Putting Logic in its Place

Formal Constraints on Rational Belief

David Christensen
University of Vermont

Clarendon Press · Oxford

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford ox2 6dp
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi Kuala Lumpur
Madrid Melbourne Mexico City Nairobi New Delhi Taipei Toronto
Shanghai
With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatermala Hungary Italy Japan South Korea Poland Portugal
Singapore Switzerland Thailand Turkey Ukraine Vietnam
Published in the United States
by Oxford University Press Inc., New York

First published 2004

British Library Cataloguing in Publication Data
Data available
Library of Congress Cataloging in Publication Data
Data available
ISBN 0-19-926325-6
1 3 5 7 9 10 8 6 4 2
Typeset by Kolam Information Services Pvt. Ltd, Pondicherry, India
Printed in Great Britain
on acid-free paper by
Biddles Ltd, King's Lynn, Norfolk

5 LOGIC, GRADED BELIEF, AND PREFERENCES

5.1 Graded Beliefs and Preferences

The suggestion that logic contributes to epistemic rationality primarily through imposing conditions on graded beliefs is a relatively new one in the history of thinking about logic. But we've already seen that the traditional approach of imposing deductive cogency on binary belief, despite its undoubted intuitive naturalness, cannot capture the way logic informs epistemic rationality. Moreover, we've seen that what is perhaps the central role that logic has traditionally been thought to play in our epistemic lives—subjecting rational belief to valid argument—may be explained not by a cogency requirement on binary belief, but instead by constraints on rational degrees of belief. For these reasons, it is worth taking seriously the possibility that logic gains its epistemic purchase on us primarily through the constraints of probabilistic coherence.

The idea that probabilistic coherence is a rational requirement—let alone the primary way that logic informs epistemic rationality—has, however, met with quite a bit of resistance. Some of the resistance stems from the impression that the mathematics of probabilistic coherence involves an unacceptable level of idealization: it just seems wrong to suppose that we accord mathematically precise probabilities to the various propositions we have beliefs about—or even, many would hold, that it would be ideal to do so. I'd like to put off discussion of the role of idealization in epistemology for now, though, to concentrate on a more fundamental source of resistance

to probabilistic coherence requirements. This source of resistance stems from the fact that proponents of probabilistic coherence have traditionally cast their arguments in a way that makes their subject matter—graded belief—seem much less like binary belief than one might at first have supposed.

Let us call the view that ideally rational degrees of belief must be probabilistically coherent "probabilism." The traditional arguments for probabilism have tried to accomplish two tasks simultaneously. The first—a quasi-descriptive or stipulative task—is to provide for some way of defining and/or measuring graded beliefs. This has seemed necessary in part because our natural way of thinking and talking about beliefs is binary; graded beliefs seem in a way more like "theoretical" entities than like common-sense objects of our everyday epistemic experience. The second task the traditional arguments have sought to accomplish is a normative one: to show that graded beliefs, so defined, should be probabilistically coherent. Both of these tasks have been accomplished by tying graded beliefs to something that is not obviously within the epistemic realm: preferences. Degrees of belief are *defined* in terms of preferences, and then intuitively rational conditions on preferences are shown to impose probabilistic coherence on these degrees of belief.

The obvious worry occasioned by such arguments is that we've strayed from the topic of epistemic (as opposed to pragmatic) rationality. And this worry is sharpened by the fact that our natural way of isolating epistemic rationality invokes a goal of something like accurate representation of the world. This has an obvious application to binary beliefs; after all, the propositions we accept can be true or false, and accurate representation of the world can naturally be thought of in terms of believing true propositions and not believing false ones. But there is no similarly obvious sense in which, say, believing a true proposition to degree 2/3 contributes to the accuracy of the agent's representation of the world.[1]

[1] This is not to say that there is no way of capturing this idea; in fact, various proposals have been advanced for characterizing the accuracy or nearness-to-the-truth of graded beliefs. James M. Joyce (1998) has even shown, for certain attractive measures

Some advocates of pragmatic approaches to graded belief have been sanguine about the thought that defining graded beliefs in terms of preferences makes them into something quite unlike the beliefs we wonder about pre-theoretically. Richard Jeffrey, for example, endorses Ramsey's idea that the state we define in terms of an agent's preferences is the agent's "belief *qua* basis of action." Jeffrey writes:

> [I am not] disturbed by the fact that our ordinary notion of *belief* is only vestigially present in the notion of degree of belief. I am inclined to think that Ramsey sucked the marrow out of the ordinary notion, and used it to nourish a more adequate view. (Jeffrey 1970, 171–2)

It seems to me, however, that this sanguinity is misplaced. For one thing, the move of defining degrees of belief in terms of an agent's preferences (as revealed in her choice-behavior) is reminiscent of the standard operationalist strategy in philosophy of science: taking one way of measuring a theoretical quantity and treating is as a definition. Bruno de Finetti, one of the founders of the preference-based approach to graded belief, is quite straightforward about his operationalist motivations in this matter. Commenting on his definition of personal probabilities in terms of betting preferences, he writes:

> The important thing to stress is that this is in keeping with the basic requirement of a valid definition of a magnitude having meaning (from the methodological, pragmatic, and rigorous standpoints) instead of having remained at the level of verbal diarrhoea . . . (de Finetti 1977, 212)

But today, operationalism and kindred approaches to theoretical magnitudes are widely seen to be misguided. And this goes not only

of accuracy, that any set of graded beliefs that violates the probability axioms can be replaced by a probabilistically coherent set that is guaranteed to be more accurate. Joyce offers this as a clearly non-pragmatic vindication of probabilism. Unfortunately, as Maher (2002) has pointed out, there are other accuracy measures that do not support this result, and the arguments in Joyce (1998) that would rule out these measures are not fully convincing. At this point, it seems to me that the jury is still out on the prospects for providing a clear accuracy-based argument for probabilism. See Fallis (2003) for useful further discussion and references related to this topic.

for physical quantities such as length and temperature, but also for psychological concepts such as pain, intelligence, and belief.

Unfortunately, the traditional arguments supporting probabilistic coherence as a norm for graded belief make explicit use of definitional connections between beliefs and preferences. This raises the question: can we support probabilistic coherence as a norm for rational degrees of confidence, without making graded beliefs into something that they are not? Clearly, the answer will depend on the way in which the preference-based definitions figure in the relevant arguments. In this chapter, I'll look more closely at the two main strands of preference-based argument that have been used to support probabilistic coherence requirements: Dutch Book Arguments, and arguments based on Representation Theorems.

5.2 Dutch Book Arguments and Pragmatic Consistency

"Dutch Book" Arguments (DBAs) are the best-known way of supporting the claim that one's graded beliefs should be probabilistically consistent. The arguments' central premise posits a close connection between an agent's graded beliefs and her betting behavior: the agent's degree of belief in a proposition P is assumed to be measurable by her preferences as they are expressed in her willingness to accept bets on P. Though the details of the betting arrangements in various DBAs differ somewhat, they all involve the agent accepting bets at the odds dictated in the intuitively natural way by her degrees of belief. For example, on the basis of my 0.75 degree of belief in my having sausages for dinner tonight, I would be willing to accept a bet at 3:1 odds that I will eat sausages, and equally willing to accept a bet at 1:3 odds that I will not have sausages.[2]

[2] In general, an agent's degree of belief in a proposition P is taken to be given by her *betting quotient q*. An agent's betting quotient for P is q if she would be indifferent between taking either side of a bet on P at odds of q: $(1-q)$. This general pattern fits

Of course, the agent's degrees of belief so measured may not obey the laws of probability—there may be no probability function that matches the agent's degree of belief function for every proposition about which the agent has a degree of belief. That will be the case if, for example, my degree of belief in P is greater than my degree of belief in (P $\vee$ Q). The DBAs show that in all such cases the agent will be willing to accept a set of bets on which she is guaranteed to lose money overall—no matter what the truth is about the matters on which the bets are made.[3]

The vulnerability to this sort of guaranteed loss is taken to indicate irrationality, and thus the lesson of the DBAs is supposed to be that ideally rational degrees of belief must conform to the probability calculus.

Now the key argumentative move—from the hypothetical vulnerability to guaranteed betting losses to constraints on rational belief—has seemed to many a *non-sequitur*. It has been pointed out, for example, that there are no clever bookies who know my degrees of belief and can compel me to wager with them. Clearly, Dutch Book vulnerability is not a real practical liability. Moreover, even if probabilistically incoherent agents were subject to real practical difficulties, it would not obviously follow that their beliefs were defective from the *epistemic* standpoint (as opposed to being merely pragmatically unwise).[4]

Defenders of the arguments have replied that the point of DBAs is not to indicate a practical problem. Rather, Dutch Book vulnerability indicates a kind of *inconsistency*. It is the inconsistency, not

the example in the text; 3 : 1 odds are the same as 0.75 : 0.25 odds. Thus, the agent is taken to have a degree of belief function that assigns a number from 0 to 1—corresponding to the agent's betting quotient—to each proposition about which she has beliefs.

[3] I will not rehearse the mathematical details of the proof that violations of the probability calculus entail Dutch Book vulnerability. The classic presentations are in Ramsey (1926) and de Finetti (1937). Prominent contemporary presentations include Skyrms (1975), Horwich (1982), and Howson and Urbach (1989).

[4] I have mentioned some representative criticisms, but there are more. For useful discussion and references to the literature, see Eells (1982), Maher (1993), Kaplan (1996), and Armendt (1993).

the likely prospect of monetary loss, that is the problem. This is an especially appealing kind of answer if one would like to see the probabilistic laws as, in Ramsey's words, "an extension to partial beliefs of formal logic, the logic of consistency" (1926, 41).

This general line of thought has considerable appeal; for although the DBAs have seemed persuasive to many, it is hard to see how they would have any force at all if their point were to reveal some practical disadvantage that came from violating the rules of probability. The suggested approach avoids seeing DBAs as crudely prudential. Rather than taking probabilistic coherence as an economically useful defense against being impoverished by transactions with improbably clever bookies, it sees probabilistic incoherence as involving structural defects in the agent's cognitive system.

On close inspection, however, the "inconsistency" that Dutch Book defenders are talking about is less parallel to standard deductive inconsistency than one might have hoped. The classic formulators of DBAs, Ramsey and de Finetti, did not simply make the assumption that certain degrees of belief could naturally be expected to lead to certain betting preferences: rather, they *defined* degrees of belief in terms of betting preferences. If degrees of belief are, at bottom, defined in terms of preferences, the inconsistency involved in having probabilistically incoherent degrees of belief turns out to be an *inconsistency of preference*. Thus, Ramsey writes:

> Any definite set of degrees of belief which broke [the laws of probability] would be inconsistent in the sense that it violated the laws of preferences between options, such as that preferability is a transitive asymmetrical relation . . . (Ramsey 1926, 41)

More recently, Brian Skyrms put the point this way:

> Ramsey and De Finetti have provided a way in which the fundamental laws of probability can be viewed as pragmatic consistency conditions: conditions for the consistent evaluation of betting arrangements no matter how described. (Skyrms 1980, 120)

Clearly, this sort of consistency of *preference* is not the sort of consistency one would initially expect to come from generalizing the notion of deductive consistency to degrees of belief.[5] Let us call this the "pragmatic consistency" interpretation of the DBAs.

It seems to me that there is something very unsatisfying about the DBAs understood in this way. How plausible is it, after all, that the intellectual defect exemplified by an agent's being more confident in P than in $(P \vee Q)$ is, at bottom, a defect in that agent's *preferences*? It is only plausible to the extent that we take seriously and literally the proposal that particular degrees of belief are defined by particular preferences—or, perhaps more precisely, that degrees of belief reduce to (or necessarily include) certain preferences. Now this proposal may not represent the considered judgment of all defenders of the pragmatic consistency interpretation of DBAs, some of whom also talk of the relation between beliefs and preferences in more ordinary causal terms. But the important point is this: for inconsistency in beliefs to *be* inconsistency of preference, certain preferences must *be* (at least a necessary part of) the beliefs.[6]

This seems at best a very dubious metaphysical view. It is true that one need not be an old-fashioned operationalist to hold that there is some constitutive connection between beliefs and preferences. Certain more sophisticated contemporary approaches to philosophy of mind—various versions of functionalism—still posit a deep metaphysical connection between beliefs and their

[5] Indeed, one might well doubt that "inconsistent" is the best word to use in describing preferences that violate transitivity, for example. Since this terminology has become established, though, I will for convenience continue to use the term in a broad and informal way.

[6] Some presentations of Dutch Book results simply assume that agents' betting preferences correspond to their degrees of belief (see Skyrms 1990). For explicit identifications/reductions/definitions of graded beliefs in terms of betting preferences, see de Finetti (1977); Ramsey (1926, 36); and Jeffrey (1965*b*, 1991). Howson and Franklin (1994) and Howson and Urbach (1989) identify an agent's degrees of belief with the betting quotients she takes to be fair (though they don't take these as entailing any willingness to bet). For interesting expressions of looser connections between beliefs and preferences, see Ramsey (1926, 30–35) and Armendt (1993, 7).

typical causes and effects (including other mental states such as preferences and, of course, other beliefs). But the causal interconnections that are said to define or constitute a belief are quite complex. They never simply require that a certain belief state necessarily give rise to certain preferences. This brings up a revealing tension in the pragmatic consistency approach to DBAs.

Suppose that beliefs are individuated—with respect to degree as well as content—by their causal roles. Then it might be that my high degree of belief that P is in a sense partially constituted by my belief's connections to, e.g., the fact that I would pay a lot of money for a ticket that is good for a big prize conditional on P's truth. But if beliefs are individuated by their causal roles, they will be individuated not only by their connections to particular betting preferences, but also by their connections to other psychological states—in particular, to other beliefs. If that is true, however, then my strong belief in P would also be partially constituted by its connections to my strong belief that $(P \vee Q)$.

This is where the tension comes in. The entire interest of taking the probability calculus as a normative constraint on belief depends on countenancing the real possibility that the second sort of connection might fail to measure up to probabilistic correctness: I might strongly believe P but not have a sufficiently strong belief in $(P \vee Q)$. But once we countenance this possibility, do we have any justification for refusing to countenance the following possibility: that I strongly believe P but do not have a sufficiently strong preference for receiving a prize conditional on P's truth? It seems to me that we do not. We have been given no reason to think that having certain appropriate betting preferences is somehow more essential to having a given belief than having appropriate other beliefs is. Thus, the interest of taking the probability calculus as a normative constraint on beliefs is predicated on countenancing the very sort of possibility—failure of a given belief to give rise to the appropriate other psychological states—that undermines the reductionism at the heart of the pragmatic consistency interpretation. An acceptable interpretation of the DBAs must acknowledge that

partial beliefs may, and undoubtedly do, sometimes fail to give rise to the preferences with which they are ideally associated.[7]

It is important to note that these considerations do not undermine the view that theorizing about degrees of belief requires that we have some fairly reliable method—or better, methods—for measuring them. Nor do they undermine the view that eliciting preferences in certain ways can provide very reliable measurements of beliefs. But they do, I think, serve to break the definitional link on which the pragmatic consistency version of DBAs depends: they undermine the oversimplified metaphysical *reduction* of beliefs to particular betting preferences.

Rejecting this sort of reduction has an important consequence for the interpretation of DBAs. The arguments' force depends on seeing Dutch Book vulnerability not as a practical liability, but rather as an indication of an underlying inconsistency. Once we have clearly distinguished degrees of belief from the preferences to which they ideally give rise, we see that inconsistency in degrees of belief cannot simply be inconsistency of preferences. If the DBAs are to support taking the laws of probability as normative constraints on degrees of belief, then Dutch Book vulnerability must indicate something deeper than—or at least not identical to—the agent's valuing betting arrangements inconsistently.

Now one possibility here is to defend what might be called a "mitigated pragmatic consistency interpretation." One might

[7] A similar problem applies to a somewhat different consistency-based interpretation of the Dutch Book results given by Colin Howson and Alan Franklin (1994) (a related approach is given in Howson and Urbach 1989, ch. 3). They argue that an agent who has a certain degree of belief makes an implicit claim that certain betting odds are fair. On this assumption, an agent with incoherent degrees of belief is believing a pair of deductively inconsistent claims about fair betting odds. Howson and Franklin conclude that the probability axioms "are no more than (deductive) logic" (p. 457). But just as a particular degree of belief may, or may not, give rise to the ideally correlated betting preferences, a given degree of belief may or may not give rise to the correlated belief about fair betting odds. Even if we take degrees of belief to justify the correlated beliefs about fair bets, a degree of belief and a belief about betting are not the same thing. Once we see the possibility of this metaphysical connection being broken, it seems a mistake to hold that the real problem with incoherent degrees of belief lies in the claims about bets with which they are ideally correlated.

acknowledge that there is no necessary metaphysical connection between degrees of belief and bet evaluations. But one might hold that there are causal connections that hold in certain ideal situations, and that in those ideal situations violations of the probability calculus are always accompanied by preference inconsistencies. One might then point out, quite rightly, that finding norms for idealized situations is a standard and reasonable way of shedding light on normative aspects of situations where the idealizations do not hold.

But this, too, is unsatisfying. If the ultimate problem with incoherent degrees of belief lay just in their leading to preference inconsistencies, then there would seem to be no problem at all with incoherent beliefs in those non-ideal cases where they did not happen to give rise to inconsistent preferences. This seems quite unintuitive: there is something wrong with the beliefs of an agent who thinks P more likely to be true than $(P \vee Q)$, even if the psychological mechanisms that would ideally lead from these beliefs to the correlated preferences are for some reason disrupted. And it would involve quite a strain to suggest that the ultimate problem with such an agent's beliefs lay simply in the fact that these preferences would, in ideal circumstances, give rise to inconsistent preferences: there seems to be something wrong with thinking that P is more likely to be true than $(P \vee Q)$, quite apart from any effect this opinion might have on the agent's practical choices or preferences. Ultimately, to locate the problem with probabilistically incoherent degrees of belief in the believer's preferences, actual or counterfactual, is to mislocate the problem.

For these reasons, I think we must reject the pragmatic consistency interpretations of the DBAs. Should we, then, give up on the DBAs themselves? Perhaps not. It seems to me that the arguments have enough initial intuitive power that it would be disappointing, and even a bit surprising, if they turned out to be as thoroughly misguided as their pragmatic interpretations seem to make them. In the next section, I'll explore the possibility of making sense of the DBAs in a fully non-pragmatic way.

5.3 Dutch Book Arguments Depragmatized

Although the relationship between degrees of belief and the evaluations of betting odds to which they often give rise may not be as close as some have thought, there is, I think, a relationship that goes well beyond the rough psychological causal pattern. Putting aside any behaviorist or functionalist accounts of partial belief, it is initially quite plausible that, in ordinary circumstances, a degree of belief in P of, e.g., 2/3 that of certainty *sanctions as fair*—in one relatively pre-theoretic, intuitive sense—a monetary bet on P at 2 : 1 odds. Intuitively, the agent's level of confidence in P's truth provides *justification* for the agent's bet evaluation—it is part of what makes the bet evaluation a reasonable one.

Let us try to make the intuitive idea a bit more precise. To begin with, let us say that an agent's degree of belief in a certain proposition sanctions a bet as fair if it provides justification for evaluating the bet as fair—i.e. for being indifferent to taking either side of the bet. Clearly, this connection depends in any given case 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. If she values the two equally, however, and values nothing else relevant in the context, she should be indifferent to taking either side of a bet, at one duck to one turnip, on the next flip of a coin she believes to be fair.

How does this general idea connect with monetary betting odds? It cannot, of course, be that any agent with 2/3 degree of belief in P is rationally obliged to agree to putting up $200 to the bookmaker's $100 on a bet the agent wins if P is true. Various factors may make it irrational for her to accept such bets. The value of money may be non-linear for her, so that, e.g., the 200th dollar would be worth less than the 17th. Or she may have non-monetary values—such as risk aversion—which affect the values she attaches to making the monetary bets. So, in general, we cannot correlate a person's degree of

belief in P with the monetary odds at which it is reasonable for her to bet on P.

In order to sidestep these issues, let us concentrate for the time being on agents with value structures so simple that such considerations do not arise. Let us consider an agent who values money positively, in a linear way, so that the 200th dollar is worth exactly the same as the 17th. And let us suppose that he does not value anything else at all, positively or negatively. I'll call this sort of being a *simple agent.* For a simple agent, there does seem to be a clear relation between degrees of belief and the monetary odds at which it is reasonable for him to bet. If a simple agent has a degree of belief of, e.g., 2/3 that P, and if he is offered a bet in which he will 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 is true but would pay him $200 if P is false. I take these as very plausible normative judgments: any agent who values money positively and linearly, and who cares about nothing else, *should* evaluate bets in this way. This suggests the following principle relating a simple agent's degrees of belief to the bet evaluations it is reasonable for him to make.

> *Sanctioning.* A simple agent's degrees of belief sanction as fair monetary bets at odds matching his degrees of belief.[8]

Degrees of belief may in this way *sanction* certain bets as fair, even if the degrees of belief do not *consist in* propensities to bet, or even to evaluate bets, in the sanctioned way. The connection is neither causal nor definitional: it is purely normative.

Now one might wonder whether this normative claim begs the present question. After all, the matching between beliefs and betting odds is the same one that emerges from expected utility theory,

[8] "Matching" here is understood in the natural way, corresponding to the betting quotients mentioned in fn. 1 above. Thus, if one's degree of belief in proposition P is r, the matching odds would be $\$r$:$\$(1-r)$. 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.

which already presupposes a probabilistic consistency requirement. But the intuitive normative connection between degrees of belief and bets need not derive from an understanding of expected utility theory; a person might see the intuitive relationship between bets and degrees of belief even if she could not begin to describe even roughly how the probability of P, Q, (P & Q), and (P ∨ Q) should in general relate to one another. Of course, there may be a sense in which our intuitions on these topics are all interrelated, and spring from some inchoate understanding of certain principles of belief and decision. But that seems unobjectionable; indeed, it is typical of situations in which we support a general formal reasoning theory by showing that it coheres with our more specific intuitions.[9]

Given this normative connection between an agent's degrees of belief and betting preferences, the rest of the DBA can be constructed in a fairly standard way. We may say that if a set of bets is logically guaranteed to leave an agent worse off, by his own lights, then there is something rationally defective about that set of bets. This general intuition may easily be applied to a simple agent in a straightforward way: 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. This yields the following principle.

> *Bet Defectiveness.* For a simple agent, a set of bets that is logically guaranteed to leave him monetarily worse off is rationally defective.

[9] It is also worth noting that even the "mitigated pragmatic consistency" interpretation of the DBA discussed above must presuppose a basic normative connection between degrees of belief and bet evaluations. On this view, degrees of belief lead causally to the correlated betting preferences in ideal circumstances. But one might ask: which circumstances are "ideal"? Why single out those circumstances in which degrees of belief lead to exactly the preferences that expected utility theory would dictate? The answer, it seems to me, is that we are intuitively committed to a certain normative relation between degrees of belief and preferences. Circumstances are "ideal" when, and because, this intuitively plausible relation obtains. If this answer is right, then what is perhaps the most controversial assumption in the non-pragmatic interpretation of Dutch Books given in the text also figures in the "mitigated pragmatic consistency" interpretation.

We now need a principle that connects the rational defectiveness in a set of bets to a rational defect in the degrees of belief that sanction those bets. But it is not generally true that, for any agent, a set of beliefs that sanctions each of a defective set of bets is itself defective. 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 is something wrong with an agent whose beliefs individually sanction bets that, if all taken together, would leave her 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 that 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.

Fortunately, we already have before us 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 the following principle is, I think, quite plausible.

Belief Defectiveness. If a simple agent's beliefs sanction as fair each of a set of bets, and that set of bets is rationally defective, then the agent's beliefs are rationally defective.

It is worth noting that the intuitive appeal of Belief Defectiveness does flow, at least in part, from some general intuition about beliefs fitting together. So one might worry that the principle's intuitive plausibility presupposes a commitment to probabilistic coherence. Maher, criticizing a related principle, raises the following sort of worry. Consider a simple agent whose degree of belief in P is 1/3, yet whose degree of belief in not-P is also 1/3, violating probabilistic coherence. Such an agent's beliefs would sanction a defective set of bets.[10] But suppose one were to claim that the agent's beliefs were not themselves defective. We could not reply, without begging the question, by claiming that beliefs should fit together in the manner prescribed by the laws of probability.[11]

Nevertheless, this sort of example does not show that the plausibility of Belief Defectiveness is somehow intuitively dependent on the assumption of a probabilistic coherence requirement. The defect in the set of sanctioned bets lies in the way they fit together. The intuition behind Belief Defectiveness is that, absent value interference effects, this failure of the bets to fit together reflects a lack of fit between the beliefs that sanctioned those bets. But saying that the plausibility of the principle depends on a general intuition about beliefs fitting together does not mean that it depends intuitively on a prior acceptance of probabilistic coherence in particular. Belief Defectiveness 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 ∨ 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 our principle to be plausible. Thus while Belief Defectiveness is certainly contestable, it seems to me intuitively

[10] His degree of belief in P would sanction a bet costing $2 if P is true, and paying $1 if P is false. His degree of belief in not-P would sanction a bet costing $2 if not-P is true, and paying $1 if not-P is false. The set of two bets is guaranteed to cost him $1.

[11] Maher's point is in his (1997), in the section criticizing Christensen (1996).

plausible, quite independently of the conclusion the DBA is aiming to reach.

With the three more philosophical premises in place, all that is needed for a DBA is the mathematical part.

> *Dutch Book Theorem.* 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, that will logically guarantee the agent's monetary loss.

The argument proceeds as follows. Suppose a simple agent has probabilistically incoherent degrees of belief. By the Dutch Book Theorem, there is a set of monetary bets at odds matching his degrees of belief which logically guarantee his monetary loss. By Bet Defectiveness, this set of bets is rationally defective, and by sanctioning, each member of this set of bets is sanctioned by his degrees of belief. Then, by Belief Defectiveness, his beliefs are rationally defective. Thus we arrive at the following.

> *Simple Agent Probabilism.* If a simple agent's degrees of belief violate the probability axioms, they are rationally defective.

This distinctively non-pragmatic version of the DBA allows us to see why its force does not depend on the real possibility of being duped by clever bookies. It does not aim at showing that probabilistically incoherent degrees of belief are unwise to harbor for practical reasons. Nor does it locate the problem with probabilistically incoherent beliefs in some sort of preference inconsistency. Thus it does not need to identify, or define, degrees of belief by the ideally associated bet evaluations. Instead, this DBA 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 rationally defective in a particularly obvious way. The fact that the diagnosis can be made *a priori* indicates that the defect is not one of fitting the beliefs with the way the world happens to be: it is a defect internal to the agent's belief system.

As set out above, the conclusion of the DBA has its scope restricted to simple agents. And this fact gives rise to a potentially troubling question: doesn't this deprive the argument 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?[12]

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 argument takes advantage of the fact that rational preferences for bets are informed jointly by an agent's values and an agent's representations of the world—her beliefs. In our thought-experiment, we consider how a certain set of beliefs would inform the betting preferences of an (imaginary) agent who cared only about one sort of thing, and cared about it in a very simple way (money is the traditional choice, but it's arbitrary; grains of sand would serve as well). This particularly transparent context allows us to see a clear intuitive connection between the set of beliefs and certain bets: given the simple values, the beliefs provide justification for evaluating the bets as fair. We show that, if the beliefs are incoherent, they would justify the imagined agent's preferring to take each of a set of bets that would logically guarantee his losing the only commodity he values. Given the agent's simple value structure, the problem with the set of bets cannot be that the costs or benefits of one bet affect the value of the costs or benefits of another. Rather, the problem is that there is no way the world could turn out that would make the set of bets work out well—or even neutrally—for the agent. In this sort of case, it seems to me that the overwhelmingly plausible diagnosis is that there is something intrinsically wrong with the representations of the world that justified the agent's preferences for these bets.

[12] This objection is similar to one considered by Kaplan, whose argument for a weakened version of 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, 43–4).

This is in part why it is important to be clear on the role that preferences play in the DBA. If the basic problem diagnosed in these cases were that the simple agent's preferences would get him into 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?"[13] For if the basic defect were located 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. Once the connection between beliefs and preferences is understood as normative rather than metaphysical, we can see that the simple agent's problematic preferences function in the DBA 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, results in Dutch Book vulnerability may even occur in agents in whom no bet evaluations, and hence no bet 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. I think that 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 did not involve simple agents might

[13] I owe this formulation of the question to an anonymous referee for *Philosophy of Science*.

have been more persuasive, the simple-agent-based example used in the depragmatized DBA above seems to me to provide powerful intuitive support for probabilism.[14]

5.4 Representation Theorem Arguments

If DBAs are the best-known ways of supporting probabilism, Representation Theorem Arguments (RTAs) are perhaps taken most seriously by committed probabilists.[15] RTAs approach an agent's beliefs and values in a more holistic way than do DBAs. The arguments begin by taking ideally rational preferences to be subject to certain intuitively attractive formal constraints, such as transitivity. They then proceed to demonstrate mathematically (via a Representation Theorem) that if an agent's preferences obey the formal constraints, they can be represented as resulting from a relatively unique[16] pair, consisting of a set of degrees of belief and a set of utilities, such that (1) the degrees of belief are probabilistically

[14] This point suggests another approach to the worry expressed in the text. If the monetary bets that 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, 97–8) 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, generalizing the DBA in terms of utiles would decrease the intuitive transparency of its premises. Insofar as the point of the argument is to provide intuitive support for probabilism, the more general argument would, I suspect, actually be less powerful.

[15] See e.g. Maher (1993, ch. 4.6) or Kaplan (1996, ch. 5).

[16] "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.

coherent, and (2) the preferences maximize expected utility relative to those beliefs and utilities. Thus, typical RTAs 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. The fact that an agent's preferences can be represented as resulting from some U and B does not show that U and B are that agent's actual utilities and degrees of belief. Typically, RTA proponents rely in their arguments on some principle positing a tight definitional or constitutive connection between an agent's preferences and her beliefs and utilities. The precise form of the principle making the connection may vary, and it may receive little philosophical comment, but the following sort of connection is taken to emerge from the argument.

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.[17] Instead, as with the DBA, I'll focus on the purported connection between the clearly epistemic and the pragmatic aspects of rationality, as summarized in the Representation Accuracy Principle. Suppose that an agent has preferences that 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—most importantly for our purposes—beliefs?[18]

Representation Accuracy posits that a particular connection holds 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 is, given human psychology, causally necessary for having

[17] Patrick Maher (1993) provides very nice explanations of—and defenses against—these objections.

[18] Lyle Zynda (2000) focuses on this aspect of the RTA; he 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.

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 provide a correct way of applying logic to degrees of belief—then the connection between preferences and beliefs will have to be a deeper one.

In fact, RTA proponents do posit deeper connections between preferences and beliefs. Like DBA proponents, they typically take degrees of belief (and utilities) to be in some sense *defined* by preferences. Taken unsympathetically, this suggests some sort of operationalism 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 that degrees of confidence play in psychological explanation. Clearly, we often explain behavior—especially in deliberate choice situations—by invoking degrees of confidence. 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." 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 operationalism, or to seeing the relevant definition as analytic or a priori. And the definition need not be the simple sort that figures in some presentations of the DBA, where an agent's degree of belief is defined in terms of very particular betting preferences. We needn't even see the agent's preferences as epistemically privileged, compared with her beliefs and utilities. Jeffrey writes:

In fact, I do not regard the notion of preference as epistemologically prior to the notions of probability and utility. In many cases we or the agent may be fairly clear about the probabilities the agent ascribes to certain propositions without having much idea of their preference ranking, which we thereupon deduce indirectly, in part by using probability considerations. The notions of preference, probability, and utility are intimately related; and the object of the present theory is to reveal their interconnections, not to "reduce" two of them to one of the others. (Jeffrey 1965*b*, 220–1)

The envisioned account of graded belief might thus be understood as a more holistic scientific definition, combining elements of conceptual refinement with empirical investigation. Beliefs turn out to be something like functional or dispositional properties of people, defined, along with utilities, by their causal connections to the agent's utilities, other beliefs, and 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. But part of what *constitutes* a given 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 among degrees of belief, utilities, and preferences. This idea has obvious connections to functionalist theories in mainstream philosophy of mind.

Nevertheless, this claim about the nature of beliefs cannot represent mere naked stipulation. If the definition 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. And it seems to me that there are grounds for doubting that the envisaged definition will pass this test.

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. 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.[19] 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

[19] Zynda's B′ is a linear transformation of B; the non-standard valuation function is tailored to compensate for this transformation; see (Zynda 2000, 8 ff.).

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 jibes with 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, 171–2). Ellery Eells (1982, 41–3) 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 pre-theoretic 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 pre-theoretic 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 will 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 her behavior 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 that help explain behavior.

Beliefs may also have less obvious non-behavioral effects. Every reputable drug study controls for the placebo effect. According to received wisdom, 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 (and one recent study suggests that this effect is far less prevalent than it is standardly taken to be).[20] But insofar as the placebo effect is real, it is not explained by any disposition of the patients to have preferences or make choices that maximize utility relative to a high probability of their having taken effective medicine.

[20] See Hrobjartsson and Gotzsche (2001). As might be expected, the study's conclusions are somewhat controversial. The authors conclude that there is no justification for using placebos therapeutically, but they do not recommend the elimination of placebos in clinical trials.

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, and 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 that 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.[21] If the connections between beliefs and preferences have the status of empirical regularities rather than definitions—if the connections are merely causal and not 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 is not surprising that the ironclad belief–preference connection posited in Representation Accuracy fails to be groundable in—or even to 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

[21] For an argument showing that functionalist accounts of mental states are fundamentally incompatible with a robust kind of scientific realism, see Derk Pereboom (1991).

had no preferences—that he had no beliefs about anything. After all, it is obvious from one's own case that one cares about some things much more than one cares about 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 common-sense psychology that, e.g., the strength of an agent's disposition 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 among them, completely mysterious. Why, for example, would the brute disposition to form preferences in a certain way correlate with feelings of fear?[22]

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 is 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 Kaplan notes (in arguing for a different point), "You have

[22] Sin yee Chan (1999) makes a parallel point about emotional states.

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 with 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.[23]

5.5 De-metaphysicized Representation Theorem Arguments

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

[23] Brad Armendt (1993) notes that in both the DBA 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" (1993, 16). 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.

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 parallel to the one exploited in the depragmatized DBA: to substitute a normative connection for a definitional or metaphysical one. In the RTA, we already assume that an ideally rational agent's preferences are consistent with one another in the ways presupposed in the obviously normative Preference Consistency principle. The present proposal is that, in addition, an ideally rational agent's preferences must cohere in a certain way with her beliefs. 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. As noted above, Representation Accuracy emerges from a more specific belief–preference connection made in the course of the RTA. In proving their results, Savage and Maher 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 key 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.[24] (Maher 1993, 192)

Now 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, which holds for all ideally rational agents.

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

> *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.[25]

This normative principle avoids the universal metaphysical commitments entailed by the definitional approach. 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

[24] The formal definition which cashes out this intuitive description is quite complex, and is premised on the agent's preferences satisfying certain conditions.

[25] 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.

consistent with one another. At the same time, the principle forges the preference–belief connection for all ideally rational agents, who are anyway the only ones subject to the RTA's desired conclusion.[26]

Suppose, then, that the RTA was formulated using a suitably precise version of Informed Preference. Of course, this sort of RTA would not support the principle of Representation Accuracy—but, as we have seen, this is as it should be. What would emerge from the reformulated RTA would be Representation Accuracy's normative analogue.

> *Representation Rationality.* If an *ideally rational* 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.

This principle, no less than the rejected Representation Accuracy, may be combined with Preference Consistency and a Representation Theorem to yield Probabilism.

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 metaphysic-

[26] A principle much like Informed Preference is endorsed by Kaplan, in the course of giving his 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 concerned primarily with the issues we've been concentrating on: 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 that avoids certain consequences of Probabilism which 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.

ally interpreted RTA: it cannot 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—which forges the basic belief–preference connection, and from which Representation Rationality ultimately derives—does not presuppose 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. The argument is also freed from its traditional entanglement with behaviorist definition or other fishy metaphysics. Moreover, this frankly normative approach to the RTA answers the question posed above: 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 justificatory rather than causal or constitutive, we need not suppose that it ever holds exactly, even in uncomplicated cases. Thus, it seems to me that the RTA may be de-metaphysicized successfully; once this is done, the argument can lend substantial support to Probabilism.

5.6 Preferences and Logic

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 assumptions that try to secure the belief–preference connections by definitional or metaphysical means. But the metaphysically intimate connections between beliefs and preferences that have been posited by proponents of preference-based arguments for probabilism sit uneasily with our pre-theoretic understanding of what belief is. This tension is surely part of what is expressed when Ramsey restricts his interest to "beliefs qua basis for action," or when Jeffrey acknowledges that our pre-theoretic notion of belief is "only vestigially present in the notion of degree of belief." It is understandable that many epistemologists have been reluctant to embrace arguments that treat belief as part of a "device for interpreting a person's preferences."

A related point concerns the status of logical norms for graded belief. Standard logical properties of propositions, and relations among them, may be used to constrain rational graded belief via the probability calculus. This is not an unnatural suggestion. But it is unnatural to suppose that the illogicality or lapse of epistemic rationality embodied in incoherent graded beliefs is, at bottom, a defect in the believer's (actual or counterfactual) preferences. Any argument that locates the irrationality of probabilistically incoherent graded belief in the believer's preferences invites the suspicion that it is addressed to pragmatic, not epistemic, rationality. It makes it seem that probabilism is doing something quite different from what deductive cogency conditions were supposed to do for belief on the traditional binary conception.

We've seen that the definitional or metaphysical connections traditionally posited to underpin the preference-based arguments must be discarded. Fortunately, this need not mean discarding the insights that lie at the bottom of the RTA and DBA. For in each case, the argument's insights can be prised apart from the unsupportable assumptions. In each case, the insights can be preserved by seeing the belief–preference connections as straightforwardly normative rather than metaphysical. Once this is done, we see that the arguments apply to beliefs that are no more essentially

pragmatic than binary beliefs have traditionally been thought to be.

On this interpretation, probabilism is nothing more than a way of imposing traditional logic on belief—it's just that this turns out to require that belief be seen in a more fine-grained way than it often has been. When we see belief as coming in degrees, and see logic as governing the degree to which we believe things, rather than as governing some all-or-nothing attitude of acceptance, probability theory is the overwhelmingly natural choice for applying logic to belief. The preference-based arguments supply natural support for this choice.

The best way of looking at both arguments is as using 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 that violate probabilism should 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 intrinsically defective too. But in both cases, the preference defects are symptomatic, not constitutive, of the purely epistemic ones.

Though the two preference-based arguments are similar, there are some interesting differences between them. The RTA's Informed Preference principle is simpler than the DBA's Sanctioning. 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; and 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 of these arguments, I think, provides probabilism with interesting and non-question-begging intuitive support. Each shows that probabilism fits well with (relatively) pre-theoretic intuitions about rationality. And that may be the best one can hope for, in thinking about our most basic epistemic principles.