Kyburg, H. 1978. Subjective probability, *Journal of Philosophical Logic* **7**, 157–80.

Lewis, D. K. 1966. An argument for the identity thesis, *Journal of Philosophy* **63**, 17–25.

MacFarland, D. 1976. Form and function in the temporal organisation of behaviour. In *Growing Points in Ethology*, ed. P. Bateson and R. Hinde, pp. 55–94. Cambridge.

McGinn, C. 1978. Mental states, natural kinds and psychophysical laws, *The Aristotelian Society Supplementary Volume* **52**, 195–220.

Peirce, C. S. 1935–56. *Collected Papers*. Vols 1–6 ed. C. Hartshorne and P. Weiss, vols 7–8 ed. A. Burks. Cambridge, Mass.

Putnam, H. 1975. *Mind, Language and Reality*. Cambridge.

Ramsey, F. P. 1978. *Foundations*, ed. D. H. Mellor. London.

Stich, S. 1978. Subdoxastic states, *Philosophy of Science* **45**, 499–518.

Thayer, H. S. 1968. *Meaning and Action: A Critical History of Pragmatism*. Indianapolis.

6 *Higher order degrees of belief*

BRIAN SKYRMS

It is hardly in dispute that people have beliefs about their beliefs. Thus, if we distinguish degrees of belief, we would not shrink from saying that people have degrees of belief about their degrees of belief. It would then be entirely natural for a degree-of-belief theory of probability to treat probabilities of probabilities. Nevertheless, the founding fathers of the theory of personal probability are strangely reticent about extending that theory to probabilities of higher order. Ramsey does not consider the possibility. De Finetti rejects it. Savage toys with it, but decides against it. I. J. Good (1965) and E. T. Jaynes (1958) put the mathematics of higher order probability to work, but remain rather non-committal about its interpretation. This reticence is, I believe, ill-founded.

I will argue here that higher order personal probabilities are legitimate, non-trivial, and theoretically fruitful. In part **I** I will defend the conception of higher order personal probabilities against charges of inconsistency, illegitimacy and triviality. In part **II** I will illustrate one aspect of their theoretical fruitfulness in connection with the question of the laws of motion for rational belief, and the relations between probability kinematics, the information theoretic approach to statistics, and conditionalisation.

I *The legitimacy of higher order personal probabilities.* The worst suspicion that has been voiced about higher order probabilities is that they lead to an actual *inconsistency*. Thus, in the development of his

I would like to thank Zoltan Domotor, Glenn Shafer, and Bas van Fraassen for letting me see copies of papers not yet published. Each of these illuminates some of the issues discussed here, and each represents a different philosophical viewpoint. I would also like to thank Richard Jeffrey for discussion of some of these ideas.

theory in terms of conditional probabilities of propositions in *Probability and the Weighing of Evidence*, ch. III, I. J. Good takes pains to exclude higher-order probabilities: 'it will be taken that the propositions *E, H, etc.* never involve probabilities or beliefs' (Good 1950: 19). With regard to this restriction he makes the following comment:

> The development of the abstract theory must follow the rules of ordinary logic and pure mathematics. Hence we could, at this stage, hardly allow the propositions E, F, H, etc. to involve probabilities . . . To what extent this restriction may be relaxed is an interesting question. If it were entirely relaxed . . . the resulting theory would have some convenience, but it would also be confusing and might even be self-contradictory (Good 1950: 20).

Good does not spell out the inconsistency that he has in mind, so we can only speculate as to the nature of the perceived danger. It is, of course possible to blunder into an inconsistency when treating propositions and propositional attitudes. Suppose one maintained that there is a set of all propositions *P*; that for any subset *S* of that set, there is a proposition to the effect that George believes just the members of *S*; that if *S* and *S′* are distinct sets, the propositions to the effect that George believes just the members of these sets respectively are distinct propositions. One would then be maintaining that there is a set *S*, whose power set can be mapped into it, which is impossible. There are various variations one can play on this. In particular, what can be done with belief can, *a fortiori*, be done with probability. The set of probability distributions over a given set of propositions is of greater cardinality than the initial set of propositions. There is some reason to believe that Good has this sort of difficulty in mind. He touches on the matter again in the next chapter. 'Perhaps the most obvious method would be to extend the meaning of the word 'proposition' so as to allow it to refer to probabilities, but this course may lead to logical difficulties', a remark which receives the following amplification in a footnote: 'it may require a 'theory of types' as in symbolic logic' (Good 1950: 41).

The moral of this story for those who wish to consider higher order probabilities is simply, 'Be careful'. We know how to avoid such contradictions. One can start with some ground level set of propositions (without any claims to exhaustiveness), and build a language–metalanguage hierarchy on top of it, adding at each level propositions about the probabilities of lower-level propositions. (I take it that this sort of idea is what is behind Good's reference to types.) This is not to say that the story is uninteresting for ontologists who wish to think in some sense about all propositions. And psychological theorists who are interested in propositional attitudes may well draw the conclusion that the hierarchy shouldn't be run up so high that the results won't fit in their subjects' heads. But the fear that considerations of probabilities of probabilities *must* involve presuppositions of the sort that led in our story to an inconsistency is groundless. (For an explicit construction of a system of higher order personal probabilities see Gärdenfors (1975).)

Another way in which probabilities of probabilities have been thought to cause logical difficulties is embodied in a paradox due to David Miller (1966). The paradox can be put as follows:

Premiss 1: $Pr(\text{not-}E) = Pr[E \text{ given that } Pr(E) = Pr(\text{not-}E)]$.
Premiss 2: $Pr[E \text{ given that } Pr(E) = Pr(\text{not-}E)] = 1/2$.
Conclusion: $Pr(\text{not-}E) = 1/2$.

Since the proof is for any proposition, *E*, we have not just an absurdity, but also an inconsistency with the rules of the probability calculus.

This paradox generated a surprising amount of discussion in the journals, but it really should be transparent to anyone who has paid attention to recent philosophy of language, for it rests on a simple *de dicto–de re* confusion. (Let us remember that the probability contexts at issue are intensional; the probability that the morning star = the evening star may not equal the probability that the morning star is the morning star.) Consider premiss 1. Its plausibility depends on the appropriate *de re* reading of the right hand expression: '$Pr[E$ given that $Pr(E) = Pr(\text{not-}E)]$'. That is, in Donellan's terminology, the embedded description '$Pr(\text{not-}E)$' is to be thought of *referentially*. If the actual probability of not-*E* has a certain value, say 3/4, then I think of the embedded description '$Pr(\text{not-}E)$' having as its sole function the designation of this value. There is nothing wrong with:

$$3/4 = PR[E \text{ given that } Pr(E) = 3/4]$$

or indeed with its generalisation:

$$a = PR[E \text{ given that } Pr(E) = a]$$

(assume that $PR[Pr(E) = 3/4] \neq 0$, so that the conditional probability is well-defined in the standard way) where '*a*' is rigid designator: that is, a name which designates the same numerical value at every point in the space. I will call this principle *Miller's principle*. Those who have followed the development of modal logic will already know that we invite no additional difficulty by universally generalising Miller's principle to:

$$\text{for any } x,\ x = PR[E \text{ given that } Pr(E) = x]$$

provided that we restrict universal specification to rigid designators of the type indicated. We shall see that Miller's principle has a genuine significance independent of Miller's paradox.

The second premiss of Miller's paradox depends on a *de dicto* reading for its plausibility. It requires that the description, '*Pr*(not-*E*)' be taken attributively rather than referentially. We are to think of it as designating at a point in the probability space the value of the random variable at that point in the probability space, not as a rigid designator of a numerical value. Likewise for the description '*Pr*(*E*)'.

It is evident, then, that Miller's paradox is simply a fallacy of equivocation. The plausibility of the first premiss depends on reading 'the probability of *E*' attributively and 'the probability of not-*E*' referentially. The plausibility of the second depends on reading them both attributively. If both are given a uniform attributive reading, and the probability of *E* is not, in fact, 1/2, then the first premiss is false, and can be derived from Miller's principle only by a fallacious universal specification.

These are the two arguments I know that allege a formal inconsistency in the higher order probability approach. I would say nothing more about formal inconsistency were it not that some reputable philosophers continue to have suspicions (if not arguments) in these directions. Though it may be a case of bringing out a cannon to swat a fly, I therefore feel obliged to point out that there is implicit in de Finetti's work a proof of formal consistency for a theory of second order probabilities: simply interpret *pr* as relative frequency probability (*i.e.* probability conditional on relative frequency. Indeed any way of explaining *pr* as "objectified" probability relative to a partition will do. See Jeffrey (1965: ch. 12).). This is not the intended interpretation, but it suffices to settle the question of consistency.

One might, however, hold that, although formally consistent, a theory of higher-order *personal* probabilities is, in some way, *philosophically* incoherent. This appears to be de Finetti's position. De Finetti adopts an *emotive* theory of probability attribution (de Finetti 1972).

> Any assertion concerning probabilities of events is merely the expression of somebody's opinion and not itself an event. There is no meaning, therefore, in asking whether such an assertion is true or false or more or less probable . . . speaking of unknown probabilities must be forbidden as meaningless.

If probability attributions are merely ways of evincing degrees of belief, they do not express genuine propositions and are not capable themselves of standing as objects of probability attribution.

De Finetti's positivism stands in sharp contrast to Ramsey's pragmatism:

> There are, I think, two ways in which we can begin. We can, in the first place, suppose that the degree of belief is something perceptible by its owner; for instance that beliefs differ in the intensity of a feeling . . . of conviction, and that by the degree of belief we mean the intensity of this feeling. This view . . . seems to me observably false, for the beliefs we hold most strongly are often accompanied by practically no feeling at all . . .
>
> We are driven therefore to the second supposition that the degree of belief is a causal property of it, which we can express vaguely as the extent to which we are prepared to act on it.
>
> . . . the kind of measurement of belief with which probability is concerned is . . . a measurement of belief *qua* basis of action, (Ramsey 1926: 71).

For Ramsey then, a probability attribution is a theoretical claim. It is evident that on Ramsey's conception of personal probability, higher order personal probabilities are permitted (and indeed required). (It is perhaps also worth noting that anyone who takes Ramsey's view of degrees of belief, and is willing to accept personal probabilities of propensities, or propensities of propensities, must also accept second-order personal probabilities, for on Ramsey's view personal probabilities *are* a kind of propensity.)

Even from de Finetti's viewpoint, the situation is more favourable to a theory of higher order personal probabilities than might at first appear. For a given person and time there must *be*, after all, a proposition to the effect that that person then has the degree of belief that he might evince by uttering a certain probability attribution. De Finetti grants as much:

> The situation is different of course, if we are concerned not with the

assertion itself but with whether 'someone holds or expresses such an opinion or acts according to it,' for this is a real event or proposition (de Finetti 1972: 189).

With this, de Finetti grants the existence of propositions on which a theory of higher order personal probabilities can be built, but never follows up this possibility.

Perhaps this is because of another sort of philosophical objection to second-order personal probabilities which, I think, is akin to the former in philosophical presupposition, though not in substance. Higher order personal probabilities are well-defined all right – so this line goes – but they are trivial; they only take on the values zero and one. According to this story, personal probabilities – if they exist at all – are directly open to introspection; so one should be certain about their values. If my degree of belief in p is x, then my degree of belief that my degree of belief in p is x will be one, and my degree of belief that my degree of belief in p is unequal to x will be zero. Put so baldly, the objection may seem a bit silly, but I will discuss it because I think that something like it often hovers in the background of discussions of personal probability. But, first, I would like to point out that this objection has a much narrower scope than the previous one. According to the view now under consideration, it is perfectly all right to postulate non-trivial personal probabilities about personal probabilities, if they are my probabilities now about your probabilities now or my probabilities now about my probabilities yesterday or tomorrow. What become trivial, according to this view, are my probabilities now about my probabilities (that I am introspecting) now.

The foregoing objection is an expression of a form of positivism which most philosophers would consider a combination of bad psychology and bad epistemology. Ramsey's pragmatism is again good medicine. If we focus on degrees of belief *qua* basis of action rather than the intensity-of-feeling notion, there is much less reason to put so much weight on introspection. (It is perhaps worth a passing remark that those philosophers who argue that personal probabilities don't exist because they can't introspect them are relying on the same positivistic preconceptions.) For a dispositional sense of belief, the status of my beliefs about my beliefs now is not so different in principle from the status of my beliefs now about my beliefs yesterday, or indeed about the status of my beliefs now about your beliefs now (although there will typically be differences in degree). In a word, the dispositional sense of belief makes sense of the possibility that someone may not *know his own mind* with certainty, and thus makes sense of this last disputed case. (See Jeffrey (1974) for a discussion of second order preferences, desires and probabilities.)

I should mention at this point that some philosophers do adopt a pragmatic, dispositional sense of belief but do so in such a rigid operationist way that they are led to have verificationist doubts about the case in question. The following argument has been made to me in conversation:

> Probability is a disposition to bet in certain ways. To test his second order degrees of belief, we must get him to bet on his first order degrees of belief. To determine the payoff on this bet we must test his first order degrees of belief. To do this we must get him to bet on ground level propositions. But the ratios at which he bets on these propositions may be distorted by his efforts to protect his previous higher order wagers.

To this objection there is both an internal and external reply. The internal reply is that we can ameliorate the bias by making the first order bets small with respect to the second order bets. The external reply is that one surely need not be so rigidly operationist as to assume that the *only* way that one can gain evidence for a degree of belief is by making a wager. The pragmatic notion of probability that Ramsey espouses in 'Truth and probability' is by no means so rigid. Ramsey thinks of personal probabilities as theoretical parts of an imperfect but useful psychological model, rather than as concepts given a strict operational definition. Ramsey's point of view is, I think, infinitely preferable to either the left-wing positivism implicit in the objection just discussed, or the right-wing positivism of the one preceding it.

There is some psychological evidence, however, which suggests that even Ramsey's modest claims of approximate truth for the theory of personal probability as a psychological theory may be overstated. Actual preferences often appear to be ill-defined, or, where defined, incoherent. Depending on how bad things really are (I will not try to evaluate that here), it may be better to stress the normative rather than the descriptive aspect of the theory of personal probability. According to this view, the theory of personal probability is a prescription for coherence, just as the theory of deductive logic contains prescriptions for consistency. It is this strand of thought that is really fundamental, I think, in 'Truth and probability' and it remains even if the average man proves more incoherent than Ramsey expected. Let us notice now that if the theory of personal probabilities is conceived of as medicine, then we need second order medicine for our second order degrees of belief just as we need first order medicine for our first order degrees of belief. Higher order personal probabilities remain a natural and indeed an inescapable part of the theory of personal probability.

I hope that in the preceding I have been able to sweep away some of the philosophical debris that has played a part in blocking the development of a theory of higher order personal probabilities. But even when one is convinced that the conception is consistent and philosophically legitimate, then the question remains as to whether they are of any special interest. Savage's brief discussion in *The Foundations of Statistics* is along these lines:

> there seem to be some probability relations about which we feel relatively "sure" as compared with others. When our opinions, as reflected in real or envisaged action, are inconsistent, we sacrifice the unsure opinions to the sure ones . . . There is some temptation to introduce probabilities of a second order so that the person would find himself saying such things as 'the probability that B is more probable than C is greater than the probability that F is more probable than G.' But such a program seems to meet insurmountable difficulties . . .
>
> If the primary probability of an event B were a random variable b with respect to secondary probability, then B would have a "composite" probability, by which I mean the (secondary) expectation of b. Composite probability would then play the allegedly villainous role that secondary probability was intended to obviate, and nothing would have been accomplished.
>
> Again, once second order probabilities are introduced, the introduction of an endless hierarchy seems inescapable. Such a hierarchy seems very difficult to interpret, and it seems at best to make the theory less realistic, not more.
>
> Finally, the objection concerning composite probability would seem to apply, even if an endless hierarchy of higher order probabilities were introduced. The composite probability of B would here be the limit of a sequence of numbers, $E_n (E_{n-1} (\ldots E_2(P_1(B)) \ldots))$, a limit that could scarcely be postulated not to exist in any interpretable theory of this sort . . .
>
> The interplay between the "sure" and "unsure" is interestingly expressed by de Finetti thus: 'The fact that a direct estimate of a probability is

> not always possible is just the reason that the logical rules of probability are useful. The practical object of these rules is simply to reduce an evaluation, scarcely accessible directly, to others by means of which the determination is rendered easier and more precise' (Savage 1972: 57–8).

In this passage, Savage appears to have two rather different motivations in mind for higher order probabilities. The first is the consideration that he begins with: that there is a second-order aspect of our beliefs, *i.e.* "sureness" about our first order beliefs, which is not adequately reflected in the first order probability distribution alone. The second is the idea that second order distributions might be a *tool* for representing vague, fuzzy, or ill-defined first order degrees of belief with greater psychological realism than a first order distribution would provide. This second motivation is implicit in the discussion of the "insuperable difficulties", and becomes even clearer in a footnote to the second edition:

> One tempting representation of the unsure is to replace a person's single probability measure P by a set of such measures, especially a convex set (Savage 1972: 58).

I think that it is very important to carefully distinguish these two lines of thought. Savage's "insuperable difficulties" are serious objections against the suggestion that second order distributions provide a good mathematical representation of vague, fuzzy, or ill defined first order beliefs. Indeed, an apparatus of second order distributions presumes more structure than conventional first order distributions rather than less, and the first order structure can be recovered as an expectation (providing we have Miller's principle: see Skyrms 1980: Appendices 0 and 1). But *however* we wish to model vague or fuzzy first order degrees of belief, we shall, given beliefs about beliefs, wish to model vague or fuzzy second order degrees of belief as well. Interval valued, fuzzy logical, and convex set representations of imprecise first order degrees of belief are not *competitors* with second order probabilities; they are aimed at a different problem. If we then return to Savage's first motivation, we find that *vis à vis* this problem, the "insuperable objections" are not objections at all. The extra structure of higher order probabilities is just what is wanted. That two second order distributions for $pr(p) = x$ can have the same mean but different variance gives us a representation of the intuitive phenomenon with which Savage broached the discussion: two people may have the same first order probabilities, but different degrees of sureness about them.

There is one further strand in the passage from Savage that invites comment. Savage speculates that the notion of sureness may give us some insight into probability *change*: 'When our opinions, as reflected in real or envisaged action are inconsistent, we sacrifice the unsure opinions to the sure ones.' One version of Savage's first objection might hold that everything that we can know about probability change is already encoded in the first order conditional probabilities, so that any second order information must be either redundant or irrelevant. Such a position rests on several questionable premisses; but there is one in particular to which I would again like to call attention. That is, that second order probabilities should only be treated *instrumentally*, *i.e.* that the relevant inputs and outputs of probability change must always be first order. Once we take the philosophical position that higher order probabilities can refer to something as real as first order probabilities, it opens up the possibility of conditionalising at a higher level, *e.g.* conditionalising on some statement about the first order probability distribution. It therefore opens up possibilities that simply do not exist as we restrict ourselves to the first order setup. I believe that these possibilities do indeed illuminate questions of probability change. I will give a brief illustration of this in the second part of this paper.

II *Higher order personal probabilities and the question of the laws of motion for rational belief.* The 'rational' in 'rational belief' refers to *coherence*. The idea of justifying the probability calculus as embodying laws of *static* coherence for degrees of belief occurred independently to Ramsey and de Finetti. Each had the idea that qualitative constraints could lead to a representation theory for probability. And each had the idea of a Dutch book theorem; a theorem to the effect that if probabilities are taken as betting quotients, then someone who violates the laws of the probability calculus would be susceptible to a system of bets, each of which he considers fair or favourable, such that he would suffer a net loss no matter what happened. A great deal turns on the significance of these theorems, and indeed this has been the subject of some philosophical dispute. I think that the way in which Ramsey states the Dutch book theorem is enlightening:

> If anyone's mental condition violated these laws, his choice would depend on the precise form in which the option were offered him, which would be absurd. He could then have a book made against him by a cunning bettor and would then stand to lose in any event (Ramsey 1926: 84).

It is clear that what is important for Ramsey about coherence, and what makes it for him a kind of consistency, is that someone who is incoherent is willing to bet on the same betting arrangement at two different rates, depending on how that arrangement is described to him. The remark about the cunning bettor is simply a striking corollary to this fundamental theorem. Thus, let the criterion of individuation of a *betting arrangement* be the schedule specifying the *net payoff* on each possible outcome. The additivity law for probability is then justified by the observation that the same betting arrangement may either be described as a bet on a disjunction of two mutually exclusive propositions, or as the upshot of separate bets on each of the two propositions. The condition that the betting arrangement be evaluated consistently, no matter which advertising brochure accompanies it, is just that the probability of the disjunction be equal to the sum of the probabilities of the disjuncts. Along the same lines, de Finetti provides a justification for the customary definition of conditional probability:

$$Pr(q \text{ given } p) = Pr(p \;\&\; q)/Pr(q)$$

via the notion of a conditional bet. A bet on q conditional on p is called off if p is false, otherwise won or lost depending on the truth value of q. Again such a conditional betting arrangement can be redescribed as the upshot of separate bets on p & q and against p, with the consequency that coherence *requires* the foregoing treatment of conditional probability.

I find these arguments very compelling. And I think that some philosophers who fail to find them compelling, fail to do so because they focus on the striking corollary about the cunning bettor rather than on the fundamental theorem. 'Must the rational man always behave,' they ask, 'as if the world were a cunning bettor, lying in wait to make a Dutch book?' Asking the question in this way appears to make the subjective theory of probability rest on a kind of methodological paranoia that is usually associated only with the theory of games. This is, I think, the wrong way to look at the question. Of course there are situations in which a little incoherence won't hurt you, just as there are situations in which a little deductive inconsistency won't hurt you. (Remember, it is Ramsey's remark that he believes each of his beliefs but believes that at least one of his beliefs is false.) Of course there are situations in which it would be too costly to remove an incoherence to be worth it, just as there are situations in which it would be too costly to remove a deductive inconsistency to be worth it. Ramsey's pragmatism is not William James' pragmatism! But this is all, I think, beside the point. At a deeper level, Ramsey and de Finetti have provided a way in which the fundamental laws of probability can be viewed as pragmatic consistency conditions: conditions for the consistent evaluation of betting arrangements no matter how described.

The question naturally arises as to whether there is any analogous coherence argument for ways of *changing* degrees of belief. Ramsey strongly suggests that he believes that there *is* such an argument for conditionalisation:

> Since an observation changes (in degree at least) my opinion about the fact observed, some of my degrees of belief after the observation are necessarily inconsistent with those I had before. We have therefore to explain how exactly the observation should modify my degrees of belief; obviously if p is the fact observed, my degree of belief in q after the observation should be equal to my degree of belief in q given p before, or by the multiplication law to the quotient of my degree of belief in pq by my degree of belief in p. When my degrees of belief change in this way we can say that they have been changed *consistently* by my observation (Ramsey 1926: 94).

but does not explicitly set out any such argument. Hacking (1967) doubts if there can be such an argument, and regards it as a serious failing of Bayesian theory that this "dynamic assumption" lacks a justification. Nevertheless, David Lewis has produced a coherence argument for conditionalisation (reported in Teller (1973). See also Freedman and Purves (1969).) I would like to give the leading idea of Lewis' argument here, so that it may be compared with the static coherence arguments. Suppose that I am about to find out whether a certain proposition, p, is true or false (*e.g.* the result of a certain experiment is about to come in); that I have a rule or disposition to change my degrees of belief in a certain way upon learning that p is true; that PR represents my degrees of belief just before learning whether p is true or not and PR_p the degrees of belief that I would have according to the rule (or disposition) upon learning that p is true. The key point is this: prior to finding out about p, *the rule or disposition to change my beliefs in a certain way upon learning p is tantamount to having a set of betting ratios for bets conditional on p.* (Someone can achieve a betting arrangement for a bet on q conditional on p with me, at the betting ratio $PR_p(q)$, just by reserving a sum of money which he will bet on q with me *after* I change my degrees of belief if p turns out true, and which he will not bet at all if p turns out false.) But we also know from de Finetti's observation that PR alone commits us to betting ratios for conditional bets in a different way, with those betting ratios being reflected in the conditional probabilities of PR (assuming $PR(p) \neq 0$). For the conditional betting ratios arrived at in these two ways to coincide, PR_p must come from PR by conditionalisation on p (*i.e.* for all q, $PR_p(q) = PR(p \,\&\, q)/PR(p)$). (Obviously, the same argument can be repeated for $PR_{\sim p}$, and for the more general case where the experimental report may consist of any one of a set of mutually exclusive and exhaustive propositions.)

This observation yields a Dutch book theorem as a corollary. If someone does not change his degree of belief by conditionalisation, then someone who knows how he does change his belief can exploit the different betting ratios for bets conditional on p to make a Dutch book conditional on p, which can then be turned into an unconditional Dutch book by making an appropriate small side bet against p.

We can only speculate as to whether Ramsey had this sort of argument in mind. But it is clear that the Lewis argument is quite in the spirit of Ramsey, and rests on the same conception of pragmatic consistency as the static consistency arguments of Ramsey. It is, I think, undeniable that it establishes a special status for conditionalisation as a law of motion for rational belief in the cases which satisfy the conditions of the argument. But what of cases in which these conditions are not met? In particular, what about those cases to which Ramsey alludes, in which observation changes *in degree* my opinion about the fact observed, but where that change is not a change to probability 1 of some observation proposition? Richard Jeffrey (1965) introduces *probability kinematics* for just this purpose. Suppose that an observational interaction autonomously changes the probability of some proposition, p, but does not change it to 1. In such a situation we might plausibly decide to take as our final probability distribution a mixture (weighted average) of the probability distribution we would get by conditionalising on p, and the probability distribution that we would get by conditionalising on not-p. Then, for any q,

$$\begin{aligned} PR_{final}(q) = {} & PR_{final}(p)\, PR_{initial}(q \text{ given } p) \\ & + PR_{final}(\sim p)\, PR_{initial}(q \text{ given } \sim p). \end{aligned}$$

Under these circumstances, we say that the final probability distribution comes from the initial probability distribution by probability kinematics on p. More generally,

> *Probability Kinematics:* Let Pr_i and Pr_f be probability functions on the same field of propositions and let $\{p_j : j = 1, \ldots, n\}$ be a partitioning of that field such that $Pr_i(p_j) \neq 0$ and $Pr_f(p_j) = a_j$. Pr_f is said to come from Pr_i *by probability kinematics on* $\{p_j\}$ iff:
>
> For all propositions, q, in the field:
> $Pr_f(q) = \Sigma_j a_j Pr_i$ (q given p_j).

This is equivalent (Jeffrey 1965: ch. 11) to:

> For all q and j : Pr_f(q given p_j) $=$ Pr_i(q given p_j).

Conditionalisation on p is a special case of probability kinematics where the partitioning consists of $[p, \sim p]$ and the final probability of p is 1. Probability kinematics takes a certain special kind of constraint on the final probability distribution as its input, the constraint as to the final probabilities of the p_js. E. T. Jaynes, the originator of the information theoretic approach to statistical mechanics, suggests a more generally applicable rule (Jaynes 1957): Maximise the relative entropy in the final distribution relative to the initial distribution subject to the stated constraints. Jaynes' maxim can be put roughly as: Be as modest as possible about the amount of information you have acquired. Several writers have recently pointed out that probability kinematics is a special case of Jaynes' rule (May & Harper 1976, Shafer 1979, and van Fraassen and Domotor, Zanotti and Graves in not yet published papers). But neither the maximum entropy rule in general nor the special case of probability kinematics has the kind of Ramsey–de Finetti justification that Lewis supplied for conditionalisation. True modesty is, no doubt, an epistemic virtue but false modesty is not, and the question is now to distinguish true modesty from false.

If only we had some proposition in our language which summed up the content of our imperfect observation, we could simply conditionalise on it. But the assumption that every observation can be interpreted as conferring certainty to some observational proposition leads to an unacceptable epistemology of the given. There is, however, another, entirely natural way in which the sorts of cases under consideration can be assimilated to conditionalisation.

Within the framework of second order personal probabilities, we can answer that in the case of probability kinematics there was, after all, something that we did learn for certain from the observation. We learned the values of the final probabilities of the members of the partition. The same remark generalises to other cases in which Jaynes' rule of maximising relative entropy relative to a set of constraints on the final distribution applies. In the higher order probability setup, we can conditionalise on statements specifying those constraints (by conditionalising on random variables on the second order probability space). I would like to proceed to discuss the relation between second order conditionalisation and the first order generalisations of conditionalisation suggested by Jeffrey and Jaynes. I will start with the case of probability kinematics, but much of what I have to say will carry over to maximum relative entropy inference as well.

I would first like to set out the formal connection between first order probability kinematics and second order conditionalisation, and then discuss the interpretation of this connection in the light of what I have said about higher order personal probabilities. First, the framework for second order probabilities:

Let L_1 consist of some field of propositions. We extend L_1 to L_2 by adding every proposition of the form: $pr(p) \in S$ where p is a proposition of L_1 and S is a subinterval of the unit interval, and closing under finite truth-functional combination. (We could iterate this process as far as you please.) I will here only discuss a probability distribution PR on L_2 with English being the language of discussion.

Suppose that $[p_i : i = 1, \ldots, n]$ is a partition such that $PR[p_i] \neq 0$ and let $[a_i : i = 1, \ldots, n]$ be numerical values such that $PR[\bigwedge_i pr(\bigwedge_i p_i) = a_i] \neq 0$. Under what conditions does conditioning on the second order proposition, $\bigwedge_i pr(p_i) = a_i$, which specifies probability values for every member of the partition, have the same effect at the first order level as probability kinematics on that partition $[p_i]$? By the characterisation (Jeffrey 1965: ch. 11) of probability kinematics on $[p_i]$ as a change which leaves the probabilities of propositions conditional on members of the partition unchanged, it follows that

> Conditionalisation on $\bigwedge_i pr(p_i) = a_i$ is equivalent at the first order level to probability kinematics on the partition $[p_i]$ if and only if:
>
> (Sufficiency Condition): PR (q given p_j and $\bigwedge_i pr(p_i) = a_i$)

= $PR(q$ given $p_j)$ for all first order propositions, q, and all elements, p_j, of the partition.

For conditionalisation on $\bigwedge_i pr(p_i) = a_i$ to also be a change which leads to each member of the partition $[p_i]$ having as its final probability the corresponding a_i, we must also have:

(GENERALIZED MILLER) : $PR(p_j$ given $\bigwedge_i pr(p_i) = a_i) = a_j$ for all elements, p_j, of the partition.

(The special case of the foregoing observation, for probability kinematics on a partition consisting of $[p, \sim p]$ is discussed in Skyrms (1980: Appendix 1).)

This bit of mathematics is open to more than one interpretation. One could, for instance, use it to argue for probability kinematics in cases in which one, by reflection, comes to know his own mind a little better. But, here, I would like to focus on the sort of interpretation for which it was designed. Here, *pr* signifies the final probabilities that are the upshot of an observational interaction. Under this interpretation it is plausible that there should be a wide range of first order propositions for which the sufficiency condition holds. Notice that here the analogy with the de Finetti decomposition breaks down. In the de Finetti setup, relative frequency plays the role of our random variable, *pr*. De Finetti shows that an infinite sequence of exchangeable trials has a unique representation as a mixture of *independent* identically distributed trials. So our sufficiency condition fails: *e.g.* consider an infinite sequence of exchangeable tosses of a coin, which is an equal mixture of two Bernoulli sequences with pr(heads) = 1/4 and pr(heads) = 3/4 respectively. Then PR[Heads on trial 2 given pr(heads on 1) = 3/4] = 3/4; PR[Heads on trial 2 given heads on trial 1] = 9/16; but PR[Heads on trial 2 given both heads on trial 1 and pr(heads on 1) = 3/4] = 3/4. The plausibility of our sufficiency condition depends on the interpretation of *pr* as degree of belief. In the de Finetti setup, sufficiency runs the other way. There we have:

$$PR(q \text{ given } pr(q) = a_i \wedge p) = PR(q \text{ given } pr(q) = a_i) = a_i \text{ (for all } i)$$

where PR is concentrated on $a_1 \ldots, a_n$ as before. And thus

$$PR\,(q \text{ given } p) = \Sigma_i a_i PR\,(pr(q) = a_i \text{ given } p)$$

(First order conditionalisations) = (Second order probability kinematics)!

But for personal probabilities one is almost tempted to think of the sufficiency condition as a methodological postulate. My degrees of belief with regard to the p_is are irrelevant to the probability of q in the presence of the truth about the world regarding the p_is (*i.e.* p_j). This, however, would be going too far. Notice that the sufficiency condition would not be plausible if we allowed second order propositions to take the place of the q. But we can think of examples of first order propositions which are highly correlated with the second order final probability propositions in question, and for these the sufficiency condition may fail too. (*E.g.* my current probability that I will sweat at the moment of arriving at my final probability, conditional on the fact that Black Bart will not really try to gun me down *and* that my final probability that he will try to kill me will be 0.999, is *not* equal to my current probability that I will sweat, conditional on the fact that he will not really try to gun me down. The sweating is highly correlated with my final degree of belief rather than the fact of the matter.) So we must make do with the more modest claim that in typical situations there is a wide range of first order propositions for which the sufficiency condition holds. When our first order language consists of such propositions (relative to the probability measure and partition in question) we shall have probability kinematics on that partition as the first order consequence of second order conditionalisation. The generalised Miller condition is also highly plausible under this interpretation, though for different reasons. Here the plausibility depends on the interpretation of *pr* as my *final* probability, after the observational interaction. Under the assumption that my final probability is to be arrived at by conditionalisation on $\bigwedge_i pr(p_i) = a_i$, and under the assumption that *pr* is to be interpreted as final probability, the generalised Miller principle says that conditional on the final probability of p_j having a certain value (and a few other things), it has that value. Of course, we can invent cases, where these assumptions do not hold. I might have reason, for instance, to believe that my final probabilities would not be reached by conditionalisation, but rather would be distorted and biased in a foreseeable way by an evil force. Contemplating that sort of final probability distribution from my antecedent, clearheaded state, I would not have probabilities $[PR]$ which exemplify Miller's principle, but rather probabilities which compensate for the projected bias in $[pr]$. The point I would like to make here is that the approach by way of higher order probabilities both shows us why probability kinematics is the right approach in a wide variety of cases, and enables us to isolate "pathological" cases in which it would give the wrong results. Furthermore, in the cases in which it is correct, it appears not merely as a successful *ad hoc* method, but rather as a special case of second order conditionalisation. As such, it inherits the force of Lewis' coherence argument.

I mentioned another way in which the generalised Miller principle might fail which leads to questions of independent interest. The Miller principle might fail if $[pr]$ were not interpreted as a final probability. Now the model that we have been studying, where the observational interaction forces *final* probabilities of the elements of a partition on you, and your final probabilities of all other sentences are then determined with reference to these, leaves something to be desired. It would be more informative if we could separate the observational input of a probability change from the final upshot. We might take as input the observational probabilities of the elements of a partition, *i.e.* the values that they would have on the basis of the observation alone – and then combine these with our prior information to get a final probability distribution. (I suggested in Skyrms (1975: 196–8) that such an analysis would be desirable, but could not see how to provide it. An observational parametrisation, at the first order level, has recently been provided by Field (1978).) Suppose, then, that we reinterpret pr as observational probability rather than final probability (which I will indicate with a subscript, o). How does this reinterpretation affect the sufficiency and generalised Miller conditions? The considerations that were adduced in favour of the sufficiency condition remain substantially unchanged; information as to the true member of the partition should typically swamp the effects of observational probabilities of members of the partition on the probability of a first order sentence, q. But under the reinterpretation, the generalised Miller principle is no longer plausible. I observe a piece of cloth by candlelight. My observational probability that it is black is 0.8; that it is purple 0.05; dark blue 0.05; other 0.1. I may, however, already have strong evidence that the cloth is blue, and perhaps also that the light is deceiving, *etc.* This prior evidence has been assimilated into my probability function PR. Under these conditions, my final probabilities for the members of the partition (black, purple, dark blue, other) will not equal the observational probabilities, but rather will be the end product of some way of weighing the observational probabilities against the other evidence. So here we can expect probability kinematics, by virtue of the sufficiency principle, but we should not expect:

$$PR(q \text{ given } \textstyle\bigwedge_i pr_o(p_i) = a_i) = \Sigma_i a_i PR(q \text{ given } p_i)$$

because of the failure of the generalised Miller principle.

prec

Suppose that p_1 and p_2 are mutually exclusive hypotheses and o is a piece of evidence. Then, by Bayes theorem, $PR(p \text{ given } o) = Pr(p)$ $PR(o \text{ given } p)/PR(o)$, so we have $PR(p_1 \text{ given } o)/PR(p_2 \text{ given } o) = [PR(p_1)/PR(p_2)][PR(o \text{ given } p_1)/PR(o \text{ given } p_2)]$. That is, the ratio of the final probabilities of the hypotheses on the evidence can be obtained by multiplying the ratio of the initial probabilities by the *likelihood ratio*. Furthermore, if we have a series of pieces of evidence, such that the pieces of evidence are independent, conditional on the hypotheses, then the likelihood ratio for the conjunction of the pieces of evidence can be obtained by multiplying through, *i.e.*:

$$\begin{aligned}&PR(p_1 \text{ given } \textstyle\bigwedge_i o_i)/PR(p_2 \text{ given } \textstyle\bigwedge_i o_i)\\ &\quad = [PR(p_1)/PR(p_2)]\pi_i[PR(o_i \text{ given } p_1)/PR(o_i \text{ given } p_2)].\end{aligned}$$

If we want to "add up" evidence instead of multiplying through, we need only take logarithms of the terms (the choice of base being a matter of convention). Following I. J. Good (1950) (who was following a suggestion of Alan Turing) we focus on the logarithm of the likelihood ratio, $\ln[PR(o \text{ given } p_1)/PR(o \text{ given } p_2)]$ as the *weight of evidence* in o for p_1 as against p_2. For a series of observations which are independent, conditional on the hypotheses, the final relative probabilities $PR(p_1 \text{ given } \bigwedge_i o_i)/PR(p_2 \text{ given } \bigwedge_i o_i)$ can be recovered from the prior relative probabilities and accumulated weight of evidence as:

$$[PR(p_1)/PR(p_2)]\exp\Sigma_i \ln[PR(o_i \text{ given} p_i)/PR(o_i \text{ given given} p_2)].$$

Shafer (1979) shows how Field's model falls out of Good's Bayesian analysis, provided that there are propositions, o_i, in the domain of $[PR]$ which represent the results of the observational interaction. Thus, the Field's formula, $PR_f(p)/PR_f(\sim p) = [PR_i(p)$ $/PR_i(\sim p)]e^{2\alpha}$, Field's parameter of observational input, α, is just 1/2 Good's weight of evidence for p as against not-p. Shafer, however, is sceptical about the existence of such propositions.

If one adopts the point of view suggested in this paper, these propositions are not to be found among the first order propositions, but rather as statements of observational probability. That is, an observational statement, o_i, will consist of a conjunction of observational probability statements, $\bigwedge_j pr_o(p_j) = a_j$. We have already seen that the generalised Miller principle typically fails for observational probabilities. The question arises whether the partition independence of the generalised Miller principle survives, *i.e.* whether we have $PR(p_j \text{ given } \bigwedge_i pr_o\ (p_i) = a_i = PR(p_j \text{ given } pr_o(p_j) = a_j)$. I think that in general we do not. Suppose that the p_i s are orange, red, pink, blue and black. Then the probability of red given an observation would plausibly depend not only on the observational probability of red but also on the distribution of observational probabilities among orange, pink, blue and black. What is more to the point is the likelihood $PR[\wedge pr_o(p_i) = a_i \text{ given } p_j]$ which evidently is typically not equal to $PR[pr_o(p_j) = a_j \text{given } p_j]$. But the foregoing example also shows that typically we cannot even factor $PR[\wedge pr_o(p_i) = a_i \text{ given} p_j]$ as $PR[pr_o(p_j) = a_j \text{ given } p_j]\ [PR(pr_o(p_j) = a_j)/PR(\bigwedge_i(pr_o(p_i) = a_i)]$. Thus, the likelihood function will in general have to take into account interactions between the members of the partition, and these interactions will depend on the partition and observational situation at issue. These complications obligingly disappear when the partition in question has only two members, $[p, \sim p]$ since the observational probability of p determines the observational probability of its denial.

Another complication affecting the representation of observation is that some observations may count for more than others, even though the distribution of probabilities is the same. An observation in the light of three candles may count for more than an observation in the light of one; an observation made while sober may count for more than one made when drunk or while under the influence of hallucinogens, *etc.* This difference must show up in the likelihood, $PR(o \text{ given } p)$, and so a sufficient representation of observations should, in such cases, be capable of representing it. The minimal way to effect such a representation would be to add a weight parameter, I, to the statement of observational probabilities, *i.e.* $o = [\bigwedge_j pr_o(p_j) = a_j \ \&\ \text{weight } pr_o(p_j) = I_j]$ (I_j will often, but not invariably, be independent of j). This concept of weight is just Savage's intuitive concept of "sureness" which applies as much to observational probabilities as any other species of degree of belief. A deeper analysis would then represent an observation as a second order probability distribution over pr_o, with our old observational probabilities emerging as expectations, and (in cases where the second order distribution is nice enough) our weight being a function of its variance. Such a representation would then move our involvement with higher order personal probabilities to level three. I do not propose to develop this in detail here, but only to make the point that higher order personal probabilities provide a rich framework for analysis of what is going on inside Field's parameter α.

I offer one brief illustration of how the likelihoods might go in some cases, with no special claims as to the value of the illustration, other than as nostalgia. I will forgo higher order observational probabilities and keep weights, and confine myself to the partition $[p, \sim p]$. I will have both observational probabilities, pr_{o_i}, and final probability, pr_f, as random variables, and will deal with probability density functions *PDF* rather than probabilities on these random variables. By Bayes' theorem:

$$PDF\ [pr_f(p) = w \text{ given } o] \propto$$
$$PDF\ [pr_f(p) = w]\ PDF\ [o \text{ given} pr_f(p) = w]$$

(where the proportionality symbol indicates multiplication by a normalising term not dependent on w). Suppose that the likelihood function has a Bernoullian distribution:

$$PDF[o_1 \ldots o_n \text{ given } pr_f(p) = w] = w^{\gamma}\ (1 - w)^{\delta}$$

where $\gamma = \Sigma_1^n\, pr_{o_i}(p)\ I_i$ and $\delta = \Sigma_1^n\, pr_{o_i}(\sim p)\ I_i$ and that the prior distribution for $pr_f(p)$ is a beta distribution:

$$PDF[pr_f(p) = w] \propto w^{\alpha-1}(1-w)^{\beta-1}$$

(a case of some special interest being that in which $\alpha = \beta = 1$ which gives the "flat" or "uniform" prior, $PDF[pr_f(p) = w] = 1$). Then by Bayes' theorem, the posterior distribution, $PDF[pr_f(p) = w \text{ given } o_1 \ldots . o_n]$, is also a beta distribution with parameters $\alpha' = \alpha + \sum_{i=1}^{n} pr_o(p) I_i$ and $\beta' = \beta + \sum_{i=1}^{n} [1 - pr_{o_i}(p)_i] \mathrm{I}_i$. That is, in this situation, where the observations are Bernoullian in the sense indicated, repeated observations never move the probability distribution function out of the family of beta distributions, but only change the values of the parameters, alpha and beta. The family of beta distributions is called the family of *conjugate prior* distributions for samples from a Bernoulli distribution (Raiffa and Schlaifer 1961: ch. 3). We take the posterior probability of p to be the posterior expectation of pr_f (this being another form of Miller's principle).

$$\begin{aligned} &PR[p \text{ given } o_1 \ldots o_n] \\ &\quad = \int w \; PDF[pr_f(p) = w \text{ given } o_1 \ldots o_n]. \end{aligned}$$

The expectation of a beta distribution is $\alpha/(\alpha+\beta)$. Therefore:

$$\begin{aligned} &PR[p \text{ given } o_1 \ldots o_n] \\ &\quad = \alpha + \sum_{i=1}^{n} pr_{o_i}(p)I_i/[\alpha + \sum_{i=1}^{n} pr_{o_i}(p)I_i] + [\beta + \sum_{i=1}^{n} pr_o(p)I_i] \end{aligned}$$

where α and β are the parameters of the prior PDF of $pr_f(p)$. In the case in which $\alpha = \beta$, we have:

$$PR[p \text{ given } o_1 \ldots o_n] = [\Sigma pr_{o_i}(p)I_i + \beta]/[\Sigma I_i + 2\beta].$$

If we consider the case where observational probabilities are always 0 or 1 and where I_i always equals 1, we have (letting N be the number of observations, $n = \Sigma pr_{o_i}(p)$, and $\lambda = \beta/2$):

$$\text{Carnap: } PR[p \text{ given } o_1 \ldots o_n] = [2n + \lambda]/[2N + 2\lambda]$$

(see Carnap (1952) and Jaynes (1958)), and taking the case of a flat prior:

$$\text{Laplace: } PR[p \text{ given } o_1 \ldots o_n] = (n+1)/(n+2).$$

We can recover the formula for $\alpha = \beta$ from Carnap if we take an observation, o_i, as equivalent to a virtual sample of size I_i and frequency pr_{o_i}. But in some situations, the likelihood may *not* be as if the observational probabilities were relative frequencies within Bernoullian samples, so I reemphasise that the foregoing is offered as an illustration rather than as methodology.

At the beginning of this section, I claimed that the points I made about the relation between probability kinematics and higher order conditionalisation would generalise to Jaynes' information theoretic approach to statistical inference (Jaynes (1957, 1979), Kullback (1959)). The extent to which the points made about probability kinematics generalise to inference maximising relative entropy is in fact quite remarkable. I will close this section with a brief sketch of their relation. I rely for this account on Halmos and Savage (1949), Kullback and Leibler (1951) and Kullback (1959). The question is there treated in a very general setting. We deal with abstract probability spaces, $\langle X,S,\mu_i\rangle$, where X is a set, S a sigma algebra of subsets

of S, and μ_i a measure on S. The relative information of μ_1 with respect to μ_2 is defined as:

$$I(1{:}2) = \int f_1(x) \ln[f_1(x)/f_2(x)] \mathrm{d}\lambda(x)$$

where f_1 and f_2 are generalised probability density functions whose existence is guaranteed under mild conditions of continuity. (We need only assume that μ_1 is absolutely continuous with μ_2. Then we can take $f_1(x)$ as the Radon–Nikodym derivative of μ_1 with respect to μ_2. $f_2(x)$ then equals 1, since it is $\mathrm{d}\mu_2/\mathrm{d}\mu_2$, so in this case $I(1{:}2)$ $=\int f_1(x)\ln[f_1(x)]\mathrm{d}\mu_2(x)$; see Kullback (1959: 28–9). We can then take the information theoretic version of statistical inference as follows: starting from an initial probability distribution, μ_2, and a constraint of the form $\int T(x)f_1(x)\mathrm{d}\lambda(x) = \theta$, and a set of eligible candidates for final probability distribution with respect to which μ_1 is absolutely continuous, infer that final probability distribution μ_2 such that among the candidates for final distribution which meet the constraint it minimises the relative information, $I(1{:}2)$. Here $T(x)$ is a measureable statistic, either real or vector valued and θ is a constant. For example: (1) suppose $T(x)$ is the characteristic function of measurable set, s; $\mu_2(s) > 0$; $\theta = 1$; and we take every μ_1 with which μ_2 is absolutely continuous as eligible. Then the final probability, μ_1^*, which minimises $I(1{:}2)$ comes from μ_2 by conditionalisation on s. (2) Just as (1), except that $\theta = a$ for some $0 \leq a \leq 1$. Then μ_1^* comes from μ_2 by probability kinematics on the partition $[s,\bar{s}]$ (see Kullback 1959: example 2.3, pp. 42–3). (3) As before except that $T(x) = \langle c_1(x) \ldots c_n(x)\rangle$, where each c_i is the characteristic function of an element, p_i, of a finite partition of X; $\mu_2(p_i) > 0$; $\theta = \langle a_1 \ldots a_n\rangle$ with $0 \leq a_i \leq 1$. Then μ_1^* comes fron μ_2 by probability kinematics on the partition $[p_i]$. (4) As (3) except that the partition may be countably infinite. Then μ_2^* still comes from μ_1 by probability kinematics, in a sense that I will explain. (5) However, the random variables are not limited to characteristic functions, or vectors whose components are characteristic functions, but may be any measurable functions, or vectors whose components are measurable functions.

To discuss the relation of probability kinematics to the maximum relative entropy (minimum relative information) rule, we need an equally general formulation of probability kinematics. Suppose that we have a countably infinite partition, $[p_i]$, such that $\mu_2(p_i) > 0$ for all p_i. Then we can say that μ_1 comes from μ_2 by generalised probability kinematics on the partition, $[p_i]$, if and only if for each element of the partition, p_i, the ratio of the posterior to the prior density at each point in p_i, $f_1(x)/f_2(x)$, is equal to the ratio of the posterior to the prior probabilities of that element, $\mu_1(p_i)/\mu_2(p_i)$. That is, the criterion of constancy of probability, conditional on members of the partition in the finite case, is simply generalised to constancy of conditional density. Given this generalised definition my remark under (4) above holds. Strictly speaking, I should add the qualification: *modulo* a set of points of measure zero in μ_2. In fact, the generalised probability distribution functions we have been using are only determined to within a set of measure zero in μ_2. This qualification should be understood to hold throughout. If $\langle X,S,\mu_1\rangle$ comes from $\langle X,S,\mu_2\rangle$ by probability kinematics on the countable partition, $[p_i]$, in the way indicated, we say that the partition is *sufficient* for $\langle\mu_1,\mu_2\rangle$ because in this case measuring the relative information with respect to the partition $\Sigma_i\mu_1\ln[\mu_1(p_i)/\mu_2(p_i)]$ gives the relative information, $I(1{:}2)$. If a partition is not sufficient, *i.e.* if μ_2 does not come from μ_1 by kinematics on that partition, then the information with respect to the partition is less than $I(1{:}2)$; measuring information relative to a insufficient partition causes loss of information (Kullback 1959: corollary 3.2, p. 16). It follows that if a constraint consists in specifying the final probabilities of each member of a countable partition each of whose members have positive prior probability, then the final probability measure comes from the initial one by minimising relative information *if and only if* it comes from the initial one by probability kinematics on that partition.

For a fully general formulation of probability kinematics, we need to remove the restrictions on the partition. Any statistic (measurable function) on a probability space *induces a partition* on that space, with the elements of the partition being the inverse images of the values of the statistic (*e.g.* if I have a probability space of baskets and the statistic 'number of eggs in', the statistic induces a partition of baskets such that two baskets are members of the same element of the partition if and only if they contain the same number of eggs. Notice that a vector valued statistic induces the partition that is the common refinement of the partitions induced by its components.) Conversely, any partition whose elements are measurable sets is induced by some statistic. (*n.b.* I have not made any limitation as to the types of values that statistics can take.) So we can have full generality if we formulate probability kinematics

relative to a statistic. Let T be a statistic, with domain X and range Y, and let R be the class of measurable subsets of Y (*i.e.* $G\in R \longleftrightarrow T^{-1}(G)\in S$). Then starting with a probability space, $\langle X,S,\mu\rangle$. and a statistic, T, we can consider an associated probability space $\langle Y,R,\upsilon\rangle$ where the measure υ is derived from μ, by $\upsilon(G) = \mu(T^{-1}(G))$. Then $\upsilon(G) = \int_G g(Y)d\gamma(Y)$ where $\gamma(G) = \lambda(T^{-1}(G))$. The associated probability space can be thought of as representing the effect of consolidating the elements of the partition induced by the statistic into single elements. Let g_1 and g_2 be the generalised probability density functions for the spaces $\langle Y,R,\upsilon_1\rangle$; $\langle Y,R,\upsilon_2\rangle$ which corresponds to $\langle X,S,\mu_1\rangle$; $\langle X,S,\mu_2\rangle$ under the statistic T. (We may take them as the Radon–Nikodym derivatives of μ_1 and μ_2 with respect to μ_2.) Then we will say that $\langle X,S,\mu_1\rangle$ comes from $\langle X,S,\mu_2\rangle$ *by generalised probability kinematics on the statistic* T (or, if you please, on the partition induced by T) if and only if $f_1(x)/f_2(x) = g_1T(x)/g_2T(x)$. This is equivalent to saying that the conditional density conditional on $T(x) = y$, *i.e.* $f(x)/gT(x)$, remains the same before and after the change, and is thus clearly the correct statement of probability kinematics for this general case. If $\langle X,S,\mu_1\rangle$ comes from $\langle X,S,\mu_2\rangle$ by generalised probability kinematics on the statistic T, we shall say that *T is a sufficient statistic*, relative to them. Just as in the countable case, we have the result that the relative information measured on the partition, $I(1{:}2,Y)$, is less than or equal to the relative information, $I(1{:}2,X)$, with equality if and only if the statistic is sufficient (Kullback & Leibler 1951: theorem 4.1; Kullback 1959: theorem 4.1). A sufficient statistic is one which loses no information. It follows from this that, if the satisfaction of a certain constraint is a function of the posterior values of $g(y)$ (*i.e.* the posterior values of the conditional expectation of $f(x)$ given that $T(x) = y$), then the posterior distribution that comes from the prior by minimising relative information subject to that constraint comes from the prior by generalised probability kinematics on T. For consider the posterior distribution of values of $g(y)$ in the minimum relative information posterior. By the hypothesis that the satisfaction of the constraint is a function of the posterior values of $g(y)$, any posterior distribution with these values satisfies the constraint. Thus the posterior which comes from the prior by probability kinematics on T with this final distribution for $g(y)$ satisfies the constraint. And, by the previous theorem, it must be the minimum relative information posterior.

But the constraints considered in the maximum entropy (minimum relative information) formalism are all of this character! Remember that the constraints all consisted in specifying the posterior expectation of a statistic:

$$\int T(x) f_1(x) \mathrm{d}\lambda(x) = \theta.$$

This can now be rewritten as:

$$\int y g_1(y) \mathrm{d}\gamma(y) = \theta.$$

So T is a sufficient statistic relative to a prior and a posterior which minimises relative information subject to the constraint that the posterior expectation of T has a certain value. (See Kullback 1959: 43–4; van Fraassen has also discussed a special case of this in a recent unpublished paper.) We have just established a theorem, which can be roughly put as:

JAYNES implies JEFFREY.

That is, if we start with a prior and move to that posterior (among those with respect to which the prior is absolutely continuous and which satisfy the constraint that a statistic T has a certain posterior expected value) which minimises relative information, then the posterior comes from the prior by generalised probability kinematics on the statistic, T. Conversely, we can say that if a posterior comes from a prior by generalised probability kinematics on T, then it minimises relative information subject to the constraint that $g(y)$ has those posterior values. In the finite or countable case, we could always put that constraint in Jaynes' form by considering a statistic T' which is a vector of characteristic functions of elements of the partition induced by T. But in the general cases it is not clear that we can always put the constraint in Jaynes' form (unless we use the same trick and countenance vectors of uncountable dimension). So it is only with some qualification that we can assert that JEFFREY implies JAYNES.

The point I made about the relation of higher order probabilities to probability kinematics carries over to this general setting. Let me explain why I called the sufficiency condition:

$$PR(q \text{ given } p_j \text{ and } \textstyle\bigwedge_i pr(p_i) = a_i) = PR(q \text{ given } p_j)$$
$$\text{for all first order } q \text{ and all } j$$

by that name. Given that the change from the initial distribution to

the final distribution is to be by conditionalisation on $\bigwedge_i pr(p_i) = a_i$, the condition can be rewritten as:

$$PR_{\mathit{final}}(q \text{ given } p_j) = PR_{\mathit{initial}}(q \text{ given } p_j)$$
$$\text{for all first order } q \text{ and all } j$$

which means that if we look at the first order probability space, we have the condition for $[p_j]$ to be a sufficient partitioning for $(PR_{\mathit{initial}}, PR_{\mathit{final}})$.

It should be clear from the foregoing discussion that we can carry all this over to abstract probability spaces and generalised probability density functions. So we will be able to say in general that second order conditionalisation results in first order probability kinematics on a partition if and only if such a sufficiency condition is satisfied.

It would be of some interest to investigate the full structure of minimum relative information inference from the same point of view. That is, we look at a second order distribution where we can conditionalise on the constraint that the final expected value of a statistic has a certain value, and identify the characteristics that the second order distribution must have for such conditionalising to coincide at the first order level with minimum relative information inference. We have some of the answers already, but not all. It is a necessary condition for minimum relative information that we have generalised probability kinematics on T, the statistic of the constraint. We have this if and only if the second order probability distribution satisfies the proper sufficiency condition. To ensure that conditionalising at the second order level leads us to a distribution in which the constraint is in fact met requires a further generalisation of the generalised Miller condition. But there will be further requirements necessary to guarantee that the final density over the elements of the partition induced by T is of the proper exponential character required to satisfy the minimum relative information principle. (See Kullback (1959: ch. 3) for derivation and discussion of the exponential solutions.)

Conclusion

If I end with more questions open than closed, I hope this will only reinforce the point that second-order personal probabilities are not only legitimate, but also theoretically interesting. They provide a perspective from which the scope of applicability of first order generalisations of conditionalisation can be assessed. Counter instances to the sufficiency condition are counter instances to both probability kinematics and the minimum relative information principle. The scarcity and peculiar nature of such counter instances explains the wide range within which probability kinematics can be plausibly applied. An analogous determination of the range of applicability of minimum relative information inference calls for further study.

University of Illinois at Chicago Circle

REFERENCES

Carnap, R. 1952. *The Continuum of Inductive Methods*. Chicago.

de Finetti, B. 1972. *Probability, Induction and Statistics*. London.

Domotor, Z., Zanotti, M. and Groves, G. (Forthcoming). Probability kinematics, *Synthese*.

Field, H. 1978. A note of Jeffrey conditionalization, *Philosophy of Science* **45**, 171–85.

Freedman, D. & Purves, R. 1969.Bayes method for bookies, *Annals of Mathematical Statistics* **40**, 1177–86.

Gärdenfors, P. 1975. Qualitative probability as intensional logic, *Journal of Philosophical Logic* **4**, 171–85.

Good, I. J. 1950. *Probability and the Weighing of Evidence*. London.

Good, I. J. 1965. *The Estimation of Probabilities*. Cambridge, Massachusetts.

Good, I. J. 1971. The probabilistic explication of information, evidence, surprise, causality, explanation and utility. In *Foundations of Statistical Inference*, ed. Godambe & Sprott. Toronto.

Hacking, I. 1967. Slightly more realistic personal probability, *Philosophy of Science* **34**, 311–25.

Halmos, P. R. and Savage, L. J. 1949. Application of the Radon–Nikodym theorem to the theory of sufficient statistics, *Annals of Mathematical Statistics* **20**, 225–41.

Hintikka, J. 1971. Unknown probabilities, Bayesianism and de Finetti's representation theorem. In *Boston Studies in the Philosophy of Science*, vol. 8. Dordrecht.

Jaynes, E. T. 1957. Information theory and statistical mechanics, *Physical Review* **106**, 620–30.

Jaynes, E. T. 1958. *Probability Theory in Science and Engineering*. Dallas.

Jaynes, E. T. 1979. Where do we stand on maximum entropy? In *The Maximum Entropy Formalism*, ed. R. D. Levine & M. Tribus, pp. 115–18. Cambridge, Massachusetts.

Jeffrey, R. 1965. *The Logic of Decision*. New York.

Jeffrey, R. 1974. Preference among preferences, *Journal of Philosophy* **63**, 377–91.

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

Kullback, S. & Leibler, R. A. 1951. On information and sufficiency, *Annals of Mathematical Statistics* **22**, 79–86.

May, S. & Harper, W. 1976. Toward an optimization procedure for applying minimum change principles in probability kinematics. In *Foundation of Probability Theory, Statistical Inference and Statistical Theories of Science*, vol. I, ed. W. L. Harper & C. A. Hooker. Dordrecht.

Miller, D. 1966. A paradox of information, *British Journal for the Philosophy of Science* **17**, 59–61.

Raiffa, H. & Schlaifer, R. 1961. *Applied Statistical Decision Theory*. Boston.

Ramsey, F. P. 1926. Truth and probability. In his *Foundations*, ed. D. H. Mellor, pp. 58–100. 1978. London.

Savage, L. J. 1972. *The Foundations of Statistics*, 2nd ed. New York.

Shafer, G. 1979. Jeffrey's rule of conditioning, *Technical Report 131*, Department of Statistics, Stanford University.

Skyrms, B. 1975. *Choice and Chance*, 2nd ed. Belmont, California.

Skyrms, B. 1980. *Causal Necessity*. New Haven, Connecticut.

Teller, P. 1973. Conditionalization and observation, *Synthese* **26**, 218–58.