

# PREFERENCE-BASED ARGUMENTS FOR PROBABILISM\*

David Christensen†‡

Department of Philosophy, University of Vermont

---

\*

†

‡ Thanks to Mark Kaplan, Hilary Kornblith, and Patrick Maher and three anonymous referees for helpful discussions or comments on earlier drafts.

## **ABSTRACT**

Both Representation Theorem Arguments and Dutch Book Arguments support taking probabilistic coherence as an epistemic norm. Both depend on connecting beliefs to preferences, which are not clearly within the epistemic domain. Moreover, these connections are standardly grounded in questionable definitional/metaphysical claims. The paper argues that these definitional/metaphysical claims are insupportable. It offers a way of reconceiving Representation Theorem arguments which avoids the untenable premises. It then develops a parallel approach to Dutch Book Arguments, and compares the results. In each case preference-defects serve as a diagnostic tool, indicating purely epistemic defects.

## **1. Introduction**

The most natural way of applying the principle of deductive consistency to degrees of belief is provided by probability theory. Probabilism, which endorses this application of traditional logic, seems to be a purely epistemic position. But the most prominent arguments for Probabilism--Representation Theorem Arguments and Dutch Book Arguments--depend crucially on positing certain connections between beliefs and preferences, which are not clearly within the epistemic domain.

Standard presentations of both sorts of argument ground the posited connections in definitional or metaphysical claims. Here, I'd like to look first at arguments based on Representation Theorems, arguing that the relevant definitional/metaphysical claims are insupportable. I'll then offer a way of reconceiving the arguments which avoids the untenable premises. Finally, I'll turn to parallel issues in the context of Dutch Book arguments for Probabilism, and compare the results. In each case, I'll argue that preference-defects serve as a diagnostic tool, indicating purely epistemic defects.

## **2. The Representation Theorem Argument (RTA)**

Representation Theorems show that if an agent's preferences obey certain intuitively attractive formal constraints, they can be represented as resulting from a relatively unique<sup>1</sup> pair, consisting of a set of degrees of belief and a set of utilities, such that (1) the degrees of belief are probabilistically coherent, and (2) the preferences maximize expected utility relative to those beliefs and utilities. But a theorem is not an argument. Typical arguments defending probabilism begin with some version of the following two principles:

**Preference Consistency:** Ideally rational agents' preferences obey constraints C.

**Representation Theorem:** If an agent's preferences obey constraints C, then they can be represented as resulting from some unique set of utilities U and probabilistically coherent degrees of belief B relative to which they maximize expected utility.

Clearly, these principles alone are not enough to support the intended conclusion. Thus standard treatments of the RTA also endorse a version of the following principle:

**Representation Accuracy:** If an agent's preferences can be represented as resulting from unique utilities U and probabilistically coherent degrees of belief B relative to which they maximize expected utility, then the agent's actual utilities are U and her actual degrees of belief are B.

Given these three principles, we get:

**Probabilism:** Ideally rational agents have probabilistically coherent degrees of belief.

Thus understood, representation theorems provide for a particularly interesting kind of argument. From a normative constraint on preferences alone, along with some mathematics and a principle about the accuracy of certain representations, we can derive a normative constraint on degrees of belief.

The mathematical meat of this argument--the representation theorem itself--has naturally received most of the attention. Of the more purely philosophical principles, Preference Consistency has been discussed much more widely. Some claim that its constraints on preferences are not satisfied by real people, and, more interestingly, that violations of the constraints are not irrational. I'll pass over this discussion for the present, assuming that the constraints are plausible rational requirements.<sup>2</sup> Instead, I'll begin by focusing on Representation Accuracy. Suppose that an agent has preferences which would accord with expected-utility (EU) maximization relative to some unique U and B. Why should we then take U and B to be her actual utilities and beliefs?<sup>3</sup>

Representation Accuracy posits a particular connection to hold among agents' preferences, utilities and beliefs. That there is, in general, some connection of very roughly the sort posited is an obvious truism of folk psychology. People do typically have preferences for options based on how likely they believe the options are to lead to outcomes they value, and on how highly they value the possible outcomes. But the cogency of the RTA requires a connection much tighter than this.

We can start to see why by noting that the purposes of the RTA would not be served by taking Representation Accuracy as a mere empirical regularity, no matter how well confirmed. For the purported empirical fact--that having probabilistically coherent beliefs was, given human psychology, causally necessary for having consistent preferences--would at best show probabilistic coherence valuable in a derivative and contingent way. After all, one might discover empirically that, given human psychology, only those whose beliefs were unrealistically simple, or only those suffering from paranoid delusions, had preferences consistent enough to

obey the relevant constraints. If a representation theorem is to provide a satisfying justification for Probabilism--if it is to show that the rules of probability are the rules of logic for degrees of belief--then the connection between preferences and beliefs will have to be a deeper one.

### **3. The RTA and the Metaphysics of Belief**

In fact, RTA-proponents do posit deeper connections between preferences and beliefs. They typically take degrees of belief (and utilities) to be in some sense *defined* by preferences. In their philosophical comments on the proposed definitions, Ramsey and some subsequent RTA-proponents have emphasized the need to *measure* degrees of belief. Taken unsympathetically, this suggests some sort of operationism or related notion of definition via analytic meaning postulates. But it seems to me that a more charitable reading of the argument is available.

Let us begin with a look at the role degrees of confidence play in psychological explanation. As RTA-proponents point out, we often explain behavior--especially in deliberate choice situations--by invoking degrees of belief. Often, these explanations seem to proceed via just the sort of principle that lies behind Representation Accuracy. We explain someone's selling a stock by an increase in his confidence that it will soon go down, assuming that his choice is produced by his preferences, which themselves result from his beliefs and utilities in something like an EU-maximizing way.

Thus we might see Representation Accuracy as supported by the following kind of thought: "The belief-desire model is central to the project of explaining human behavior. Degrees of belief are posited as working with utilities to produce preferences (and hence choice behavior). The law connecting beliefs and utilities to preferences is that of maximizing EU. So

beliefs are, essentially, that which, when combined with utilities, determine preferences via EU-maximization.” Thus Ramsey gives this preliminary definition of degree of belief:

I suggest that we introduce as a law of psychology that [a person’s] behaviour is governed by what is called the mathematical expectation; that is to say that, if  $p$  is a proposition of which he is doubtful, any goods or bads for whose realization  $p$  is in his view a necessary and sufficient condition enter into his calculations multiplied by the same fraction, which is called the ‘degree of belief in  $p$ ’. We thus define degree of belief in a way which presupposes the use of the mathematical expectation.<sup>4</sup>

Patrick Maher, in a sophisticated recent defense of the RTA, writes:

I suggest that we understand probability and utility as essentially a device for interpreting a person’s preferences. On this view, an attribution of probabilities and utilities is correct just in case it is part of an overall interpretation of the person’s preferences that makes sufficiently good sense of them and better sense than any competing interpretation does. ... [I]f a person’s preferences all maximize expected utility relative to some  $p$  and  $u$ , then it provides a perfect interpretation of the person’s preferences to say that  $p$  and  $u$  are the person’s probability and utility functions. (Maher 1993, 9)

This approach toward defining degrees of belief by preferences need not be fleshed out by any naive commitment to operationism, or to seeing the relevant definition as analytic or a priori.

It might rather be thought of as a scientific definition, combining elements of conceptual refinement with empirical investigation. Beliefs turn out to be functional or dispositional properties of people, defined, along with utilities, by their joint causal connections to preferences. On such a view, the fact that a strong belief that a stock will go down produces a strong preference to sell it is neither an analytic truth nor a mere empirical regularity. Part of what *constitutes* an agent's having a strong belief that the stock will go down is precisely her disposition (given the usual utilities) to prefer selling the stock. Thus there is a metaphysical or constitutive connection between degrees of belief, utilities, and preferences. This idea has obvious connections to functionalist theories in mainstream philosophy of mind.

Of course, this claim about the nature of beliefs cannot represent mere naked stipulation. If it is to have relevance to epistemology, the entities it defines must be the ones we started wondering about when we began to inquire into rational constraints on belief.

One worry we might have on this score is that the EU-based definition offered by RTA-proponents is not the only one that would fit the somewhat vague intuitions we have about, e.g., the stock-selling case. Suppose we have an agent whose preferences fit the constraints and can thus be represented as resulting from coherent beliefs  $B$  and utilities  $U$ . Lyle Zynda argues that there will be another belief-function  $B'$  which is probabilistically incoherent, yet which may be combined with  $U$  (non-standardly) to yield a valuation function fitting the agent's preference ordering equally well.<sup>5</sup> Zynda concludes that the RTA can be maintained, but that we must justify our choice of  $B$  over  $B'$ . Endorsing Maher's view that probabilities and utilities are "essentially a device for interpreting a person's preferences," he favors taking a less-than-fully realistic view of beliefs, on which our choice of  $B$  over  $B'$  can be made on frankly pragmatic

grounds.

It seems to me, however, that the RTA-proponent faces complexities beyond those revealed by Zynda's example. For our question is not merely whether the proposed definition uniquely satisfies our intuitions about deliberate choice cases. We want to know how closely this definition fits our intuitive concept in general. Let us look, then, a bit more broadly at the pre-(decision-)theoretic notion of strength of belief.

To begin with, it is obvious that anyone can tell by quick introspection that she is more confident that the sun will rise tomorrow than that it will rain tomorrow. But it is not at all clear that this aspect of our common notion is captured by the envisioned definition. And, in fact, some RTA-proponents have considered this sort of worry. Ramsey, dubious of measuring degrees of belief by intensity of introspected feeling, saw his definition as capturing "belief qua basis of action," arguing that even if belief-feelings could be quantified, beliefs as bases of action were what was really important (1926, 171-172). Ellery Eells (1982, 41-43) also supports seeing beliefs as dispositions to action by developing Ramsey's criticism of measuring degrees of belief via feelings of conviction.

This discounting of the introspective aspect of our pretheoretic notion is not an unreasonable sort of move to make. If a common concept is connected both to quick identification criteria and to deeper explanatory concerns, we do often override parts of common practice. Thus we might discount introspectively-based claims about degrees of belief if and when they conflict with the criteria flowing from our explanatory theory. This move is made more reasonable by the fact, emphasized by some RTA-proponents, that our introspective access seems pretty vague and prone to confusion.

But the general worry--that the preference-based definition leaves out important parts of our pretheoretic notion--is not this easily put aside. For one thing, it seems clear that even within the realm of explaining behavior, degrees of belief function in ways additional to explaining preferences (and thereby choice-behavior). For example, we may explain someone coming off well socially on the basis of her high confidence that she will be liked. Or we may explain an athlete's poor performance by citing his low confidence that he would succeed.

Examples like this can be multiplied without effort. And it does not seem that anything involving choice between options, or, really, any aspect of preferences, is being explained in such cases. Rather, it is an important psychological fact that a person's beliefs--the way she represents the world--affect the way she behaves in countless ways that have nothing directly to do with the decision theorist's paradigm of cost-benefit calculation.

Moreover, degrees of belief help explain much more than behavior. We constantly invoke them in explanations of other psychological states and processes. Inference is one obvious sort of case: we explain the meteorologist's increasing confidence in rain tomorrow by reference to changes in her beliefs about the locations of weather systems. But beliefs are also universally invoked in explanations of psychological states other than beliefs (and other than preferences). We attribute our friend's sadness to her low confidence in getting the job she's applied for. We explain a movie character's increasing levels of fear on the basis of his increasing levels of confidence that there is a stranger walking around in his house. The connections between beliefs and other psychological states invoked in such explanations are, I think, as basic, universal and obvious as the central connections between beliefs and preferences which help explain behavior.

Other non-behavioral effects of beliefs are less obvious, though equally real. Every reputable drug study controls for the placebo effect--the fact that people's confidence that they are taking effective medicine reliably causes their conditions to improve, often in physiologically measurable ways. The exact mechanisms behind the placebo effect are unclear, but one thing is clear: this effect of beliefs is not explained by any disposition of the patients to have preferences or make choices which maximize utility relative to a high probability of their having taken effective medicine.

Thus it turns out that the RTA-proponents' problem with accommodating introspective access to our degrees of belief represents the tip of a very large iceberg. True, degrees of belief are intimately connected with preferences and choice behavior. But they are also massively and intimately connected with all sorts of other aspects of our psychology (and perhaps even physiology). This being so, the move of settling on just one of these connections--even an important one--as definitional comes to look highly suspicious.

This is not to deny that beliefs may, in the end, be constituted by their relations to behaviors and other mental states--by their functional role in the agent. But even functionalists have not limited their belief-defining functional relations to those involving preferences; indeed, it is hard to see any independent motivation for doing so. And if the preference-explaining dispositions are only parts of a much larger cluster of dispositions which help to constitute degrees of belief, then it is hard to see how Representation Accuracy, or Maher's claim quoted above, can be maintained. After all, a given interpretation of an agent's degrees of belief might maximize expected-utility fit with the agent's preferences, while a different interpretation might fit much better with other psychological-explanatory principles. In such cases of conflict, where

no interpretation makes all the connections come out ideally, there is no guarantee that the best interpretation will be the one on which the agent's preferences accord perfectly with maximizing EU. And if it is not, then even an agent whose preferences obey Preference Consistency may fail to have probabilistically coherent degrees of belief. Thus it seems that even if we take a broadly functionalist account of degrees of belief, Representation Accuracy is implausible.

Moreover, it is worth pointing out that the assumption that beliefs reduce to dispositional or functional states of any sort is highly questionable. The assumption is clearly not needed in order to hold, e.g., that preferences give us a quite reliable way of measuring degrees of belief, or that beliefs play a pervasive role in explaining preferences and other mental states and behaviors. Beliefs can enter into all sorts of psychological laws, and be known through these laws, without being reductively defined by those laws. They may, in short, be treated as typical theoretical entities, as conceived of in realistic philosophy of science.<sup>6</sup> If the connections between beliefs and preferences have the status of empirical regularities rather than definitions--if the connections are causal rather than constitutive--then the RTA would fail in the manner described above. It would be reduced to showing that, given human psychology (and probably subject to extensive *ceteris paribus* conditions) coherent beliefs do produce rational preferences. This is a long way from showing that coherence is the correct logical standard for degrees of belief.

In retrospect, perhaps it's not surprising that the ironclad belief-preference connection posited in Representation Accuracy fails to be groundable in--or even cohere with--a plausible metaphysics of belief. Degrees of belief are not merely part of a "device for interpreting a person's preferences." Beliefs are our way of representing the world. They come in degrees because our evidence about the world justifies varying degrees of confidence in the truth of

various propositions about the world. True, these representations are extremely useful in practical decisions; but that does not reduce them to mere propensities to decide. After all, it seems perfectly coherent that a being could use evidence to represent the world in a graded manner without having utilities or preferences at all!

Such a being would not be an ordinary human, of course. But even among humans, we can observe differences in apparent preference intensities. (Clearly, intersubjective comparisons are difficult, but that hardly shows that intersubjective differences are unreal.) I don't think that we would be tempted to say, of a person affected with an extreme form of diminished affect--a person who had no preferences--, that he had no beliefs about anything. After all, it's obvious that one cares about some things much more than others. One can easily imagine one's self coming to care less and less about more and more things. But insofar as one can imagine this process continuing to the limit, it does not in the slightest seem as if one would thereby lose all beliefs.

One might object that a preferenceless being would still have *dispositions* to form EU-maximizing preferences, in circumstances where it acquired utilities. But what reason would we have to insist on this? Given the being's psychological makeup, it might be impossible for it to form utilities. Or the circumstances in which it would form utilities might be ones where its representations of the world would be destroyed or radically altered.

The suggestion that having a certain degree of belief reduces to nothing more than the disposition to form preferences in a certain way should have struck us as overly simplistic from the beginning. After all, it is part of commonsense psychology that the strength of an agent's disposition, e.g., to prefer bets on the presence of an intruder in the house, will be strongly

correlated with the strength of the agent's disposition to feel afraid, and with the strength of his disposition to express confidence that there's an intruder in the house, etc.. The view that identifies the belief with just one of these dispositions leaves the other dispositions, and all the correlations between them, completely mysterious. Why, for example, would the brute disposition to form preferences in a certain way correlate with feelings of fear?<sup>7</sup>

This point also makes clear why it won't do to brush the problem aside by claiming only to be discussing a particular sort of belief, such as "beliefs qua basis of action." It's not as if we have one sort of psychological state whose purpose is to inform preferences, and a separate sort of state whose purpose is to guide our emotional lives, etc.. As Mark Kaplan notes (in arguing for a different point), "You have only one state of opinion to adopt--not one for epistemic purposes and another for non-epistemic purposes" (1996, 40). What explains the correlations is that they all involve a common psychological entity: the degree of belief.

Degrees of belief, then, are psychological states that interact with utilities and preferences, as well as other aspects of our psychology, and perhaps physiology, in complex ways, one of which typically roughly approximates EU-maximization. Whether we see the connection between the preference-dispositions and beliefs as partially constitutive (as functionalism would) or as resulting from purely contingent psychological laws (as a more robust realism might) is not crucial here. For neither one of these more reasonable metaphysical views of belief can support Representation Accuracy. If this is correct, then it becomes unclear how a Representation Theorem, even in conjunction with Preference Consistency, can lend support to Probabilism.<sup>8</sup>

#### 4. A De-metaphysicized RTA

Representation Accuracy asserted that whenever *any* agent's preferences maximized EU relative to a unique U and B, the agent's actual utilities and beliefs were U and B. The suspicious metaphysics was needed to ensure the universality of the posited preference-belief connection. But the RTA's conclusion does not apply to all agents--only to ideally rational ones. Thus the purpose of the RTA could be served without commitment to the preference-belief connection holding universally. It would be served if such a connection could be said to hold for all *ideally rational* agents.

Now one might well be pessimistic here--after all, if agents in general may have degrees of belief that do not match up with their utilities and preferences in an EU-maximizing way, why should this be impossible for ideally rational agents? The answer would have to be that the EU-maximizing connection is guaranteed by some aspect of ideal rationality. In other words, the source of the guarantee would be in a *normative*, rather than a metaphysical, principle.

This basic idea is plausible enough: An ideally rational agent's preferences are not only consistent with one another in the ways presupposed in the obviously normative Preference Consistency principle. In addition, they must also cohere in a certain way with the agent's degrees of belief. Of course, we cannot simply posit that such an agent's preferences maximize EU relative to her beliefs and utilities. Expected utility is standardly defined relative to a probabilistically coherent belief function. So understood, our posit would blatantly beg the question: if we presuppose that ideal rationality requires maximizing EU in this sense, then the rest of the RTA, including the RT itself, is rendered superfluous.

Nevertheless, I think that a more promising approach may be found along roughly these

lines. Let us begin by examining the basic preference-belief connection assumed to hold by RTA-proponents such as Savage (1954) and Maher. In the course of proving their results, they first define a “qualitative probability” relation. This definition is in terms of preferences; it is at this point that the connection between preferences and beliefs is forged. The arguments then go on to show how (under specified conditions) a unique quantitative probability function corresponds to the defined qualitative relation.

Maher explains the definition of qualitative probabilities intuitively as follows: “We can say that event  $B$  is more probable for you than event  $A$ , just in case you prefer the option of getting a desirable prize if  $B$  obtains, to the option of getting the same prize if  $A$  obtains.”<sup>9</sup> And it seems to me that there is something undeniably attractive about the idea that, in general, when people are offered gambles for desirable prizes, they will prefer the gambles in which the prizes are contingent on more probable propositions. However, in light of the arguments above, we should not follow Savage and Maher in taking this sort of preference-belief correspondence to *define* degrees of belief. In fact, we should not even assume that the connection holds *true* for all agents (or even for all agents whose preferences satisfy the RTA’s constraints on preferences). Instead, we may take this sort of preference-belief connection to be a *normative* one--a connection that holds for all ideally rational agents.

Seen as a claim about the way preferences *should* connect with beliefs, the normative connection posited in the RTA would amount to something like this:

**Informed Preference:** An *ideally rational* agent prefers the option of getting a desirable prize if

$B$  obtains to the option of getting the same prize if  $A$  obtains, just in case  $B$  is more

probable for that agent than  $A$ .<sup>10</sup>

Informed Preference avoids the universal metaphysical commitments of Representation Accuracy. We may maintain such a principle while acknowledging the psychological possibility of a certain amount of dissonance between an agent's degrees of belief and her preferences, even when those preferences are consistent with one another. At the same time, Informed Preference forges the preference-belief connection for all ideally rational agents, who are anyway the only ones subject to the RTA's desired conclusion.<sup>11</sup>

Suppose, then, that the RTA was formulated using a suitably precise version of Informed Preference, understood not as a definition, but as a normative requirement. The RTA thus understood would presuppose explicitly a frankly normative connection between beliefs and preferences, something the RTA as standardly propounded does not do. Such an argument will thus need to be in one way more modest than the metaphysically interpreted RTA: it will not purport to derive normative conditions on beliefs in a way whose only normative assumptions involve conditions on preferences alone.

Still, strengthening the RTA's normative assumptions in this way does not render it question-begging, as simply assuming EU-maximization would have. The intuitive appeal of Informed Preference does not derive from any explicit understanding of the principles of probabilistic coherence. The principle would, I think, appeal on a common-sense level to many who do not understand EU, and who are completely unaware of, e.g., the additive law for probabilities.

Thus understood, the RTA still provides an interesting and powerful result. From

intuitively appealing normative conditions on preferences alone, along with an appealing normative principle connecting preferences with beliefs, we may derive a substantial normative constraint on beliefs--a constraint that is not obviously implicit in our normative starting points. Thus it seems to me that the RTA can be freed from any entanglement with fishy metaphysics; and once this is done, The RTA can lend substantial support to Probabilism.<sup>12</sup>

### **5. The Depragmatized DBA: Criticisms and Corrections**

The RTA is cousin to a less formal argument for the same conclusion: the Dutch Book Argument (DBA), which shows that a person who accepts bets at odds set in the natural way by her degrees of belief will accept a set of bets guaranteeing her loss--unless her beliefs are probabilistically coherent. The DBA has widely been criticized as too pragmatic to ground a purely epistemic constraint. Some attempts to answer this criticism have held that the beliefs producing Dutch Book vulnerability must involve preference inconsistency, and hence that the vulnerable agent does suffer from a logical, and not just a practical, problem. But this line requires seeing preferences and degrees of belief as being related in definitional or metaphysical ways very similar to those criticized above.<sup>13</sup>

Attempts have been made to give clearly non-pragmatic versions of the DBA which do not connect beliefs and preferences in this way. Maher has rejected all such attempts, arguing that the DBA should be abandoned in favor of his version of the RTA.<sup>14</sup> Although we have seen reason to reject Maher's version of the RTA, it is worth examining how our de-metaphysicized RTA compares to the depragmatized DBA, in light of Maher's criticisms of the latter. I will concentrate on the informal version of the DBA given in Christensen 1996, which proceeds

roughly as follows:

1. If an agent's degrees of belief violate the probability axioms, then there is a set of monetary bets, at odds matching those degrees of belief, which will inflict on the agent a sure monetary loss.<sup>15</sup>
  2. An agent's degrees of belief *sanction as fair* monetary bets at odds matching her degrees of belief.
  3. If a set of purportedly fair betting odds allows construction of a set of bets at those odds which will inflict on the agent a sure monetary loss, the set of betting odds is defective.
  4. If an agent's beliefs sanction as fair each of a set of betting odds, and that set of betting odds is defective, then the agent's beliefs are defective.
- C. If an agent's degrees of belief violate the probability axioms, they are defective.

The main difference between this and standard versions of the DBA is in premise 2. While other versions see a definitional or metaphysical connection between an agent's degrees of belief and her bet-evaluations, the de pragmatized DBA sees only a normative connection. An agent's degrees of belief are taken to "sanction as fair" certain betting odds. Sanctioning the odds as fair is to be understood as providing them with *ceteris paribus* justification.

Unfortunately, Christensen 1996 does not specify what is supposed to be covered in the *ceteris paribus* condition. It is, however, clearly intended at least to exclude cases in which money is valued in a non-linear way, or in which the agent has other values, such as risk-aversion, which would enter into evaluating the bets.

Maher represents the above argument in much more mathematical terms, and then develops two main lines of criticism. The less interesting of these, to my mind, is one that depends in part on a feature Maher introduces into his formal interpretation.<sup>16</sup> Maher in effect reconstrues the third and fourth premises above so that they apply to single bets sanctioned by the agent's beliefs, rather than to sets of bets. He then proves that the argument thus construed only works if one adds a further premise, roughly as follows:

M: If an agent's beliefs sanction each of two bets as fair, then they also sanction the mathematical sum of the two bets as fair.

Maher then asks how M could be supported. He notes that violations of M would occur in certain situations where the additive law for probabilities was violated, but points out that it would be question-begging to invoke the additive law in support of M.

The short answer to this objection would be that supporting M is not required here, since M is not part of the above argument. Still, it will be worth seeing whether an analogue of this question poses a problem for an M-free version of the DBA. I'll return to this matter below, after the argument has been reformulated in light of Maher's second line of criticism.

As Maher notes, "sanctioning odds as fair" is explained as constituting a *ceteris paribus*

justification for evaluating odds as fair. But absent a more careful spelling out of the ceteris paribus conditions, it is impossible to evaluate the plausibility of premise 2.<sup>17</sup> Moreover, the argument above holds that if an agent's beliefs sanction as fair a defective set of odds, then those beliefs are defective. But if the "defectiveness" in the odds is just due to failure of the ceteris paribus conditions, then, as Maher correctly points out, it's far from clear that any defect need be found in the beliefs, which, after all, provide justification for those odds only contingent on the ceteris paribus conditions obtaining.

Given the difficulties with the ceteris paribus-dependent notion of sanctioning as fair, it seems to me that the DBA is best served if we avoid ceteris paribus conditions altogether, and reformulate the de pragmatized DBA accordingly. The reformulation I have in mind does not require changing the first premise of the above-given formulation, so our new premise 1' may simply repeat premise 1 above:

1'. If an agent's degrees of belief violate the probability axioms, then there is a set of monetary bets, at odds matching those degrees of belief, which will inflict on the agent a sure monetary loss.

But the second premise, which depends on the ceteris paribus clause, does need revision, in two ways. First, let us apply the premise--and thus the argument as a whole--explicitly to the sort of agent who would have satisfied the sort of ceteris paribus clause gestured at in Christensen 1996. In particular, let us concentrate on what I'll call a *simple agent*: one who values money positively, in a linear way, and who does not value anything else.

Second, let us take the concept of *sanctioning as fair* in a way that is not subject to any implicit *ceteris paribus* clause. Sanctioning as fair is an informal, intuitive normative connection between an agent's beliefs and her preferences concerning betting odds. An agent's degree of belief in a certain proposition sanctions betting odds as fair iff it provides justification for evaluating bets at those odds as fair--i.e. for being indifferent to taking either side of bets at those odds. Clearly, this connection depends on the agent's values. If an agent values roast ducks more than boiled turnips, her belief that a coin is unbiased will not sanction as fair a bet in which she risks a roast duck for a chance of gaining a boiled turnip on the next coin flip.

Putting these two ideas together, let us ask how a simple agent should evaluate possible betting odds. It seems to me that if a simple agent has a degree of belief of, e.g.,  $2/3$  that P, and if he's offered a bet in which he'd win \$1 if P is true and lose \$2 if P is false, he should evaluate the bet as fair. The same would hold of a bet that would cost him \$100 if P but would pay him \$200 if not-P. I take this as a very plausible normative judgement: any agent who values money positively and linearly, and who cares about nothing else, *should* evaluate bets in this way. This suggests the following reformulation of the DBA's second premise:

2'. A simple agent's degrees of belief sanction as fair monetary bets at odds matching his degrees of belief.

In thinking about the third premise of the de pragmatized DBA, the notion of a set of betting odds being "defective" needs explanation. It seems plausible in general to say that a set of concurrently offered betting odds is defective if it makes possible a set of bets whose payoffs

would be logically guaranteed to leave the agent worse off, according to the agent's own values.

Applying this general idea to the case of the simple agent give us the following variant of premise 3 above:

- 3'. For a simple agent, a set of concurrently offered betting odds that allows construction of a set of bets whose payoffs are logically guaranteed to leave him monetarily worse off is defective.

Since the simple agent cares solely and positively about money, a set of bets that is guaranteed to cost him money is guaranteed to leave him worse off, by his own lights. A set of concurrently offered betting odds which allows the construction of such a set of bets is, I think, defective in a fairly straightforward intuitive sense.

Premise 4 is more problematic.<sup>18</sup> It asserts that, for any agent, beliefs which sanction a defective *set* of betting odds are themselves defective. There is, I think, something undeniably attractive about this idea, but as it stands, it is too simple a connection to hold in general. The reason for this stems from an obvious fact about values: in general, the values of things are dependent on the agent's circumstances. Right now, I would put quite a high value on obtaining a roast duck, but if I already had a roast duck in front of me, obtaining another would be much less attractive. This phenomenon applies to the prices and payoffs of bets as much as to anything else; thus there can be what one might call *value-interference* effects between bets. The price or payoff of one bet may be such that it would alter the value of the price or payoff of a second bet. And this may happen in a way that makes the second unfair--even though it would have been

perfectly fair, absent the first bet. Because of such value-interference effects, it is not in general true that there's something wrong with an agent whose beliefs individually sanction bets that, if all taken together, would leave the agent worse off.

Of course, insofar as value-interference effects are absent, the costs or payoffs from one bet will not affect the value of costs or payoffs from another. And if the values which make a bet worth taking are not affected by a given factor, then the acceptability of the bet should not depend on that factor's presence or absence. Thus in circumstances where value-interference does not occur, bets that are individually acceptable should, intuitively, be acceptable in combination. In such cases, it seems to me that a principle like premise 4 is quite plausible.

Fortunately, we have before us already a model situation in which value-interference is absent: the case of the simple agent. The simple agent values money linearly; the millionth dollar is just as valuable as the first, and so the value of the costs and payoffs from one bet will not be diminished or augmented by costs or payoffs from another. Thus premise 4 may be replaced by:

4'. If a simple agent's beliefs sanction as fair each of a set of betting odds, and that set of betting odds is defective, then the agent's beliefs are rationally defective.

It is worth recalling that we are now interpreting *sanctioning as fair* as providing justification without the need for ceteris paribus conditions holding. So we need not worry about the possibility raised by Maher--that the defectiveness in the betting odds was due merely to failure of the ceteris paribus conditions, and thus not indicative of defectiveness of the

sanctioning beliefs. But we might now wonder about the question Maher raised with respect to his principle M. Consider a simple agent whose degree of belief in P was  $1/3$ , but whose degree of belief in not-P was also  $1/3$ . Such an agent's beliefs sanction a set of betting odds which are defective in our sense. Thus someone might challenge 4' by claiming that, nevertheless, the agent's beliefs are not rationally defective. And we could not respond by assuming that the agent's beliefs should obey the additive law for probabilities; that would beg the question.

This sort of example does not, however, show that the plausibility of 4' is somehow intuitively dependent on the assumption of additivity. True, the intuitive appeal of 4' is based at least in part on some general intuition about beliefs fitting together. In Maher's sort of case, the defect in betting odds lies in the way the odds sanctioned by different beliefs fit together. The intuition behind 4' is that, absent value-interference effects, this failure of the odds to fit together reflects a lack of fit between the beliefs that sanctioned those odds. But saying that the plausibility of 4' depends on a general intuition about beliefs fitting together does not mean that 4' depends intuitively on a prior acceptance of the additive law for probabilities. Premise 4' would, I think, appeal intuitively to people who were quite agnostic on the question of whether, when A and B are mutually exclusive, the probability of  $(A \vee B)$  was equal to the sum of the probability of A and the probability of B. The idea that beliefs should fit together *in that particular way* need not be embraced, or even understood, in order for a general fitting-together requirement along the lines embodied in premise 4' to be plausible. Thus while 4' is certainly contestable (as are 2' and 3'), it seems to me intuitively plausible quite independently of the conclusion the DBA is aiming to reach.

The conclusion of the revised DBA must, of course, be restricted to simple agents:

C'. If a simple agent's degrees of belief violate the probability axioms, they are rationally defective.

The resulting argument preserves the distinctive feature of the de pragmatized DBA. It does not aim at showing that probabilistically incoherent degrees of belief are unwise to harbor for practical reasons. Nor does it identify, or define, degrees of belief by the ideally associated bet-evaluations. It aims to show that probabilistically incoherent beliefs are rationally defective by showing that, in certain particularly revealing circumstances, they would provide *justification* for bets that are defective in a particularly obvious way. As in the de-metaphysicized RTA, the connection between beliefs and evaluative attitudes is normative rather than causal or constitutive.

Unlike the RTA, this DBA has its scope restricted to simple agents. And this fact gives rise to a potentially troubling question: does the restricted scope of the DBA deprive it of its interest? After all, it is clear that there are not, and have never been, any simple agents. What is the point, then, of showing that simple agents' beliefs ought to be probabilistically coherent?<sup>19</sup>

The answer to this question is that while the values of simple agents are peculiarly simple, the point of the DBA is not dependent on this peculiarity. The point of the DBA is to support the claim that a probabilistically incoherent set of beliefs is rationally defective. The DBA illustrates and illuminates the defect by setting the incoherent beliefs in the context of an agent with simple values. This context allows us to see a clear intuitive connection between the agent's beliefs and certain monetary betting odds: given these simple values, the beliefs provide justification for evaluating certain monetary odds as fair. Moreover, the context is free of the

problem of value-interference, and thus constitutes a situation in which bets that are individually fair should be fair when taken together. When, in such a context, a set of odds sanctioned as fair turns out to allow bets that taken together are clearly unfair to the agent--which guarantee his loss--we are given reason to believe that there was something wrong with the beliefs that sanctioned those odds.

It is important to see how the simple agent cases are being used here. If the basic problem diagnosed in these cases were that the simple agent's preferences would get her in trouble, or even that the simple agent's preferences were themselves inconsistent, then one might well ask "Why is the correct conclusion that the degrees of belief are irrational per se, rather than that it is irrational to have incoherent beliefs if you are a simple agent?"<sup>20</sup> For if the basic defect were in the simple agent's preferences, then it would be unclear why we should think that the problem would generalize to agents with very different preference structures. But the basic defect diagnosed in the simple agent is not a preference-defect. In severing the definitional or metaphysical ties between belief and preferences, the depragmatized DBA frees us from seeing the basic problem with incoherent beliefs as a pragmatic one, in any sense.<sup>21</sup> The simple agent's problematic preferences are functioning here merely as a diagnostic device, a device that discloses a purely epistemic defect.

Thus the lesson of the depragmatized DBA is not restricted to simple agents. Nor is it restricted to agents who actually have the preferences sanctioned by their beliefs. (In fact, the defect that, in simple agents, is illuminated by Dutch book vulnerability may even occur in agents in whom no evaluations, and hence no evaluation inconsistencies, are present.) The power of the thought experiment depends on its being plausible that the epistemic defect we see so clearly

when incoherent beliefs are placed in the value-context of the simple agent is also present in agents whose values are more complex. To me, this is quite plausible. There is no reason to think that the defect is somehow an artefact of the imagined agent's unusually simple value structure. So although an equally clear thought-experiment that didn't have to posit so simple an agent might have been more persuasive, the simple-agent-based example used in the DBA above seems to me to provide powerful intuitive support for probabilism.<sup>22</sup>

## **6. Conclusion: Preferences and Probabilism**

Both the RTA and the DBA attempt to support Probabilism by exploiting connections between an agent's degrees of belief and her preferences. Both arguments have traditionally been tied to unsupportable assumptions which try to secure these connections by definitional or metaphysical means. But in each case, the argument's insights can be prised apart from the unsupportable assumptions by seeing the belief-preference connections as straightforwardly normative rather than metaphysical.

I would suggest that the best way of looking at both arguments is as using these connections between beliefs and preferences purely diagnostically. In neither case should we see the argument as showing that the defect in incoherent beliefs really lies in the affected agent's preferences. Nor should we even see the problem as consisting in the beliefs' failure to *accord* with rational preferences. Beliefs are, after all, more than just a basis of action. The defect inherent in beliefs which violate probabilism should, I think, be seen as primarily epistemic rather than pragmatic. The epistemic defect shows itself in pragmatic ways, for a fairly simple reason. The normative principles governing preferences must of course take account of the

agent's information about how the world is. When the agent's beliefs--which represent that information--are intrinsically defective, the preferences informed by those defective beliefs show themselves defective too. But in both cases, the preference-defects are symptomatic, not constitutive, of the purely epistemic ones.<sup>23</sup>

Though the arguments are similar, there are also interesting differences between them. The RTA's Informed Preference principle is simpler than the betting-odds sanctioning principle in the DBA's second premise. The RTA also applies directly to any rational agent. But the RTA depends on some fairly refined claims about conditions on rational preferences, claims that some have found implausible. The DBA, though it applies directly only to simple agents, does not require taking the RTA's preference-consistency principles as premises.

I suspect that different people will quite reasonably be moved to different degrees by these two arguments. Thus although I would reject Maher's suggestion that the DBA should be abandoned in favor of the RTA, I don't see much point in trying to form very precise judgments about the arguments' relative merits. Neither one comes close to being a knockdown argument for Probabilism, and non-probabilists will find contestable assumptions in both. But each one, I think, provides probabilism with interesting and non-question-begging intuitive support. And that may be the best one can hope for, in thinking about our most basic principles of rationality.

## REFERENCES

- Armendt, Brad (1993), "Dutch Books, Additivity, and Utility Theory", *Philosophical Topics* XXI, 1: 1-20.
- Chan, Sin yee (forthcoming), "The Standing Emotions", *Southern Journal of Philosophy*.
- Christensen, David (1996), "Dutch-Book Arguments Depragmatized: Epistemic Consistency for Partial Believers", *Journal of Philosophy* XCIII, 9: 450-479.
- Eells, Ellery (1982), *Rational Decision and Causality* (New York: Cambridge).
- Hellman, Geoffrey (1997), "Bayes and Beyond", *Philosophy of Science* 64, 2: 191-221.
- Howson, Colin and Peter Urbach (1993), *Scientific Reasoning: The Bayesian Approach*, 2nd. ed. (Chicago: Open Court).
- Joyce, James M. (1998), "A Nonpragmatic Vindication of Probabilism", *Philosophy of Science* 65, 4: 575-603.
- Kaplan, Mark (1996), *Decision Theory as Philosophy* (New York: Cambridge).
- Maher, Patrick (1993), *Betting on Theories* (New York: Cambridge).
- (1997), "Depragmatized Dutch Book Arguments", *Philosophy of Science* 64, 2: 291-305.
- Pereboom, Derk (1991), "Why a Scientific Realist Cannot Be a Functionalist", *Synthese* 88: 341-358.
- Ramsey, Frank P. (1926), "Truth and Probability", in his *The Foundations of Mathematics* (Totowa, NJ: Littlefield, Adams, 1965).
- Savage, Leonard J. (1954), *The Foundations of Statistics* (New York: John Wiley & Sons).
- Vineberg, Susan (1998), "Can the Dutch Book Argument Really be Depragmatized?" APA Central Division.

Zynda, Lyle (2000), "Representation Theorems and Realism about Degrees of Belief",  
*Philosophy of Science* 67: 45-69.

## FOOTNOTES

1. “Relatively” unique because, e.g., different choices of a zero point or unit for a utility scale might work equally well. Different representation theorems achieve different sorts of relative uniqueness. For present purposes, I’ll put aside worries about the way particular versions of the RTA deal with failure of absolute uniqueness. Since the issues raised below would arise even if absolute uniqueness were achieved, I’ll write as if the theorems achieved true uniqueness.
2. Patrick Maher (1993) provides very nice explanations of--and defenses against--these objections.
3. Lyle Zynda (2000) focuses on this aspect of the RTA; Zynda calls it “The Reality Condition”. My overall sketch of the RTA is very similar to Zynda’s, though my conclusions diverge quite widely from his.
4. See Ramsey 1926, 174. This definition assumes utilities have already been measured. Ramsey later gives a more sophisticated final definition which is freed of this assumption, but it is cast in more technical terms which make its philosophical motivation less clear.
5. Zynda’s  $B'$  is a linear transformation of  $B$ ; the non-standard valuation function is tailored to compensate for this transformation. See Zynda 2000, 8ff.
6. For an argument showing that functionalist accounts of mental states are fundamentally incompatible with robust scientific realism, see Derk Pereboom’s (1991).
7. Sin yee Chan (forthcoming) makes a parallel point about emotional states.
8. Brad Armendt (1993) notes that in both the Dutch Book Argument and the RTA, the

connections between beliefs and preferences may be challenged. But he holds that the move of defining beliefs in terms of preferences is inessential. The RTA's assumption about the belief-preference connection applies in "uncomplicated cases where EU is most appropriate," (16) and the Dutch Book Argument's betting scenarios provide a helpful illustration of such a situation.

This point of Armendt's seems correct. But acknowledging that the belief-preference connection actually holds only in certain cases threatens to undermine the RTA. We are left needing a reason for thinking that the situations in which the belief-preference connection does hold are normatively privileged. Otherwise, it is hard to see why a result that applies to these cases--that preference-consistency requires probabilistic consistency--would have any general normative significance. The next section attempts to provide just such a reason.

9. Maher 1993, 192. The definition is premised on the agent's preferences satisfying certain conditions.

10. This is, of course, an informal statement. Like Maher's informal definition above, it must be understood as applying only when certain conditions are met.

11. A principle much like Informed Preference is endorsed by Kaplan, in the course of giving decision-theoretic argument for a weakened version of Probabilism which Kaplan terms "Modest Probabilism":

... you should want to conform to the following principle.

**Confidence.** For any hypotheses P and Q, you are more confident that P than you are that

Q *if and only if* you prefer (\$1 if P, \$0 if ~P) to (\$1 if Q, \$0 if ~Q). (1996, 8)

Kaplan presents Confidence not as a definition, but as a principle to which we are committed (under suitable conditions) by reason.

Kaplan's book is not primarily concerned with the issues of this paper; he is concerned to present an alternative to the Savage-style RTA which is much simpler to grasp, and which yields a weaker constraint on degrees of belief, a constraint which avoids certain consequences of Probabilism Kaplan finds implausible. But while Kaplan does not discuss his departure from Savage's definitional approach to the connection between preferences and degrees of belief, his argument for Modest Probabilism exemplifies the general approach to RTA-type arguments advocated here.

12. This approach to the RTA thus answers the question posed above (fn. 8): how would a result that held in only special situations support a general normative requirement? On the approach advocated here, since the posited preference-belief connection is normative rather than causal or constitutive, we need not suppose that it ever holds exactly, even in uncomplicated cases.

13. This is argued in greater detail in Christensen 1996.

14. See Howson and Urbach 1993; Christensen 1996; and Hellman 1997 for non-pragmatic versions of the DBA. See Maher 1997 for criticisms.

15. "Matching" here is understood in the natural way: if one's degree of belief in proposition P is  $r$ , the matching odds would be  $r:(1-r)$ . Thus if my degree of belief in P is  $3/4$ , a bet I'd win if P were true, and in which I put up my 75¢ to my opponent's 25¢, would be at matching odds, as would a bet in which I put up \$3 to my opponent's \$1.

16. Maher acknowledges that he has "taken some liberties with what [Christensen] actually wrote," but claims to be giving a "maximally charitable formulation." (1997, 301)

17. Susan Vineberg (1998) also notes the vagueness of sanctioning as understood above.

18. The problem described below is essentially the one developed in Maher 1993, 96, though he

describes it differently. Thanks to Maher for pointing out to me the need to deal with it.

19. This objection is similar to one considered by Kaplan, whose argument for Modest Probabilism incorporates the same assumptions about the agent's values. My answer is in part along lines roughly similar to Kaplan's; see Kaplan 1996, 43-44.

20. I owe this formulation of the question to an anonymous referee.

21. Thus I would reject view suggested in Armendt 1993, that the flaw Dutch Book vulnerability reveals in an agent's beliefs "is that they are tied to inconsistency of the kind Ramsey suggests: an *inconsistent evaluation* of a single option under different descriptions. ... The idea is that the irrationality lies in the inconsistency, when it is present...". (3, my emphasis) On the view I've been defending, the irrationality lies in the beliefs, not the evaluations.

22. This point suggests another approach to the worry expressed in the text. If the money-based bets which figured in the simple-agent DBA were replaced by bets that paid off in "utiles" instead of dollars, the argument could be rewritten without the restriction to simple agents. (The idea here is not that the bets would be paid monetarily, with amounts determined by the monetary sums' utilities relative to the agent's *pre-bet* values; as Maher (1993, 97-98) points out, this would not solve the problem. The idea is that a bet on which an agent won, e.g., 2 utiles would pay her in commodities that would be worth 2 utiles at the time of payment. Because of value-interference, a proper definition of the payoffs might have to preclude bets being paid off absolutely simultaneously, but I don't see this as presenting much of a problem.)

Nevertheless, this generalization of the DBA would decrease the intuitive transparency of its premises. Insofar as the point of the argument is to provide minimally technical intuitive support for Probabilism, the more general argument would, I suspect, actually be less powerful.

23. It is also worth noting that there are powerful arguments for Probabilism which are completely independent of considerations involving preferences. James M. Joyce (1998) develops a notion of accuracy for graded beliefs, and uses it to support Probabilism.