BRIAN SKYRMS

UPDATING, SUPPOSING, AND MAXENT

I. INTRODUCTION

Updating subjective belief to assimilate a given bit of information and supposing what the world would be like were that bit of information true, are distinct mental processes for which distinct rules are appropriate. A Warrenite asked to update on the piece of information that Oswald didn't kill Kennedy would come to the conclusion that someone else did; but when asked to suppose what the world would be like had Oswald not killed Kennedy will not suppose that someone else would have. The difference is often marked in ordinary language by the distinction between indicative and subjunctive mood. The Warrenite will assert: "If Oswald didn't kill Kennedy, then someone else did" but deny: "If Oswald hadn't killed Kennedy, then someone else would have".[1]

Given an initial probability measure and a constraint on possible final probability measures one moves to a final probability by the rule of MAXENT if one chooses from among the final probabilities which satisfy he constraint, the one which has minimum information (or equivalently) maximum entropy relative to the intitial probability. This rule was introduced as "the principle of minimum discrimination information" by Kullback and Leibler[2] and as the rule of maximum entropy by Jaynes.[3] It has found application in a wide variety of fields,[4] but its logical status remains a matter of controversy.

Some supporters of MAXENT go so far as to give it the status of a principle of Bayesian *logic*, on a par with additivity of probability or Bayes' rule of conditioning.[5] Some of its detractors claim that it is almost inconsistent with Bayesian methodology.[6] Much of the debate appears to proceed on the assumption, tacit or explicit, that MAXENT is an *inductive* rule, i.e. as a rule for *updating* subjective probabilities.[7] I want to suggest that this is the wrong way to look at MAXENT. Properly viewed, MAXENT is a rule for *stochastic hypothesizing*; a rule for *supposing*.

Theory and Decision **22** (1987) 225–246.

In Section II, I will discuss dynamic coherence requirements for Bayesian updating. The framework will be broad enough to cover cases in which Bayes' rule of conditionalization does not directly apply, including cases in which MAXENT is construed as a rule of updating. In Section III, I will introduce the Stalnaker logic of supposing and adapt it to the case where the possible situations are stochastic; i.e. statistical models. Section IV will develop a few technical facts about the relation of MAXENT, exponential families and sufficient statistics. With the stage thus set, the next two sections will analyze MAXENT alternatively as an updating rule and as a supposing rule. Section V will find it wanting as a generally valid rule for Bayesian updating for essentially the reasons put forward by Shimony and his coworkers. Section VI will argue that MAXENT properly applied as a Bayesian *supposing* or *hypothesizing* rule is perfectly legitimate.

II. UPDATING SUBJECTIVE PROBABILITY

Bayesians update subjective probabilities by *Bayes' Rule: Condition on the evidence!* That is, when presented with new evidence, e, which has positive prior probability, revise yout subjective probabilities such that:

$$\text{NEWPR}(q) = \text{OLDPR}(q \mid e) = \text{OLDPR}(q \,\&\, e)/\text{OLDPR}(e)$$

Why use Bayes' Rule for such situations, rather than some other? It is a necessary and sufficient condition for *dynamic coherence* that one do so.

Thus, suppose that you have an initial probability at time t_1; that there is a partition, E, of your probability space each of whose members have positive initial probability; that then you are to be told the true member of the partition and are to move to a final probability at time t_2 by some epistemic rule for updating probability. A bettor who knows your rule can make a finite number of bets with you at time t_1 according to your probabilities at t_2. He will be said to make an unconditional dynamic dutch book against you if he has a strategy which always leaves him at t_2 with a system of bets whose net gain to him is positive in every possible situation. *You are not open to an unconditional dynamic dutch book if and only if your epistemic rule is Bayes' rule.*

The foregoing updating model presumes that two things are given: (1) a learning situation in which the input is just the information that the true state of affairs is in some designated member of the partition, E, and (2) a initial probability distribution which gives the relevant conditional probabilities: $\Pr(q \mid e)$. The model may not apply in many cases of interest because we do not have the nice package of (1) and (2). What sort of theory is possible if these assumptions are weakened?

Non-Bayesian statisticians (Fisher, Neyman) question (2). Suppose we have a statistical model with an observation space, X, with chances depending on a parameter, θ. In Bayes' method, we assume a prior over the parameter space, calculate $\mathrm{pr}(\theta \mid x)$ by Bayes' theorem, and update on our observation by Bayes' rule. But what if we don't have a prior on the parameter space? Suppose we have some rule assigning probabilities to the parameter values after an observation. Call this rule *chance coherent* if it always assigns posterior probabilities which are immune from a *dutch book in chance*, i.e. a finite system of bets which have a negative chance expectation for every value of θ. Cornfield (1969) and Freedman and Purves (1969) show that our updating rule is chance coherent just in case it coincides with Bayes' method applied to some prior or other over the parameter space.

Epistemologists (Austin, Sellars, Jeffrey) question 1. Various modified versions of the learning situation are possible. Can we say anything general enough to cover them all. Suppose you have at time t_1 an initial probability and will update it at time t_2 to a probability revised in the light of some sort of learning experience. Suppose that you have at time t_1 initial probabilities over one's possible revised probabilities. Assume for the moment that one has only a finite number of possible revised probabilities, each with positive initial probability. A bettor can bet with you at times t_1 and t_2. If we give the learning situation only this much structure, what can we say about dynamic coherence? It is a necessary and sufficient condition for coherence, that your initial probability, pr_1, satisfy principle M:[8]

$$M: \quad \mathrm{pr}_1(q \mid \mathrm{pr}_2 = \mathrm{pr}^*) = \mathrm{pr}^*(q)$$

If we give the learning situation additional structure, M remains necessary for coherence but may not be sufficient. Notice that in the presence of M the move to an enlarged probability space with pr_2 as a random variable

formally embeds our "black box" learning situation in a conditioning model.

It appears that Bayes' rule of conditioning has greater scope in the theory of updating subjective probability than some of its critics have been willing to concede. Very substancial generalizations of the setting for updating carry the consequence that dynamic coherence requires embeddability in a conditioning model.

III. STATISTICAL SUPPOSITION

Supposing as a branch of logic has been developed as the theory of subjunctive conditionals. This may seem rather remote from the subject matter at hand, but I urge the reader to be patient. Subjunctive conditionals were a problem for logical empiricists; a puzzle knot in the hands of Nelson Goodman; and only became a branch of logic when Stalnaker (1968) cut the knot. Stalnaker's idea was to introduce a *selection function*, f, which maps an ordered pair $\langle w,s \rangle$ where w is a possible world and s is a supposition onto a world w'. The idea is that according to that selection function, starting in world (or situation) w and supposing s takes you to w'. Then, a subjunctive conditional "If s were the case, q would be" is true in w just in case q is true in $f\langle w,s \rangle$; false otherwise. Stalnaker required a selection function to have certain properties: (i) supposition s indeed holds in $f\langle w,s \rangle$; (ii) If s holds in w, then $f\langle w,s \rangle = w$; (iii) If s' holds in $f\langle w,s \rangle$ and s holds in $f\langle w,s' \rangle$, then $f\langle w,s \rangle = f\langle w,s' \rangle$. The second and third conditions are motivated by the idea that $f\langle w,s \rangle$ should be the most similar world to w in which s holds. (There is a fourth condition which requires impossible presuppositions to take one to an impossible world where everything is true, but the treatment of impossible presuppositions need not concern us here.) Stalnaker then studied the logic of subjunctive conditionals that holds for every such selection function. Stalnaker's account was generalized by Lewis who challenges the assumptions of existence and uniqueness for "the world minimally different from W in which S holds", and suggests an extension of the Stalnaker semantics to the cases where these assumptions fail.

How can these ideas apply to statistics? Let us shrink the grandiose philosophical notion of a possible world to the more modest one of a possible situation, and let us make that situation stochastic. Then what

we have is a chance distribution over some outcome space. A set of chance distributions over some outcome space X, indexed by some parameter space W – i.e. a statistical model – is a natural domain of application for a stochastic Stalnaker selection function. Stochastic hypotheses will be taken as consistent constraints on the chances (which may undetermine the chance distribution). The game here is to find some interesting sense of "the chance distribution most similar to w which satisfies constraint S".

Note that in this setting the difference between *supposing* and *updating* is mathematically clearcut. In a typical Bayesian updating situation one is uncertain about the chances, and so ones subjective probability distribution on the outcome space is a mixture of the possible chance distributions. Updating is an operation which typically takes one from one point in the interior of the convex closure of the chance distributions to another; supposing moves from one chance distribution to another.

IV. MAXENT

Let us start with the simplest case, where our outcome space, X, contains only a finite number of points, $x_1, x_2, \ldots x_n$. Then the *entropy* of a probability, P, on this space is:

$$-\sum_i P(x_i) \log P(x_i)$$

and the *information* is the negative of the entropy. The minimum information or maximum entropy probability is the one which makes the states equiprobable: $P(x_i) = 1/n$.

Suppose that one has some requirements about what the probability on this space should look like in the form of what expectations it should give to some random variables. These constraints might very well underdetermine the probability measure. In the absence of any further information or disaderata about what the probability should look like, it might seem natural to choose among the probabilities satisfying the constraints that which has the minimum information.[9] This is the rule MAXENT, suggested by Jaynes (1957) and others.

For a simple example, consider an outcome space with just three points, x_1, x_2, x_3. You can think of these as the outcome of the roll of a three sided die. Consider the random variable $f(x_i) = i$, (the number of spots

showing on the die). The MAXENT probability gives $P(x_i) = 1/3$, and $E(f) = 2$. Choosing the MAXENT probability under the constraint that $E(f)$ have a different value has the following results (courtesy of E. T. Jaynes' Basic program, MAXENT 1.16):

$E(f)$	$P(x_1)$	$P(x_2)$	$P(x_3)$
1	1	0	0
0.1	0.907833	0.084333	0.007834
0.2	0.826297	0.147407	0.026297
0.3	0.751567	0.196866	0.051567
0.4	0.681867	0.236267	0.081867
0.5	0.616204	0.267592	0.116204
0.6	0.553972	0.292055	0.153972
0.7	0.494780	0.310440	0.194780
0.8	0.438371	0.323257	0.238271
0.9	0.384586	0.330829	0.284586
2.0	0.333333	0.333333	0.333333
2.1	0.284586	0.330829	0.384586
2.2	0.238372	0.323257	0.438370
2.3	0.194780	0.310440	0.494780
2.4	0.153972	0.292055	0.553972
2.5	0.116204	0.267592	0.616203
2.6	0.081867	0.236267	0.681867
2.7	0.051567	0.196866	0.751567
2.8	0.026297	0.147407	0.826296
2.9	0.007834	0.084332	0.907834
3.0	0	0	1

These results are plotted in Figure 1. Notice that this family of probabilities is not closed under mixing. E.g. an equal mixture of $\langle 1,0,0 \rangle$ and $\langle 0,0,1 \rangle$ is $\langle 1/2,0,1/2 \rangle$ but that is not in the family.

To extend these ideas to the general case, the notion of information needs to be generalized to the Kullback-Leibler discrimination information. Suppose that we start with a prior probability, P, and move to a posterior Q which satisfies certain constraints. For a countable space, W, the discrimination information in Q with respect to P is:

$$I(Q,P) = \sum_W Q(w) \log [Q(w)/P(w)]$$

The finite sample space is a special case with P making the points equiprobable.

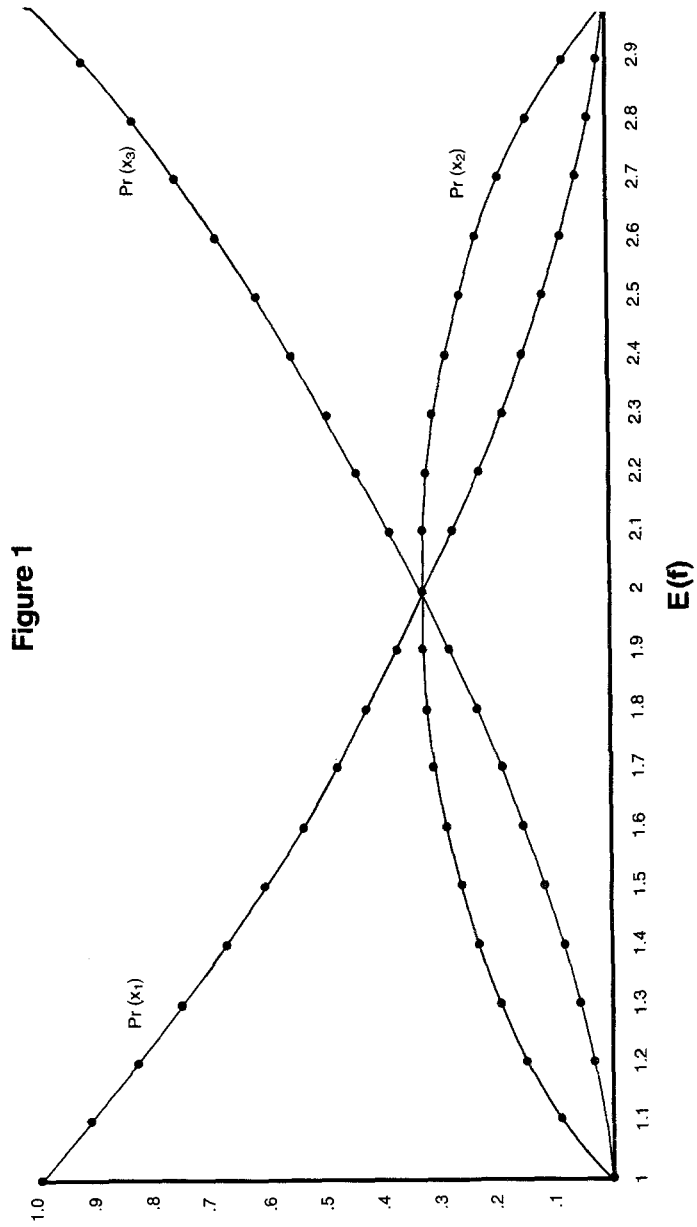
Pr (x3)
Pr (x2)
Pr (x1)
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
1
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
E(f)

Figure 1

More generally, let $\langle W, X_1, \mu_1 \rangle$ be a probability space, with μ_1 being our initial probability measure. Let μ_2 and m be probability measures on this space such that μ_1 and μ_2 are both absolutely continuous with respect to m. Then the Radon-Nikodym derivatives $P = d\mu_1/dm$ and $Q = d\mu_2/dm$ exist and the discrimination information in μ_2 with respect to μ_1 is:

$$I(\mu_2, \mu_1) = \int_W Q(w) \log [Q(w)/P(w)] \, dm$$ [10]

The principle of minimizing this quantity subject to constraints, was put forward and extensively studied by Kullback and Leibler (1951); Kullback (1959).

The notion of a chance supposition or constraint in the most general form imaginable would be just a set of possible chance probability measures. If the constraint is a *convex* set, then if a MAXENT solution exists, it is unique since $I(Q,P)$ is strictly convex in Q. Constraints taking the form of the specification of the desired expectation of a random variable specify such a convex set. Topological conditions on the constraint set which guarantee the existence of a MAXENT solution are given in Csizar (1975).[11]

Consider simple constraints consisting of the specification of the expectation of a random variable, $E(f) = a$; where the MAXENT solution exists. For fixed f, letting a vary *the solutions form an exponential family for which f is a sufficient statistic passing through the initial probability* P. This family has m density:

$$P(x) \exp [k f(x)]/N(k)$$

(Here P is the m density of the initial probability. If we let $m = P$, it is unity. k is adjusted to give the value of a required by the constraint. N is a normalizing factor.) If a member of the constraint set has this density it is the maxent solution. Moreover, if μ_1 is the initial probability and μ_2 is the MAXENT solution, then for any probability, m, in the constraint set:

(MDI) $\quad I(m, \mu_1) = I(m, \mu_2) + I(\mu_2, \mu_1)$

MAXENT solves the problem of selecting a member of the constraint set by using the initial probability and the statistic of the constraint to generate in a canonical way a statistical model which contains just one member of the constraint set. Essentially the same thing happens in the

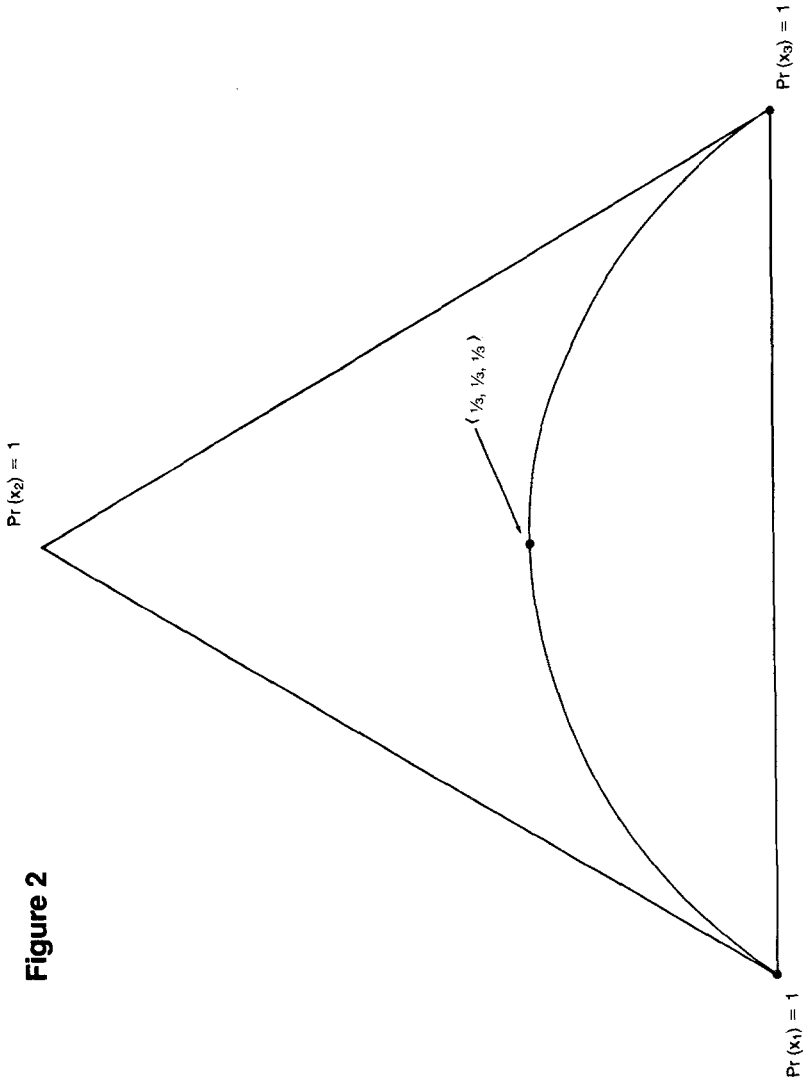

Figure 2

general case.[12] The exponential family of MAXENT solutions for the three sided die is graphed in Figure 2.

V. MAXENT AS BAYESIAN UPDATING

As we say in Section II, *coherence* gives a special place in the theory of updating subjective probability to Bayes' rule of conditioning. Bayes' rule is a special case of MAXENT. Let the random variable I_c be the indicator function which takes the value 1 in the set c and the value 0 outside c. Let c be a set with non-negative initial prior probability. And let the constraint be $E(I_c)=1$. Then the MAXENT solution is the same as that gotten by updating by Bayes' rule, i.e. conditioning on c. This follows immediately from Kullback's theorem (noted in Section III) that the statistic of the constraint is a sufficient statistic for the exponential family of MAXENT solutions. This means that probabilities conditional on c must be the same in all members of the family, including the initial probability. Sufficiency together with the fact that in the final probability $E(I_c)=1$ determine that the final probability comes from the initial one by conditioning on c.

Kullback's sufficiency theorem also leads immediately to the result that Richard Jeffrey's (1965) generalization of Bayes' rule is also a special case of MAXENT. Suppose there is a finite partition, $\{p_i\}$, each of whose members have positive probability. Jeffrey says that a final probability comes from an initial probability by *probability kinematics on the partition* $\{p_i\}$, just in case the probabilities conditional on members of the partition remain the same during the change. The MAXENT solution for a finite number of constraints of the form $E[Ip_i] =$ Finalprobability (p_i) is just the change by probability kinematics which results in those final probabilities for members of the partition. Jeffrey's rule has a connection with dynamic coherence, although it is a little more delicate than that of Bayes' rule because of the relativity to a partition.[13] One way of putting the matter is by embedding Jeffrey's rule in a conditioning model as in Section II. That is, take the initial "small" probability space and enlarge it by adding final probability as a random variable. Say that $\{p_i\}$ is subjectively sufficient for belief change if in the initial probability, $\mathrm{PR}\,[r \mid \wedge_i \mathrm{pr}_f(p_i)=a_i \,\&\, p_j]=\mathrm{PR}\,[r \mid p_j]$ for all r in the small space and all members of the partition. Then a necessary condition for dynamic

coherence is that it be by belief change by probability kinematics on all partitions subjectively sufficent for belief change.[14]

The connection between MAXENT and the updating rules of Bayes and Jeffrey have led to speculation that MAXENT may be a generally valid rule for updating subjective probability.[15] Much of the critical discussion of the MAXENT rule has also cast it in this role.[16] A consideration put forward in 1971 by Friedman and Shimony shows that it is difficult to maintain this point of view.

Consider the case of the three sided die which we have been using as a simple illustrations. Suppose that you are a MAXENT updater and that your initial subjective probabilities are 1/3 for each side. The desired final expected number of spots, $E_f(g) \in [1,3]$, is for you information that you will somehow acquire before updating. Consider your initial probability over possible values of the desired expectation. They are tantamount to initial probabilities over your possible final probabilities since for you the MAXENT rule associates a unique final probability. It might occur to you to take the flat prior (with respect to Lebesque measure) on [1,3]. But this choice would be *dynamically incoherent!* The prior probability would not equal the expectation of posterior probability, and a dynamic dutch book could be made against you. All right, you needen't make that application of the principle of insufficient reason. What do your initial probabilities on the value of the desired expectation need to be in order to escape the dynamic dutch book? You must concentrate probability 1 on the desired expectation being 2! This is the only way in which the initial expectation of the final probability of 2 spots can equal the initial probability of 2 spots, because the final probability of 2 spots under the MAXENT revision rule is a strictly concave function of the desired expected number of spots, taking its maximum at the initial probability of 2 spots, 1/3. This is easily seen in Figures 1 and 2. But this is just the case in which under MAXENT the initial probability is not revised. It appears that the MAXENT updater in this case can only be coherent if he believes with probability one that the rule will not lead to any substantive belief revision. This is hardly a desideratum for a rule for updating subjective probability.

There have been attempts to discount the Shimony-Friedman example, but I do not think that they are successful. Williams (1980) claims that the rule does not violate static coherence:

According to the present interpretation, the probabilities emerging from the principle of minimum information are not conditional probabilities associated with the prior distribution but unconditional probabilities of a new and entirely different distribution, unrelated to the prior distribution by the normal 'synchronic' probability calculus. This is to be understood even in the case corresponding to Bayesian conditionalization. If this is accepted, objections of the type raised by Friedman and Shimony (1971) are not applicable.

I think that the discussion of dynamic coherence in Section II shows that this response to the example is inadequate. It might be argued that you might not have probabilities over the possible values of the constraint; but if such values are incoming data, I see no reason why a Bayesian should not be able to have probabilities on them. It might be argued that the constraint isn't data, but rather something quite different. In a sense I think that this is correct, but this is really to give up maintaining that MAXENT is a rule for Bayesian updating and to assert that it is a rule for something else.

Notice that the Friedman-Shimony example applies to a wide range of "minimal revision" rules for updating; not just MAXENT. Any rule which provides a solution satisfying the constraint must agree with MAXENT for $E(g)=1$ and $E(g)=3$. Any rule which gives a unique minimal revision must agree with MAXENT on $E(g)=2$ since it takes no revision to satisfy this constraint. If in addition the rule makes a monotonic transition in final probability of 2 spots from $E(g)=2$ to the extremes, the Friedman-Shimony reasoning applies.

In fact, let us suppose that for whatever reason, your initial probabilities for the desired value of the constraint are concentrated on the values 1,2,3. Why shouldn't they be? Then any minimal revision rule which calls for no revision if the constraint is actually satisfied will be subject to the Friedman-Shimony analysis.

To understand what is happening, it is instructive to embed the problem in several alternative Bayesian settings:

EXAMPLE 1 (*Determinism*). You are sure that the die is a trick die which will always come up the same way, but are unsure which is the favored side with probabilities 1/3,1/3,1/3. A friend will conduct a large number of independent trials (to all intents and purposes infinite) and report to you the sample mean. Your initial probabilities for $E(g)=1,2,3$ are

1/3,1/3,1/3. If he reports $E(g)=2$, your final probability for two spots should be 1, rather than the 1/3 prescribed by MAXENT.

EXAMPLE 2 (*Determinism or No Information*). As above, but there is a one in four chance that your friend will forget to perform the experiment, in which case he will also report $E(g)=2$. Then your initial probabilities for $E(g)=1,2,3$ are respectively 1/4,1/2,1/4. A report of $E(g)=2$. has ambiguous significance. Upon recieving such a report you should change your final probability of 2 spots showing to 2/3.

EXAMPLE 3 (*Uncertain Chance*). You believe that the die is a chance device with uncertain chance, and your initial probabilities for the chances are given by the measure uniform with respect to Lebesgue measure in Euclidian space (see Figure 3). The data is a report of the true chance expectation. The data $E(g)=2$ should lead you to revise your probability upward to 1/2.

In Examples 1–3 the report $E(g)=2$, although compatible with your present probability is nevertheless grounds for Bayesian revision. MAXENT gives different results in these cases because it interprets $E(g)=2$ as "no news" and no reason for revision. In extreme case in which MAXENT is compatible with conditioning, i.e. where the initial probability of $E(g)=2$ is one, the data that $E(g)=2$ really is no news in the sense of Bayes' rule. The probabilities conditional on it must be the same as the unconditional probabilities. But this is hardly the typical setting for Bayesian updating. MAXENT is not a generally valid updating rule.

VI. MAXENT AS SUPPOSITION

Suppose we have a given chance model; for instance the equiprobable chance probability on the three sided die, and want to hypothesize about the chance probability which satisfies a certain constraint and is in some interesting sense most similar to our given chance model. To do this systematically, we would like something like a Stalnaker selection function for chance models. It is just this that MAXENT gives us, at least for certain well behaved hypotheses.

In the general case, a constraint set of probabilities will be considered

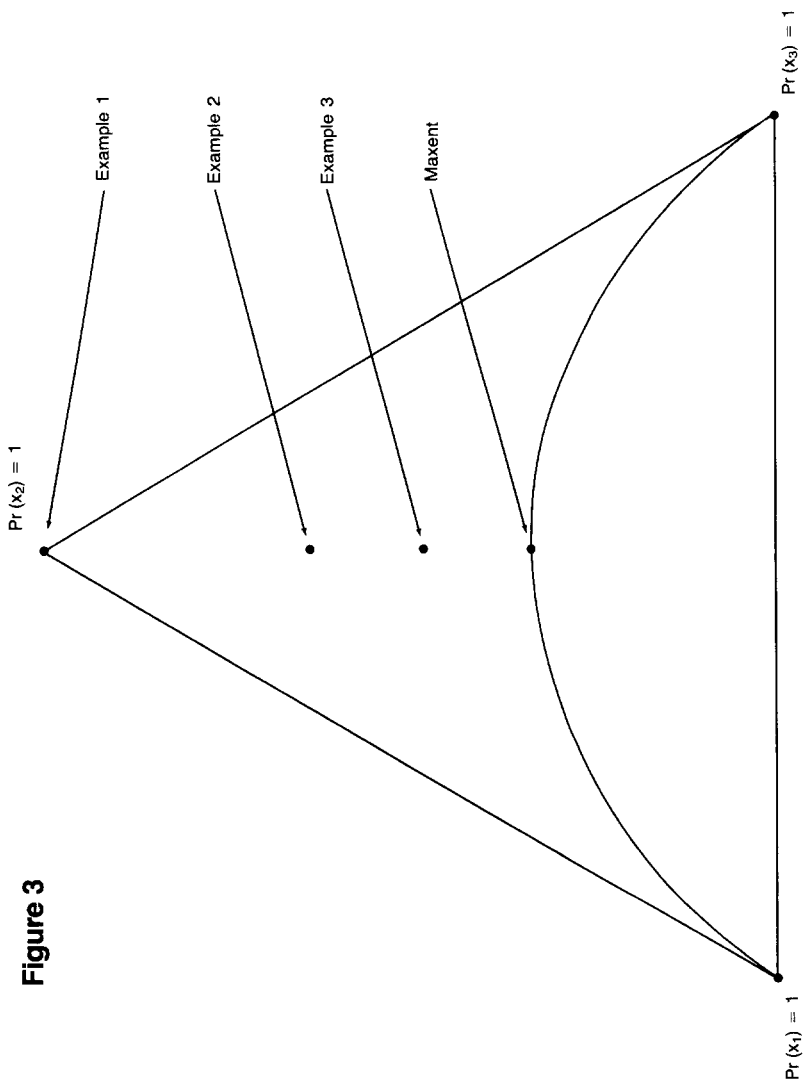

Figure 3

a well-behaved hypothesis with respect to an initial probability if it is a convex, closed set of probabilities absolutely continuous with respect to the initial probability. For such hypotheses the MAXENT solution exists and is unique, so on the domain of well-behaved hypotheses, the MAXENT method determines a MAXENT SELECTION FUNCTION.

It is easy to see that the MAXENT selection function is STALNAKER, i.e. it satisfies Stalnaker's three conditions listed in Section III:

(i) By definition of the MAXENT solution it is in the constraint set, so $f(w,s) \in s$.

(ii) Since the information in the initial probability with respect to itself is minimal and the Kullback-Leibler $I(\mathrm{pr}_f,\mathrm{pr}_i)$ is strictly convex in pr_f, it follows that if pr_i satisfies the constraint, it is the MAXENT solution. i.e. If $w \in s$ then $f(w,s) = w$.

(iii) Show that $f(w,s) \in s'$ and $f(w,s') \in s$ then $f(w,s) = f(w,s')$: By hypothesis and the MDI equality of Section IV, we have both:

$$I(f\langle \mathrm{pr}_i,s'\rangle,\mathrm{pr}_i) = I(f\langle \mathrm{pr}_i,s'\rangle, f\langle \mathrm{pr}_i,s\rangle) + I(f\langle \mathrm{pr}_i,s\rangle,\mathrm{pr}_i)$$

and

$$I(f\langle \mathrm{pr}_i,s\rangle,\mathrm{pr}_i) = I(f\langle \mathrm{pr}_i,s'\rangle, f\langle \mathrm{pr}_i,s'\rangle) + I(f\langle \mathrm{pr}_i,s'\rangle,\mathrm{pr}_i).$$

Since all terms are nonnegative:

$$I(f\langle \mathrm{pr}_i,s\rangle, f\langle \mathrm{pr}_i,s'\rangle) = I(f\langle \mathrm{pr}_i,s'\rangle, f\langle \mathrm{pr}_i,s\rangle) = 0$$

By strict convexity, $f\langle \mathrm{pr}_i,s\rangle = f\langle \mathrm{pr}_i,s'\rangle$ as required.

Notice that it is just these properties which qualigy MAXENT as defining a Stalnaker selection function for well-behaved hypotheses which caused trouble for it as a method of Bayesian updating. (In the examples where the initial probability was concentrated on $E(2) = 1,2,3$, the trouble can be gotten from Conditions (i) and (ii) alone.) Any stochastic Stalnaker selection function will get into Friedman-Shimony difficulties if it is applied as a rule for updating subjective probabilities.

This has a certain general significance, because there is a whole family of minimal revision rules which can be made to yield Stalnaker selection functions for chance models, several of which have been considered as possible rules for mechanical updating of subjective probability. One can minimize the variational distance, the Hellinger distance, etc.[17] Each of these should be thought of as defining a different selection function for well-behaved hypotheses or suppositions rather than as rules for updating subjective probability.

VII. SUPPOSING AS CONDITIONING

MAXENT gives us one selection function among many possible ones. Is there anything specially interesting about this one? There is a deep connection with the concept of *sufficiency*, and with conditioning after all in the one limiting case in which MAXENT is consistent with conditioning.

As Jaynes (1979) notes,[18] the MAXENT solution gives us the exponential families as the Darmois-Koopman-Pitman theorem. Consider the exponential family: $\mathrm{PR}(X) = \mathrm{P}(X) \exp [a\ T(X)]/N$. T is a sufficent statistic. Furthermore, for multiple IID trials the sum $\sum_i T(X_i)$ is a sufficient statistic. Darmois-Koopman-Pitman is the converse. If the sum is a sufficient statistic, then we have the functional equation:

$$\mathrm{PR}_a[T(X_1)]\,\mathrm{PR}_a[T(X_2)]\ldots = \mathrm{PR}_a[T(X_1) + T(X_2) + \ldots]$$

which under mild regularity conditions has the exponential solution.[19]

If the members of the family are the physical probabilities, then in a typical case of uncertainty about the true physical probabilities, degree of belief will be a mixture of the members of the family. In the product space if the physical probabilities make the trials independent, then the degrees of belief will be exchangeable. deFinetti's theorem shows how to go the other way. An exchangeable sequence of random variables has a unique representation as a mixture of independent ones. If in the degree of belief probabilities T is a sufficient statistic such that the sum is a sufficient statistic for multiple trials, then Koopman-Pitman-Darmois can be combined with deFinetti to characterize the extreme points; the "physical probabilities" which are implicit in the degree of belief probabilities. They consist of exponential families of the statistic.[20] This means that for a subjectivist who regards chances or physical probabilities as artifacts of the deFinetti representation theorem: *if T is for him such an additive sufficient statistic then MAXENT applied to constraints of the form $E_f(T) = b$ is for him a way of moving from one possible physical probability to another.*

Typically Bayesian conditioning is applied to update subjective probability by moving from one non-trivial mixture of possible physical probabilities to another. As we saw in Section V, MAXENT fails to be embeddible in a conditioning model in such contexts. Suppose, however,

that your subjective probability is concentrated on one possible physical probability; e.g. you are sure that the three sided die is fair and multiple trials are independent. You now expect with limiting probability one that the average number of spots showing in a long series of trials will be 2. Suppose that you now get the information that this empirical average on a long sequence of trials *including trial i* was different from 2, say b. Conditional on the evidence, your probabilities of outcomes on trial 1 will change. Taking the limit of these probabilities as the number of trials goes to infinity, we get the probability distribution that is given by applying MAXENT to the constraint $E_f(T)=b$. More generally, under suitable regularity conditions if T is a sufficient statistic, for an IID sequence of random variables conditioning on an empirical average $1/n \sum_i T(X_i)=b$ gives in the limit the same result as applying MAXENT to $E_f T(X_1)=b$.[21] Conditioning on a biased mean of a sufficient statistic can be used to give us a supposing or hypothesizing rule; a way of moving from one statistical hypothesis to another. *MAXENT gives us a selection function with the remarkable property of agreement with the rule of conditioning on a biased mean of a sufficient statistic.*

VIII. CONCLUSION

The philosophical controversy concerning the logical status of MAXENT may be in large measure due to the conflation of two distinct logical roles: (1) A general inductive principle for updating subjective prohabilities (2) a supposing rule for moving from one chance probability to another. When judged under standards of dynamic coherence appropriate to (1), MAXENT is found wanting. When judged in terms of the logic appropriate to (2) MAXENT yields for convex closed constraint sets a reasonable selection function with interesting connections with sufficiency and conditioning. Indeed it is just the features of MAXENT which make it appropriate for (2) which make it inappropriate for (1). MAXENT can be thought of as part of Bayesian logic. But it is part of the logic of supposition rather than the logic of induction.[22]

NOTES

[1] The example is due to Adams (1970). See also the discussion in Lewis (1976).

[2] Kullback and Leibler (1951); The rule is extensively studied in Kullback (1959).

[3] Jaynes (1957, 1963, 1967, 1974, and 1980).
[4] E.g. statistical mechanics (Jaynes, 1957; image enhancement Frieden, 1972).
[5] Jaynes, Shore and Johnson (1980), Williams (1980), Cheesman.
[6] Friedman and Shimony (1971), Shimony (1073), Dias and Shimony (1981), Shimony (1985), Seidenfeld (198–).
[7] At least insofar as one can tell from the discussion.
[8] The dutch book is constructed in Goldstein (1983) and van Fraassen (1984). For further discussion of principle M and its generalizations see Gaifman (forthcoming) and Skyrms (1987a,b).
[9] Uniqueness is guaranteed by strict convexity of the entropy as a function of P.
[10] Or, letting $m=P$, $\int_W Q \log Q \, dP$.
[11] The constraint set being a convex set closed in the topology of variational distance guarantees the existence of a MAXENT solution.
[12] See Kullback (1959); Csizar (1975). In particular the minimum discrimination information equation, (MDI) holds in general for the maxent solution it exists. We will use this fact in Section VI.
[13] There is an extensive discussion in Skyrms (1987).
[14] See Skyrms (1980a,b) and Good (1981).
[15] E.g. Williams (1980) "the purpose of the principle is to assist in the rational modification of beliefs." (p. 132, ftnt 1). See also Shore and Johnson (1980); Domotor (1980); Cheesman (1983).
[16] van Fraassen (1980, 1981); Shimony (1973); Dias and Shimony (1981); Friedman and Shimony (1971).
[17] For a quick survey see Diaconis and Zabell (1982) Sections 5 and 6. The point applies generally to minimization of any f-divergence in the sense of Csizar (1967). Diaconis and Zabell show that the variational distance and Hellinger distances are both f-divergences. See also May and Harper (1976).
[18] "An interesting fact which may have some deep significance as not yet seen, is that the class of maximum entropy functions is, by the Pitman-Koopman theorem, identical with the class of functions admitting sufficient statistics." Jaynes (1979), p. 87.
[19] There is a whole family of theorems of this nature. One can consider multidimensional sufficient statistics, and one can consider refinements of the regularity conditions. See Koopman (1936), Hipp (1974).
[20] See Freedman (1962) and Diaconis and Freedman (1981).
[21] See van Campenhout and Cover (1981); Tjur (1974); Zabell (1974).
[22] Research partially supported by N.S.F. grant SES–8605122.

REFERENCES

Adams, E.: 1962, 'On rational betting systems', *Archive für mathematische Logik und Grundlagenforschung* **6**, 7–29, 112–128.
Adams, E.: 1970, 'Subjunctive and indicative conditionals', *Foundations of Language* **6**, 89–94.
Adams, E.: 1975, *The Logic of Conditionals*, D. Reidel, Dordrecht.
Armendt, B.: 1980, 'Is there a Dutch book, argument for probability kinematics?', *Philosophy of Science* **47**, 583–588.
Cheesman, P.: 1983, 'A method of computing generalized Bayesian probability values for

expert systems', *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, Karlsruhe, West Germany, 198–202.

Cheesman, P.: 1985, 'In defense of probability', *Proc. AAAI-85*, 1002–1009.

Csizar, I.: 1975, 'I-divergence geometry of probability distributions and minimization problems', *Annals of Probability* **3**, 146–158.

Dawid, A. P. and Stone, M.: 1972, 'Expectation consistency of inverse probability distributions', *Biometrika* **59**, 486–489.

Dawid, A. P. and Stone, M.: 1973, 'Expectation consistency and generalized Bayes inference', *The Annals of Statistics* **1**, 478–485.

de Finetti, B.: 1937, 'La prévision: ses lois logiques, ses sources subjectives', *Annales de l'Institut Henri Poincaré* **7**, 1–68. Tr. as 'Foresight: Its logical laws, its subjective sources', in *Studies in Subjective Probability*, ed. H. E. Kyburg, Jr. and H. Smokler (Huntington, N.Y.: Krieger, 1980).

Dempster, A. P.: 1967, 'Upper and lower probabilities induced by a multivalued mapping', *Annals of Mathematical Statistics* **38**, 325–339.

Dempster, A. P.: 1968, 'A generalization of Bayesian inference', *Journal of the Royal Statistical Society, Series B* **30**, 205–249.

Diaconis, P. and Zabell, S.: 1982, 'Updating subjective probability', *Journal of the American Statistical Association* **77**, 822–830.

Diaconis, P. and Zabell, S.: 1983, 'Some alternatives to Bayes' rule', Technical Report 205, Dept. of Statistics, Stanford University.

Diaconis, P. and Freedman, D. A.: 1985, 'Partial exchangeability and sufficiency', *Statistics: Applications and New Directions*, ed. J. Gosh and J. Roy (Indian Statistical Institute: Calcutta), pp. 205–236.

Dias, P. M. C. and Shimon, A.: 1981, 'A critique of Jaynes' maximum entropy principle', *Advances in Applies Mathematics* **2**, 172–211.

Domotor, Z.: 1980, 'Probability kinematics and the representation of belief change', *Philosophy of Science* **47**, 384–403.

Dynkin, E.: 1978, 'Sufficient statistics and extreme points', *Annals of Probability* **6**, 705–730.

Freedman, D. A.: 1962, 'Invariants under mixing which generalize de Finetti's theorem', *Annals of Mathematical Statistics* **33**, 916–923.

Freedman, D. A. and Purves, R. A.: 1969, 'Bayes method for bookies', *The Annals of Mathematical Statistics* **40**, 1177–1186.

Frieden, B. R.: 1972, 'Restoring with maximum likelihood and maximum entropy'', *Journal of the Optical Society of America* **62**, 511–518.

Friedman, K. and Shimony, A.: 1971, 'Jaynes's maximum entropy prescription and probability theory', *Journal of Statistical Physics* **3**, 381–384.

Gaifman, H.: forthcoming, 'A theory of higher order probabilities', in *Probability and Causation: Proceedings of the Irvine Conference*, ed. Harper, W. and Skyrms, B., D. Reidel, Dordrecht, Holland.

Gibbard, A.: 1981, 'Two recent theories of conditionals', in *Ifs*, ed. Harper *et al.*, D. Reidel, Dordrecht.

Goldstein, M.: 1983, 'The prevision of a prevision', *Journal of the AMerican Statistical Association* **78**, 817–819.

Good, I. J.: 1950, *Probability and the Weighing of Evidence*, Hafner, New York.

Good, I. J.: 1963, 'Maximum entropy or hypothesis formation, especially for multidimensional contingency tables', *Annals of Mathematical Statistics* **34**, 911–934.

Good, I. J.: 1981, 'The weight of evidence provided by uncertain testimony or from an uncertain event', *Journal of Statistical Computation and Simulation* **13**, 56–60.

Hacking, I.: 1967, 'Slightly more realistic personal probability', *Philosophy of Science* **34**, 311–325.

Heath, D. and Sudderth, W.: 1972, 'On a theorem of de Finetti, oddsmaking and game theory', *Annals of Mathematical Statistics* **43**, 2072–2077.

Heath, D. and Sudderth, W.: 1978, 'On finitely additive priors, coherence, and extended admissibility', *Annals os Statistics* **6**, 333–345.

Hipp, C.: 1974, 'Sufficient statistics and exponential families', *Annals of Statistics* **2**, 1283–1292.

Hunter, D.: 1986, 'Uncertain reasoning using maximum entropy inference', in *Uncertainty in Artificial Intelligence*, ed. L. K. Konol and J. F. Lemmaer, Elsevier Science Publishers, Amsterdam.

Jaynes, E. T.: 1957, 'Information theory and statistical mechanics', *Physical Review* **106**, 620–630.

Jaynes, E. T.: 1963, 'Information theory and statistical mechanics', in *Statistical Physics*, ed. Ford, Benjamin, N.Y., pp. 181–218.

Jaynes, E. T.: 1967, 'Foundations of probability theory and statistical mechanics', *Delaware Seminar in the Foundations of Physics*, ed. Bunge, Springer, Berlin, 77–101.

Jaynes, E. T.: 1974, *Probability Theory with Applications in Science and Engineering*, Department of Physics Washington Univeristy, St. Louis.

Jaynes, E. T.: 1980, 'Where do we stand on maximum entropy', in *The Maximum Entropy Formalism*, ed. Levine and Tribus, MIT Press, Cambridge, 15–118.

Jeffrey, R.: 1965, *The Logic of Decision*, McGraw Hill, N.Y. (2nd ed., University of Chicago Press: Chicago, 1983).

Jeffrey, R.: 1968, 'Probable knowledge', in *The Problem of Inductive Logic*, ed. Lakatos, North Holland, Amsterdam.

Kemeny, J.: 1955, 'Fair bets and inductive probabilities', *Journal of Symbolic Logic* **20**, 263–273.

Koopman, B. O.: 1936, 'On distributions admitting a sufficient statistic', *Transactions of the American Mathematical Society* **39**, 399–409.

Kullback, S. and Liebler, R.: 1951, 'On information and sufficiency', *Annals of Mathematical Statistics* **23**, 8–102.

Kullback, S.: 1959, *Information Theory and Statistics*, Wiley, New York.

Kupperman, M.: 1958, 'Probabilities of hypotheses and information-statistics in sampling from exponential class populations', *Annals of Mathematical Statistics* **29**, 571–574.

Lane, D. and Sudderth, W.: 1983, 'Coherent and continuous inference', *The Annals of Statistics* **11**, 114–120.

Lehman, R.: 1955, 'On confirmation and rational betting', *Journal of Symbolic Logic* **20**, 251–262.

Lewis, David: 1976, 'Probabilities of conditionals and conditional probabilities', *Philosophical Review* **85**, 297–315.

Lewis, David: 1973, *Counterfactuals*, Blackwell, Oxford.

May, S. and Harper, W.: 1976, 'Toward an omtimization procedure for applying minimum change principles in probability kinematics', *Foundations of Probability Theory, Inductive Inference, and Statistical Theories of Science* **1** (Reidel: Dordrecht), pp. 137–166.

Ramsey, F. P.: 1931, 'Truth and probability', in *The Foundations of Mathematics and Other Essays*, ed. R. B. Braithwaite (N.Y.: Harcourt Brace), and in *Studies in Subjective Probability*, ed. H. Kyburg and H. Smokler (Huntington, N.Y.: Krieger, 1980).

Seidenfeld, T.: 'Entropy and uncertainty', forthcoming in *Philosophy of Science*.

Shafer, G.: 1976, *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, N.J.

Shafer, G.: 1981, 'Jeffrey's rule of conditioning', *Philosophy of Science* **48**, 337–362.

Shimony, A.: 1955, 'Coherence and the axioms of confirmation', *Journal of Symbolic Logic* **20**, 1–28.

Shimony, A.: 1973, 'Comment on the interpretation of inductive probabilities', *Journal of Statistical Physics* **9**, 187–191.

Shore, J. and Johnson, R.: 1980, 'Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy', *IEEE Transactions in Information Theory It-26* **1**, 26–37.

Skyrms, B.: 1980a, *Causal Necessity*, Yale University Press, New Haven, Appendix 2.

Skyrms, B.: 1980b, 'Higher order degrees of belief', in *Prospects for Pragmatism*, ed. D. H. Mellor, Cambridge University Press, Cambridge.

Skyrms, B.: 1983, 'Zeno's paradox of measure', in *Physics, Philosophy, and Psychoanalysis*, ed. R. S. Cohen and L. Laudan, D. Reidel, Dordrecht, Holland, 223–254.

Skyrms, B.: 1984, *Pragmatics and Empiricism*, Yale University Press, New Haven, Conn.

Skyrms, B.: 1985, 'Maximum entropy inference as a special case of conditionalization', *Synthese* **63**, 55–74.

Skyrms, B.: (1987a), 'Dynamic coherence and probability kinematics', in *Philosophy of Science* **54**, 1–20.

Skyrms, B.: (1987b), 'Dynamic coherence', in *Advances in the Statistical Sciences VII Foundations of Statistical Inference*, ed. I. B. MacNeill and G. Umphrey, D. Reidel, Dordrecht, 233–243.

Skyrms, B.: (1987c), 'Coherence', in *Scientific Inquiry in Philosophical Perspective*, ed. N. Rescher, University of Pittsburgh Press, Pittsburgh, 225–242.

Skyrms, B.: (forthcoming), 'The value of knowledge', in *Justification Discovery, and Evolution of Scientific Theories*, ed. C. Wade Savage, University of Minnesota Press, Minneapolis.

Stalnaker, R. C.: 1968, 'A theory of conditionals', *Studies in Logical Theory*, ed. N. Rescher, Blackwell, London.

Teller, P.: 1973, 'Conditionalization and observation', *Synthese* **26**, 218–258.

Teller, P.: 1976, 'Conditionalization, observation, and change of preference', in *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, ed. W. Harper and C. Hooker (D. Reidel, Dordrecht-Holland), pp. 205–253).

Tjur, T.: 1974, *Conditional Probability Distributions*, Lecture Notes 2, Institute of Mathematical Statistics, University of Copenhagen, Sections 36, 37.

van Campenhout, J. and Cover, T.: 1981, 'Maximum entropy and conditional probability', *IEEE Transactions on Information Theory IT-27*, 483–489.

van Fraassen, B.: 1980, 'Rational belief and probability kinematics', *Philosophy of Science* **47**, 165–187.

van Fraassen, B.: 1981, 'A problem for relative information minimizers in probability kinematics', *British Journal for the Philosophy of Science* **32**, 375–379.
van Fraassen, B.: 1984, 'Belief and the will', *Journal of Philosophy* **81**, 235–256.
Williams, P. M.: 1980, 'Bayesian conditionalization and the principle of minimum information', *British Journal for the Philosophy of Science* **31**, 131–144.
Zabell, S.: 1974, 'A limit theorem for conditional expectations with applications to probability theory and statistical mechanics', PhD. Dissertation, Harvard University Cambridge, Massachusetts.

Department of Philosophy,
University of California at Irvine,
Irvine, CA 92717,
U.S.A.