Review Article

THE WORK OF E. T. JAYNES ON PROBABILITY, STATISTICS AND STATISTICAL PHYSICS*

An important contribution to the foundations of probability theory, statistics and statistical physics has been made by E. T. Jaynes. The recent publication of his collected works provides an appropriate opportunity to attempt an assessment of this contribution.

1 INTRODUCTION

Most scientists are interested in solving particular problems. To do so they use whatever methods and theoretical structures seem to them appropriate. This is 'normal science' in the sense of Kuhn [1962] and as he says, 'One of the reasons why normal science seems to progress so rapidly is that its practitioners concentrate on problems that only their own lack of ingenuity should keep them from solving', (*ibid.* p. 37). Those of us who are concerned with our job prospects and publication lists avoid carefully the conceptually difficult problems associated with the foundations of our subject. Of course we are interested in why things are as they are and we shall be enthusiastic about new methods which give insight into this. The use of the renormalisation group in statistical mechanics (see *e.g.* Pfeuty and Toulouse [1977]) is a good example of this. By exposing rather clearly the fundamental role of the correlation length in critical behaviour it gives a new understanding of phase changes. The whole theory is, however, clearly seen to be within the established structures of statistical mechanics. The foundations of science are at a deeper level than this and they are an object of concern in two more or less distinct ways. The first is brought to the fore when we have two or more substantially different theories competing for acceptance. This 'revolutionary' situation is neither Jaynes' concern nor the subject of this paper. The second, which is, is when the interpretation of one or more of the concepts of an established theory are a matter of contention.

Given a collection of magnetic dipoles interacting with certain prescribed forces any competent solid state physicist will know the rules for obtaining the magnetism and susceptibility of the system in equilibrium at a particular

* Review of E. T. JAYNES [1983]: *Papers on Probability, Statistics and Statistical Physics*. Edited by R. D. Rosenkrantz. D. Reidel Publishing Company. US $49.50. Pp. xxiv + 434. We are grateful to Harvey Brown, Kenneth Denbigh, Udi Makov and Oliver Penrose for their valuable criticisms of this paper.

temperature. To calculate these he will use the equilibrium probability distribution for the orientations of the dipoles, although in his final answer there need be no hint of this probabalistic element of the theory. He need not, and probably will not, choose to assign any particular meaning to the probabilities which he uses. Even in the works of those who made the greatest contribution to the foundations of statistical mechanics it is not always clear how they viewed the probabilities which they used.[1] The major contribution of Jaynes is in the proposition and development of one particular view of probability theory in relation to statistical mechanics. The publication of his collected papers (Jaynes [1983])[2] is of interest not only because of the intrinsic interest of his approach but also as a record of the development of his ideas over the twenty-four years represented by these publications.

2 THE JAYNES MAXIMUM ENTROPY METHOD

Jaynes' early papers are concerned with statistical mechanics. From the outset he rejects a frequentist view of probability and regards the probability distributions of statistical mechanics as expressions of human ignorance. This has led to much misunderstanding and misrepresentation of his ideas, sometimes abetted by his rather hyperbolic way of expressing himself. This has been particularly evident in discussions of entropy. His remark (*CP*, 86) that 'entropy is an anthropomorphic concept, not only in the well-known statistical sense that it measures the extent of human ignorance as to the microstate. *Even at the purely phenomenological level entropy is an anthopomorphic concept.* For it is a property not of the physical system but of the particular experiments you or I choose to perform on it', is often quoted. As Denbigh [1981] helpfully suggests: 'There is no need to bring in "you or I"; that last sentence could equally well have been written "It is a property of the variables required to specify the physical system under the conditions of the particular experiment".' Whether or not Jaynes would regard this alternative phraseology as an acceptable alternative to his own we do not know. What, however, is clear is that he has, over the years, become increasingly aware of the hazards of misunderstanding and has made every effort to develop and consolidate his position. In doing so he has developed his work on Baysian theory to the point where statistical mechanics is seen as no different from any other situation in which there is a need to make predictions for a system about which there is a degree of uncertainty. 'Predictive statistical mechanics' is Jaynes' chosen name for his approach to statistical mechanics. For him it 'is not a physical theory, but a form of statistical inference' (*CP*, 416). Rather than asking the question: 'How does the system behave?' which he takes to be the usual purpose of a physical

[1] For a discussion of this see *e.g.* Lavis [1977].

[2] Henceforth referred to as *CP*. For convenience, all references to Jaynes' work are to *CP*, although most material is derived from work previously published.

theory (*CP*, 416) he chooses the question: 'Given the partial information we do, in fact, have what are the best predictions we can make of observable phenomena?' (*CP*, 416).

For Jaynes a probability distribution (in statistical mechanics or anywhere else) is neither completely subjective, in the sense of Ramsey or de Finetti, nor is it a property only of the system under investigation. Rather it is an attribute, both of the system and of the information we have or the observations we choose to make. The element of objectivity in his approach is contained in the assertion (*CP*, 117) that 'in two problems where we have the same prior information we should assign the same prior probabilities'. A rational and agreed procedure is, therefore, needed for the derivation of the probability distribution subject to the constraints provided by the given information. The key to this is the idea of uncertainty. Suppose we have a system whose states are given in terms of the discrete-valued random variable x with range $\{x_1, x_2, \ldots, x_n\}$. Let p_i be the probability of the event $x = x_i$. Jaynes argues that a reasonable measure of the uncertainty in the distribution $\{p_i\}$ must satisfy the conditions:

(i) It should be a non-negative continuous function of the variables $\{p_i\}$.

(ii) When $p_i = 1/n$, for all i, it should be a monotonically increasing function of n.

(iii) It should satisfy a composition law consistent with the additivity of the probabilities of mutually exclusive events $x = x_i$.

Shannon has shown (see Shannon and Weaver [1949]) that the information entropy

$$S_I = -\sum_i p_i \ln p_i \tag{1}$$

is the unique function, apart from a positive multiplicative constant, which satisfies these criteria. Jaynes takes this as his measure of uncertainty (*CP*, 8) remarking that it 'agrees with our intuitive notions that a broad distribution represents more uncertainty than does a sharply peaked one'. He then adopts the principle that the best probability distribution is that for which S_I is a maximum subject to any constraints imposed by the information we have about the system since (*CP*, 9) 'it is uniquely determined as the one which is maximally non-committal with regard to missing information'. If this argument is accepted then it is mathematically simple to derive the appropriate distribution.[1] This method is applied to equilibrium and non-equilibrium statistical mechanics and to the general problem of statistical prediction. The rest of this paper is divided into three sections concentrating

[1] The cases where the system is quantal, with non-orthogonal wave-functions, or where the underlying random variable x is continuous are also considered by Jaynes. Equation (1) is then generalised to

$$S_I = -\mathrm{Trace}[\hat{p} \ln \hat{p}] \tag{1a}$$

on Jaynes' work in these areas followed by a fourth section containing our concluding discussion.

3 EQUILIBRIUM STATISTICAL MECHANICS

Jaynes takes as typical (*CP*, 9–10) the case in which we have a system with known energy spectrum $\{E_1, E_2, \dots, E_n\}$ and the information we are given is the average value ε of the energy. Taking E as the random variable x of section 2 and equating ε with the expectation value $\langle E \rangle$ we have the constraint

$$\varepsilon = \sum_i p_i E_i. \tag{2}$$

We use the undetermined multiplier λ to maximise S_I subject to the condition (2). This gives

$$p_i = \exp(-E_i\lambda)/Z(\lambda) \tag{3a}$$

where

$$Z(\lambda) = \sum_i \exp(-E_i\lambda) \tag{3b}$$

and λ is given from (2) and (3a) by

$$\varepsilon = -\mathrm{d}\ln Z/\mathrm{d}\lambda. \tag{4}$$

Equations (2) and (3) correspond closely to those of the usual canonical distribution. If we write $S_e = k(S_I)_{\max}$ and $T = 1/(\lambda k)$, where k is Boltzmann's constant, then it is straightforward to show that, in this case, equations are produced of the form of the basic thermodynamic relations with S_e and T playing respectively the roles of thermodynamic entropy and temperature. In Jaynes' words (*CP*, 54) the 'maximum-entropy formalism leads automatically to definitions of quantities *analogous* to those of thermodynamics' and we agree with him when he continues: 'This is, of course, as far as any mathematical theory can go; no amount of mathematics can prove anything about experimental facts.'

It is convenient to remain with this simple example while discussing a number of problems which have been raised, or we wish to raise, in relation to Jaynes' method. These can be broadly classified into three types: I. Experimental, II. Conceptual and III. Mathematical.

and

$$S_I = -\int \rho(x) \ln[\rho(x)/m(x)]\,\mathrm{d}x \tag{1b}$$

respectively, where in (1a) $\hat{p}$ is a probability density operator (*CP*, 20–5) and in (1b) $\rho(x)$ is a probability density function with $m(x)$ the corresponding probability density function for the situation of complete ignorance (*CP*, 283). Jaynes' partial solution to the problem of determining $m(x)$ is by means of the invariance group of the system. This problem is discussed in more detail in section 5.

I. The *experimental* problems can be characterised by asking two questions about the information which Jaynes uses to constrain his maximisation of the information entropy:

(a) How is it collected?
(b) Is it necessarily numerical?

For the system described above we have given a quantity $\langle E\rangle$ which is construed as the expectation value of the energy. Two methods are described by means of which it may be obtained. In the first of these, a sample average is taken over a number of measurements. This is then treated as the expectation value $\langle E\rangle$. Jaynes himself asks (*CP*, 269): 'Is there not an element of arbitrariness in this?' and then proceeds to answer his own question in the following way (*CP*, 269–71). (i) If we decide to use maximum entropy based on the expectation value of E then we know in advance that the final distribution will be of the form of equations (3). (ii) Regarding (3) as a class of distributions parametrised by λ it can be shown (and he shows it) that the maximum likelihood estimate of λ is given by (4). Accepting Jaynes' theory this argument has force. The second method described for obtaining $\langle E\rangle$ seems to us rather less convincing. As he says (*CP*, 14), 'In practice a measurement of energy is rarely part of the initial information available; it is the temperature that is easily measured.' To explain how a measurement of temperature is processed to give a value for $\langle E\rangle$ he begins in a rather familiar way. He supposes that the system of interest (called σ_1) is in interaction with a second system σ_2, which plays the role of a heat bath, the total energy E of $\sum = \sigma_1 + \sigma_2$ being assumed to be the sum of the energies E_1 and E_2 of σ_1 and σ_2 respectively. Two pieces of information are given or inferred, (i) all states of $\sum$, with given value of E, are 'equivalent' (*CP*, 14). (ii) the expectation value $\langle E_2\rangle$ of the energy of σ_2 is assumed to be derivable from the temperature of the heat bath. The argument for the derivation of the canonical distribution for σ_1 is not given in detail, but could be reconstructed in the spirit of his method.[1] Our reservations about this programme concern the information content of (ii). Jaynes asserts (*CP*, 14) that 'a *thermometer* is a heat bath σ_2 equipped with a pointer which reads its average energy. The scale is, however, calibrated so as to give a number T, called the *temperature*, defined by $1/T = dS_2/d\langle E_2\rangle$.' This definition of a thermometer must be taken in conjunction with a preceding sentence which asserts that, for a heat bath 'the entropy S_2 of the maximum-entropy distribution for given $\langle E_2\rangle$ is a definite monotonic function of $\langle E_2\rangle$'. We find some difficulty in interpreting these remarks. If the second one is intended to mean that $S_2 = S_2(\langle E_2\rangle)$ is a *known* function, then it must have been obtained by some observations. If they are

[1] One possible way to do this would be to (a) Use (i) and the maximum-entropy principle to show that the marginal probability $p_1(i)$ of σ_1 in state i with energy E_i is of the form $p_1(i) \propto \omega_2(E - E_i)$, where ω_2 is the density of states of σ_2. (b) Use (ii) and the maximum-entropy principle to obtain the marginal probability $p_2(j)$ of σ_2 in canonical form. (c) Invert the partition function of σ_2 in the limit of large σ_2 using the saddle-point method (Kubo [1965]) to give ω_2.

measurments of E_2then this surely defeats the object of the exercise (which is to formulate the programme in terms of temperature measurements). If they are measurements of temperature then maximum-entropy cannot be used with respect to $\langle E_2\rangle$. Alternatively one might suppose that he means us to understand that a measurement of temperature implies the existence of a well-defined but *unknown* $\langle E_2\rangle$ the value of which is given only at the end of the derivation of $S_2(\langle E_2\rangle)$. If this is the case, he might simply modify his usual discussion of entropy maximisation, subject to a given value of the expectation value of the energy, and say that a measurement of temperature is equivalent to knowing that $\langle E\rangle$ exists with a specific but unknown value and that, only at the end of the analysis, is the value of $\langle E\rangle$ revealed.

These problems may be resolved if we could be clear what his answer is to our question (b) given above. The problem can be sharpened by examing the following situation. Suppose we have the system with n energy levels described above and we see (possibly through a glass screen or similar device) that it is immersed in a heat bath. We are unable to make any measurements on the system to obtain numerical information (other than knowing n, the number of energy levels). In this situation would Jaynes (i) use the maximum-entropy formulation to obtain the uniform distribution

$$p_i = 1/n, \quad i = 1, \ldots, n \tag{5a}$$

with entropy

$$S_e = k(S_I)_{\max} = k \ln(n) \tag{5b}$$

which, for known n, would allow him to make definite predictions? Or would he (ii) infer from the presence of the heat bath that $\langle E\rangle$ is well-defined and use maximum-entropy to derive the canonical distribution given by (3) with $\lambda = 1/(kT)$ and with

$$S_e = \frac{\partial}{\partial T}\{kT \ln Z(1/(kT))\}. \tag{6}$$

In the latter case he would not be able to make any numerical predictions since T is unknown.

II. The *conceptual* aspect of Jaynes' programme which has caused most discussion and dispute concerns the status of entropy. The subjective-objective controversy about entropy is an old problem, but it is posed in a particularly clear way by Jaynes. Given that his analogue (*CP*, 54) for experimental entropy is obtained by maximising the information entropy, it is clear that it is a function, not only of the physical system, but of the information provided by the experiments. Jaynes recounts a conversation with Uhlenbeck (*CP*, 237–8), preceding any of his papers, in which he argues for his point of view. We are not told the precise words he used to put his case, but if they were in the spirit of the remarks quoted in section 2, above, then Uhlenbeck's reaction is easy to understand. There are obviously some genuine elements of disagreement between Jaynes and his critics, but it seems that his choice of language sometimes exacerbates the situation and leads to the appearance of a wider gulf that actually exists.

In an attempt to clarify the situation, let us look at it in the following way: The object of interest for statistical mechanics is a system $\mathscr{S} = \{\mathscr{S}_M, \mathscr{S}_T\}$, where $\mathscr{S}_M$ denotes the qualities of the system at the micro or atomic level and $\mathscr{S}_T$ denotes the qualities of the system at the macro or thermodynamic level. About such a system we have a certain amount of information $\mathscr{I} = \{\mathscr{I}_M, \mathscr{I}_T\}$. Part of this information will be qualitative (*e.g.* $\mathscr{I}_M$ will include details of the types of molecules, whether they are polar *etc.* and $\mathscr{I}_T$ will tell us whether the system is contained within rigid walls or subject to external forces, whether it is thermally isolated or in a heat bath *etc.*) and part may be numerical, obtained from measurements of quantities whose qualitative role we know. Finally we can imagine that we have a model $\mathscr{m} = \{\mathscr{m}_M, \mathscr{m}_T\}$ of the system. Given this situation the question is: *What is it that has entropy*?

The distinction between $\mathscr{S}_T$ and $\mathscr{S}_M$ is, it appears, the distinction made by Jaynes (*CP*, 85) between a 'thermodynamic system' and a 'physical system' respectively. He argues that there is no notion of the entropy of a physical system. That is, that the entropy $S(\mathscr{S}_M)$ of $\mathscr{S}_M$ is not well-defined. In this we agree with him, as would, we imagine, most people. The crucial question is whether there is a quantity $S(\mathscr{S})$, the entropy of the system as a whole. Jaynes maintains that there is not. He supports his position by arguing that 'we can always introduce as many new degrees of freedom as we please' (*CP*, 86). If this assertion were true then there would be no way of obtaining a 'complete' set of degrees of freedom for the system and thus no way of obtaining the true entropy. Hidden in this argument there seems to be a confusion between two distinct sets of variables. There is the set v_1 of variables describing the degrees of freedom and there is the set v_2 of parameters on which the Hamiltonian of the system depends. Since, on any analysis of statistical mechanics, the entropy is a sum or integral over the points of phase space it will not be a function of the variables v_1. It will, of course, be a function of the variables v_2. The number $N(v_1)$ of members of v_1 is in general large; the number $N(v_2)$ of members of v_2 may be large or small and greater or less than $N(v_1)$. For a perfect gas of n identical particles $N(v_1) = 6n$ whereas $N(v_2) = 3$, the only members of v_2 being n, the mass m of a particle and the volume V of the container. If all the particles have different masses then $N(v_1) = 6n$ and $N(v_2) = n+2$. By introducing suitable interparticle interactions one may increase $N(v_2)$ so as to exceed $N(v_1)$. This does not lead to the collapse of the concept of entropy. It simply increases the technical difficulty of its computation. Jaynes' discussion of the strain tensor (*CP*, 85–6) in which he resolves it into a complete set of orthogonal functions seems to be a case in which the number of members of v_2 is increased. We do not see how this leads to a collapse of the concept of entropy nor to an approach to the point 'where we control the location of each atom independently' (*CP*, 86), since the variables v_1, which specify those locations will not appear in the entropy. The part of Jaynes' argument with which we agree can be expressed in terms of the contrast between the system $\mathscr{S}$ and its model $\mathscr{m}$. It is almost inevitably true that $\mathscr{m}$ will be much simpler than $\mathscr{S}$ and degrees of freedom present in $\mathscr{S}$ will be absent in $\mathscr{m}$. The extent to which the entropy $S(\mathscr{m})$ of $\mathscr{m}$ is a good representation of $S(\mathscr{S})$ will depend on whether, during a particular process, the part of $S(\mathscr{S})$ associated with degrees of freedom absent in $\mathscr{m}$ changes significantly. This is the point made by Denbigh [1981] in his discussion of nuclear entropies and the possible presence of unknown isotopes. From the fact that it must be conceded that $\mathscr{m}$ is a simplification of $\mathscr{S}$, involving the possible ignorance of important features of $\mathscr{S}$, it does not follow that $\mathscr{S}$ contains the potentiality for an infinite proliferation of degrees of freedom. Jaynes would, we suppose, reject this mode of discussion, since he makes very little explicit use of models. For him the important quantity is the information and his entropy is $S(\mathscr{I})$, the entropy of $\mathscr{I}$. One may, however, construe the information as a background model $\mathscr{m}$ together with measured values for the parameters contained in $\mathscr{m}$. In this sense $S(\mathscr{I})$ will be a numerical value of $S(\mathscr{m})$ and the discussion above, about the relationship between $S(\mathscr{m})$ and $S(\mathscr{S})$, goes through for $S(\mathscr{I})$ and $S(\mathscr{S})$, given that one believes, as we do, in the existence of $S(\mathscr{S})$. This belief, although of course initially a matter of philosophical choice, could be undermined by a valid argument, which showed that $S(\mathscr{S})$ could not, even in principle, be properly defined. Jaynes' attempt to do this, described above, does not seem to us to achieve this end.

III. A *mathematical* problem associated with Jaynes' method was first raised by Friedman and Shimony [1971]. They considered the system, described above, with energy spectrum $\{E_1, \ldots, E_n\}$. Supposing that the background information B contains 'no information about the system other than its structure (which determines the set of possible states)' they obtain the prior probabilities $p_i = P(E = E_i \mid B)$ by maximising the information entropy S_I, given by (1), subject only to the normalisation condition. This, of course, results in the uniform distribution (5a). They then suppose that D_ε is the 'evidence that the posterior expectation value of E is ε'. This is used as a constraint when maximising S_I to produce the new distribution for E given by (3a). This, however, is the Bayes posterior distribution of E, with uniform prior, if and only if the conditional probability density function $\rho(D_\varepsilon \mid B)$ for D_ε given B is

$$\rho(D_\varepsilon \mid B) = \delta(\varepsilon - \bar{E}) \tag{7a}$$

where

$$\bar{E} = n^{-1} \sum_i E_i. \tag{7b}$$

Friedman and Shimony find this an unacceptable condition. Their position is that any distribution which professes to reflect ignorance cannot reasonably make such positive predictions. Their replies (Shimony [1973]; Friedman [1973]) to Hobson [1972], Tribus and Montroni [1972] and Gage and Hestenes [1973] illustrate their difficulty more clearly. They accept the proof by Hobson that equations (7) are a necessary consequence of Bayes' theorem if the sequence of trials used to obtain ε are independent, but they question this assumption of independence. Friedman gives the following argument: 'If for a very large value of m the first m trials strongly favor a proper subset of the energy spectrum, it is very likely that subsequent trials will do so as well. Thus in general the probability that the $(m+1)$th trial will yield a particular energy state depends on whether it is conditioned by just our background knowledge, or by our background knowledge and also our knowledge of the first m trials. Hence the trials are not statistically independent.' This argument seems to us to be false. It is certainly true that our *knowledge* of the distribution will change during the sequence of trials and we may even choose to reassess our assignment of probabilities to the various outcomes, but our supposition of independence will not be affected. A similar error is made by Shimony when he argues that 'the outcomes of the various trials are surely linked in the sense that a single die [in the example which he considers] is used throughout, so that any weighting which influences the outcome of one toss also influences the outcomes of the others'. Later he reinforces his point by arguing that 'while the data B do not assert a linkage [between?] trials, they do not preclude one; they are simply mute on the question. To justify Hobson's conclusion, one would have to say that on the bare background data B, the probability of the existence of a linking mechanism which would cause a statistical deviation from $\langle f \rangle$ [$\langle E \rangle$ in our notation] is zero.' The justification for the assumption of independence is surely Occam's Razor. We must assume *some* covariance structure. In the absence of definite evidence to the contrary we use the simplest model available, that of independence. Only if this model does not satisfactorily fit the data is there a need to include more parameters.

Jaynes dismisses the example of Friedman and Shimony by pointing out (*CP*, 250) that, if D_ε is a statement about a probability distribution on the sample space $S = \{E_1, \ldots, E_n\}$, then it can be used as a constraint when maximising entropy but not as a conditioning statement in Bayes' theorem, since it is not a statement about an event in S. On the other hand, if D_ε is a statement defining an event on the sample space S^m, of m trials, then it can be used as a conditioning statement for Bayes' theorem but not as a constraint when applying the maximum entropy procedure to events in S. For him the question: 'What is the probability of D_ε given the background information?' which is implicitly asked by Friedman and Shimony, is meaningless. The only question we can ask is, 'What is the probability that a sample of size m will give a mean of ε?'. It would seem that some of the difficulty arises from the failure to appreciate that $\{p_i\}$ are parameters of the prior distribution not of the sampling distribution (see *CP*, 118).

However, this argument can also be turned on Jaynes, since he uses sample estimates as constrains when maximising entropy. In these circumstances, the maximum-entropy procedure is a model building method. We wish to fit a multinomial distribution to some data. Jaynes' method is to maximise entropy subject to the values of certain 'physically meaningful' expectations. In orthodox terms he adds parameters to his model until a satisfactory fit is achieved. That is, he employs a Forward Selection method rather than a Backwards Elimination method which Friedman and Shimony would presumably prefer. This point applies to his general approach to statistical problems and will be dealt with in greater detail in section 5.

4 NON-EQUILIBRIUM STATISTICAL MECHANICS

According to Penrose [1979]: 'The ultimate aim of non-equilibrium statistical mechanics is to derive laws describing the macroscopic behaviour of systems not at equilibrium, starting from the microscopic laws of motion', and his review of the progress in this area makes it clear that there is no general agreement that this aim has been achieved. A crucial problem here is to find the appropriate definition for entropy. The natural choice for this is the Gibbs or 'fine-grained' entropy $S_G = kS_I$, where S_I is given by (1a) or (1b) according as the system is quantal or classical and the probability density operator or function in now dependent on the time t.[1] In either case the probability distribution satisfies a conservation condition known as Liouville's equation, and it can be shown (see *e.g.* Jancel [1969]) that S_G remains constant as the system evolves in adiabatic conditions. This is usually regarded as a problem, since it is a consequence of the second law of thermodynamics that there is an increase in entropy during an irreversible adiabatic process. A variety of ways of avoiding this difficulty have been proposed. These consist mainly of replacing the probability distribution by either some course-grained or time-averaged distribution, or by some marginal distribution governed by a master equation (for a review see Penrose, *ibid.*). Although these methods achieve the desired increase in entropy, their physical meaning is not always very clear.

Jaynes' approach to the problem is very different. His 'goal is not to "explain irreversibility" but to predict observable facts' (*CP*, 2) and he asks: 'What probability assignment to microstates correctly describes the state of knowledge which we have, in practice, about a non-equilibrium state?' (*CP*, 109). After some years attempting to introduce 'new and more complicated principles' (*CP*, 110) Jaynes finally concluded that the procedures, which he had used for equilibrium systems, remained valid for non-equilibrium. The only generalisation needed was to provide for the possibility that information was collected over a time interval. In his initial attempt to tackle the

[1] The equation (1) for S_I can be regarded as that of a system for which the density operator is diagonal in the energy representation, a condition which does not necessarily persist in time.

problem Jaynes considered a system which began in equilibrium, was subject to an adiabatic change, and was then allowed to return to equilibrium. He then computed the entropy for only the initial and final equilibrium states (*CP*, 82–5). His later work (*CP*, 292) and that of other authors using the same method (Robertson [1966]; Hobson and Loomis [1968]) indicates a relaxation of these restrictions and we shall present our description of his procedure in the more general way. For simplicity we use the language of quantum statistics but it should be borne in mind that the method applies equally well to classical systems (Hobson and Loomis [1968]).

Consider a system undergoing an adiabatic change, which has a set of time-dependent observables $\{\hat{\Omega}_1(t), \hat{\Omega}_2(t), \ldots, \hat{\Omega}_m(t)\}$, (of which one will normally be the Hamiltonian $\hat{H}(t)$). Suppose measurements are made of these observables at time t_0 with results $\{\omega_1(t_0), \ldots, \omega_m(t_0)\}$. We obtain the density operator $\hat{p}_0(t_0)$ which maximises S_I, given by (1a), subject to the constraints

$$\omega_k(t_0) = \text{Trace}\,[\hat{p}_0(t_0)\hat{\Omega}_k(t_0)], \quad k = 1, \ldots, m \tag{8}$$

and the experimental entropy $S_e(t_0)$ is taken to be equal to the fine-grained entropy

$$S_e^{(0)}(t_0) = -k\,\text{Trace}\,[\hat{p}_0(t_0)\ln\hat{p}_0(t_0)] \qquad (9)^1$$

At some later time t the density operator evolves to $\hat{p}_0(t)$ and we predict the values $\{\omega_1(t), \ldots, \omega_m(t)\}$ of the observables using

$$\omega_k(t) = \text{Trace}\,[\hat{p}_0(t)\hat{\Omega}_k(t)], \quad k = 1, \ldots, m. \tag{10}$$

A new density operator $\hat{p}(t)$ is now calculated by maximising S_I subject to the constraints

$$\omega_k(t) = \text{Trace}\,[\hat{p}(t)\hat{\Omega}_k(t)], \quad k = 1, \ldots, m \tag{11}$$

and the experimental entropy is then taken equal to

$$S_e(t) = -k\,\text{Trace}\,[\hat{p}(t)\ln\hat{p}(t)]. \tag{12}$$

Since

(i) $S_e^{(0)}(t)$ is invariant under the time evolution ($S_e^{(0)}(t) = S_e^{(0)}(t_0)$),
(ii) both $\hat{p}(t)$ and $\hat{p}_0(t)$ satisfy the constraint conditions (10),
(iii) $\hat{p}(t)$, but not necessarily $\hat{p}_0(t)$, maximises S_I subject to the constraints,

we have

$$S_e^{(0)}(t_0) \leqslant S_e(t). \tag{13}$$

This is the essence of Jaynes' procedure for demonstrating increasing entropy during an adiabatic change using the maximum-entropy method.

[1] In the case where we have only one observable $\hat{\Omega}(t)$ it has been shown by Hobson [1967] that $\hat{p}_0(t_0)$ is diagonal in the $\hat{\Omega}(t_0)$ representation.

Let us, however, consider an increasing sequence of times $\{t_0, t_1, t_2, \ldots\}$. If we use (10) to predict the values $\{\omega_1(t_j), \ldots, \omega_m(t_j)\}$ of the observables at time t_j and the constraints (10) to derive $S_e(t_j)$, then, of course,

$$S_e^{(0)}(t_0) \leqslant S_e(t_j). \tag{14}$$

It does not, however, follow that $S_e(t_j) \leqslant S_e(t_{j+1})$. The entropy is not necessarily a monotonically increasing function. This consequence of the method has been recognised by Robertson [1966]. He considers the situation where the system is in equilibrium for $t < 0$ and the observables (in this case only the Hamiltonian) are time-dependent only for the period $0 < t < t'$. He asserts, without proof, that, if the system settles down to thermodynamic equilibrium for $t \gg t'$ then $S_e(t) \leqslant S_e(\infty)$ for $t > t'$. 'That is, when the system settles down to equilibrium after being disturbed, the entropy will equal the maximum value it attains while, with time-independent Hamiltonian, the system was approaching equilibrium' (Robertson, *ibid.*). Although this result is weaker than the thermodynamic law of entropy increase for an isolated system, it may well be the strongest result which can be derived from statistical mechanics. The situation is, however, less clear when we consider adiabatic processes. Here we may envisage the following situation: Suppose, for some increasing sequence of times $\{t_0, t_1', t_1, t_2', t_2, \ldots\}$, the system is isolated and in equilibrium for $t < t_0$ and isolated for $t_j' < t < t_j, j = 1, 2, \ldots$, but that it is subjected to adiabatic changes in the time intervals $t_j < t < t_{j+1}', j = 0, 1, \ldots$. Suppose also that the intervals (t_j', t_j) are sufficiently long for the system to attain equilibrium. In Jaynes original discussion of the problem he showed (*CP*, 82–5) that $S_e^{(0)}(t_0) \leqslant S_e(t_1)$. Now, however, we have a sequence of adiabatic changes from equilibrium to equilibrium, over the intervals (t_j, t_{j+1}) and, as indicated above, it is not possible, on this analysis, to show that $S_e(t_j) \leqslant S_e(t_{j+1})$, only that $S_e^{(0)}(t_0) < S_e(t_j)$. The important distinguishing feature of t_0 is not just that it represents the end of a time interval during which the system was in equilibrium but also that it is a time *at which the observables were measured*. One way of achieving the result $S_e(t_j) \leqslant S_e(t_{j+1})$ would be to use the evolved form $\hat{p}_j(t_{j+1})$ of $\hat{p}_j(t_j)$ to calculate $\omega_k(t_{j+1})$ rather than $\hat{p}_0(t_{j+1})$. The difficulty with this, as can be seen from numerical calculations with simple examples, is that entropy curve becomes dependent on the time sequence chosen. In any case this alternative would probably have no appeal for Jaynes for whom the special status of t_0, illustrating as it does the subjective nature of his entropy in its relationship to knowledge, would probably be no problem.

5 STATISTICAL PREDICTION

Most critics of Baysian statistics have concentrated their attack on the concept of the prior distribution. Their objections have been of one or both of the following two types:

(i) In many circumstances it is possible to define prior probabilities only if it is accepted that probability measures (or can measure) 'degree of belief'. This, of course, raises the wider issue of debate between 'frequentists' and 'subjectivists' concerning the nature of probability.

(ii) Even if prior probabilities are meaningful there is no way we can justify any particular choice of prior distribution. Since the choice of prior will affect the conclusions drawn from the data there is the danger that two workers could legitimately produce contradictory answers from the same experiment.

In Jaynes' view, the answer to the first objection depends on the answer to the second. His paper 'Confidence intervals *vs* Baysian intervals' (*CP*, 151–89) is written to show that the Baysian methodology provides sensible answers to statistical problems far more easily that Orthodox methods. He claims (*CP*, 183) that 'one can produce any number of examples, at first sight quite innocent-looking, in which use of confidence intervals or orthodox significance tests leads to absurd or dangerously misleading results', and he goes on to say that to the best of his knowledge 'nobody has ever produced an example where the Baysian method fails to yield a reasonable result'. He recognises that the simplicity of the Baysian approach will not convince the sceptics unless some rational procedure for determining prior probabilities is found and, as we have seen, his solution to this problem is the introduction of the maximum entropy criterion. The 'objective' prior distribution is the one which maximises the entropy subject to any constraints supplied by the prior information. While, however, this usually suffices to provide an answer, when the prior distribution is discrete, it is not enough when the prior is continuous. When we extend the concept of entropy to the continuous case we find that S_I is given, in terms of a probability density function $\rho(x)$ by (1b), where $m(x)$ is an as yet arbitrary function. We are left with the problem of the choice of $m(x)$. If we determine the distribution which maximises S_I subject only to the normalisation condition then we find that

$$\rho(x) = m(x)\left[\int m(x)\,\mathrm{d}x\right]^{-1}. \qquad (15)$$

'Except for a constant factor, the measure $m(x)$ is also the prior distribution describing "complete ignorance" of x' (*CP*, 125). In other words, the problem of determining $m(x)$ is simply the problem of defining "complete ignorance".

Bayes, applying the principle of indifference, suggested the use of a uniform prior in circumstances of ignorance. However, as Jaynes points out (*CP*, 125), 'Bayes' rule has the obvious difficulty that it is not invariant under a change of parameters, and there seems to be no criterion telling us which parametrization to use.' So, for example, if we take a prior that is uniform in x, this will not be the same as a prior that is uniform in x^3. This failing must be one shared by all choices of prior. We cannot hope to have invariance of form under all parameter changes. However, Jaynes expects that the form of the prior should be invariant under those changes which merely convert the original problem into an equivalent one. As he says (*CP*, 144): 'It is dangerous to apply this principle [of indifference] at the level of indifference between *events*.... However, the principle of indifference may, in our view, be applied legitimately at the more abstract level of indifference between *problems*.' Thus Jaynes defines the 'ignorance prior' as the distribution which is form invariant under a certain group of transformations.

An excellent example of this method is found in his paper: 'The well-posed problem' (*CP*, 133–48). This work deals with Bertrand's problem concerning straight lines drawn 'at random' intersecting a circle. The task is to find the probability that the chord formed by the line is longer than the side of the inscribed equilateral triangle. There are various ways of defining the term 'at random' in this context and they lead to different solutions to the problem. The standard response to this is to say that the problem is 'ill-posed' since 'at random' is undefined. Jaynes tackles the problem by recognising that nothing is said about the size or position of the circle. Any consistent solution must, therefore, use a distribution which is invariant under such transformations. This requirement leads to the conclusion that 'at random' can mean only that the distance between the centre of the chord and the centre of the circle is uniformly distributed. This method has wide applicability leading to the general principle that 'every circumstance left unspecified by the statement of a problem defines an invariance property which the solution must have if there is to be any definite solution at all' (*CP*, 144). There are, however, problems when so little information is given that no solution is possible. In terms of transformation groups the difficulty is not that the problem is underdetermined, but rather that it is over-determined, since no distribution exists which satisfies all the requirements.

Jaynes seems to have found a satisfactory method of defining a prior distribution which measures 'complete ignorance'. If 'testable' prior information is available[1] then this can be incorporated into the prior by using the information as a constraint when maximising entropy. Often, however, it is known that the prior information is not exact and there seems to be, in general, no way of introducing this uncertainty into the methodology. Jaynes accepts that the prior information 'might, for example, be only the guess of an idiot' (*CP*, 272) but he goes on to say that 'nevertheless, that is the number given to us, and our job is not to question it, but to do the best we can with it. This may seem an inflexible, cavalier attitude; I am convinced that nothing short of it can ever remove the ambiguity of *what is*

[1] Information is said to be testable if, given any proposed distribution $\{p_i\}$, we can determine unambiguously whether the information agrees or disagrees with $\{p_i\}$ (*CP*, 240).

the problem? that has plagued probability theory for two centuries.' Surely this is wrong? Prior information must be examined to see if it is reliable enough to be used. Otherwise we shall find ourselves in the situation described in objection (ii), above; two workers with the guesses of idiots can validly arrive at different priors and so at different conclusions. Jaynes' only advice in these circumstances (*CP*, 272) is that if the information on the reliability of the orginal information is testable then this can be incorporated by adding a new constraint and then maximising entropy. He goes on to say (*CP*, 272–3) that 'of course, whenever information of this kind is available it should in principle be taken into account in this way. I would "hold as self-evident" that for any problem of inference, the ideal toward which we should aim is that *all* the relevant information we have ought to be incorporated explicitly into the equations.'

The method of maximising entropy is not used by Jaynes only to find a prior distribution which satisfies testable information. He also uses it as a general model building tool in situations where we can observe directly the variable whose distribution we wish to model. The best practical example of this type is Jaynes' analysis of the Wolf Dice data. The Swiss astronomer Rudolph Wolf (1816–1893) conducted an experiment in which a red and a white die were tossed together 20,000 times. As we should expect, using the Chi-squared test, the dice showed no sign of dependence, but there is strong evidence that for neither die are the outcomes equally likely. This is where previous studies ended, with a comment that Wolf wasted his time with badly made dice. Rowlinson [1970] uses the constraint of the observed mean score, as recommended by Jaynes (*CP*, 41–45) in applying the principle of maximum entropy. This leads to a lack of fit with the observed frequencies and he comments that 'what is clearly wrong with the indiscriminate use of this rule, and of the older rules from which it stems, is that they ignore the physics of the problems. Until we know something of the mechanics of the dice . . . we can say almost nothing about p_i. In response to this Jaynes considers the physical causes of bias in the white die (*CP*, 330). 'The two most obvious are (1) a shift of the center of gravity due to the mass of ivory excavated from the spots, which being proportional to the number of spots on any side, should make the quantity $f_1(i) = i-3.5$ have a nonzero expectation; and (2) errors in trying to machine a perfect cube, which will tend to make one dimension (the last side cut) slightly different from the other two. It is clear from the data that Wolf's white die gave a lower frequency for the faces (3, 4); and therefore that the (3–4) dimension was undoubtably greater than the (1–6) or (2–5) ones. The effect of this is that the function $f_2(i) = -2$, if $i = 3, 4$; $+1$ otherwise, has a nonzero expectation.' The observed averages of these two quantities were then used to form constraints before maximising entropy. If only the first constraint is applied then the Chi-squared test provides clear evidence of lack of fit. However, when both constraints are used then the Chi-squared statistic is just significant (the critical level is 0.025). Thus the two obvious causes are sufficient to explain all but a fraction of the bias. 'To assume a further very tiny imperfection [the (2–3–6) corner chipped off] we could make even this discrepancy disappear; but in view of the great number of trials one will probably not consider the result as sufficiently strong evidence for this' (*CP*, 332).

The philosophy underlying this example is completely different from that used when finding prior probabilities. Here the various causes of bias are used to construct models which are then tested against the data. If the fit is not satisfactory then clearly new constraints are required, which can then be investigated. This is simply a standard modelling approach. In fact, as was mentioned above the maximum entropy distribution can be regarded as the Maximum Likelihood distribution of a class in which $\ln(p_i)$ is a linear function of the lambdas (these being the Lagrange multipliers of the maximum entropy method). The use of such log-linear models for multinomial distributions is well-established in the study of multiway contingency tables. Indeed most computer packages which allow log-linear models with Poisson distribution errors[1] will perform the calculations needed to follow Jaynes' method. Jaynes' extension of such models to other examples of the multinomial distribution is particularly valuable, especially since, as the Wolf's Dice problem shows, we can incorporate our theory of the mechanism into the model. A weakness of the method, which has already been mentioned, is that Jaynes seems to employ only a Forward Selection procedure, in which parameters are added until a satisfactory fit is found. This has the disadvantage that no parameter is removed after it has entered the equation. There is nothing, however, in the method which prevents us adopting the more flexible approach that is usually used in linear modelling. Notice also that no attempt is made to use all the information in the data, i.e. the frequencies of each outcome.[2] In particular, information on the reliability of the observed averages is not included, since this is unnecessary. It follows from the theory of Maximum Likelihood Estimation that the variance-covariance matrix of the estimates of the parameters is found from the second derivatives of the logarithm of the likelihood function.

It seems that Jaynes has two different attitudes to the constraints when maximising entropy:

(i) When the variable can be observed directly then the constraint must be physically meaningful, and is tested against the data. This approach will surely be accepted by all and Jaynes' technique will be a valuable addition to the statistician's 'tool-box'.

(ii) When the variable can be observed only indirectly and a prior distribution is required, then the constraints can be anything which is given. Presumably, the assumption is that, when the sample is used

[1] *E.g. GLIM* (Generalised Linear Interactive Modelling), available from *NAG* Algorithms Ltd.

[2] Since, if they were used, then the method would equate the probabilities with the observed relative frequencies.

> to 'update' the prior by calculating the posterior distribution, then the effect of any false assumptions should be negligible. This is fine if the experiment is reasonably informative, but, if prior information is so unimportant, why not use the ignorance prior instead? On the other hand, if prior information is important then we must be prepared to justify anything we use to form our constraints. Since, otherwise, we shall be in the absurd situation of being able to draw contradictory conclusions from an experiment.

Jaynes gives a convincing technique for finding prior distributions which represent ignorance, and of incorporating certainly known prior information into the prior. But if the information is known not to be completely sure, then we have a problem. This is dealt with convincingly by Jaynes only when direct observations are possible.

6 CONCLUSIONS

All scientists have philosophical prejudices, more euphemistically called philosophical presuppositions. One of the useful roles of the philosophy of science is to make these explicit rather than implicit and to help scientists see how they affect the choice and construction of acceptable theories. We admit that our prejudices are towards the view that the role of science is to ask the question 'What is nature really like?' or, in particular cases, 'How does this system behave?' It is, therefore, clear from what has gone before that we are not entirely in sympathy with Jaynes' more restricted aim of simply making satisfactory predictions for a system. That having been said, we find his contribution to the scientific debate both valuable and stimulating and we have sympathy with him in his feeling (see *e.g. CP*, 115, 149–150) that he has in many cases been treated less than fairly. It would seem that his method has in some cases been dismissed after an insufficiently careful analysis, because of the philosophical prejudices of the critic rather than because of the method used to tackle the particular problem under discussion.[1]

Our aim in writing this paper has been to attempt to exemplify the kind of difficulties which we suppose others may have in reading Jaynes' work. We should value a fuller explanation of the points raised. It may be that the philosophical divide would still prevent an acceptance of his complete position but as he says (CP, 112): 'If you do not like my philosophy, but you find that the formalism, nevertheless, does give useful results then I am quite sure that you will be able to invent some *other* philosophy by which that formalism can be justified!' If the publication of his collected papers goes some way towards this end and towards a fuller appreciation of his work it will be fully justified.

D. A. LAVIS and P. J. MILLIGAN
Chelsea College, University of London

[1] One of the writers (Lavis [1977]) would plead guilty to this charge.

REFERENCES

DENBIGH, K. G. [1981]: *Chemistry in Britain*, **17**, p. 168.

FRIEDMAN, J. [1973]: *J. Stat. Phys.*, **9**, p. 265.

FRIEDMAN, J. and SHIMONY, A. [1971]: *J. Stat. Phys.*, **3**, p. 381.

GAGE, D. W. and HESTENES, D. [1973]: *J. Stat. Phys.*, **7**, p. 89.

HOBSON, A. [1967]: *J. Chem. Phys.* **46**, p. 1365.

HOBSON, A. [1972]: *J. Stat. Phys.*, **6**, p. 189.

HOBSON, A. and LOOMIS, D. N. [1968]: *Phys. Rev.*, **173**, p. 285.

JAYNES, E. T. [1983]: *Papers in probability, statistics and statistical physics, ed.* R. D. Rosenkrantz. Reidel.

JANCEL, R. [1969]: *Foundations of classical and quantum statistical mechanics.* Pergamon.

KUBO, R. [1965]: *Statistical mechanics.* North-Holland.

KUHN, T. S. [1962]: *The structure of scientific revolutions.* Chicago.

LAVIS, D. A. [1977]: *Brit. J. Phil. Sci.*, **28**, p. 255.

PENROSE, O. [1979]: *Rep. Prog. Phys.*, **42**, p. 1937.

PFEUTY, P. and TOULOUSE, G. [1977]: *Introduction to the renormalisation group and to critical phemomena.* Wiley.

ROBERTSON, B. [1966]: *Phys. Rev.*, **144**, p. 151.

ROWLINSON, J. S. [1970]: *Nature*, **225**, p. 1196.

SHIMONY, A. [1973]: *J. Stat. Phys.*, **9**, p. 187.

TRIBUS, M. and MONTRONI, H. [1972]: *J. Stat. Phys.*, **4**, p. 227.