BAS C. VAN FRAASSEN

BELIEF AND THE PROBLEM OF ULYSSES AND THE SIRENS[1]

> ...my dear friends, we shall behave like people who have fallen in love but realize that their passion is not beneficial. They force themselves to stay away from their loved one....
>
> (Plato, *Republic* X, 607e)

This is surely a bit of Socrates' famous irony. He draws the analogy to explain how his friends should regard poetry as they regretfully banish it from the ideal state. But lovers were no more sensible then than they are now. The advice to banish poetry, undermined already by Plato's own delight and skill in drama, is perhaps undermined still further by this evocation of a 'sensible' lover who counts love so well lost.

Yet Socrates' image is one of avowed rationality and prudence. The sensible lover imitates the older literary example of Ulysses' tying himself to the mast. (The example belongs therefore to the class of problems treated in Elster (1979)). Both this lover and Ulysses foresee that under certain possible future conditions, their opinions, values and preferences will or would differ from what they are now, in a very definite fashion.

To what extent is such foresight possible? Correspondingly (when we do not claim foreknowledge) to what extent is such opinion reasonable, rational, coherent, or consistent in some suitably broad sense? It is not easy to understand exactly what is possible or even logically permissible in this respect. In an earlier paper, "Belief and the Will", I argued for a principle ("Reflection") to govern such deliberation. Here I will both generalize the treatment of opinion in "Belief and the Will" and respond to criticism. Critical examples mainly resembled the story of Ulysses who foresaw a period of dysfunction (at the sound of the sirens) in his epistemic and/or doxastic future. Other criticism focused on the model

Philosophical Studies **77:** 7–37, 1995.

of opinion used (precise numerical subjective probability) and on the merits of Dutch Book arguments. The present argument will not rely on Dutch book arguments and strategies, and the Reflection principle will be formulated so as to apply also to vague opinion.

1. PERSONAL PROBABILITY AS PROPOSITIONAL ATTITUDE AND AS MODE OF SELF-ATTRIBUTION

The general discussion of opinion about our own opinion is severely constrained by certain discoveries in the philosophy of language. In the terminology of Russell's lectures on logical atomism, "I believe that A" attributes to the speaker a *propositional attitude*, which is a relation to a proposition [that A]. But sentences of form "I believe that I am F" resist construal of this sort: I may say this and refuse to say "I believe that BvF is F", although "I am F" expresses in my mouth the same proposition as "BvF is F". Without entering upon the arguments and complexities, I submit that the correct conclusion is David Lewis' (1979): propositional attitudes are special cases of self-attribution of properties. The act of self-attribution is basic. The special case of "I believe that A" is to be constructed as "I believe myself to be such that A".

This point applies mutatis mutandis to other epistemic attitudes. Those expressed in the form "It seems likely to me that A" for example can be construed as "I seem likely to myself to be such that A", instantiating the form "I seem likely to myself to be F", an attenuated mode of self-attribution. I shall take it for granted here that when we discuss personal probability we shall have in mind first of all a simple fragment of language in which all sentences have the form $P(\ldots) = x$, which may be read as "It seems to degree x to me that I am such that ...". The ellipsis may include further occurrences of the personal pronoun, and specifically I shall write $P(p_t(A) = x) = y$ which may be read as "It seems likely to me to degree y that at time t it will (did) seem likely to me to degree x that I am such that A". The phrase "that I am such that A" may indeed generally be replaced by "that A" in the reading without loss. The important point however is that the sentence as a whole is in the first person singular present tense indicative, while of course no such

restriction applies to the contained sentence A. This will be crucial to my interpretation of this discourse: the function of the sentence $P(\ldots) = x$ as a whole is to *express* the epistemic attitude of self-attribution, while the contained $p_t(\ldots) = x$ *states* the fact that (at t) I have this attitude. In English, of course, both roles are played by the same words.

Probabilism in epistemology comes in many flavors these days, and to facilitate discussion I must indicate several more differences from other writers on the subject. I do not take for granted that we update our opinion by conditionalizing on certain propositions taken as evidence. This raises the question what conditions of coherence constrain persons who do not have Bayesian conditionalization as their epistemic policy. Much of the discussion in this paper loses its point if that question is moot. More fundamentally, I also reject a certain kind of probabilist foundationalism which holds that a person's total state of opinion is (or should be) functionally determined by the set of propositions which receive probability 1 (or by some set of propositions which are subjectively privileged in some other way). This view (which may be associated with C. I. Lewis' slogan that nothing can be probable unless something is certain) would also eliminate some of our difficulties. It seems to me to be mistaken, but I will not argue that here.

2. THE LIMITS OF DUTCH BOOK ARGUMENTS AND THEIR HEURISTIC USE

To give the reader a brief and user-friendly access to the previous literature, I want to explain here the basic principle of diachronic Dutch bookmaking. Immediately afterward I will explain why I do not want to rely on Dutch Book arguments any more, nor discuss rationality in the terms they set, though they have a very specific heuristic value.

Diachronic Dutch bookmaking, stripped to its bare bones so to speak, uses what I shall call "double trades". They lend themselves easily to nefarious purposes. Here is the simplest sort of example. A little boy, Pierino, has two sorts of goods, blocks and marbles. He values these in ratio 3:1. That is, he regards with indifference the prospect of any exchange of blocks for *three* times that many marbles, or vice versa. To say that he regards it with indifference means that he values the

two asset situations – before and after such a contemplated exchange – equally. (The best analogue for us is currency trading – dollars and yen for example.) But in addition he knows, and so do we, that a year from now he will value blocks and marbles in the opposite ratio, indifferent to any exchange of marbles with *three* times as many blocks, or vice versa.

How should we regard this boy? How should Pierino regard himself? As fully rational, as functioning well when it comes to valuing and knowing? Let me define a *double trade* as a sequence of two transactions:

sell X for kY at t

buy X for mY at t+1

In our present example one such double trade would be: Pierino sells me *three* marbles for *one* block now, and a year from now he buys *three* marbles (back) for *nine* blocks. This double trade has net effect *zero* on his stock of marbles, but net loss of *eight* blocks. The story implies that Pierino regards the first transaction with indifference, and *knows* that he will regard the second transaction with indifference at its appointed time. That is very puzzling, but it is good for me, since I am making 800% profit, and he doesn't seem to mind his clear and absolute loss.

There are two common reactions to this sort of reasoning. The first is to say that Pierino knew he would get his marbles' worth, because he enjoyed playing with the extra block for the duration of the year; that was worth the loss of eight blocks overall. On that supposition, however, one of the story's premises is contradicted: that at the outset he was indifferent to the 1 block/3 marbles exchange. Instead he preferred the actual exchange to its opposite, and would not have sold 1 block for 3 marbles then. The second reaction is that Pierino knows that he will stop the double trade midway: realizing that there would be an absolute loss, he is sure he will refuse to buy the three marbles back. But that contradicts the other premise of the story, namely that he is sure that he will be indifferent to such exchanges at the later time.

Perhaps it is not really possible for a person to have the sort of knowledge (foresight) here attributed to Pierino. If not, Pierino may still be able to have opinions about what his values could be a year

from now. Imagine that he foresees a number of possible changes in his exchange ratio of blocks to marbles, and perhaps even ranks some of these as more or less likely than others. Today one block is worth three marbles to him, and he foresees that a year from now it will be worth x marbles to him ($x = a_1, a_2, \ldots$ and can be a fraction). Imagine finally that x may lie inside the interval [a,b], but not outside, as far as his current opinion goes. Then there will be double trades which leave him with inevitable net losses unless $a \leq 3 \leq b$. For example, if 3 is actually less than a, Pierino could sell us one block now for three marbles, and then with equanimity buy one block (back) a year from now for $x \geq$ a marbles, thus having a net loss of at least $a - 3$. So we arrive at the following principle for Pierino:

> [If I am not to be indifferent to entering upon double trades which in my opinion are certain to incur sure loss, then] my current evaluation of good G as equal to k$ must lie in the range spanned by the possible valuations of G in terms of $ which I may have at later time t, as far as my present opinion is concerned.

This principle is to be understood in the manner indicated above: valuing G as equal to k$ means regarding with indifference a change of current assets consisting in the replacement of any multiple of G with an equal multiple of k$ (fair exchange ratio).

I have not talked about bets at all so far. But bets are a straightforward instance of goods. Suppose we talk about bets with payoffs and costs all listed in 1973 dollars, say, to be able to disregard inflation. You offer either to buy from me or sell me a bet on A, which may or may not occur the day after tomorrow, with payoffs $1 if A and $0 if not-A. My present probability for A equals x, so today my fair price for this bet equals $x. Suppose now that I foresee various possibilities for what my probability for A will be tomorrow, when you might think of completing a double trade. Let a be the highest number such that I am sure that it will below my tomorrow's probability for A, and b the lowest which I am sure will be above that probability. Then I am sure that tomorrow my fair buying price for A will be in the interval [a,b]. I am vulnerable to a double trade with sure loss if and only x is outside that interval. So if Pierino is going to have bets among his goods, and they are to be treated as the other goods discussed above, then we have for him the corollary principle:

[If I am not to be indifferent to entering upon double trades which in my opinion are certain to incur sure loss, then:] my current probability for event A must lie in the range spanned by the possible probabilities for A I may come to have at later time t, as far as my present opinion is concerned.

This too must be understood in the sense explained in the preceding text: the interval in question is bounded by the supremum of numbers which I am now sure will be below my probability then, and the infimum of those which I am now sure will be above it. This is phrased in such a way that we allow for someone who has *vague* (imprecise) probabilities; whereof more later.

Finally, just to round off the discussion, what goes for values and for probabilities here, goes also for expectation values for random variables in general. Here is an unusual bet: it will pay, the day after tomorrow, in hundreds of 1973 dollars, the numerical equivalent of the number of inches of rainfall in Indianapolis on that day. If I offer to trade this bet with you, your fair buying and selling price for it will be your current expectation value for that quantity. It had better be in the interval spanned by what you are now sure your possible expectation values for that quantity may be tomorrow, if you are not to be indifferent to embarking on double trades which you are certain will incur sure loss.

I think that I have now covered the main uses of diachronic Dutch Book arguments. As we will see in the discussion below, the principles found here tell Pierino to obey what in "Belief and the Will" I call the Reflection Principle. But I do not any longer think that these sorts of arguments provide a good basis for discussion. They rest on a quite possibly over-simple decision theory model which may be of very limited direct applicability. While there are well known ways to extend the model's applicability, it seems to me that a discussion of opinion and policies for opinion revision should not be saddled with the complexities of another domain, if avoidable. I would insist only on the heuristic value of arguments about betting, which recreate in the gambler's sharply defined microcosm all the great issues of human existence.

3. COHERENCE FOR OLD-FASHIONED FORECASTING POLICIES

In this section I am going to look at Pierino's grandfather Pierone, so-called because of his size. His opinions are of the old fashioned kind: he simply believes certain propositions. But he does have a policy for changing his opinion. Pierone distinguishes between the topics he has opinions about, and the possible inputs – deliverances of experience – which he treats as relevant for them. An example might be input from observations which he marks oracularly ("Red at night, shepherds' delight!") and his opinions simple yes/no beliefs about the weather to come: "It will be a wet winter", "There will be an average amount of hail but lots of snow, or else no snow but freezing dry weather", and so forth.

If he were to describe his policy he would outline a number of possible alternative states of affairs $A_1, \ldots, A_n, \ldots$, and say that at any moment his opinion will rule out some of those as not being the actual one, leaving a disjunction of the others. Then he has rules, or rather (like in chess for example) specifications of permissible moves: given current opinion B and input J, it is permissible to change his opinion to B′. More than one such possible change may be permissible; then he can freely choose among those. Allowed for instance might be audacious deletions of disjuncts from B and/or prudent additions to it of new disjuncts. The input J may include a register of past experiences and of new theoretical innovations as well as current observations.

Is there a requirement on what his policy should be like, in analogy to the principles we found for Pierino? To find one, we need to be able to talk about what an opinion may change to over a finite interval of time; I'll assume we get one input per unit of time. Let us call the *n-descendants* of belief B those possible beliefs B′ for which there is some input sequence $\langle J_m \rangle$ of length n, such that there is an epistemic history which is in accord with Pierone's policy and which has B as first member, $\langle J_m \rangle$ as successive inputs, and B′ as n^{th} member.

In analogy with what we had for Pierino, let us say that Pierone's policy satisfies the *Dogmatic Reflection Requirement* exactly if for each integer n, each possible belief B implies the disjunction of all its

n-descendants. That is, B implies the disjunction of all opinions it can permissibly be changed to in n moves – where n is any finite number.

Is this a requirement which Pierone's policy should satisfy? Suppose that for a certain opinion B, and a certain length n, B does not imply the disjunction of its n-descendants. That means of course that some disjunct A_k of B is missing from each n-descendant of B. We infer specifically that each permitted epistemic history which starts with B "loses" A_k in at most n steps. The situation is this:

> There are possible circumstances – namely that A_k is the case – in which opinion B is true but inevitably, regardless of which possible inputs are experienced, that true opinion gives way to one that is false n units of time later, for anyone following this policy.

I take it this is a design flaw in the policy.

The policy could be defended by saying that A_k could not be true, or that it is a basic assumption built into the policy (and a belief adopted by anyone who follows it) that A_k cannot be true. To say that, however, is to admit that if the policy had been written without listing A_k among the possible states of affairs in the first place, it would have been from its own point of view equivalent, and then of course it would have satisfied the Dogmatic Reflection Requirement. Indeed, the defence can be read as saying that it only looks as if this policy violates the Requirement: B does imply the disjunction of its n-descendants, for B is unchanged, as far as genuine content is concerned, if A_k is deleted from it. (Similar remarks apply if the defence is that B could have been arrived at only through inputs that rule out A_k.) So although a specific such policy can be defended, the defence amounts to an endorsement of the Requirement.

4. COHERENCE FOR PROBABILIST FORECASTING POLICIES

I do want to get back again to the more sophisticated sort of person, whose opinions admit of degrees. So let us turn to the golden mean in this family, Piero, son of Pierone and father of Pierino. (These are all men, with glaring defects.) Piero is a professional, one who is professionally engaged in fashioning and purveying expert opinion.[2] Like his father Pierone before him, he is a weather forecaster; he makes up forecasts for the 6 p.m. news on the basis of forty-eight hours of data each day.

He does not really give precise probabilities; if he says e.g. "There is an 80% probability of rain", we take it that he means something like "Rain seems *approximately* 4 times as likely as not", or that his probability is 80% give or take 5%. Suppose he also updates that forecast on the basis of new data which came in during the night, for the next morning's 8 a.m. news. He does not pretend to have an exact inductive scheme, and he succeeds only partially in making his procedures explicit. But to the extent he can do this, he writes them up in a training manual for meteorologists.

But imagine now how defects in Piero's procedures could become apparent. For simplicity suppose that Piero announces his probability of 50% for rain tomorrow. Imagine that I consider this 6 p.m. forecast, its current 48 hour data basis, and the procedures for updating during the night. Suppose now that on the basis of his taxonomy for nocturnal data and manual of procedures I find that, on every scenario, the update will lead to a chance of rain update *below* 20% by tomorrow morning's 8 a.m. news. There is no question that this would be defect. There would have been an exactly similar defect if all scenarios overnight would lead to a chance of rain update above 80%.

It is quite another thing, of course, to say exactly what the defect is. That depends on the prior question: "What is the point of having, forming, and cultivating opinion anyway?" In practice, opinions are evaluated in many ways. Again for simplicity, suppose Pierone simply announced rain if Piero's probability for rain was above 2/3, no rain if it was below 1/3, and just said "some chance of rain" otherwise. In his case we would point out that Pierone would be a good forecaster if his definite announcements were usually true, and if it rained roughly somewhere between 1/3 and 2/3 of the time when his announcement was indefinite. But if we look at a large number of days (evening and next morning) showing the sort of defect described above, we would necessarily find that he was not a good forecaster by that criterion. Suppose for example that we evaluate his performance on a sample of 100 days, in which in the evening he announces rain, and in the morning he just says "a chance of rain". Then either his definite announcements in that sample were false more than a third of the time, or his indefinite announcements were followed by rain more than two thirds of the time. This could also

happen to a very good forecaster on a random sample, but here we are envisaging the case in which the initial data plus the instruction manual determine that it be so: on *that* sample, *his* performance will necessarily be bad in that way.

The evaluation I just described is rather crude; it only pays attention to how often the forecaster is "right". It is a familiar point that being right is not the only thing we take into account. (This point can be made in terms of a measure of calibration; see Seidenfeld (1985) and references therein.) Specifically, we also want forecasters to be informative; if they say something indefinite too often, they are not of much use as far as input for my own decision making is concerned. Let us not assume anything about how exactly I make my decisions, except that I treat the latest weather forecast as a relevant factor. Normally, I prefer to postpone my decision about whether to carry an umbrella tomorrow till after the 8 a.m. news. If I cannot do that, I will settle for the 6 p.m. forecast now; but normally that will be the best available estimate of what the 8 a.m. forecast will be. In this story, however, the forecaster's own procedures undermine the status of the 6 p.m. news as putatively the best available estimate – they entail a deficiency in one definite direction.

So I advise the forecaster to revise his procedures. Should he agree with me on that? If not, what does he think he is doing? What is the point?

Notice that I am not giving a Dutch Book argument. Certainly I could dramatize the situation by doing so: I could design a series of bets to offer the forecaster evenings and mornings. These bets would siphon money out of his pockets if he were willing to buy and sell them at the odds effectively posted by his own forecasts. But I won't. There is no need. The forecaster's integrity, his role as a professional, is incompatible with equanimity about what has come to light. Hence the principle which I suggest he should embrace:

> *General Reflection Principle.* My current opinion about event E must lie in the range spanned by the possible opinions I may come to have about E at later time t, as far as my present opinion is concerned.[3]

This principle, the same as was found for Pierino, is written so as to apply also to someone who may have only vague opinions.

The argument for this Principle was admittedly again based on a very specific paradigm: someone like meteorologist Piero, with assumed concern for his own professional role as weather forecaster (similar to the assumed self-concern for Pierino). It is clear therefore (as perhaps it was already for Pierone) that we are not dealing with criteria of rationality in terms simply of momentary states of opinion (although that is what the Reflection Principles apply to, as putative criteria of coherence). The argument purports to show that violation of this Principle is a symptom, within the current epistemic state, of a deeper defect: that the person holding this opinion cannot regard him or herself as following a rational policy for opinion change.

5. GENERAL REFLECTION IS IMPLIED BY CONDITIONALIZATION

Some (though not I myself) take as paradigm of rationality the ideal Bayesian agent, who has opinion in the form of precise numerical probabilities, and changes it solely by Conditionalization on evidence. *Such an agent automatically satisfies the General Reflection Principle*: starting with subjective probability P now he will have at time t one of the functions P (.|E(i,t)) where E(i,t) is a possible evidence scenario between now and t. Because E(i,t) is a partition (disjoint and exhaustive), probability theory entails that P(A) is a convex combination of, hence lies in the interval spanned by, the numbers P(A|E(i,t)). This remark can be taken as partial support for Bayesian Conditionalization, or conversely, for the General Reflection Principle; I'll leave that to the reader.

It is wonderfully remarkable and disturbing that all the criticisms directed at the Reflection Principle were not already previously raised. What was more salient in the literature than the Bayesian principle that the ideal epistemic subject updates his opinion by Conditionalization? As we have just seen, the one implies the other.[4] My own interest in Reflection derives largely from the conviction that rationality does not require a policy of conditionalizing on evidence as sole epistemic move.

Reflection is a weaker condition on epistemic policies, a partial answer to the question: if not Conditionalization, then what?

6. GENERAL REFLECTION IMPLIES SPECIAL REFLECTION; RELATION TO MOORE'S PARADOX

In contrast to an earlier version (1984), I wrote the General Reflection Principle here so as to apply also directly to cases of vague opinion. A person may not have a precise numerical subjective probability for rain. He or she may still consider rain more likely than not, or consider it at least twice, but no more than three times, as likely as not. Then we also say that his or her lower and upper probability for rain are 2/3 and 3/4, or that his or her probability interval is [2/3, 3/4], and call this subjective probability indeterminate or *vague*.[5]

What does the General Reflection Principle entail for the person whose opinion does take the form of sharp numerical subjective probabilities? I propose here to focus on a special sort of epistemic subject: (1) his opinion is "sharp" in this sense, (2) he can express propositions of about what his opinions are and will be, (3) at any given time t his opinion about what his *current* opinions are is entirely correct, (4) the general form in which he can express his opinion is in terms of expectation value, and (5) that he satisfies the General Reflection Principle. The expectation value of quantity (random variable) f relative to probability measure p is defined by $E(p,f) = \Sigma\ p(i)f(i)$ where quantity f takes possible value f(i) with probability p(i). We must resist here the idea that value is utility, or indeed that it is meant in any sense except "value of a function or parameter" – expectation as such has nothing to do with preferences, but is simply one (very general) form of opinion.

Applying the General Reflection Principle to this form, we find the following on the assumption that I am a subject of the above special type. Let P denote my current probability and p_t my probability at later time t. Define: quantity q takes value 0 if $p_t(A) \neq x$, while if $p_t(A) = x$, then q has value $(1 - x)$ if A is true and $(-x)$ if A is false. Then clearly my expectation value for q must be *zero* at t; I assume here that at t, I will be able to express, and so know, what my current opinion about A

is.[6] So by Reflection, my present expectation value for q is also *zero*. This entails, via a very brief calculation,[7] that:

> *Special Reflection Principle.* $P(A|p_t(A) = x) = x$ when defined.

This conditional form, for precise numerical probability, is the Reflection principle of my 1984 article, here shown as a corollary to the general principle. We can also deduce at once, on the supposition that I am sure that I will have some precise probability for A at t, that:

$$P(A) = \Sigma\, xP(p_t(A) = x)$$

i.e. P(A) equals my current expectation value of $p_t(A)$.[8]

Finally, let us note the formal connection, at least, between Moore's Paradox and the Reflection Principle. If we try to generalize Moore's sentence schema to a probabilistic form, we arrive at:

> It seems certain [likely, very likely] to me that: A and it seems unlikely to me that A;

or less qualitatively:

> It seems likely to me to degree y that (A and it seems likely to me to degree x that A):
>
> $P(A \,\&\, p(A) = x) = y$

where the number x is lower than the number y and "p" describes present (current) opinion. The synchronic form of the Special Reflection Principle (t = now) is violated unless y is less than or equal to x (since $P(A \,\&\, p_t(A) = x) = xP(p_t(A) = x)$ which is less than or equal to x). Another way to put this, perhaps more perspicuous if less informative, is that the Reflection Principle does not allow you to conditionalize on $(A \,\&\, p(A) = x)$ if $x < 1$, although it does allow you to give it a positive probability. Of course, the Reflection Principle says all this as well with "p" replaced by "p_t", for future times t, while Moore's Paradox pertains only to current opinion.

7. INITIAL CHALLENGES: CONFIDENCE, MEMORY, AND MODESTY

Ulysses appears to violate the Reflection principles. He currently values the song of the sirens but not above personal safety; he believes that this value judgment will be reversed on closer acquaintance. More to the point: he presently believes that it will be disastrous to steer the ship toward the rocks, but believes that, while hearing the siren song, he will have the opinion that steering the ship toward the rocks will lead to peace and happiness. The *Odyssey* does not tell us to read the episode in either of these ways, for it is told as if Ulysses has no inner life at all. But we readers take it so.

If Ulysses is rational and prudent, and violates Reflection, then Reflection is not a good principle and should not constrain our epistemic and valuational policies. We must distinguish very strictly here between a *defence* of Reflection (either Ulysses is not rational or he does not violate Reflection) and a *revision*. I intend here to defend it. But before facing the major challenges of the Ulysses type, I shall discuss more initially similar problem cases which I think the Reflection Principle can more easily take in stride. They concern (over-)confidence, memory, and modesty. By itself, the General Reflection Principle does not entail great confidence in my future opinion. I may well believe that I shall be affected by overconfidence or underconfidence, and that there may be chance factors in the selection of evidence which will typically lead me to a definite opinion, but in an unreliable way. That is fine: such thoughts *widen* the range of possible future opinions I foresee. The Principle forbids only opinion which is at odds with *any and all* opinions I think I may come to have at a certain future time.[9]

Expectation values will in general be vague when probabilities are vague. To see the difference vagueness makes to Reflection, consider my present belief that croissants are healthy. I have just read a highly convincing article about it, and my present subjective probability for this is high, let us say 0.98 (or perhaps it is vague, 0.98 plus or minus 0.2). On the other hand, I know that this is the sort of thing I seem to change my mind about very often. The trouble with a very high probability now is that if I admit many other future probabilities, most of them will be lower – and so, unless the lower ones look *very* unlikely, those future

probabilities will average out comparatively low as well. But then, if I am sure that I will have some precise probability at later time t for Q = [*croissants are healthy*], my current expectation value for its future probability will be low as well – in conflict with Reflection, given my current very high probability.

To be realistic, however, I also envisage quite a lot of ways in which I may be vague on the subject at that time. Indeed, if I realize later that my opinion in such matters is subject to much fluctuation, it seems more likely that I will be vague on it then. In such cases there may be only a lower bound: my total opinion on the matter then may be summed up by the judgment that croissants being healthy is no less than K times as likely as not (which sets no upper limit; if K = (1/9) the interval is [0.1,1]). If these sorts of cases play an important role among the envisaged possible future opinions, they may suffice to satisfy Reflection.[10] If the plausible examples of this sort satisfy Reflection, it need not worry us if some variant would violate the principle; to be good, a counterexample cannot be too contrived or unrealistic.

The above example bears some resemblance to a possibly more difficult one: William Talbot's (1991) example, essentially, that I am now sure I had a croissant for breakfast today, but that a year from now, having forgotten a good deal, it will seem unlikely to me that I had a croissant on this particular day. Memory loss is not a feature generally accommodated in probabilist models of opinion, but does seem to characterize real subjects! Indeed we might here consider the project of rational and efficient memory management, with policies for discarding memories likely to be of too little use to reward retention.[11] I will ignore subjects who have no interest in rational opinion and memory management at this point, though I will return to them in the last section.

The project of memory management is actually a problem for any probabilist account, for by the theorem of total probability, my probability that I have a croissant today or any other day (such as yesterday or tomorrow) has to be a certain weighted average of my probabilities of having done so conditional on the various opinions I may have about that one year hence. Let J be the set of future opinions that I give positive probability of coming about, and let A_i state that I had a croissant on day i = yesterday, today, tomorrow, ... Then $P(A_i)$ must, on pain of

incoherence, equal the sum of the factors $P(p_t(A_i) = x)P(A_i|p_t(A_i) = x)$ with x assigned by some member of the set J.

There seem to me to be three ways of not violating the probability calculus here. One, of course, is the Reflection Principle. The second is to regard one's future opinions in these matters to be totally irrelevant to the truthvalues of their topics of concern: $P(A_i|p_t(A_i) = x)$ is simply $P(A_i)$. In that case one may well ask how there is any motive for forming these future opinions at all, or to see any point in them. And the third is to think of one's future opinions as having some value as opinion, but to ensure synchronic coherence by pre-establishing such harmony among the numbers that the theorem is not violated. If discarding memory is thought of as a work saving device, it is then probably self-defeating.

There is it seems to me a very simply way of dealing with this: a year from now I should say, when asked about this, that I have definite opinions about the rate I was eating croissants per week or month in that earlier time, but no opinion (i.e. totally vague opinion) about any particular day therein. This will automatically satisfy Reflection of course.

Finally, do some concepts force a conflict with Reflection? Modesty would, for example, if it were a conceptual truth that the more [less] I believe that I am modest, the less [more] likely I am to be modest.[12] If that were so, the Reflection principle in conditional form would seem to be violated at once. But in fact, this is a badly constructed concept of modesty, for it would lead to the argument: for any number x below 50%, if I believe only to degree x that I am modest, then it is likely to some degree y strictly between x and 50% that I am modest. Similarly for any number x above 50%. By iteration, I (or anyone) would have to think myself (or oneself) to be exactly as likely to be modest as not!

8. THE DEATH AND DISABILITY DEFENCE

So Reflection is certainly less easily violated than it has sometimes been thought to be. Yet Ulysses may still make its defence seem quixotic anyway. There is one case in which the Reflection principles simply do not apply: if I truly believe that at later time t I shall be dead. Death

itself is not the crucial notion: perhaps we shall have opinions and values after death, or perhaps we have none in a coma or after severe brain damage short of death. The crucial point is that Reflection most certainly needs no revision for the time when, according to me, I shall have no (values and/or) opinions at all. There I can satisfy Reflection principles vacuously.[13]

Referring to examples of Brian Skyrms (1987c), Richard Jeffrey (1988) has proposed an amendment to Reflection principles that exploits analogies to death. Some envisaged transitions are classified by the person himself as "not reasonable" (Jeffrey's term), and Refection is restricted to foreseen future states of mind resulting from "reasonable" transitions. When the examples are of a truly pathological sort (Skyrms' wrath of Khan, who will send a mindworm to infest me, was the earliest, I think; Christensen's "pharmaceutical fiction" of psychedelic Kool-Aid the most recent), it would be more perspicuous to use "pathological" rather than "not reasonable." But however phrased, Jeffrey's proposal was not for a defence, but for a revision. Is that really necessary?

There is a continuum of troublesome cases from the clearly pathological (as classified by the subject) to the clearly endorsed as reasonable (by the subject). At one extreme we have mind-death: there are no opinions or values then. The principles need no amendment there. At the other extreme is the weather forecaster who has tried so hard to codify a good policy for updating his forecasts. To him or her, the principles give good advice for revising the policy. The question is therefore only whether we need a revision for cases around the middle: fear, fatigue, alcohol, sirens' songs and lovers' infatuation, to name but a few.

The question is not only what foreseen transitions – ways of changing my mind – I, the subject, classify as pathological or reasonable. The question is also what I am willing to classify as future opinions of mine. When I imagine myself at some future time talking in my sleep, or repeating (with every sign of personal conviction) what the torturer dictates or the hypnotist has planted as post-hypnotic suggestion, am I seeing this as myself expressing my opinions as they are then? I think not. But what if I now picture myself speaking confidently of my ability to drive home, at the end of a meal garnished with an aperitif, a bottle

or two of Bordeaux or Beaujolais, rum tart, and a snifter of brandy? Or in the middle of that meal?[14]

Slippery slope arguments shouldn't carry all that much weight. A borderline case of a rational subject won't furnish a clear case of a rational violation of Reflection. But I think we must agree with Jeffrey that the Death defence does not remove all our problems – we must make some room for Disability as well. Still it is important that there is such a defence (not a revision) for those cases in which it applies. In Reflection, I refer to the range of genuine values and opinions which are genuinely mine, at the relevant future time, as far as my present opinion of that future allows – not to anything I do not classify as such. We turn now to a different defence.

9. THE INTEGRITY DEFENCE

If you have charge of the departmental cookie fund, and I ask whether you will divert some of it to your own pocket, you'll certainly be offended. If you answer me at all, you'll say *no*! But suppose I continue and ask, what if you were to begin to notice strange, unusual inclinations in yourself to move some of the money from the jar into your pocket? The issue for you, at this point, is whether I am continuing to call into question your collegial and financial integrity, or am doing something quite different. You may find it difficult to take me seriously; but what would or should you say if you did take me seriously?

Let us bring the corresponding question about epistemic integrity into focus, by considering a Good Scientist – you yourself perhaps – who is part of a team which has just made a major announcement. You and your team have established something about materialism in philosophy, namely

materialism is due to a dietary deficiency.[15]

Imagine moreover that you are not a materialist now, but that you have also found out that your diet will have that precise deficiency for the next five years. There is a *prima facie* case for the belief that you will have, five years hence, the opposite opinion about materialism from

the opinion you have now. Do you now violate Reflection, and indeed, violate it on strictly scientific grounds?

But the question of how you will husband your opinions over time is not like the question of whether your hair will turn grey. I think we all know what you, this scientist, should say: "Forewarned as I am, and as no one before us could be, I shall take good care to change my mind about materialism only for good reasons, and not in an irrational fashion." The question about how you will revise your opinion is in the first instance a question about your integrity as epistemic agent. The first reply must be to express your commitment to follow only epistemic policies which you can endorse.[16]

Is this unrealistically optimistic about what lies within one's power? Not necessarily. Consider those chilblain sufferers, who know that they can reliably predict rain when their chilblains hurt. Their forecasts reflect the degree of pain, and the correlations found in their experience so far. But if their chilblain condition deteriorates, they "re-calibrate": they begin to predict rain only after a certain greater amount of pain than before. None of this may be due to careful deliberation and calculation. They let nature condition their expectations, but do not allow their first inclination to predict rain to qualify automatically as their considered opinion, when they have reason to think that the correlation is changing. Similarly with the scientist in our example: he will possibly feel more sympathy for materialism, but not automatically equate his inclinations with his considered opinion.

What if the team establishes that the dietary deficiency interferes also with this task of carefully re-calibrating ourselves, with the ability to form a considered opinion distinct from spontaneous credulity? Again, the first response should be: "Forewarned as I am, I shall take good care. . . ." But clearly the effect could be so strong as to obstruct his proper functioning. In that case, he will no longer be able to formulate a considered opinion, but be at the mercy of strong impulses which he himself classifies as irrational. He will not be in control. From his present point of view, his future behavior will then be a sad parody of epistemic activity. The Death or Disability defence comes into play.

The complete defence of Reflection principles consists therefore in a dilemma, to be faced by any putative counterexample.[17] Integrity

requires me to express my commitment to proceed in what I now classify as a rational manner, to stand behind the ways in which I shall revise my values and opinions. It is on this basis that I rely with confidence on my future opinion, to the modest extent of satisfying the Reflection principle. But integrity pertains to how I shall manage what is in my power. My behavior, verbal or otherwise, is no clue to opinions and values when it does not bespeak free, intentional mental activity. Of course, on the other hand, such activity could be both free and intentional, but careless, mistaken, and fallible in ways that do not make me irrational. When such deficiencies do not lead me in a *predictable and foreseen direction* away from truth, however, foreseeing them does not violate Reflection. It is part of rationality to take the predictable and foreseen into account; that is all.

10. GENERALIZING MOORE'S PARADOX

What sort of status can such a principle as Reflection have at all? My defence in terms of one's integrity as an epistemic agent is very far removed from the usual considerations of gain and loss. I say that opinion which violates Reflection is incoherent. When someone's opinion is in accord with Reflection, this shows up in his or her judgments, which include all those of the form found in my statement of the Reflection Principle, or at least in the absence of contrary judgments.

By what is meant by *incoherence* in this context? Since I do not want to rest the defence on Dutch Book considerations, incoherence is not defined here as vulnerability to such betting schemes. We must revert to its root meaning, which is something like a notion of inconsistency sufficiently broadened so as to be applicable to personal probability judgments. I will argue as follows: the notion of inconsistency which we must broaden, so as to arrive at the relevant probabilist concept, is one found only at the level of pragmatics and not at the level of semantics.

In discussion of subjective probability, it seems to me, the expression of opinion (though first person present tense indicative sentences) is often confused with autobiographical statements of fact (made by means

of those very same sentences). There is no such distinction of linguistic roles to be made for future or past tenses, or for third-person sentences. It is exactly in Moore's paradox and, according to the present argument, in the Reflection Principle that the distinction becomes crucial. It seems to me therefore that the correct notion of probabilistic incoherence must take its inspiration from the notion of inconsistency made manifest by Moore's paradox:

> [MOORE] P, and I do not believe that P
>
> [e.g.] It will rain tomorrow, and I do not believe that it will rain tomorrow

A statement of form [MOORE] might well be true if I said it; indeed it is true for every replacement of P with a true statement which I do not believe. So it is not inconsistent in the sense of "unsatisfiable", "incapable of being true", which is the semantic notion of inconsistency. But I cannot have a coherent state of opinion which I could express by a statement of form [MOORE].

This is paradoxical: how could I agree that I could not coherently believe something which I can clearly see *could* be true? Another way to make our discomfort visible is to look at the situation both from the point of view of the speaker and from that of his neighbor. Your neighbor agrees that you cannot coherently believe that (P and you do not believe that P) though he himself fully believes that P and that you do not believe that P. So it appears that he also thinks that your opinion would be more correct or accurate if you did believe that!

In each case, I think, our discomfort derives from the insistent conviction that *consistent* must coincide with *satisfiable*, and *coherent* with *possibly correct*. But look again at the second way of the previous paragraph: your neighbor does not in fact think that your opinion would necessarily be in a better state if you simply added the true conjunction that you lack. Suppose that all your full beliefs are jointly satisfiable and coherent by every other criterion and they include that P and that Peter does not believe that P. Now you learn the true fact that you are Peter, and behold, if you add this fact to your set of beliefs then you become incoherent. But everyone including your neighbor would agree

that you should not simply add what you learn, but revise your previous opinions accordingly.

We can try to isolate this point by the observation that a true proposition can be simply added to one's beliefs unless some of the prior beliefs are false. That is not strictly speaking correct: if I only have true belief and in fact none of them are beliefs about beliefs at any point in my actual life, the addition of *that* proposition still yields an unsatisfiable set of beliefs. The case of probabilistic judgment gives more difficulty here independently: there is no direct analogue to truth but only "fit" to the facts which admits of degree. The nearest to "simply adding" a proposition is conditionalizing on it, and this changes the probabilities in many places. Only in exceptional cases does it leave the prior opinion as part of the posterior.

This bears on Reflection as follows. Suppose someone says: my probability for A, given that it will be x, equals x. We onlookers say to ourselves that it will indeed later be x but because of the sirens' song, and perceive a conflict between this judgment and the judgment that the sirens' song produces unreliable probabilities. So we might say: the speaker's opinion would improve if he added: my future probability for A will be x, and not-A. However, as I have argued above, he cannot simply "add" this: faced with our putative knowledge about what will happen, he must either (1) insist that he will falsify our claim, (2) give up on the enterprise of rationally managed opinion, or (3) conclude that at this later time he will not have any genuine opinion on the matter. In the last case he will deny what we want him to add. In the second case, this "giving up" – a condition in which he can lay no claim to coherence – will manifest itself in judgments which signal what is wrong with him, i.e. judgments that violate Reflection.

Let us return here to the wider epistemological context broached in the first section of this paper: are there any ways in which my knowledge and opinion of myself can systematically differ from my knowledge and opinion about others, or about people in general? In his book *Freedom of the Individual* Stuart Hampshire argued that sometimes I know what I am going to do because I intend to do it. *How* do I know in such a case? The answer cannot be that I know what I will do because I know that I intend to do it. That cannot be the right answer, exactly because I know too

much else. To give an example, suppose I ask someone "Will you marry me?" and she answers "I fully intend to marry you, and statistics show that such an intention, when not hampered by circumstances beyond one's control, is followed by marriage in 78% of all cases." Then I have a problem. It is not that I doubt what she said, necessarily. Perhaps I believe she is fully sincere and has correctly expressed (and brought to "thetic" awareness) her intention; and perhaps I have also read those very same statistical studies. Then I believe both conjuncts, and hence the conjunction. The problem seems to be rather than when she brings her intention to awareness, the "Hampshire effect" goes away: she no longer knows what she is going to do.

This is not a problem peculiarly about knowledge. The problem of inference to what I will do, from the premise that I *believe* myself to have a certain intention, is certainly worse. If Hampshire is right at all, it must be because I sometimes know (am sure, believe) that I will do something, but not on any basis which I treat as premise for inference. It does not at all follow from this that we have privileged access to a special source of factual information regarding our own future action. There is another possibility. The point may be rather that certain statements, in my mouth, would be inconsistent, but in a *broad* sense of inconsistency.

A longer example may help to relate Moore's paradox more closely to Hampshire's discussion of intention. In Sartre's discussion of bad faith, one sort is called the "project of sincerity." Imagine that a new friends turns out after a few days to have been lying to you. Imagine that when you confront him with this, he says "Yes, I knew that you would soon notice.... I may as well own up: I am a liar. I lie constantly, and even though I realize that it is wrong, and I realize I could do better and often tell myself that I should, the fact is that I won't. I had better admit right now that I will keep lying to you...." That person is in bad faith. The problem here is not that a person couldn't know of some pathological condition resulting in a behavioral disorder. If that were present, his statement would definitely be false, because it included "[I realize that] I could do better." The problem with what he said, its quasi-logical or broadly logical oddity also remains whether we take his words as expressing putative knowledge or only belief about

himself. The answer as a whole is incoherent in a certain way, though not logically inconsistent in a narrow sense.

Given that the Reflection Principle takes first person form, the appearance of arrogance – a sort of trust in one's own opinion one would certainly not have in someone else's – may be a logical illusion. The Moore Paradox is closely allied to the Preface Paradox. Suppose you write a book, intending to assert everything you write. Now you write the preface, and feel inclined to say modestly that your book too, like all preceding ones, will contain some errors. Then you have a dilemma. You can write that preface, but if you do so, you cease to *assert* what follows: you are now merely holding up the body of the book as a text which is predominantly true but false overall. Should you instead still wish to assert the body of the book, you cannot write in the preface anything that contradicts it (no matter what linguistic level you choose for discourse). If you are logically precluded from saying A, and you want to say something definite of that order, you must say not-A. In this case, you would have to say: unlike in other people's books, everything that follows is true. But better not: readers may not realize that you are merely, and modestly, saying what logic forbids you to deny.

So looked at, the foresight involved in intention is assimilated to something quite different from strict logical inference. Normally, we can infer B from A exactly if the conjunction of A and not B is inconsistent. Yet if I am asked whether the truth of P is to be inferred – or even whether I should infer it – from my belief that P, then I will balk. But I will balk here only because I think my answer would be tacitly universalized by the audience in a certain way – as if I were claiming infallibility. The balking does not mean that I cannot see the incoherence or inconsistency, broadly construed, in a state of opinion – if there can be one – correctly expressed by [MOORE]. If I had so constructed my opinions that (implicitly perhaps) their expression would require me both to assert P and deny belief that P, I would have failed in my task as epistemic agent – I would have missed the point of what I was doing.

My contention about Reflection is that if my opinion satisfies these Principles then that is because my intentions, my commitments, which constitute me as an epistemic agent, are reflected in my current opinion – in the way that intentions are naturally so reflected.

11. CASUISTRY FOR EPISTEMIC FRAILTY AND SIN

There is quite a lot about epistemology that is not going to be captured in simple formal principles. My defence of Reflection here implicitly concedes that I can envisage myself violating it: I may say, yes, I do expect to be sufficiently in control of myself to manage my opinion rationally, by my own lights, but I prefer to give in to the temptation not to do so. This is exactly analogous to saying, for example, that one does regard cheating on one's taxes (alimony, child support, ...) as immoral by one's own standards, and can afford not to cheat, yet is going to do so nevertheless. In some of these cases, at least, the culprit is not such an awful person even if s/he is quite ready to forgive her- or himself; in other cases, s/he is. We do not give an absolute overriding priority to epistemic integrity in daily life. To manage our opinion rationally in all respects is not our categorical imperative. Not only to err, but to be irrational, is human; it is normal, and *ceteris paribus*, permissible. An overly strict insistence on epistemic integrity may be, like foolish and needless consistency, a hobgoblin of little minds.[18] But all of this applies equally to "ordinary" coherence, in just the way that it does to Reflection.

Recognizing this, however, leaves us with two important tasks: to understand exactly what form violations of Reflection can take, and to see what advice one might be able to give oneself *on the supposition that* one is violating it: casuistry for epistemic frailty and sin.

The first analogy to explore is with the Good Samaritan paradox in deontic logic. Such examples as "The poor ought to be succored" need to be construed carefully so that it does not entail too much (we can't succor the poor unless there are some, and *ought* implies *can*; so there ought to be poverty?). There are primary principles (e.g. that there ought to be no poverty); if they are violated, we must attend to principles conditioned on their violation. If those in turn are violated, for whatever reason, there will be further deontic advice: if there are poor, and they are not going to be succored, they ought at least not to be exploited, and so forth.

As epistemic example, consider (without endorsement) the principle that one's subjective probability should not be incoherent. Among the

violations, we can certainly still make distinctions of bad and worse.[19] Similarly, I think we can distinguish more and less severe cases among states of opinion violating Reflection. This requires systematic description of what lies between, so to speak, two classes of probability functions defined on a given domain (which includes propositions about one's own future opinion): the large class which the person regards as not unreasonable if we suspend the principle of Reflection, and its subclass of functions which do satisfy that principle. Let us call the former PROB and the latter REFL. I am going to assume that PROB is closed under conditionalization and mixing (convex combination).

For numerically precise opinion (to which I will restrict our attention here), any violation of Reflection can be produced via Moore's paradox. Call B a *Moore proposition* for a reflective person with probability function P if $P(-|B)$ violates some epistemic principle; setting $B = (A \,\&\, p_t(A) = 0.5)$ gives us an example. The problem to solve is this: given that $C = (p_t(A) = x)$ and $P(C) = a$, how can we produce from P a new function P' such that $P'(C) = a$ but $P'(A|p_t(A) = x) = y$? Answer: let $B = (A \,\&\, C)$ and $B' = (-A \,\&\, C)$, and set $P'(-) = ayP(-|B) + a(1 - y)P(-|B') + (1 - a)P(-|-C)$. Note that P' is a Jeffrey conditionalization (i.e. mixture of simple conditionalizations) of P, and that B and B' are Moore propositions for this person.

This gives us a clue to a relevant division of cases. Call REFL-1 the set of functions produced from REFL by conditionalizing on arbitrary propositions. REFL itself is not closed under conditionalization, but REFL-1 is. In general however, neither is closed under Jeffrey conditionalization. Call REFL-2 the set formed from REFL by Jeffrey conditionalization on arbitrary finite partitions.[20] This set is closed under that operation (and the closure of REFL-1 thereunder is the same set). In general REFL-2 will still be only a proper subset of PROB. There are obviously a number of further technical questions to be raised, to most of which I do not yet have answers. We may be able to subdivide REFL-1, for example, by some precise version of: if A and B are two non-equivalent Moore propositions for you, the state of opinion reached by conditionalization on A&B is a more severe violation of Reflection than the one reached by conditionalizing on A alone.

This is a very modest beginning to something I think is important for epistemology: to focus not solely on the constraints of rationality, but also to set about charting the seas of irrationality. This may leave proponents and critics of Reflection finally with a disagreement only over the label of rationality. I think of our epistemic activity as having a point of its own, independent of the ways in which opinion interacts with other factors in evaluating and deciding, and therefore also its own integrity. Others may insist, however, that reason does not require epistemic integrity, and that only the severe sanctions of profit and loss could support putative constraints of rationality. But if we agree to explore what lies on either side of the demarcation, we will have a cooperative enterprise in common regardless of where we draw that line.

Bibliographical Note

When I was writing "Belief and the Will" (1984) I did not realize that the statistician Michael Goldstein (1983, 1985) formulated essentially the same argument for an equivalent principle of iterated expectation. Brian Skyrms had offered an account of opinion about opinion in his "Higher Order Degrees of Belief" (1980) and he extended the theory of diachronic coherence in a number of subsequent publications (1987a; 1987b; 1990, Ch. 5). Other important discussions of the scope and limits of diachronic coherence include J. H. Sobel (1987, 1990). It is not possible to give an exhaustive account of the relevant literature here (see also additional references in the text above), but the most far-reaching treatment so far is undoubtedly Haim Gaifman (1988). This paper also introduces the general idea of the *expert function*, on which I drew both in section IV above and in Chapter 8 of my (1989).

While the above literature is not uncritical, there has also grown up a considerable body of more severe criticism of this approach, some of which (as I noted) I have come to accept. Especially valuable to me were discussions by F. Schick, I. Levi, W. H. Harper, D. K. Lewis, R. Jeffrey, A. Plantinga, and R. Foley. The core criticisms were already presented in essence by Brian Skryms and William Harper at the first presentation of "Belief and the Will" at the New Jersey Regional Philos-

ophy Society in 1983; for a balanced assessment see especially Skyrms (1987a, 1987b) and Jeffrey (1988). Critical papers to which I have reacted in the text above include W. J. Talbot (1991) (first presented at the American Philosophical Association, Eastern Division, 1987); this is also discussed in Bacchus, Kyburg, and Thalos (1990). Two recent critical articles, referred to repeatedly above, are by David Christensen (1991) and by Patrick Maher (1992). Christensen says "bluntly, there are cases in which satisfying ... Reflection would be downright stupid" (p. 230). Maher tempers his critique with the point that arguments for Bayesian conditionalization are also *ipso facto* arguments for Reflection, and with a decision theoretic argument for satisfying Reflection under certain conditions.

NOTES

[1] While self-contained, this paper is a sequel to my "Belief and the Will" (1984); see the *Bibliographical Note* at the end of this paper for a survey of the relevant literature since then and supplementary references. I have meanwhile benefitted greatly from Brad Armendt's commentary and Richard Foley's "How should future opinion affect current opinion?", both presented at a symposium at the APA Central Division, April 1993.

[2] It is not a coincidence that I choose here an "expert" as example; I am indebted here to Gaifman (1988).

[3] "Opinion" here covers both probability and expectation. Semantic and set-theoretic paradoxes threaten if such a principle is left with the range of applicability unrestricted. I will come back to that below.

[4] The same may be remarked for the more liberal ideal introduced by Richard Jeffrey if we construe it as follows: there is at least one partition E(i,t) all of whose members have positive probability now and the foreseen possible opinions at future time t include the Jeffrey Conditionalizations on this partition. Jeffrey (1988) discusses the relation between Reflection and such "sufficiency" conditions.

[5] Here I can state the exact model I have in mind for the subject's form of opinion. There is a fixed field of measurable sets ("propositions"), and his state of opinion at any given time can be summed up as a set of vague expectation value judgments for associated measurable functions ("random variables"). This is a guard against semantic paradoxes, and sufficiently general here. For discussion and references concerning vague opinion, see van Fraassen (1989; 1990, 153–156 and 193–194).

[6] Accordingly, such examples as Christensen's "I believe to degree 0.95 that I have no beliefs to a degree greater than 0.90" do not fall within the scope of this paper. For a

satisfactory treatment of opinion for which the noted assumption does not hold, I refer to the papers by Gaifman and Skyrms.

[7] If $E(P,q) = 0$ and we abbreviate $[p_t(A) = x]$ to X then $0 = (1 - x)P(X \& A) - xP(X \& -A) = P(X \& A) - x[P(X \& A) + P(X \& -A)] = P(X \& A) - xP(X)$ hence $P(X \& A) = xP(X)$.

[8] As Brad Armendt (1993) has correctly pointed out, the stories about Pierone and Piero are only about subjects who may or may not have opinions about their own opinion. It might be argued that the factual propositions $[p_t(A) = x]$, being in first person future tense, can bring unexpected problems for the general reasoning of the preceding sections. I grant this possibility, but submit that if their first person character is problematic for this general reasoning, that will support my exploitation of this character in the "Integrity defence" presented below.

[9] Similarly the Reflection Principle for those with precise numerical opinion easily allows the introduction of a factor of over and underconfidence U(s,t) such that $P(A|p_t(A) = x \text{ and } U(s,t)) = sx$. The expectation value of the parameter s will then equal 1 if Reflection in satisfied.

[10] It won't if my current probability equals 100%, and I have a lower bound above zero for having a future opinion definitely below that – at least not if we construe vague opinion in accord with Reflection to be what is common to a set of models of opinion constrained to being precise and also in accord with Reflection. The limitations of subjective probability models around the edges (100% and 0%) are well known, and I hope to investigate them elsewhere.

[11] I want to thank Elijah Millgram and Sarah Buss of my department for stimulating conversation on this subject.

[12] The suggestion that such "anti-correlation" characteristics might yield violations of Reflection was made by David Christensen.

[13] This is obviously so for the principle in its conditional form. The general formulation may need adjustment, depending on one's policy on non-referring descriptions, for the case in which I am *certain* I shall have no opinions at the later time. Usually I will not totally rule out that the mind-death will occur; I will just not count its manifestations as expressing opinions I then have.

[14] In Maher's central example, the person is sure that he will be overconfident after ten drinks, and does think of that overconfidence as truly his own, rejecting what I here call the "death defence". Personally I regard this example as an unrealistic (though cliché) fiction: in reality we are neither totally sure of the effect of the drink (which might be underconfidence or depression at least sometimes), nor do we classify the effect of ten drinks as so unlike more extreme cases. If a cartoon-like fiction is classified as irrational, I do not regard that as a serious counterexample to the theory. But there is indeed a slippery slope to examples which are less idiosyncratic and yet not clearly stopped by the death defence.

[15] 1986: graffiti in the men's room in the Princeton philosophy department, the washroom with the spider.

[16] This also answers Maher's argument that if you want to satisfy Reflection, you will want to take mind-bending drugs, which will make you feel unaccountably sure e.g. about which horse will win. That is similar to the argument that if you want to maximize expected utility, you should increase your subjective probability for the outcomes which you prefer. If you already classify certain ways of changing your opinion as likely to lower your calibration, or as incompatible with your integrity as epistemic actor, you are already committed not to use them.

[17] Christensen's example of the Messiah complex seems to me also to founder on this dilemma. My integrity demands no commitment that I will not reach the conclusion that I am the Messiah, but that I will not do so on insufficient grounds; a possible future in which I cannot carry out this commitment reduces to the previous case of his "pharmaceutical fiction".

[18] This admission does not give cogency to the sort of example Maher provides ("a superior being who gives eternal bliss to all and only those who are certain that pigs can fly"), in which integrity is overruled by the profit motive – I regard that as an abdication from reason (just as overruling moral integrity by considerations of gain and loss is immoral), for motives that are hardly admirable, even if understandable.

[19] Suppose the domain of P is the Boolean algebra generated by A, B, C, D. If P is additive on a subalgebra generated by three of them, that is better than if it is additive only on subalgebras generated by at most two of them.

[20] As example, suppose that REFL has only one member, and PROB is exactly its closure under conditionalization and mixing. Then REFL-1 is not closed under mixing, where REFL-2 = PROB.

REFERENCES

Bacchus, F., H. E. Kyburg Jr., and M. Thalos (1990) 'Against Conditionalization', *Synthese* 85, 475–506.

Bernardo, J. M. et al. (eds.) (1985) *Bayesian Statistics 2* (Amsterdam: Elsevier Science Publishers).

Christensen, D. (1991) 'Clever Bookies and Coherent Beliefs', *Philosophical Review*, 229–247.

Dunn, J. M. and A. Gupta (eds.) (1990) *Truth or Consequences* (Dordrecht: Kluwer Academic Publishers).

Elster, J. (1979) *Ulysses and the Sirens* (Cambridge: Cambridge University Press).

Gaifman, H. (1988) 'A Theory of Higher Order Probabilities', in Skyrms and Harper (1988), 191–219.

Goldstein, M. (1983) 'The Prevision of a Prevision', *Journal of the American Statistical Association* 78, 817–819.

Goldstein, M. (1985) 'Temporal Coherence', in Bernardo et al. (1985), 231–248.

Jeffrey, R. (1988) 'Conditioning, Kinematics, and Exchangeability', in Skyrms and Harper (1988), 221–255.
Lewis, D. K. (1979) 'Attitudes De Dicto and De Se', *Philosophical Review* 88, 513–543; reprinted in his *Collected Papers* I (New York: Oxford University Press, 1983).
Maher, P. (1992) 'Diachronic Rationality', *Philosophy of Science* 59, 120–141.
McNeill, I. B. and G. Umphrey (eds.) (1987) *Advances in the Statistical Sciences, II* (Dordrecht: Reidel Pub. Co.).
Mellor, D. H. (ed.) (1980) *Prospects for Pragmatism* (Cambridge University Press).
Savage, C. Wade (ed.) (1987) *Justification, Discovery, and the Evolution of Scientific Theories* (Minneapolis: University of Minnesota Press).
Seidenfeld, T. (1985) 'Calibration, Coherence, and Scoring Rules', *Philosophy of Science* 52, 274–294.
Skyrms, B. 'Higher Order Degrees of Belief', in Mellor (1980), 109–137.
Skyrms, B. (1987a) 'Dynamic Coherence', in McNeill and Umphrey (1987), 233–243.
Skyrms, B. (1987b) 'Dynamic Coherence and Probability Kinematics', *Philosophy of Science* 54, 1–20.
Skyrms, B. (1987c) 'The Value of Knowledge', in Savage (1987).
Skyrms, B. (1990) *The Dynamics of Rational Deliberation* (Harvard University Press).
Skyrms, B. and W. H. Harper (eds.) (1988) *Causation, Chance, and Credence* 1 (Dordrecht: Kluwer Academic Publishers).
Sobel, J. H. 'Evidential Bearings, Rational Updates, and Dutch Strategies', ms. 1990.
Sobel, J. H. (1987) 'Self-doubts and Dutch Strategies', *Australasian Journal of Philosophy* 65, 56–81.
Talbot, W. J. (1991) 'Two Principles of Bayesian Epistemology', *Philosophical Studies* 62 (1991), 135–150.
van Fraassen, B. (1984) 'Belief and the Will', *Journal of Philosophy* 81, 235–256.
van Fraassen, B. (1989) *Laws and Symmetry* (Oxford: Oxford University Press).
van Fraassen, B. (1990) 'Figures in a Probability Landscape', in Dunn and Gupta (1990), 345–356.

Department of Philosophy
Princeton University
Princeton, NJ 08544
USA