ 1’s or a string of alternating 1’s and 0’s could be expressed by equations using far less bits. This account gives roughly the same results as does mine for which is the simpler of two hypotheses relating the same properties, where their relative simplicity turns on the simplicity of the mathematical equation relating these properties. But he gives no rules for comparing hypotheses relating different properties, that is, hypotheses using different vocabularies. Thus the formulae $F = \frac{Gmm^1}{r^2}$ for the gravitational force and $F \frac{\alpha ee^1}{r^2}$ for

the electrostatic force have the same mathematical form but relate different physical variables. And so do ‘all emerealds are green’ and ‘all emeralds are grue’. Solomonoff’s account needs to be supplemented by some such criterion as my criterion of the greater simplicity of more readily observable properties. A formula $\underline{x} = \underline{y}$ could hide a very complicated relationship if it needs a great number of obsrvations and complicated extrapolation therefrom to detect the value of $\underline{x}$.

I suggest that among hypotheses of equal scope equally able to yield the data, we judge those which satisfy criteria such as mine or those of Solomonoff better to have higher prior probability. We see that from the fact that we judge such hypotheses to be more probable than hypotheses which satisfy the criteria less well on evidence predicted by all of them. Such judgments concern comparative prior probability, but in some cases we can judge that some hypothesis has some exact probability on certain evidence, which in turn will enable us to make more precise the criteria of prior porbability (and - we may hope - help us

to see what is the correct way to apply the principle of indifference in at least some of the paradoxical cases).

It is crucial to distinguish the sense of simplicity which I have been discussing, on which simplicity is a criterion for choice among theories of equal scope from the sense of 'simplicity' in which a theory being simpler than another one just is it having greater scope than the other. It was Popper, more than anyone, who championed an understanding of this kind. He began by equating simplicity with degree of falsifiability. He wrote: 'The epistemlogical questions which arise in connection with the concept of simplicity can all be answered if we equate this concept with degree of falsifiability.[25] He claimed that the 'empirical content' - in my terminology, the scope of a theory - increases with its degree of falsifiability'.[26] He compared[27] four theories of heavenly motion, 'all orbits of heavenly bodies are circles' (p), 'all orbits of planets are circles' (q), 'all orbits of heavenly bodies are ellipses' (r), and 'all orbits of planets are ellipses' (s). He claimed that since planets are only one kind of heavenly body, and circles are only one kind of ellipse, p ruled out more possible states of affairs than did the others, and q and r each ruled out more states of affairs than s. p was thus easier to falsify than, for example, q - because an observation of any heavenly body, not just a planet, not moving in a circle would suffice to falsify it; and for that reason p told you more, had greater scope, than q. For similar reasons p had greater scope than r and s; q and r each had greater scope than s.

Now there may be a lot to be said for having theories simpler in this sense. Big claims are theoretically more important than small ones; and if they can be falsified easily, at any rate some progress can often be made. Theories with great scope are, however, as such, as I have already noted and as Popper boldly proclaimed, more likely to be false than theories with small scope. And there is no point in taking the trouble to falsify theories that are almost certainly false anyway. It is at this point that simplicity in a different sense comes in, as a criterion of probable truth. In my terminology a theory that a planet moves in a circle, for example, does not as such have greater simplicity than the theory that it moves in an ellipse; it just has less free parameters (a circle being an ellipse with a particular eccentricity, zero), and thus has greater scope. The theory that it moves in a circle, however, may well be simpler in my sense than the theory that it moves in an ellipse of a particular non-circular shape (where both have the same number of free parameters). The issue of how 'simplicity' should be understood arises in the papers in this volume of both Sober and Howson. Sober operates with a Popperian understanding of simplicity, and points out (p.13) that for 'nested models' where a proposition p entails a proposition q (because, for example, p gives constant values to free parameters of q), then - it follows from the axioms of the calculus - p cannot have a greater probability than q on the same evidence. Howson points out (p. 17) that the interest in simplicity

as a criterion of higher prior probability is an interest in it as a criterion for comparing incompatible (and so non-nested) hypotheses.

Some philosophers who are unwilling to allow that there are any general domain-indifferent criteria of prior probability hold (in effect) that within wide domains of which we have some experience, we know which hypotheses have some 'plausibility'. We know the kind of factors that might be work, and so have a finite number of hypotheses with which we can operate. In such cases we may perhaps in effect asume all 'plausible' hypotheses to have equal prior probabilities, and then accumulate evidence which is more probable given some hypotheses than others and so - by Bayes's Theorem will raise the posterior probabilities of the former over the latter. And if the number of hypotheses being considered is small, it might seem that it will not often greatly matter (within limits) how one allocates prior probabilities, since evidence in the form of a large collection of statistical data may quickly give a high posterior probability to a hypothesis whose prior probability is low. However, the nagging question remains as to the criteria by which our prior experience of the field leads us to select certain hypotheses and not others as having 'plausibility'.

It rather looks as if 'plausibility' is just another name for 'probability', and that - the claim is - contingent background evidence about a wide domain (k*)* gives different degrees of prior probability to different hypotheses. But now that we have used up all our contingent evidence (putting it into k) , only a priori criteria can tell us which hypotheses (among the infinite number of logically possible ones compatible with and able to predict with equal probability what has happened in the wide domain) have which degrees of prior probability. Unless there are true intrinsic probabilities, there cannot be true contingent prior probabilities nor any true posterior probabilities at all. Either no scientific conclusion about which hypothesis or prediction is more probably true than any other, has any objective warrant; or there are correct a priori criteria of intrinsic probability.

This issue of how to determine prior probabilities affects the very simplest situation in statistics, where the experimenter is testing between two hypotheses; the hypothesis that some factor makes a difference to something and the hypothesis that it doesn't, e.g. that smoking increases the risk of cancer, or that it doesn't. The investigator collects statistics of those who develop cancer among two large samples of the population - those who smoke and those who don't; and finds - say - that the proportion of those who develop cancer is much greater in the first sample. But maybe there is some other explanation of this, than that smoking increases the risk of cancer. Maybe all in the first group are underfed, or have heart disease, factors known to be relevant to the

occurrence of disease. But the two samples are chosen so that each sample has an equal proportion of those affected by these conditions. Yet there will always be an infinite number of properties other than being smokers, which all members of the first sample will have, and no members of the second sample will have - 'living in houses numbered ... (followed by the numbers of the houses in which members of the first sample live)', or 'being members of a sample chosen by such-and-such a process'. (Perhaps the process of creating the sample via random number tables conduces to cancer.) And so on. Hypotheses that such factors are at work are thought implausible - rightly so. But we have never tested hypotheses of innumerable such implausible kinds in this kind of domain. It looks as if our reasons for regarding them as implausible must come from domain - indifferent considerations, that is from a total world-view that includes the view that (in the absence of strong positive evidence to the contrary) it is very improbable that the kind of factor cited affects disease. And since a very complicated world-view which held that the factors at work in our particular domain are very different from those at work in other domains would be perfectly compatible with all our observations so far, my view is that it is only a priori criteria of prior probability including simplicity (in my sense) which can justify our preference for a world-view that certain kinds of factors are at work in all domains. The issue of whether science needs or can have a priori criteria of prior probability is a central issue for the contributors to this volume, and in their different ways all the contributors to the original symposium - Sober, Howson, Dawid, and Earman - deny that we can. They would, I think, all also hold that as a matter of fact most of us only take certain hypotheses seriously and then (either by giving them equal prior probabilities or by allowing empirical evidence to discriminate between them in some way not governed by the calculus), we can use the apparatus of the probability calculus to give them different degrees of posterior probability in the light of evidence. My own view expressed in this introduction, as well as the view of Solomonoff, is that there are a priori criteria of prior probability and these allow us to ascribe all intrinsic probabilities to all hypotheses.

Elliott Sober claims (p. 3) that a proposition does not have a prior probability until we have 'empirical information' about 'the process at work' which will bring it about that the proposition is true or false; only in such circumstances can the Bayesian apparatus be applied. 'Empirical information' may lead us to hold that some hypotheses in the field are 'plausible' and others are not. But there are no rules for mapping this informal talk of plausibility onto formal talk of probability. In the absence of such empirical information hypotheses can be compared only in respect of their 'likelihoods', that is, in respect of how probable it is that you would find the evidence you do given the different hypotheses; not 'all epistemological concepts that bear on empirical inquiry can be

understood in terms of the probabilistic relationships described by Bayes's Theorem' (p.1). Colin Howson thinks otherwise - he holds that the axioms of the calculus and so Bayes's Theorem govern all relations of epistemic support between propositions - but only in the form of a consistency constraint. They limit the values you can consistently give to one probability, given the values which you ascribe to other probabilities. With probability logic, as with deductive logic, (p. 25) 'what you put in as a premise will be at least as fallible and conjectual as what you get out as a validly derived conclusion'. This is, of course, a form of subjective Bayesianism. He goes on to defend the Bayesian account of statistical inference against the non-Bayesian accounts of R. A. Fisher, and of J. Neyman and E. S. Pearson. (Colin Howson's detailed treatment of statistics inevitably makes his paper somewhat more technical than the other papers in this volume). Sober claims that an important recent theorem of the probability calculus - Akaike's Theorem - does provide some rationale for prefering simpler theories (in his Popperian sense of theories with fewer adjustable parameters) to others; Howson denies the relevance of this theorem.

Philip Dawid and John Earman both take for granted Bayes's Theorem and assume that we can derive prior probabilities from empirical data, without discussing the extent to which a priori considerations enter into the derivation. They then show the consequences of applying Bayes's Theorem to two areas of inquiry. Philip Dawid considers how juries should use to weigh evidence in criminal trials. John Earman considers the kind of evidence which would show that probably a miracle had occurred (in the sense of a 'violation' or 'non-repeatable exception' to a law of nature). He sets his answer in the context of the sophisticated eighteenth-century debate about to weigh 'eyewitness testimony' to the occurrence of some event, against 'uniform experience' that events of the kind in question do not happen - to which debate Bayes's eighteenth-century theorem provided a crucial contribution.

## APPENDIX : Countable Additivity and Probabilities of 1 and 0

Axiom 3 is known as the axiom of finite additivity. It may be generalized into what is known as the axiom of countable additivity, in order to allow for infinite disjunctions, to read:

$P(\bigvee a_i|r) = \sum_i P(a_i|r)$, where $\bigvee a_i$ is the proposition that a member of the set of propositions $a_i$ is true, and $\sum_i P(a_i|r)$ is the sum of the probabilities of each of the propositions $a_i$ on evidence r, no more than one of which can be true given r. This plausible axiom will, however, give rise to contradicion, unless we allow infinitesimal numbers. For consider an infinite set of propositions $a_i$ each of which is equally probable and one of which must be true given r. Take r for example, as the proposition that the length of my desk is some particular exact rational number of metres between 2 metres and 3 metres, e.g. exactly 2.00467 metres; and let each $a_i$ ascribe a different such value to the length. If we say that the prior probability of each such length has some finite value greater than zero, however small, it will follow that the sum of an infinite number of such values will be infinite; and so it would follow from the Principle of Countable Additivity that the probability that the desk has a length between 1 and 2 metres is infinite. Yet there cannot be a probability greater than 1 (which by axiom 2 is the probability, on any evidence, of a tautology). But, if we attribute a value of 0 to the prior probability of each possible value of the length, the probability that the length will lie between 1 and 2 metres will (by the Principle of Countable Additivity) be 0 - contrary to what is stated by r.

No contradiction is generated, however, if we adopt a mathematics of infinitesimal numbers, in which there are an infinite number of such numbers greater than 0 but less than any finite number. Such a mathematics, called non-standard analysis, was developed by Abraham Robinson (see Abraham Robinson, Non-Standard Analysis, North-Holland Publishing Co., 1966). This allows us to attribute the same infinitesimal value to each of an infinite number of prior probabilities, which sum conjointly to 1. If we do not adopt non-standard analysis, not merely will we be unable to calculate the probability of an infinite disjunction from the probabilities of each of its disjuncts and conversely; but we shall still have a problem about what to say about the probability of each of an infinite number of equiprobable disjuncts, one only of which is true. If we attribute a finite value to it, then however small that value is, the more general Principle of Finite Additivity will have the consequence that the probability of the disjunction of some very large finite number of disjuncts will again be greater than 1. So we would have to attribute to each disjunct the probability 0. But that would involve saying that such a disjunct was just as certainly false as a self-contradiction. That seems implausible. There is some chance of winning in a fair lottery with an infinite number of tickets! The use of infinitesimals allows us to distinguish between the probability of a proposition whose negation is entailed by its evidence, and one whose negation is not so entailed but

which has a value less than any finite value greater than 0. The use of infinitesimals also allows us to distinguish between the probability of a proposition entailed by its evidence, and the probability of a proposition not entailed by its evidence, but which has a value greater than any finite number less than 1. (For the example the proposition that my desk has a length between 2 and 3 metres other than 2.00467 metres). Without using infinitesimals, P (h□e) = 1 can only be read as 'e makes h certain' in a sense of 'certain' which fails to distinguish between the two cases.

The use of infinitesimals allows us to make an analogous distinction for physical probability. For example, consider an indeterministic world in which a point particle has an equal physical probability of landing at any of the infinite number of points on a screen. We need to distinguish the probability of its landing at a particular point from the physical impossibility of it landing at all. By arguments analogous to these for logical probability, the probability of the former will be less than any finite number. If we allow infinitesimals, we can say that it has an infinitesimal value, whereas what is physically impossible has a probability of 0.

## NOTES

* Some sections of this Introduction correspond closely to passages in my Epistemic Justification, Clarendon Press, 2001, where these issues are discussed at far greater length, and I put forward my own more definite views about them. I am grateful to the Oxford University Press for permission to reuse this material.

1. Rudolf Carnap, Logical Foundations of Probability, Chicago: Universty of Chicago Press, 1950, chapter 2.

2. More precisely - to say that the proportion of A's which are B in an infinite class of A's (taken in a certain order) is p, is to say that for every finite number $d > 0$, there is some number of A's $n_d$ such that for any $n > n_d$, where there n A's (that is, the first n A's in the order), r of which are B, $p + d > {}^{r}/_{n} > p-d$. It follows from this definition that there will not be a probability of an A being B, for all infinite classes of A's, but only for those in which there is such a limiting value p.

3. London: John MacMillan and Co.

4. Second revised English edition, London: George Allen and Unwin, 1957.

5. Los Angeles: University of California Press, 1949.

6. See Karl Popper 'the Propensity Interpretation of Probability', British Journal for the Philosophy of Science 10 (1960), 25-42. Propensity theory was also the subject of one of the two lectures, 'A World of Propensities: Two New Views of Causality' in Popper's last book A World of Propensities, Bristol: Thoemmes, 1990. Popper supposes that (even where no Quantum effects are involved) the outcomes of most macroscopic events, such as the outcome of a certain toss of a coin, are produced by propensities other than 1 or 0. I see no good reason to suppose that. The outcome of each toss of the coin is normally virtually pre-determined by the exact angle of toss, force imparted to the coin etc.; and so the propensity of the set-up to produce heads is either 1 or 0.

7. See Appendix to this Introduction and especially its final paragraph for the

difficulty in representing physical necessity by a physical probability of 1, and physical impossibility by a physical probability of 0; and how it can be overcome.

8. For example, David Lewis holds (on the whole, with occasional qualifications and doubts) that the 'chance' (his substitute for physical probability) of a particular event just is the statistical probability of an event of that kind over all time (the kind being picked out in terms of the categories used in the best scientific theory). See his 'A Subjectivist's Guide to Objective Chance' (in his Philosophical Papers, vol. II, Oxford: Oxford University Press, 1986) and the later 'Humean Supervenience Debugged' (Mind, 103 (1994), 473-90).

9. London: Macmillan.

10. 'Logical probability' is not a fully satisfactory name for this kind of probability. I use it because it has been often used in the past to designate a priori objective evidential probability, but I do not assume that the value of every such probability (that is, of P(h□e) for each h and e) is a 'truth of logic'. For there may be a priori truths which are not truths of logic.

11. For main papers by these two writers see (ed.) H. E. Kyburg and H. E. Smokler, Studies in Subjective Probability, New York: J. Wiley and Sons, 1964.

12. A. N. Kolmogorov, Foundations of Probability Theory (first published in German, 1933), English edition, Chelsea Publishing Co., 1950.

13. See Appendix.

14. Atomic propositions are signified by individual lower-case letters, such as 'q' or 'r', denoting world states Q and R. Events which are not total world states can be denoted by disjunctions of atomic propositions.

15. See the alternative calculus developed by Fetzer and Nute, presented in J. H. Fetzer, Scientific Knowledge: Causation, Explanation and Corroboration, Boston Studies in the Philosophy of Science, vol. 69, D. Reidel, 1981, pp 283-6.

16. For a fuller formal treatment of a somewhat different alternative semantics for interpreting the axioms in terms of physical probability, see C. S. I. McCurdy, 'Humphreys's Paradox and the Interpretation of Inverse Conditional Propensities', Synthese 108 (1996), 105-26.

17. But not total agreement. There are writers who have proposed rival axiom systems for evidential support. L. J. Cohen, for example, has written three main books in defence of a rival axiom system. See, for example, his The Probable and the Provable, Oxford: Clarendon Press, 1977.

18. Reprinted in this volume as Chapter 7.

19. Though not in Colin Howson's view - see his paper in this volume, p. 1.

20. This terminology, deriving from R. A. Fisher, is widely current in statistical literature. I call it misleading because in ordinary language 'likelihood' means the same as 'probability'; and P(e|h & k) is the probability of *e*, not of *h*.

21. These two principles are two different ways of spelling out what David Lewis called 'The Principal Principle'. See his 'A Subjectivist's Guide to Objective Chance' in his Philosophical Papers, vol. II, Oxford: Oxford University Press, 1986.

22. J. M. Keynes, A Treatise on Probability, p. 45. Sober gives a similar paradox in his paper in this volume.

23. I give my own detailed of simplicity in my Epistemic Justification, Oxford: Clarendon Press, 2001.

24. R. J. Solomonoff, ‘A Formal Theory of Inductive Inference’, Information and Control, 7 (1964), 1-22.

25. K. R. Popper, The Logic of Scientific Discovery, London: Hutchinson, 1959, p. 140.

26. Ibid., p. 113.

27. Ibid., p. 122.