

Partial Belief and Expert Testimony

by

Rachael Briggs

B.A., Philosophy
Syracuse University, 2003

SUBMITTED TO THE DEPARTMENT OF LINGUISTICS AND
PHILOSOPHY IN PARTIAL FULFILLMENT FOR THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN PHILOSOPHY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY


FEBRUARY 2009

© 2009 Massachusetts Institute of Technology. All rights reserved.


Signature of author: _____

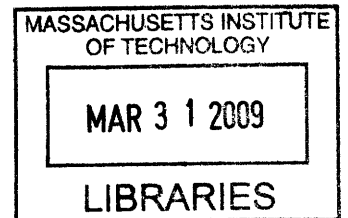
Department of Philosophy
December 2, 2008

Certified by: _____


Robert Stalnaker
Professor of Philosophy
Thesis Supervisor

Accepted by: _____


Alex Byrne
Professor of Philosophy
Chair of the Committee on Graduate Students



ARCHIVES

PARTIAL BELIEF AND EXPERT TESTIMONY

by

Rachael Briggs

Submitted to the Department of Linguistics and Philosophy
on December 2, 2008 in Partial Fulfillment for the Requirements
for the Degree of Doctor of Philosophy in Philosophy
at the Massachusetts Institute of Technology

ABSTRACT

My dissertation investigates two questions from within a partial belief framework: First, when and how should deference to experts or other information sources be qualified? Second, how closely is epistemology related to other philosophical fields, such as metaphysics, ethics, and decision theory?

Chapter 1 discusses David Lewis's "Big Bad Bug", an argument for the conclusion that the Principal Principle—the thesis that one's credence in a proposition A should equal one's expectation of A 's chance, provided one has no inadmissible information—is incompatible with Humean Supervenience—the thesis that that laws of nature, dispositions, and objective chances supervene on the distribution of categorical properties in the world (past, present, and future). I map out the logical structure of the Big Bad Bug, survey a range of possible responses to it, and argue that none of the responses are very appealing.

Chapter 2 discusses Bas van Fraassen's Reflection principle—the thesis that one's current credence in a proposition A should equal one's expected future credence in A . Van Fraassen has formulated a diachronic Dutch book argument for Reflection, but other authors cite counterexamples to Reflection that appear to undermine the credibility of diachronic Dutch books. I argue that a suitably qualified version of Reflection gets around the counterexamples. I distinguish between Dutch books that reveal incoherence—like the diachronic Dutch book for conditionalization—and Dutch books that reveal a type of problem I call self-doubt. I argue that violating Reflection is a type of self-doubt rather than a type of incoherence.

Chapter 3 argues that the halfer and thirder solutions to Adam Elga's Sleeping Beauty problem correspond to two more general approaches to *de se* information. Which approach is right depends on which approach to decision theory is right. I use Dutch books and scoring rules to argue that causal decision theorists should favor the approach that corresponds to thirding, while evidential decision theorists should favor the approach that corresponds to halving.

Thesis Supervisor: Robert Stalnaker
Title: Professor of Philosophy

CONTENTS

1. <i>The Anatomy of the Big Bad Bug</i>	7
1.1 Stage Setting	7
1.1.1 Humean Supervenience	7
1.1.2 The Principal Principle	9
1.2 The Big Bad Bug	11
1.2.1 The Big Bad Bug: a Formal Version	12
1.3 Alternatives to PP	13
1.3.1 Restricting PP	13
1.3.2 The New Principle	15
1.3.3 A More General Principal Principle?	18
1.3.4 The General Recipe	19
1.3.5 Taking Stock	20
1.4 Undermining Futures	21
1.4.1 Four Worries about Undermining	21
1.4.2 Redefining Chance	24
1.5 Conclusion	26
2. <i>Distorted Reflection</i>	27
2.1 Introduction	27
2.2 Dutch Books	27
2.3 Reflection	30
2.4 Counterexamples to Reflection	31
2.4.1 Future irrationality	31
2.4.2 Memory loss	31
2.4.3 Apparent memory loss	32
2.4.4 Future misleading evidence	32
2.4.5 Current misleading evidence	32
2.4.6 Changes in epistemic standards	32
2.5 Responses to the Counterexamples	33
2.6 Qualified Reflection	35
2.7 Distorted Reflection	36
2.8 Back to the Counterexamples	38
2.8.1 Future irrationality (LSQ)	38
2.8.2 Memory loss (spaghetti)	40
2.8.3 Apparent memory loss (Shangri La)	40
2.8.4 Future misleading evidence (Persi)	40

2.8.5	Current misleading evidence (the small-town scientist) . . .	42
2.8.6	Changes in epistemic standards (William James University)	42
2.9	Back to the Dutch Book	43
2.10	Conclusion	46
3.	<i>Putting a Value on Beauty</i>	51
3.1	Stage Setting	51
3.1.1	The <i>de dicto-de se</i> distinction	51
3.1.2	<i>De dicto</i> Confirmation Theory	53
3.1.3	Sleeping Beauty	54
3.2	Rules for Halfers and Thirder	56
3.2.1	The Halfer Rule	57
3.2.2	The Thirder Rule	57
3.2.3	Weak Indifference	58
3.3	Dutch Books	59
3.3.1	A Constraint on Dutch Books	60
3.3.2	Hitchcock's Dutch book	61
3.3.3	A New Dutch Book	64
3.3.4	Why Everyone Should Bet at Thirder Odds	64
3.3.5	Who Bets At Thirder Odds?	66
3.4	Scoring Rules	68
3.4.1	Measuring Inaccuracy	68
3.4.2	Revising the Concept of Expected Inaccuracy	70
3.4.3	Average or Total? Halfer or Thirder?	71
3.5	Stability	73
3.6	Conclusion	77
4.	<i>Conclusion</i>	79

1. THE ANATOMY OF THE BIG BAD BUG

1.1 *Stage Setting*

One fundamental schism in the philosophy of science involves the status of nomological properties: does the universe at bottom contain any laws, dispositions, or objective chances? Everyone agrees that our *talk* is loaded with nomological presuppositions, and that observation reports are theory-laden. But philosophers disagree about whether nomological talk can in principle be analyzed away. The debate can be cast as an argument about *Humean Supervenience*: roughly, the thesis that nomological facts supervene on non-nomological facts.

The debate over Humean Supervenience has a frustrating tendency to end in stalemate, with each side clinging to a different set of intuitions. Closer attention to David Lewis's 'Big Bad Bug', [Lewis, 1986a, 1994], may help advance the discussion. The Big Bad Bug is an argument to the effect that objective chances fail to supervene on non-nomological matters of fact. If sound, the Bug is disastrous for Humean Supervenience. Once a Humean has swallowed *sui generis* chances, why should she strain at *sui generis* laws and dispositions?

I argue that the Big Bad Bug is a stronger argument than many defenders of Humean Supervenience have realized. In the remainder of part 1.1, I lay the groundwork for the Bug by explaining Humean Supervenience and the Principal Principle in more detail. In part 1.2, I state the Bug informally and rephrase it as a formally valid argument. In parts 1.3 and 1.4, I map out possible objections to the Bug's premises, and argue that none of them work. I conclude that Humeans need to start taking the Bug more seriously.

1.1.1 *Humean Supervenience*

Humean Supervenience (henceforth HS) is thesis that:

HS Laws of nature, dispositions, and chances supervene on the distribution of categorical properties in the world (past, present, and future).

Two words of clarification: First, the term 'categorical' is meant to exclude nomologically loaded properties such as dispositions. Second, my version of HS differs from the thesis that Lewis calls 'Humean Supervenience' (Henceforth LHS). According to LHS, the properties that subvene the laws are local as well as categorical—they are intrinsic to point-sized regions or particles. In light of counterexamples from both quantum mechanics [Maudlin, 2007] and classical mechanics [Robinson, 1989], LHS looks untenable. But Lewis's skepticism

about irreducibly non-local properties is separable from his skepticism about irreducibly nomological properties, and I am interested chiefly in the latter.

HS plays a crucial role in motivating the ‘best system analysis’ of laws, originated by Mill [1911] and developed by Ramsey [1978] and Lewis [1986b, 1994]. The best-system analysis is a way of cashing out the idea that laws, dispositions, and chances are nothing over and above regularities in the overall distribution of categorical properties. It states that the laws of nature are the theorems of whichever deductive system best summarizes the truth (or at least most of the truth) about the past, present, and future distribution of categorical properties, while balancing the competing theoretical virtues of simplicity, strength, and fit. Dispositions and chances are then defined in terms of laws. Since the best-system analysis is well-developed, popular, and motivated by HS, it is treated in the literature as the Bug’s primary target.

Since HS is usually presented in terms of laws, while chances play the central role in the Big Bad Bug, I will briefly clarify my background assumptions about the relationship between laws and chances. For the remainder of the paper, I adopt the general framework set forth in [Lewis, 1986b]. Although Lewis treats this framework as part of the BSA, it is compatible with a wide variety of views about the metaphysics of chance, including frequentism and non-Humean propensity theories.

I assume that objective chances attach to propositions, such as the proposition that a coin lands heads on particular toss, the proposition that a coin lands heads on one particular toss and tails on the succeeding toss, or the proposition that in the whole of history, at least one horse becomes Consul to the Senate. I take chances to be defined at a world w and a time t , so that, for example, the chance that a given uranium atom decays before t varies both according to possible world it inhabits (it is higher in worlds where the atom is placed in a reactor) and according to the time (it increases when other atoms in the reactor begin to decay).¹ I use P_t to indicate the objective chance function at time t (omitting the subscript where it would be unnecessary). I use P_{tw} to rigidly designate the objective chance function at time t in world w (never omitting the subscript). I assume that “what’s past is no longer chancy” (1986, 93): for any proposition A about the history of w up to and including t , $P_{tw}(A)$ is 1 if A is true, and 0 otherwise.

I assume that at every world there is a correct theory of chance, equivalent to a set of what Lewis [1986b, 94] calls history-to-chance conditionals. These are conditionals such that:

- (1) The consequent is a proposition about chance at a certain time.
- (2) The antecedent is a proposition about history up to that time; and further, it is a complete proposition about history up to that

¹ In addition to chances defined at a world and a time, there may be chances that are defined at a world timelessly (e.g., the chances attaching to initial conditions in statistical mechanics and Bohmian mechanics). Loewer [2004] points out that Humeans as well as anti-Humeans can countenance timeless chances. As long as a Humean account is committed to the existence of time-indexed chances, adding timeless chances will not change its susceptibility to the Bug.

time, so that it either implies or else is incompatible with any other proposition about history up to that time. It fully specifies a segment, up to the given time, of some possible course of history. (3) The conditional is made from its consequent and antecedent not truth-functionally, but rather by means of a strong conditional operation of some sort.

I use T_w to indicate the correct theory of chance at w (omitting the subscript where it would be unnecessary) and P_T to indicate the objective chance function (at t , in w) according to T . I assume that the history-to-chance conditionals that make up the correct theory of chance have the status of laws.

Finally, I assume that chances are defined not just over the outcomes of individual probabilistic trials, but over arbitrary Boolean combinations of these outcomes. Thus, if there is a well-defined chance that a particular coin lands heads, and a well-defined chance that a particular die lands six, then there is a well-defined chance that the coin lands heads and the die lands six (in this case, presumably the product of the chances of the individual outcomes).

We have good reason to grant this assumption, since it explains certain common applications of probabilistic reasoning. Numerous statistical distributions (binomial, Poisson, Gaussian, etc.) are explained by the claim that the trials in question are stochastically independent—that their joint outcome probabilities are equal to the product of their individual outcome probabilities. This presupposes that the joint outcome probabilities exist. Explanations of probabilistic patterning (such as those found in statistical-mechanics), often rely on the claim that the world’s actual history is objectively likely. Again, this presupposes that world histories have objective chances.

Given Lewis’s framework of assumptions, we can show that HS contradicts PP. Although the assumptions are metaphysical, and although HS is a metaphysical thesis, the Bug springs from HS’s epistemic consequences. Thau [1994, 495] points out that if the laws and the chances supervene on the distribution of categorical properties, then a certain type of skeptical scenario is impossible—the universe cannot be globally misleading with respect to the chances. In other words, HS entails the following ‘Justified Certainty’ thesis [Thau, 1994, 495]:

JC: An observer who knew the outcomes of all the probabilistic trials in the universe would be in a position to know the chances with certainty.

JC contradicts the Principal Principle. The Bug is therefore dangerous to any theory that entails JC, even if it does not entail HS. Though I will continue to treat the best-system analysis as the Bug’s primary target, its secondary targets may include actualist frequentism, some forms of quasi-realism, and instrumentalist theories on which the chances are determined by facts about which betting rules have the best actual results.

1.1.2 The Principal Principle

The Principal Principle [Lewis, 1986b, 87]—henceforth PP—states:

Let C be any reasonable initial credence function. Let t be any time. Let x be any real number in the unit interval. Let X be the proposition that the chance, at time t , of A 's holding equals x . Let E be any proposition compatible with X that is admissible at time t . Then

$$C(A|XE) = x$$

E can be thought of as some rational agent's total evidence at t . PP says that given evidence E , the agent's credence in A conditional on the proposition that A 's chance is x should be x . If E specifies that A 's chance is x , the agent's credence in A will equal x . Otherwise, her credence in A will be a mixture of its (epistemically) possible chances, where the strength of each component is determined by the agent's credence that it is A 's chance. Roughly, then, PP states that an agent's credence in a proposition will equal her expectation of its chance, provided she begins with a rational initial credence function, updates by conditionalizing, and has only admissible information.

Two phrases require unpacking: 'initial credence function' and 'admissible'. Lewis [1986b, 88] makes two implausible claims about the initial credence function—that it can be specified prior to all evidence, and that it affords nonzero probability to every possible world. Both commitments look burdensome; the first seems to make Lewis's theory inapplicable to human beings, who don't have credence functions that are temporally or epistemically prior to their evidence, and the second requires a theory of infinitesimals, since there are (presumably) more than countably many possible worlds.

Luckily, there are ways to assuage this worry. For instance, one might claim that insofar as a person is reasonable, her belief state can be modeled as though it were the result of conditionalizing a fictitious 'reasonable initial credence function' on her total evidence. Alternatively, one might relax the concept of an initial credence function so that a credence function counts as 'initial' just in case it fulfills some criterion (e.g., being held before a certain time, by an agent who has no information which she takes to be inadmissible). Lewis does not pursue either of these options, presumably because he wants to state PP in the most general terms possible [see 1986b, 86].

For now, I will assume that there is some way of cashing out the concept of a reasonable initial credence function so that PP comes out unequivocally true. (In Chapter 4, the conclusion of my dissertation, I suggest a way of pursuing the second option in the above paragraph.) This cashing-out may involve some loss of generality, but for my purposes, having a completely general version of PP is unimportant. If HS is incompatible with any true chance-credence principle, then HS is in trouble. The weaker the principle, the worse the trouble.

What about 'admissible'? According to the earlier Lewis [1986b, 92], a proposition is admissible just in case it is "the sort of information whose impact on credence about outcomes comes entirely by way of credence about the chances of those outcomes". The later Lewis [1994], following Thau [1994, 500], notes that admissibility is always relative to the proposition A : the fact that it will

rain in London tomorrow may be admissible with respect to propositions about tomorrow's lottery outcomes, but is certainly inadmissible with respect to itself. Therefore, I will assume E is admissible with respect to A if and only if E is the sort of information whose impact on credence about A comes entirely by way of credence about $P(A)$.

Given the right readings of 'initial credence function' and 'admissible', PP captures something crucial about the relationship between chance and credence. Credence must be answerable to chance in roughly the way that PP requires. True, Lewis's definition of 'rational initial credence function' is unrealistic. Nonetheless, a property that was incapable of satisfying any version of PP—even a version that employed a more realistic definition of 'initial credence function'—would hardly deserve the name 'chance'.

1.2 The Big Bad Bug

The Big Bad Bug purports to show that HS is incompatible with PP. Since the Bug relies heavily on the concept of an undermining future, I will begin by explaining what an undermining future is. Suppose that HS is true. Then the laws are fixed by the distribution of categorical properties (past, present, and future) while the chances at a time t are fixed by the laws together with the facts about history up to t . If t is early enough, then the laws—and hence the chances—will be fixed not just by t 's past, but also by t 's future. Thus, different futures compatible with t 's past will fix different chances at t . Though only one of these futures is actual, other, counterfactual futures may have nonzero chance. If one of these counterfactual futures were to come about, the chances at t would be different—in particular, the chance of the alternative future itself might be different. Futures with this peculiar property are undermining. More formally, F is an undermining future at t in w just in case

- (a) $P_{tw}(F) > 0$, and
- (b) where H_{tw} is a complete description of w 's history up to and including t , F is incompatible with the conjunction $(H_{tw} \ \& \ P_t(F) = P_{tw}(F))$.

Without a substantive theory about what fixes the chances, it is impossible to pinpoint specific undermining futures, but Lewis Lewis [1994, 482] gives the following rough example.

[T]here is some minute present chance that far more tritium atoms will exist in the future than have existed hitherto, and each one of them will decay in only a few minutes. If this unlikely future came to pass, presumably it would complete a chancemaking pattern on which the half-life of tritium would be very much less than the actual 12.26 years. Certainly that's so under simple frequentism, and most likely it's so under the best-system analysis as well.

Together with PP, the very possibility of undermining futures entails a contradiction. Informally, the argument runs as follows: Right now, there is some undermining future F with nonzero chance. By the Principal Principle, it is rational to place nonzero credence in F , conditional on the current facts about chance. But F is incompatible with the current facts about chance; in a world where F obtains, those facts do not. Therefore, it is rational to place zero credence in F , conditional on the current facts about chance. Contradiction!

1.2.1 The Big Bad Bug: a Formal Version

Let C be any rational initial credence function.

Let F be an undermining future at time t in world w .

Let $P_{tw}(F)$ rigidly designate F 's objective chance at t in w .

Let E_F be the conjunction of H_{tw} with the proposition that F 's objective chance at t is $P_{tw}(F)$.

- | | | |
|----|---|---|
| 1) | $C(F E_F) = P_{tw}(F)$ | Consequence of PP and the definition of E_F |
| 2) | $P_{tw}(F) > 0$ | Definition of undermining futures |
| 3) | $\Box(E_F \supset \neg F)$ | Definition of undermining futures |
| 4) | $\Box(E_F \supset \neg F) \supset C(F E_F) = 0$ | Suppressed premise |
| 5) | $C(F E_F) > 0$ | 1, 2, substitution of identicals |
| 6) | $C(F E_F) = 0$ | 3, 4, modus ponens |
| 7) | Contradiction! | 5, 6 |

I'll assume that both the rules used in the formal version of the Bug—substitution of identicals and modus ponens—are uncontroversially valid. What of the Bug's premises? 1 follows straightforwardly from PP and the definitions of the relevant terms. 2 and 3 are motivated by HS. In the previous section, I argued that if HS is true, then there is at least one undermining future F . By the definition of 'undermining future', F satisfies 2. I take the modal operator in 3 to mean 'it is knowable *a priori* that...', rather than 'necessarily...'. On this reading, 3 follows from JC (a consequence of HS). Finally, I take 4 to follow from a plausible claim about the nature of *a priori* truths: that a rational agent should assign them credence 1. Where $C(E_F) > 0$, 4 follows straightforwardly. So as long as there is one reasonable initial credence function that assigns nonzero credence to E_F , the argument goes through.

Thus, the main assumptions underlying the Bug are as follows: first, that PP is a correct norm of rationality; second, that if HS is true, then there exists an undermining future F that satisfies premises 2 and 3; and third, that a rational person should assign credence 1 to all *a priori* truths. Since I know of no reason for disputing the third assumption, I will take it for granted. This leaves the first two assumptions open to criticism. I devote one part of the paper to defending each. Part 1.3 examines a range of alternatives to PP, none of which, I argue, can be given sufficient motivation. Part 1.4 discusses attempts to shed doubt on the phenomenon of undermining, none of which, I argue, is successful. I conclude that the Bug remains a serious challenge for Humeans.

1.3 Alternatives to PP

Several authors argue that PP should be either restricted or replaced with a chance-credence that is similar in spirit, but avoids the Bug. Part 1.3 discusses these proposals. In section 1.3.1, I argue against a proposal by Hoefer [1997] to restrict PP so that it only applies to relatively weak values of A , and offer an example of a situation where the restricted version PP generates contradictions. In section 1.3.2, I examine the New Principle, a candidate replacement for PP proposed independently by Lewis [1994] and Hall [1994]. I argue that Hall's arguments in favor of the New Principle fall short of being convincing. In section 1.3.3, I consider a proposal by Roberts [2001] to replace PP with a principle he calls GPP. I argue that there is no way to motivate GPP without PP, and so no reason to accept Roberts' proposal. Finally, in section 1.3.4, I examine a proposal by Ismael to replace PP with a principle she calls the General Recipe. I argue that the General Recipe contradicts Bayes' theorem, and therefore should be discarded.

1.3.1 Restricting PP

Thau [1994] points out that Humeans can escape the Bug by restricting PP. PP has a built-in escape clause: it applies only in cases where E is admissible (with respect to A). But if HS is true, then in general, information about A 's chance is also information about the future—it entails that no undermining will occur. Therefore, information about A 's chance is inadmissible, and PP does not apply. This strategy gets around the Bug, but only at the cost of rendering PP virtually useless.

Most authors, on noticing that Thau's restriction cripples PP, urge Humeans to adopt some new principle relating chance to rational credence. Hoefer [1997], however, claims that Humeans can keep PP, provided they restrict its scope to cases where A is too weak to be an undermining future. Unfortunately, Hoefer's restriction on PP does not kill the Bug, but merely sweeps it under the rug. Information about chance is generally inadmissible even with respect to relatively weak propositions about the future.

Intuitively, information about chances is inadmissible because it rules out undermining futures. But even when A is relatively weak, more undermining futures may be compatible with the proposition that $P(A) = x$ than with its negation. More formally, the explanation runs as follows. Let A be a proposition about the future at t in w , let x denote $P_{tw}(A)$, and let X be the proposition that $P(A) = x$. A can be expressed as a disjunction of propositions about the future which (in conjunction with facts about history up to t) are strong enough to fix A 's chance. Some of these propositions will entail X (call these propositions the G s, and their disjunction Γ), while the rest will entail $\neg X$ (call these propositions the F s, and their disjunction Φ). I'll assume that the relevant disjunctions can be formulated so that each F and each G receives nonzero credence. If information about A 's chance is admissible with respect to A , then by PP,

$$C(A|X) = x$$

In other words,

$$C(\Phi \vee \Gamma|X) = x$$

$$C(\Phi|X) + C(\Gamma|X) = x$$

Each of the F s is incompatible with X , so Φ is incompatible with X . Thus,

$$C(\Gamma|X) = x$$

But recall that Φ and G are incompatible, and if X is true, both of them have nonzero objective chance. You ought reasonably to believe that $P(\Phi \vee \Gamma) > P(\Gamma)$. By PP,

$$C(\Gamma|X) < x$$

Contradiction!

Since my discussion of GPP so far has been fairly abstract, it may help to close this section with an example, adapted from Hoefer [1997], that shows why restricting PP won't work. Imagine that at t , the universe is nearing its end. No theories have yet been undermined, but you are about to toss two coins. (Let the possible joint outcomes be denoted HH , HT , TH , and TT .) Right now, two theories are possible: the theory that both coins have chance $3/4$ of landing heads (call it $T_{3/4}$), and the theory that both coins have chance $1/4$ of landing heads (call it $T_{1/4}$). If HH ensues, $T_{1/4}$ will be undermined. Right now, you place credence $1/2$ in each theory of chance. Suppose that PP applies to every proposition A which is too weak to be an undermining future. By PP,

$$C(TT \vee TH|T_{1/4}) = 3/4$$

$$C(TH \vee HT|T_{1/4}) = 3/8$$

$$C(HH \vee TT|T_{1/4}) = 5/8$$

(Though it may look odd, the last application of PP is legitimate— $HH \vee TT$ is not an undermining future. Although one of its disjuncts is incompatible with $T_{1/4}$, the entire disjunction is not.) Since $T_{1/4}$ and HH are incompatible,

$$C(T_{1/4}|HH) = 0$$

By the probability calculus,

$$C(TT|T_{1/4}) = 5/8$$

$$C(TH|T_{1/4}) = 1/8$$

$$C(HT|T_{1/4}) = 1/4$$

By PP,

$$C(HT \vee TT|T_{1/4}) = 3/4$$

But

$$\begin{aligned} C(HT|T_{1/4}) + C(TT|T_{1/4}) &= 1/4 + 5/8 \\ &= 7/8 \end{aligned}$$

Your credences violate the probability calculus! A proposition that is too weak to be an undermining future may still be the disjunction of an undermining future with something else. This is enough to cause trouble for PP.

My example involved a short future, consisting of only two probabilistic trials. But this was in no way crucial. The problem turned on the fact that propositions which are not undermining futures still stand in logical relations to undermining futures. Increasing the number of future probabilistic trials won't eliminate this feature of the example. Restricting PP is not enough to eliminate the Bug. If Humeans are to revise PP, they must revise it more drastically.

1.3.2 The New Principle

Both Lewis [1994] and Hall [1994] propose that PP be replaced with a new principle, which is consistent with HS, but which captures something of PP's spirit. The New Principle (henceforth NP) states that where C is any reasonable initial credence function, H_{tw} is a complete description of world w 's history up to and including time t , and T_w is the complete theory of chance at w :

$$NP \quad C(A|H_{tw}T_w) = P(A|T_w)$$

PP stated that a rational agent's credence in a proposition equals her expectation of its chance, unless she possesses inadmissible information. NP states that a rational agent's credence in a proposition, conditional on any complete theory of chance, equals her expectation of its conditional chance, given the same theory, unless she possesses information about the future. In effect, NP replaces unconditional chances with conditional ones. The problem with PP involved undermining—the fact that the correct theory of chance assigns itself a chance of less than one. The solution is to conditionalize the objective chance function on the theory of chance.

At first glance, NP might seem objectionably unwieldy. One can sometimes hazard an educated guess at $P(A)$, but how on Earth does one calculate $P(A|T_w)$? Fortunately, PP serves as a reasonable approximation to NP for ordinary purposes. If information about A 's chance were perfectly admissible with respect to A , PP would be perfectly true. If HS is true, however, information about A 's chance is inadmissible with respect to A , since it tacitly carries information about whether A will occur. But the weaker A is, the less information it carries about the correct theory of chance, and so the less information the chances carry about it. As long as A describes a relatively small part of the world relative to the entire future, information about A 's chance will be close to admissible, and NP will approximate PP.

Still, NP requires a positive as well as a negative defense. While PP seemed to capture our actual inductive practices, NP looks revisionary. Hall [2004] argues that appearances are misleading, and that NP correctly captures our inductive practices. I will argue that Hall's defense of NP is fatally flawed, since it relies on metaphysical assumptions that are directly at odds with HS. From a Humean perspective, NP is unmotivated.

Hall's argument turns on the distinct role of chance in deliberation. He claims that chance is a sort of impersonal epistemic authority—a 'guru'—regarding propositions about the future. We treat the chance function as we would treat the credence function of an expert—we adjust our partial beliefs to its verdicts. But how should we adjust our beliefs? That depends on why chance is an expert.

Chance might be a 'database-expert', whose primary value lies in the amount of information it possesses. An eyewitness to a crime is a database-expert: her testimony is valuable because she knows more about the crime than other observers. On the other hand, chance might be an 'analyst-expert', whose primary value lies in its skill at evaluating evidence. A good advice columnist is an analyst-expert: her testimony is valuable not because she knows more about (say) her readers' mothers-in-law than they do, but because she is better than her readers at evaluating evidence. Different types of experts call for different epistemic responses.

If your guru is a database-expert, you should obey the following rule, where C is your credence function and G is your guru's credence function:

Database rule $C(A|G(A) = x) = x$

But you should avoid obeying this stronger variant, at least for unrestricted values of E :

Database+ $C(A|G(A) = x \ \& \ E) = x$

E might be a piece of information that is relevant to A but unknown by your guru. If you want to obey database+, you must add the restriction that E is admissible, i.e., that your guru has already taken into account its bearing on A (if any).

If your guru is an analyst-expert, database+ won't get you far. Nothing guarantees that analyst-experts are especially well-informed. You are likely to possess relevant information that an analyst-expert lacks, i.e., inadmissible information. You would do better to obey the following rule, where E is your total evidence:

Analyst rule $C(A|G(A|E) = x) = x$

The analyst rule tells you to adopt the credences your guru would adopt, were she in your epistemic position.

As it stands, the analyst rule is not quite right. You probably think that in order to count as an analyst-expert, your guru must fulfill some condition. Suppose she qualifies as an expert by your lights only if certain important predictions she makes hold good in the future—that is, only if certain propositions

which she deems likely really do come about. She herself may not be certain she fulfills your condition. There are some futures that she deems highly unlikely, but grants nonzero credence. If one of these futures were to come about, then in your eyes, her authority might be undermined. So if you accept the analyst rule and know what your guru believes, you cannot completely trust her. Where Q is the proposition that your purported guru satisfies your condition on expertise, $G(Q|E) < 1$. This is odd.

Things get odder if you're trying to decide which of several purported experts to trust. (In this case, let G be the credence function of the real guru, whose identity is unknown.) Suppose you think someone counts as a guru if only certain propositions which she deems likely really do come about. And suppose you have two candidate gurus: one whose credence is tightly focused on a small number of worlds, and one whose credence is spread out over a larger number of worlds. The first guru assigns a very low credence to the proposition that the future will unravel in a way she deems unlikely; the second guru assigns a higher credence to the proposition that the future will unravel in a way she deems unlikely. By your lights, the second guru's credence function is more self-undermining. Should this affect your confidence in their expertise?

Of course not. You should first use your current evidence evaluate the credentials of each purported expert, then use their predictions to adjust your beliefs about everything else. The way to isolate the part of a purported expert's belief state that bears on 'everything else' is to conditionalize her credence function on Q , the proposition that she satisfies the condition on expertise. You should really obey this rule:

$$\text{Analyst+ } C(A|QE) = G(A|QE)$$

Suppose that at t , both you and your guru know the complete history of the world up to and including t . Then $G(A|QE) = G(A|Q)$. Where H is a proposition specifying the complete history of the world up to t , analyst+ can be rewritten as follows:

$$\text{Analyst+ } C(A|HQ) = G(A|Q)$$

Note that if the supposed guru is completely confident in her own expertise—that is, if $G(A|Q) = G(A)$ for all A —then database+ and analyst+ coincide.

By now, the meaning of all this should be clear. Database+ is just PP in slightly different notation. Analyst+ is just NP in slightly different notation. Hall contends that since chance is an analyst-expert, we should obey NP. If the correct theory of chance assigns itself probability 1, then this will amount to obeying PP as well. If not, PP is derived from an erroneous concept of chance, and should be given up.

Hall's defense of PP invites two worries. The first worry involves the quantity $P_t(A|T)$, which appears on the right-hand side of NP. Is this quantity really well-defined? If it is, it seems it should equal $P_t(A)$. But we've seen that this is impossible if HS is true (at least for most values of A). Hall heads off the problem by invoking the assumption that at any time t , the chance function

is defined over arbitrary Boolean combinations of experimental outcomes. Actually, Hall needs an additional assumption—that T is among these arbitrary Boolean combinations. But given HS, the additional assumption is plausible: any proposition strong enough to fix the outcomes of all future stochastic trials should be strong enough to fix the theory of chance. Thus, the theory of chance is equivalent to some disjunction of some of the atoms in the Boolean algebra over which the chance function is defined.

The second and more serious worry involves Hall’s conception of chance: why is chance an analyst-expert, rather than a database expert? Hall might respond by calling attention to chance’s role in deliberation. The chance function ‘knows’ only propositions that has probability 1, but an agent who knows every proposition that has probability 1 can still derive epistemic benefit from considering the chances. Chance is not valuable just because it possesses more information than us—it’s not a database expert. Furthermore, we can derive epistemic benefits from considering chance’s verdicts conditional on propositions it doesn’t know. This makes chance less like an eyewitness (whose unconditional credences are all that matter) and more like an advice columnist (whose claims about hypothetical situations are just as valuable as her claims about actual situations).

But Hall’s picture of chance is smuggling in metaphysical assumptions. If HS is true, then in addition to ‘knowing’ about the past, the chance function ‘knows’ about the distribution of outcomes among future trials. Its epistemic advice depends reliably on future patterns; if the future were different, then it would give different advice. True, the chance function may not assign probability 1 to the propositions on which its advice depends. But this makes chance like a tentative eyewitness. We are justified in trusting the chance function only because it is sensitive to information about the future that we ourselves lack. Thus, although consistent with HS, NP is poorly motivated. Hall’s attempt to it a more comprehensive grounding leans on a decidedly un-Humean metaphysics of chance.

1.3.3 *A More General Principal Principle?*

So far, I have considered two unsuccessful attempts to emend PP. I turn now to a third suggestion by Roberts [2001]. Roberts [2001, S104] claims that PP is a special case of a more general rule given in [Lewis, 1986b, 87]:

$$\text{GPP } C(A|E) = \sum_x C(P(A) = x|E)x \text{ (Roberts, 2001, S104).}$$

Not only is PP a special case of GPP, says Roberts: it is a useless special case. Since no one is ever in a position to know for certain what the chances are, no one should ever conditionalize on propositions about chance. Roberts [2001, S105] restricts GPP’s application to cases where E is a piece of “evidence that is possible for finite empirical cognizers to obtain”. Since human beings are finite empirical cognizers, there is no point in extending our chance-credence principles beyond our capabilities.

But PP does not instruct anyone to conditionalize on propositions about chance. In PP, the admissible proposition E represents a piece of possible evidence, but the conjunction EX need not. The point of conditional credences is to allow speculation about which credences would be reasonable if certain propositions about chance were true. This as-if reasoning is useful to agents already endowed with partial beliefs about the chances. An agent's credences should be mixtures of her as-if credences conditional on different theories of chance, where the amount contributed by each theory depends on her credence in it.

As-if reasoning is also important in reasoning about chances themselves. In order to calculate the degree to which a proposition A about the outcomes of probabilistic trials counts as evidence toward a theory of chance T , given certain background information E , it is necessary to compare the likelihood $C(A|TE)$ with the likelihood $C(A|T'E)$ (where T' is some alternative hypothesis). But comparing likelihoods will only work where $C(A|TE)$ is well-defined.

Furthermore, Roberts is wrong in claiming that GPP is a more general version of PP. GPP is a consequence of PP—not the other way around. The probability calculus entails that

$$C(A|E) = \sum_x C(P(A) = x|E)C(A|P(A) = x)$$

PP states (in slightly altered notation), that for every x between 0 and 1,

$$C(A|E \& P(A) = x) = x$$

provided E is admissible. Thus,

$$C(A|E) = \sum_x C(P(A) = x|E)x$$

This is GPP.

One could consistently accept GPP while denying PP, provided one was willing to deny that $C(A|P(A) = x) = x$. But why deny that? It would be bizarre if $C(A|P(A) = x) = y$ for some $y \neq x$. The only other option is to claim that $C(A|P(A) = x)$ is undefined. As I have pointed out, however, this option is undesirable, since conditional probabilities play a crucial role in as-if reasoning. Without PP, there is no reason to accept GPP.

1.3.4 The General Recipe

Ismael claims that Roberts was on to something, but that he has not gone far enough. She argues that “rather than making the Principal Principle more complex to avoid inconsistency, we should simplify it. It was a mistake to conditionalize on the theory of chance in the first place”

Ismael makes an effort to account for the role of chance in as-if reasoning. She distinguishes reasoning within the scope of a theory of chance from reasoning about which theory of chance is correct. The right rule for reasoning within the scope of a theory is:

PP_{uncond} always and everywhere adjust credence to chance, no matter what historical information you possess; $C(A|H) = P(A)$

The right rule for reasoning about theories is:

Ignorance principle where you're not sure about the chances, form a mixture of the chances assigned by different theories of chance with weights determined by your relative confidence in them.

Ismael combines these rules to create a replacement for PP, which she calls the General Recipe. Where C is a rational credence function, a_T is the believer's credence in T , and P_T is the objective chance function according to T , the General Recipe states:

$$\text{General Recipe } C(A) = \sum_T a_T P_T(A)$$

Ismael acknowledges that the General Recipe is sketchy. It "is only as good as an agent's procedures for determining the values of the a_T s", and she has said little about what these procedures might be. I think there is an unacknowledged problem here. One natural way to update the a_T s is to perform an experiment, observe its outcome A , consider how well each of the T s predicts A , and use Bayes' theorem to adjust one's credences in the T s accordingly. The relevant instance of Bayes' theorem states:

$$\text{Bayes' theorem } C_A(T) = (C(T)/C(A))C_T(A)$$

If this instance of Bayes' theorem is to be useful, then $C_T(A)$ must be well-defined. What the Bug shows, however, is that there is no straightforward way of defining it. $C_T(A)$ can't (in general) be $P_T(A)$, since A may be an underminer. It can't be 0 in the case where A is an underminer and $P_T(A)$ otherwise, or we run into the troubles of section 1.3.1. Ismael's solution requires quietism about $C_T(A)$. But the Humean can't use Bayes' theorem unless she says something about $C_T(A)$.

What I've said about Bayes' theorem will generalize any method of updating the a_T s that requires comparing the likelihood $C_T(A)$ with the likelihood $C_{T'}(A)$ for some alternative hypothesis T' . Where likelihoods are undefined, comparison of likelihoods is impossible.

1.3.5 Taking Stock

Restricting PP is not enough to stave off the Bug, but none of the alternatives to PP is very well-motivated. NP can't be justified without drawing on an un-Humean metaphysics of chance; GPP can't be motivated without PP; and the General Recipe contradicts Bayes' theorem. Humeans will have to look elsewhere for solutions to the Bug. If my gloss on the Bug is right, this will mean challenging the claims about undermining futures presupposed in premises 2 and 3.

1.4 Undermining Futures

If HS is true, are there really undermining futures of the sort that satisfy premises 2 and 3? Several arguments from the literature can be construed as shedding doubt on the existence of such futures. (My treatment of these arguments does not quite line up with that of their authors, whose formulations of the Bug differ from mine. But I have tried to interpret the authors as charitably as possible, given my concerns and my formulation of the Bug.) In section 1.4.1, I run through four arguments that undermining futures are impossible. I argue that none of them work.

A less obvious way to shed doubt on undermining futures is to redefine chance so that they no longer exist. In section 1.4.2, I discuss a proposals by Schaffer [2003] and Arntzenius and Hall [2003] to do just that. I argue that while these proposals provide a logically consistent way around the Bug, their theoretical costs are excessive.

1.4.1 Four Worries about Undermining

First worry Ismael points out that even by the lights of HS, any theory of chance is compatible with any proposition about a finite course of future events, unless one adds the proviso that there are no other future events. Although Ismael does not mention it explicitly, her claim raises a worry: can any purported undermining future satisfy clause (b) in the definition of undermining futures, or premise 3 of the Bug? If the laws of nature together with facts about history entail that the universe has some fixed, finite extension in space and time, then the worry is easily quelled. (Note that the scope of the quantifier is important here. The worry is quelled if there is some finite extension e such that history and the laws together entail that the universe's extension is less than or equal to e . It is not quelled if history and laws together entail merely that the universe's extension is finite.) In fact, undermining futures are possible whenever there is a positive chance that the universe has some fixed finite extension (though again, note the point about quantifier scope). Ismael's objection requires assumptions about the extent of the universe to which Humeans are not entitled (at least not without further argument).

Second worry Vranas [2002] claims that Humean Supervenience is contingent and *a posteriori*. If Vranas is right, then no purported undermining future can satisfy clause (b) in the definition of undermining futures, or premise 3 of the Bug. Vranas [2002] cites Lewis [x 1986a, 1994, 474] in support of the claim that Humean Supervenience is contingent. I claim that HS is necessary. Despite appearances to the contrary, Lewis and I need not disagree, for we mean different things by the phrase 'Humean Supervenience'. My HS is the thesis that the laws, dispositions, and objective chances supervene on the distribution of categorical properties; Lewis's LHS is the thesis that the laws, dispositions, and objective chances supervene on the distribution of categorical properties which are intrinsic to points, or point-sized occupants of points [see Lewis, 1994, 474].

I grant that if true, LHS is contingent and *a posteriori*. The world might have contained properties that are instantiated at point-tuples instead of points, and it might have contained temporally bilocated objects. But HS seems necessary and *a priori*. If facts about laws can somehow be constructed out of facts about categorical properties, then why suppose there are worlds containing sui generis laws? Wouldn't such laws be redundant, given that we could construct the more prosaic supervenient kind out of categorical properties? This suggests that HS is necessary. Furthermore, the motivation for believing HS—a Humean skepticism about necessary connections between distinct entities—suggests that HS is *a priori*.

In order to protect Humean Supervenience from undermining, Vranas needs more than the claim that LHS is contingent and *a posteriori*: he needs the claim that every version of Humean Supervenience is contingent and *a posteriori*. Although I see no contradiction in upholding this claim, neither do I see any reason for upholding it.

Third worry In the same paper, Vranas offers another argument for the conclusion that HS does not entail the existence of undermining futures. He claims (though not in so many words) that HS is compatible with the negation of JC. Even if HS is *a priori*, says Vranas, the way in which chance supervenes on the categorical facts may not be (2002, 160). Therefore, where F is a purported undermining future (at t in w) it need not be *a priori* that F is incompatible with the laws in H_{tw} & T_w . If Vranas is right, then there may be no proposition F that satisfies clause (b) in the definition of undermining futures, or premise 3 of the Bug.

Vranas is technically right: HS does not entail JC. But it would be extremely odd to believe HS without JC, since JC helps motivate HS. HS is attractive partly because it avoids a kind of metaphysical queerness—the idea “that there are more things in heaven and earth than physics has dreamt of” [Lewis, 1994, 475]—and partly because it avoids a kind of epistemic queerness—the idea that chance is a ‘primitive whatnot’ unrelated to rational credence [Lewis, 1994, 484-485]. A being who knew everything about the distribution of categorical properties and understood the norms of partial belief would know which partial beliefs were rational (or at least, she would know enough to discard some partial beliefs as positively wrong). How could such a being fail to know which theory of chance was correct (or at least, which theories to rule out), given the constitutive link between chance and credence? I grant that the Bug is no a problem for Humeans who reject JC, but I expect that such Humeans will be extremely rare.

Fourth worry Hofer [2007] claims that Humeans should reject the background assumptions of section 1.1.1. He proposes new framework, which he takes to be more in keeping with the role of chance in the special sciences. Humeans might hope that this new framework provides an independently motivated escape from the Bug. The proposed changes to the Lewis framework can be divided into two classes—those involving the relationship between chance and time, and those

involving the domain of the chance function.

Hoeyer [2007, 554-555] objects to three of Lewis's claims about the relationship between chance time: that chances are indexed to times; that propositions about the past have either chance 0 or chance 1; and that the theory of chance is set of history-to-chance conditionals. Hoeyer argues that Humeans should treat chances as indexed to a world alone. He illustrates his proposal with the following example [2007, 554]: Suppose that at noon yesterday, I flipped a fair coin. Then

my coin flip at noon yesterday was an instance of a chance setup with two possible outcomes, each having a definite objective chance. It was a chance event. The chance of heads was $1/2$. So $1/2$ is the objective chance of A . It still is; the coin flip is and always was a chance event. Being to the past of me-now does not alter that fact, though as it happens I now know A is false.

What's past, then, may well be chancy. There is no need for history-to-chance conditionals—Humeans can use the information about chances without knowing anything about history.

Hoeyer's proposal has its merits: it simplifies the Humean theory of chance, and it provides a way of reconciling Humean chance with determinism. But it presents no obstacle to the Bug. We can easily redescribe the undermining phenomenon of section 1.4 in Hoeyer's terms. Let an undermining proposition be any proposition F such that, where $P_w(F)$ rigidly designates the objective chance function of F at world w :

- (a) $P_w(F) > 0$, and
- (b) F is incompatible with the proposition that $P(F) = P_w(F)$.

Presumably, there are undermining propositions even if the chance function is not time-indexed. There is some minute (timeless) chance that all of the tritium atoms in the universe decay much faster than they actually will. But were the tritium atoms to decay much faster than they actually will, then the chancemaking pattern of the universe would be different. This is the sort of undermining that generates the Bug. If we simply set P equal to P_w and E_F equal to the proposition that $P(F) = P_w(F)$ for some undermining future F , all four premises of the Bug hold true in Hoeyer's new framework.

Hoeyer's second set of amendments to Lewis's framework deals with the domain of the chance function. Contrary to Lewis's assumptions, says Hoeyer [2007, 562], chances are defined only "conditional on the instantiation of a chance setup". Furthermore, "[t]he domain over which the [chance] function ranges may be quite limited" [2007, 563]. Might these changes provide a way out of the Bug?

I fear not. Even if the domain of the chance function is limited, it should include arbitrary Boolean combinations of experimental outcomes, for the reasons given in section 1.1.1. We can assign chances to the Boolean combinations of experimental outcomes by treating collections of experiments as composite chance

setups, and treating the totality of all probabilistic trials—past, present, and future—as one huge compound chance setup. Let B be a complete description of this compound chance setup, and let a conditionally undermining proposition be any proposition F such that

- (a) $P_w(F|B) > 0$, and
- (b) F is incompatible with the conjunction $B \& P(F|B) = P_w(F|B)$.

Suppose we set P equal to P_w and E_F equal to the conjunction $(B \& P_w(F|B) = x)$ for some undermining future F . Then we are almost ready to generate the Bug. We need one more ingredient—a version of PP that accommodates Hoefer’s conditional chances. Any such version of PP should entail the following analogue of premise 1 in the Bug:

$$1^* C(F|E_F) = P_w(F|B)$$

By the definition of an undermining future, the following analogues of premises 2 and 3 hold true:

$$2^* P(F|B) > 0$$

$$3^* \Box(E_F \supset \neg F)$$

We can keep the original version of premise 4.

$$4. \Box(E_F \supset \neg F) \supset C(F|E_F) = 0$$

But premises 1*-3*, together with premise 4, lead to a contradiction. The new Hoefer framework is just as Bug-infested as the old Lewis framework.

1.4.2 Redefining Chance

Not everyone who objects to undermining futures tries to argue against their existence. Schaffer [2003] and Arntzenius and Hall [2003] point out that one can eliminate undermining futures simply by reinterpreting the word ‘chance’. These authors begin by pointing out that the New Principle (of section 1.3.2) is consistent with HS. Furthermore, if ordinary chance satisfies both NP and HS, then one can define a type of ‘chance’ that satisfies both PP and HS. Therefore, PP and HS are jointly consistent, given the right definition of ‘chance’.

Suppose the ordinary chances (call them L-chances) supervene, and suppose they satisfy NP. Let the L*-chances be the L-chances conditional on the theory of chance, so that

$$L^*(A) = L(A|T_w)$$

Suppose L-chance satisfies NP, i.e., the correct norm for L-chance is

$$C(A|H_{tw}T_w) = L(A|T_w)$$

By the definition of L*-chance,

$$C(A|H_{tw}T_w) = L^*(A)$$

One should place credence x in A conditional on any theory T_w that assigns L*-chance x to A . Therefore,

$$C(A|H_{tw} \& L^*(A) = x) = x$$

Assuming that information about history before t is always admissible at t ,

$$C(A|L^*(A) = x) = x$$

Thus, where E is any piece of admissible information,

$$C(A|L^*(A) = x \& E) = x$$

This is just PP in slightly different notation. Hence L*-chance satisfies PP. L*-chance also satisfies HS, since it was defined in terms of L-chance. If ‘chance’ means ‘L*-chance’, then PP and HS are compatible.

Unfortunately, L*-chance seems ill-suited to fill the role of chance in everyday reasoning. Two physically identical experimental setups may exhibit different L*-chances; in Schaffer’s terms, L*-chances are unstable. If the L-theory treats multiple trials of an experiment as drawings from an urn with replacement, then the L*-theory will treat them as drawings from an urn without replacement. In addition to being unstable, L*-chances are unlawful. It is the L-chances, and not the L*-chances, that appear in the laws.

Schaffer argues that L*-chances come close to being stable and lawful. L*-chances are stable in the sense of being exchangeable. For any time t and any trials a and b that take place after t on physically identical experimental setups, the L*-chance at t of a ’s producing an outcome is equal to the L*-chance at t of b ’s producing the same outcome. L*-chances are also approximately lawful: near the beginning of a universe with a long history, they will approximate the L-chances. And in a weaker sense, L*-chances are exactly lawful: only indeterministic worlds contain L*-chances.

But these analogues of stability and lawfulness are not enough to ground our ordinary epistemic practices. The assumption that chances are stable is important partly because it makes facts about chance epistemically accessible. If the chances are stable, then one doesn’t need to know what time it is in order to know something about the chances; the laws and the local facts will often suffice. If the chances are unstable, one must locate oneself in the history of the universe before coming to any conclusions. True, exchangeability is important for deriving statistical predictions, but exchangeability is not good enough. It is important that chance be exactly lawful so that knowledge about laws and local circumstances will translate into knowledge about chances. I am inclined to agree with the conclusion of Arntzenius and Hall [2003, 178]: “Lewis is misguided when he writes of [the Principal Principle] that it captures ‘all we know about chance’ ([1980], p. 86).” Perhaps no property could qualify as chance unless it satisfied PP, but no property could qualify as chance if it were as unstable and unlawful as L*-chance.

1.5 *Conclusion*

Escaping the clutches of the Big Bad Bug is harder than it looks, and harder than many of the Bug's critics have realized. So far, Humeans have failed to motivate a viable alternative to PP or formulate a compelling challenge to the undermining futures that appear in premises 2 and 3. The ball is in the Humeans' court. If HS is to remain viable, they must find a convincing objection to the Big Bad Bug. So far, no truly convincing objections have been offered.

2. DISTORTED REFLECTION

2.1 *Introduction*

Counterexamples to Bas van Fraassen's Reflection principle are sometimes thought to raise trouble for the diachronic Dutch book defense of conditionalization. Reflection states that an agent should treat her future self as an expert, or, roughly, that her current credence in any proposition A should equal her expected future credence in A . Van Fraassen [1984] uses a diachronic Dutch book to argue that Reflection is a norm of coherence. But Reflection is vulnerable to numerous counterexamples. These counterexamples seem to spell trouble either for Dutch books in general or for diachronic Dutch books in particular.

I argue that we should accept Reflection, but weaken it with an escape clause. I show that as long as an agent satisfies the escape clause, synchronic probability axioms alone guarantee that she will satisfy Reflection. I also show that as long as an agent comes close to satisfying the escape clause, synchronic probability axioms alone guarantee that she will come close to satisfying Reflection.

What about the Dutch book? I argue that contrary to a common misconception, not all Dutch books dramatize incoherence—some dramatize a less blame-worthy sort of epistemic frailty that I call 'self-doubt'. The distinction between Dutch books that dramatize incoherence and those that dramatize self-doubt cross-cuts the distinction between synchronic and diachronic Dutch books. I explain why the Dutch book for conditionalization reveals true incoherence, while the Dutch book for Reflection reveals only self-doubt.

2.2 *Dutch Books*

For Bayesians, probabilities represent subjective degrees of certainty, and probability axioms are logical norms governing belief. An agent can believe a proposition A to degree 1 (certainty), degree 0 (certainty that A is false), or any degree in between (uncertainty about A 's truth value, with a bias in one direction or the other). Her overall doxastic state can be represented as a credence function which maps propositions to real numbers in the unit interval. In a simple full belief model, by contrast, there are only two possible attitudes: belief and doubt. (There is no need to count disbelief as a third attitude, since disbelief in A is equivalent to belief in $\neg A$.) An agent's overall doxastic state in the full belief model can be represented as the set of propositions she believes, or equivalently, as a function that assigns 1 to all propositions in the set and 0 to

all other propositions.

In both Bayesian and full belief models, agents' overall doxastic states are subject to norms of consistency. An agent in a full belief model should avoid believing two logically incompatible propositions at the same time. Likewise, an agent in a Bayesian model should conform to the following norms of coherence, where Cr is her credence function at a particular time:

Non-negativity $Cr(A) \geq 0$, for all A .

Normalization $Cr(T) = 1$, where T is the necessary proposition.

Finite Additivity $Cr(A \vee B) = Cr(A) + Cr(B)$, where A and B are incompatible.

In addition to these synchronic norms, which govern relationships between beliefs at a time, agents are bound by diachronic norms, which govern relationships between initial beliefs, new information, and later beliefs. A simple full belief model might include a diachronic norm to the effect that an agent who starts out believing ($E \supset A$) for some E compatible with her beliefs and learns E should, on pain of incoherence, come to believe A . An analogous norm for Bayesians is:

Conditionalization $Cr_E(A) = Cr(A|E)$

where Cr is the agent's initial credence function, E is a proposition to which she initially assigns nonzero credence, $Cr(A|E)$ is equal to $Cr(A \& E)/Cr(E)$, and Cr_E is the agent's credence function after she has learned E (and nothing stronger).

The three synchronic axioms, together with conditionalization, are supposed to have normative force. But why should agents conform to that set of norms, and not some other? One explanation relies on Dutch book arguments (henceforth DBAs). I will take a DBA to be any argument of the following form (where Cr is some agent's credence function):

1. If $Cr(A) = p$, then the agent's credences condone buying or selling, for an arbitrary sum of money $\$Sp$, a ticket which entitles the buyer to $\$S$ out of the seller's pocket if A is true, and nothing otherwise.
2. If Cr violates some purported norm N , then the agent's credences condone entering into a Dutch book—that is, a set of bets which ensure that she suffers a net financial loss. (Inferred from premise 1.)
3. If an agent's credences condone entering into a Dutch book, then she is incoherent.
4. Therefore, any agent who violates N is incoherent.

Premises 1 and 3 are somewhat controversial. Spelling out what it means for a credence function to condone a set of betting odds is a delicate matter. Likewise, cashing out an appropriate notion of coherence is difficult. These are important issues, but for present purposes, I'm going set them aside, and simply assume tht some sense can be made of 'condone' and 'incoherent'.

The synchronic norms can be defended using synchronic DBAs, where the Dutch book consists of several simultaneous bets. Conditionalization must be defended using a diachronic DBA, where the Dutch book consists of bets made at different times [Lewis, 1999, Teller, 1973]. Some authors accept synchronic DBAs, but are skeptical of diachronic DBAs. Their skepticism is usually motivated by problems surrounding the DBA for Bas van Fraassen's principle of Reflection.¹ I will argue that the philosophically important distinction is not between synchronic and diachronic DBAs, but between DBAs that reveal incoherence and those that reveal a different and less serious type of epistemic defect.

Before I proceed, a brief note about bets is in order. I'll denote individual bets using tables whose left hand column lists states of the world, and whose right-hand column pairs each state with the agent's net gain or loss, should that state eventuate. Thus, a bet that costs the buyer \$ n and pays her \$ m just in case A is true will be written:

$$\begin{array}{ll} A & \$m - n \\ A & \$ - n \end{array}$$

A conditional bet which costs the buyer \$ n , pays her \$ m if $A \& B$, and is called off if $\neg B$, will be written:

$$\begin{array}{ll} A \& B & \$m - n \\ \neg A \& B & \$ - n \\ \neg B & \$0 \end{array}$$

The diachronic DBA for conditionalization can be written as follows. (Note: The Dutch book requires two assumptions—first, that the agent's possible evidence propositions form a partition—that is, any two possible evidence propositions are incompatible, and their disjunction is a tautology—and second, that she has no chance of mistaking her evidence—that is, if $Cr(E) = 1$ after she updates, then E is true, and if $Cr(E) = 0$ after she updates, then E is false.)

Let Cr be the agent's initial credence function.

Let Cr_1 be her credence function at t_1 , after she updates.

Let $Cr(A|E) = n$

Let $Cr_E(A) = r$

Let $Cr_1(E) = d$ for $0 < d < 1$

¹ Usually, but not always. Levi [1987] and Maher [1992] argue that diachronic Dutch books are ineffective because agents will always see them coming and avoid placing the initial bets. For a compelling reply to Levi and Maher, see [Skyrms, 1993].

Bets 2.1 and 2.2 are made before the agent learns whether E .

$$\begin{array}{ll} A \ \& \ E & \$1 - n \\ \neg A \ \& \ E & \$ - n \\ \neg E & \$0 \end{array} \quad (2.1)$$

$$\begin{array}{ll} E & \$(d - 1)(r - n) \\ \neg E & \$d(r - n) \end{array} \quad (2.2)$$

Bet 2.3 is made at t_1 if and only if the agent has learned E .

$$\begin{array}{ll} A & \$r - 1 \\ \neg A & \$r \end{array} \quad (2.3)$$

No matter what happens next, the buyer's net gain is $\$d(r - n)$. If $\neg E$, she wins $\$d(r - n)$ on bet 2.2, and no money changes hands on bets 2.1 or 2.3. If E , she wins a total of $\$(r - n)$ on bets 2.1 and 2.2 and wins $\$(d - 1)(r - n)$ on bet 2.3, again for a total of $\$d(r - n)$. For an agent who violates conditionalization, either $r > n$ or $r < n$. In the first case, the set of bets favors the buyer; in the second case, it favors the seller; and in either case, it constitutes a Dutch book.

2.3 Reflection

Van Fraassen [1984] argues for a norm he calls Reflection, using a diachronic DBA similar to the above argument for conditionalization. Where Cr_0 is an agent's credence function at time t_0 , Cr_1 is her credence function at some later time t_1 , and r rigidly designates a real number, Reflection states:

$$\text{Reflection } Cr_0(A|Cr_1(A) = r) = r$$

To understand what this means, it will help to imagine an agent who satisfies Reflection. Such an agent treats her future self as an expert about all propositions. If she is certain that her future self believes A to degree r , then she believes A to degree r . If she is uncertain about her future credence in A , then she sets her current credence in A equal to a mixture of her possible future credences, where the strength of each ingredient in the mixture depends on how likely it is (by her lights) to reflect her future degree of belief. Thus, her current credence in A is equal to her expected future credence in A . Finally, if she is a conditionalizer, then for any proposition A and real number r , if she learns that her later degree of belief in A will be r (and learns nothing stronger), then her new degree of belief in A will be r .

Anyone who violates Reflection is vulnerable to a Dutch book—provided that the proposition A for which $Cr_0(A|Cr_1(A) = r) \neq r$ satisfies three other assumptions. First, $Cr_0(A|Cr_1(A) = r)$ must be well-defined at t_0 . Second, the agent must be disposed to update in a way which ensures that $Cr_1(A)$ continues to be well-defined at t_1 . Finally, $Cr_0(Cr_1(A) = r)$ must be greater than zero. Where these three assumptions are satisfied, the following set of bets constitutes a Dutch book.

Let $Cr_0(A|Cr_1(A) = r) = n$

Let $Cr_0(Cr_1(A) = r) = d$ for $0 < d < 1$

Bets 2.4 and 2.5 are made at t_0 .

$$\begin{array}{ll} A \& Cr_1(A) = r & \$1 - n \\ \neg A \& Cr_1(A) = r & \$ - n \\ Cr_1(A) \neq r & & \$0 \end{array} \quad (2.4)$$

$$\begin{array}{ll} Cr_1(A) = r & \$(d - 1)(r - n) \\ Cr_1(A) \neq r & \$d(r - n) \end{array} \quad (2.5)$$

Bet 2.6 is made at t_1 if and only if $Cr_1(A) = r$.

$$\begin{array}{ll} A & \$r - 1 \\ \neg A & \$r \end{array} \quad (2.6)$$

No matter what happens next, the buyer's net gain is $\$(d(r - n))$. (Formally, the proof is exactly analogous to the proof in the Dutch book for conditionalization.)

2.4 Counterexamples to Reflection

Numerous authors have proposed counterexamples to Reflection. In this section, I divide the counterexamples into six types, and provide an instance of each type. My taxonomy roughly follows that of Bovens [1995].

2.4.1 Future irrationality

Christensen [1991] suggests the following example. The drug LSQ makes people believe to degree .99 that they can fly by flapping their arms. At t_0 , you become certain that you will take LSQ before t_1 . You deduce that at t_1 , you will place credence .99 in the proposition (call it F) that you can fly. Thus, $Cr_0(Cr_1(F) = .99) = 1$. By Reflection, $Cr_0(F)$ should be .99. This is clearly the wrong advice; your taking LSQ is not evidence that you can fly.

2.4.2 Memory loss

Talbott [1991] suggests the following example. At t_0 , you are eating a dinner of spaghetti and meatballs. You expect to forget this by t_1 , but you'll remember that t_0 was your dinner time. You'll also remember that you eat spaghetti for dinner 10% of the time. Where S is the proposition that you eat spaghetti at t_0 , $Cr_0(Cr_1(S) = .10) = 1$. Reflection advises you to set $Cr_0(S)$ equal to .10. But $Cr_0(S)$ should be much higher—at t_0 , your senses report the presence of spaghetti, and you should trust your senses.

2.4.3 *Apparent memory loss*

Arntzenius [2003] suggests the following example. A group of monks has elected to escort you to the city of Shangri La. The monks choose the route based on the outcome of a fair coin flip. If the coin lands heads, you will travel by the mountains; if tails, by the sea. If you travel by the mountains, you will arrive at Shangri La with glorious memories of the mountains. If you travel by the sea, your memories of the sea will be removed, and replaced with glorious memories of the mountains. At t_0 , you find yourself on the mountain path with the monks. You recognize that at t_1 , after you've arrived, you will place credence $1/2$ in the proposition that you traveled by the mountains. Thus, where M is the proposition that you travel by the mountains, Reflection advises you to set $Cr_0(M|Cr_1(M) = .5)$ equal to $.5$. But $Cr_0(M)$ should be 1 —again, you should trust your senses at t_0 .

2.4.4 *Future misleading evidence*

Maher [1992] suggests the following example. You are 90% certain that your friend Persi, a magician, knows the outcome of a fair coin toss. You also know that Persi is preternaturally eloquent, and can persuade you to grant credence 1 to the proposition that he knows the outcome of the coin toss. Where H is the proposition that the coin lands heads, Reflection demands that you set $Cr_0(H|Cr_1(H = 1))$ equal to 1 . This is bad advice. Right now, you surely know better than to place so much trust in Persi's testimony!

2.4.5 *Current misleading evidence*

Bovens [1995] suggests a version of the following example. You are a scientist at a small-town university. At t_0 , you believe yourself to have strong evidence for some groundbreaking hypothesis H . You know that most small-town scientists come to (justifiably) doubt their putative discoveries after three months, so you suspect that you will soon come to (justifiably) doubt H . Reflection advises you to decrease your current credence in H accordingly; thus, $Cr_0(H|Cr_1(H) = .01) = .01$. Accordingly, the higher $Cr_0(Cr_1(H) = .01)$ is, the lower $Cr_0(H)$ should be.² But surely this is wrong: expecting evidence against H is not the same as possessing evidence against H . Until the contrary evidence arrives, you should stand by your hypothesis.

2.4.6 *Changes in epistemic standards*

I propose the following example. At t_0 , you are deciding whether to enrol in the Ph.D. program at William James University, where all the professors are

² I am *not* claiming that Reflection requires you to increase your credence in D based on your knowledge about other small-town scientists. Reflection is perfectly compatible with your becoming more confident in $\neg D$ based on your knowledge of other small-town scientists. But once you have increased your credence in D , Reflection requires you to lower your credence in H accordingly.

voluntarists about belief. You are agnostic about God's existence, but you believe to degree .90 that if you are immersed in James University's voluntarist environment, you will become a theist. Where G is the proposition that God exists, Reflection tells you that $Cr_0(G|Cr_1(G) = .90) = .90$. But this is the wrong advice; you shouldn't treat your enrolment in William James University as evidence for God's existence.

2.5 Responses to the Counterexamples

How can advocates of DBAs reconcile the argument for Reflection with the apparent counterexamples? One common explanation is that diachronic DBAs, unlike synchronic DBAs, are unsound. Christensen [1991] defends a particularly persuasive version of this view. Vulnerability to Dutch books, he claims, reveals inconsistent beliefs. An agent who is susceptible to a synchronic Dutch book has inconsistent beliefs at a particular time—that is, some of the beliefs she holds at t_0 are inconsistent with other beliefs she holds at t_0 —while an agent who is susceptible to a diachronic Dutch book has inconsistent beliefs across time—that is, her beliefs as of t_0 are inconsistent with her beliefs as of t_1 . The first sort of inconsistency is problematic: it is like simultaneously believing A and $\neg A$. The second sort of inconsistency is perfectly acceptable: it is like believing A and then coming to believe $\neg A$, or like one person's believing A and a second person's believing $\neg A$. Sets of beliefs (or pairs of credence functions) held at different times are not the sorts of things that ought to be coherent.

Although tempting, this line of reasoning is misguided: pairs of credence functions held at different times are indeed the sorts of things that ought to be coherent. Bayesian decision theory should serve as a logic of inference and planning. But some sort of diachronic intrapersonal coherence is necessary for inference and planning; if an agent is to conduct her epistemic life correctly, then her earlier and later selves should be more coherent than two different people. The sort of diachronic coherence in question should not be so strong as to demand that agents never change their beliefs. But it should be strong enough to bar agents from adopting belief revision policies which lead to changes that are senseless or insupportable by their current lights. The type of coherence demanded by conditionalization is just right: strong enough to constrain the agent, but not strong enough to paralyze her.

Another possible response to the counterexamples is to accept the validity of diachronic DBAs, but claim that Reflection is only a defeasible norm. Green and Hitchcock [1994, 307] suggest that Reflection correctly describes the beliefs of "Rational Joe", a mildly idealized agent who "avoids obvious pitfalls, such as offering 2 to 1 odds on both outcomes of a coin toss", and whose credences are usually appropriately grounded in his evidence. Occasionally he errs by over- or underestimating the support that his evidence lends a hypothesis. But whenever he finds himself making an error, he immediately corrects it. Furthermore, he expects to remain rational in the future.

Rational Joe does not believe himself to be infallible. Let A be any propo-

sition. At t_0 , Rational Joe may grant some credence to the hypothesis that at t_1 he overestimates the support that his evidence lends to A . If he credits this hypothesis at t_0 , however, then at t_0 he will also credit the hypothesis that at t_1 he underestimates the support that his evidence lends to A . Rational Joe does not expect himself to make systematic errors in weighting his evidence—his expected degree of underestimation equals his expected degree of overestimation. So even though he is not certain at t_0 that his t_1 credence in A will be right, he is certain at t_0 that his *expected* t_1 credence in A is right. In other words, modest idealization though he is, Rational Joe obeys Reflection.

Still, modest idealization though he is, Rational Joe is considerably more ideal than most of us. We often notice epistemic vices in ourselves which (we suspect) we are incapable of correcting. Agents who lack Rational Joe's epistemic virtues would be unwise to emulate his adherence to Reflection—just as agents who lack the moral virtues of angels would be unwise to emulate the angels' policy of never apologizing for past mistakes. The Dutch book shows that Rational Joe obeys Reflection; it does not show that we should.

As it stands, Green and Hitchcock's proposal is somewhat unsatisfying. None of the counterexamples to Reflection requires you to suffer an epistemic mishap—you needn't take LSQ, embrace the wrong epistemic standards, or be deceived by misleading evidence. It is enough that you grant nonzero credence to the hypothesis you will suffer, or are currently suffering, an epistemic mishap. How is can it be wrong of you to acknowledge your own fallibility? Besides, Green and Hitchcock's proposal leaves the status of the diachronic DBA somewhat obscure. Does the DBA break down whenever the obligation to obey Reflection breaks down? If so, then where and why? If not, then isn't the DBA awfully easy to circumvent? Mightn't it be equally easy to circumvent the DBA for conditionalization?

A third line of thought suggests that Reflection is perfectly acceptable, provided it is somehow qualified. You should treat your future self as an expert, provided you expect to receive veridical evidence and respond to it in a rational manner [Hall, 1999, 680], or provided you know that your beliefs will improve over time [Evnine, 2007, 94], or provided you trust both your future memory and your future judgment [Elga, 2007, 480]. This line of thought is suggestive, but it does not directly address the question of where the DBA goes wrong. Besides, one might worry whether there is any formal way of cashing out the qualification on Reflection.

My account combines the best aspects of the three accounts surveyed. Like Christensen, I will reject the diachronic DBA for Reflection, but unlike Christensen, I will accept the intelligibility of diachronic constraints on coherence. Like Green and Hitchcock, I will claim that Reflection describes the credence functions of certain ideal agents, but unlike Green and Hitchcock, I will try to pinpoint exactly where the idealization goes wrong. Like Hall, Evnine, and Elga, I will suggest a qualifying clause for Reflection, but unlike these authors, I will explain how this clause, cashed out in formal terms, relates to the failure of the Dutch book.

I begin by formulating a Qualified Reflection principle, which I argue cap-

tures the intuitive ideas put forth by Hall, Evnine, and Elga. Unlike the original Reflection principle, Qualified Reflection follows from the Kolmogorov axioms (together with some plausible idealizing assumptions). I then formulate a Distorted Reflection principle which approximates Reflection even when the agent violates the escape clause in Qualified Reflection. Finally, I explain the crucial flaw in van Fraassen's DBA.

2.6 Qualified Reflection

Before cashing out the qualifications on Reflection, I will make three idealizing assumptions. First, I will assume that the agent is a perfect introspector—in other words, that $Cr_0(Cr_0(A|B) = r) = 1$ if and only if $Cr_0(A|B) = r$. (Note that this assumption entails the special case of Reflection where $Cr_0 = Cr_1$.) Second, I will assume that the agent's possible evidence propositions—that is the propositions that might serve as her total evidence between t_0 and t_1 —form a partition $\{B_1, B_2, \dots, B_n\}$. Third, I will assume that the agent can reasonably be certain that conditionalization is the right updating procedure. Not every agent satisfies these assumptions, but agents who do are bound by the following principle:

Qualified Reflection $Cr_0(A|Cr_1(A) = r) = r$, provided that for all $B \in \{B_1, B_2, \dots, B_n\}$,

- (i) $Cr_0(Cr_0(A|B) = Cr_1(A|B)) = 1$ and
- (ii) $Cr_0(B|Cr_1(B) = 1) = 1$

Qualified Reflection follows from the Kolmogorov axioms, together with my three idealizing assumptions.³

³ Proof: By the probability calculus,

$$\begin{aligned} Cr_0(A|Cr_1(A) = r) &= \frac{Cr_0(A \& Cr_1(A)=r)}{Cr_0(Cr_1(A)=r)} \\ &= \frac{\sum_{B:Cr_1(A|B)=r} Cr_0(A|Cr_1(B)=1)Cr_0(Cr_1(B)=1)}{\sum_{B:Cr_1(A|B)=r} Cr_0(Cr_1(B)=1)} \end{aligned}$$

By (ii),

$$Cr_0(A|Cr_1(A) = r) = \frac{\sum_{B:Cr_1(A|B)=r} Cr_0(A|B)Cr_0(B)}{\sum_{B:Cr_1(A|B)=r} Cr_0(B)}$$

And by (i),

$$\begin{aligned} Cr_0(A|Cr_1(A) = r) &= \frac{\sum_{B:Cr_0(A|B)=r} Cr_0(A|B)Cr_0(B)}{\sum_{B:Cr_0(A|B)=r} Cr_0(B)} \\ &= \frac{\sum_{B:Cr_0(A|B)=r} rCr_0(B)}{\sum_{B:Cr_0(A|B)=r} Cr_0(B)} \\ &= \frac{r \sum_{B:Cr_0(A|B)=r} Cr_0(B)}{\sum_{B:Cr_0(A|B)=r} Cr_0(B)} \\ &= r \end{aligned}$$

Clauses (i) and (ii) capture the intuitive ideas suggested by Hall, Elga, and Evidine.⁴ An agent satisfies (i) just in case she is certain that she will update rationally (as far as A is concerned) and she satisfies (ii) just in case she is certain that she will update on veridical evidence—in other words, just in case she satisfies Hall’s criterion. Expecting to conditionalize is a matter of trusting one’s future memory and one’s future judgment—Elga’s criterion. And an update which results from conditionalizing on veridical evidence is an epistemic improvement—Evidine’s criterion.

2.7 *Distorted Reflection*

Qualified Reflection is a useful piece of advice. But there is a catch: Qualified Reflection applies only in cases where the agent satisfies (i) and (ii), and such cases are hard to come by. Few agents are certain that they will keep conditionalizing in the future. If you entertain even the slightest suspicion that you will fail to conditionalize, Qualified Reflection gives you no advice at all.

Luckily, the problem admits of a simple solution. Even if you don’t satisfy (i) and (ii) perfectly, you may come close. And as long as you come close to satisfying (i) and (ii), you should come close to obeying Reflection. I will distinguish two axes along which an agent might fall short of perfect conformity to (i) and (ii), and indicate the proper response to movements along each axis.

Some future beliefs are more likely than others to be the result of conditionalizing on veridical evidence. If an agent believes some hard-to-verify scientific hypothesis to degree .45, this might be the result of veridical observation and reasonable updating; if she believes the same hypothesis to degree 1, this is more likely to be the result of over-confidence caused by some error. The first axis measures the number of values of r such that the agent expects a t_1 credence of r in A to be the result of conditionalization on veridical evidence, so that for all $B \in \{B_1, B_2, \dots, B_n\}$,

(iii) $Cr_0(Cr_1(A|B) = r | Cr_0(A|B) = r) = 1$ (where defined) and

(iv) $Cr_0(B | Cr_1(B) = 1 \ \& \ Cr_1(A) = r) = 1$

(iii) and (iv) state that the agent is certain at t_0 that if $Cr_1(A) = r$, she’s behaved like a conditionalizer (at least with respect to A) and updated on veridical evidence. If she satisfies (iii) and (iv) for some particular value of r , then she should obey Reflection for that value of r . (Notice that if she satisfies (iii) and (iv) for all values of r relative to A , Cr_0 , and Cr_1 , then she also satisfies (i) and (ii).)

⁴ Weisberg [2007] proves a very similar result: Conditionalization follows from my first two assumptions, the assumption that the agent is certain she will remain a perfect introspecter in the future, and (i). My result is slightly stronger than Weisberg’s, since Weisberg’s third assumption entails both my third assumption and (ii), but not vice-versa. Van Fraassen [1995] argues that conditionalization entails Reflection, but Weisberg shows that van Fraassen’s argument rests on a conflation between the opinions an agent might arrive at in the future and the opinions she thinks she might arrive at in the future.

The second axis measures the agent's expected departure from conditionalization on veridical evidence, on the hypothesis that $Cr_1(A) = r$. Expected departure from conditionalization depends both on the strength of the agent's (conditional) expectation that she will fail to conditionalize on veridical evidence (given that $Cr_1(A) = r$) and on the magnitude of the failure she expects. We might cash this out as follows.

Let $\{W_1, W_2, \dots, W_m\}$ be a set of doxastic alternatives such that each $W \in \{W_1, W_2, \dots, W_m\}$ is the conjunction of some $B \in \{B_1, B_2, \dots, B_n\}$ with $Cr_1(A) = r$. The agent's expected departure from conditionalization on veridical evidence (with respect to A , Cr_0 , and Cr_1) conditional on the proposition that $Cr_1(A) = r$ can then be defined as:⁵

$$D_r = \frac{\sum_{W \in \{W_1, W_2, \dots, W_m\}} (r - Cr_0(A|W)Cr_0(W))}{Cr_0(Cr_1(A) = r)}$$

D_r measures the degree to which an agent expects r to be an overly optimistic credence in A , relative to the support her future evidence lends A . (Where she expects r to be an overly pessimistic credence in A , D_r will be negative.)

D_r lets us adjust Reflection to account for expected failures of conditionalization. If Reflection requires an agent to set her credence in A equal to her expected later credence in A , then the following rule requires her to set her credence in A equal to her expectation of the later credence she would place in A , were she to conditionalize on veridical evidence:

$$\text{Distorted Reflection } Cr_0(A|Cr_1(A) = r) = r - D_r$$

To the extent that the agent approaches conformity to (i) and (ii) along either of the two axes, Distorted Reflection comes close to capturing Reflection. The closer D_r is to 0, the closer $Cr_0(A|Cr_1(A) = r)$ is to r . And the more values of r for which D_r is close to 0, the more values of r for which Distorted Reflection approximates Reflection. Perfect conformity to (i) and (ii) is rare, so agents are rarely required to conform perfectly to Reflection. But an agent who comes close to satisfying (i) and (ii) should come close to obeying Reflection.

Although Distorted Reflection always approximates Reflection when the agent is confident that she will come close to conditionalizing, it may approximate Reflection even when the agent is confident (or certain) that she will severely fail to conditionalize. D_r may be close to zero either because all its terms are close to 0, or because the sum of its positive terms is close in absolute value to the sum of its negative terms. The second alternative secures something like Reflection, but not because the agent is particularly confident in her future judgment—only because she has no useful information about how she will fail to conditionalize. On the second alternative, the fact that the agent satisfies Reflection is a mere lucky coincidence. Green and Hitchcock seem to

⁵ This definition relies on two assumptions—first, that $Cr_0(A|W)$ is well-defined for each W —else D_r is undefined—and second, that the agent knows the value of $Cr_0(A|W)$ for each W —else the agent does not know that $r - Cr_0(A|W)$ is the degree to which she will depart from conditionalization if W obtains.

have exactly this sort of coincidence in mind when they discuss Rational Joe, who thinks himself as likely to underestimate the weight of his evidence as to overestimate it.

What is the status of Distorted Reflection? It is simply a consequence of the probability calculus.⁶ Even so, Distorted is useful in roughly the way Bayes' theorem is useful: it expresses a hard-to-calculate quantity in terms of easier-to-calculate parts. Typically, given any evidence proposition $B \in \{B_1, B_2, \dots, B_n\}$, the agent's t_1 beliefs will be irrelevant to A —in other words, $Cr_0(A|W)$ will equal $Cr_0(A|B)$ for the B that serves as a conjunct in W . Calculating $Cr_0(W)$ is generally straightforward, and the value of r is simply stipulated.

2.8 Back to the Counterexamples

Together with Qualified Reflection, Distorted Reflection can be used account for the counterexamples to the original Reflection principle.

2.8.1 Future irrationality (LSQ)

Perhaps you think LSQ works by making you believe F to a higher degree than conditionalizing on your t_1 evidence warrants, so that for some or all of the $B \in \{B_1, B_2, \dots, B_n\}$, $Cr_0(F|B) < Cr_1(F)$. In this case, your suspicion that you will take LSQ causes you to violate (i). On the other hand, perhaps you think LSQ works by making you believe false propositions that would justify you in believing F to degree .99. In this case, your suspicion that you will take LSQ causes you to violate (ii). In either case, Qualified Reflection does not require that $Cr_0(F|Cr_1(F) = .99) = .99$.

So you needn't obey Reflection in the LSQ case. What should you do instead? Suppose there is no possible t_1 evidence that (by your current lights) could justifiably raise your credence in F to .99, so that there is no $B \in \{B_1, B_2, \dots, B_n\}$ such that $Cr_0(F|B) = .99$. Furthermore, suppose you think

⁶ Proof:

$$\begin{aligned}
 D_r &= \frac{\sum_{W \in \{W_1, W_2, \dots, W_m\}} (r - Cr_0(A|W))}{Cr_0(W)/Cr_0(Cr_1(A)=r)} \\
 &= \sum_{W \in \{W_1, W_2, \dots, W_m\}} \frac{(r - Cr_0(A|W))}{Cr_0(Cr_1(A)=r)} - \sum_{W \in \{W_1, W_2, \dots, W_m\}} \frac{Cr_0(A|W)Cr_0(W)}{Cr_0(Cr_1(A)=r)} \\
 &= \frac{rCr_0(Cr_1(A)=r)}{Cr_0(Cr_1(A)=r)} - \frac{Cr_0(A \& Cr_1(A)=r)}{Cr_0(Cr_1(A)=r)} \\
 &= r - Cr_0(A|Cr_1(A) = r)
 \end{aligned}$$

Distorted Reflection states

$$Cr_0(A|Cr_1(A) = r) = r - D_r$$

Substituting the value just derived for D_r , we see that Distorted Reflection is equivalent to the following obvious theorem of the probability calculus:

$$Cr_0(A|Cr_1(A) = r) = r - (r - Cr_0(A|Cr_1(A) = r))$$

that no matter which evidence proposition is true, the claim that you believe F to degree .99 at t_1 will have no bearing on whether you can fly— $Cr_0(F|Cr_1(F) = .99 \& B) = Cr_0(F|Cr_1(B) = 1)$ for each $B \in \{B_1, B_2, \dots, B_n\}$. By the definition of expected departure from conditionalization on veridical evidence,

$$D_r = \frac{\sum_{W \in \{W_1, W_2, \dots, W_m\}} (.99 - Cr_0(F|W)) Cr_0(W)}{Cr_0(Cr_1(A) = .99)}$$

Since the proposition that $Cr_1(F) = .99$ screens F off from each $B \in \{B_1, B_2, \dots, B_n\}$,

$$D_r = \frac{\sum_{W \in \{W_1, W_2, \dots, W_m\}} (.99 - Cr_0(F)) Cr_0(Cr_1(A) = .99)}{Cr_0(Cr_1(A) = .99)} = (.99 - Cr_0(F))$$

By Distorted Reflection,

$$\begin{aligned} Cr_0(F|Cr_1(F) = r) &= .99 - (.99 - Cr_0(F)) \\ &= Cr_0(F) \end{aligned}$$

Learning that $Cr_1(F) = .99$ should have no impact on your credence in F . Intuitively, this is right; the fact that you will believe F to degree .99 is irrelevant to F 's truth or falsity.

On the other hand, suppose that there is some evidence proposition B_x (highly unlikely by your t_0 lights) that could justify your believing F to degree .99. (Perhaps B_x involves God descending from the clouds and asserting in a booming voice that you can fly.) You are certain that you will update by conditionalizing, that if you do not take LSQ, your evidence will be veridical, and that if you take LSQ, you will believe B_x whether or not it is true. I will elaborate the example as follows (where Q is the proposition that you take LSQ):

$$\begin{aligned} Cr_0(F|B_x) &= .99 \\ Cr_0(F|B) &= .001 \text{ for all } B \in \{B_1, B_2, \dots, B_n\} \neq B_x \\ Cr_0(Cr_1(F = .99)|\neg Q) &= .001 \\ Cr_0(Cr_1(F) = .99|Q) &= 1 \\ Cr_0(\neg B_x \& Cr_1(F) = .99) &= Cr_0(\neg B_x \& Q) \\ Cr_0(B_x \& Cr_1(F) = .99) &= Cr_0(B_x) \end{aligned}$$

By the definition of D_r ,

$$\begin{aligned} D_r &= \frac{(.99 - .99)Cr_0(B_x) + (.99 - .001)Cr_0(\neg B_x \& Cr_1(F) = .99)}{Cr_0(Cr_1(F) = .99)} \\ &= \frac{.989(Cr_0(\neg B_x \& Cr_1(F) = .99))}{Cr_0(Cr_1(F) = .99)} \\ &= \frac{.989Cr_0(\neg B_x \& Q)}{Cr_0(Cr_1(F) = .99)} \\ &= \frac{.989Cr_0(\neg B_x \& Q)}{Cr_0(\neg B_x \& Q) + Cr_0(B_x)} \end{aligned}$$

On the assumption that B_x and Q are evidentially irrelevant to one another, the value of D_r depends on only two factors: $Cr_0(Q)$, and $Cr_0(B_x)$. As $Cr_0(Q)$

increases, D_r increases; the more certain you are that you will take LSQ, the less you should increase your confidence in F upon learning that $Cr_0(F) = .99$ (all other things being equal). As $Cr_0(B_x)$ increases, D_r decreases; the more certain you are that God will actually tell you that you can fly, the more you should increase her confidence in F upon learning that $Cr_0(F) = .99$. Again, both results seem right.

2.8.2 Memory loss (spaghetti)

You expect to place credence .10 in S at t_1 , even though conditionalizing on your total evidence (no matter what it will be) should lead you to be much more confident in S than that. Thus, for any $B \in \{B_1, B_2, \dots, B_n\}$, $Cr_0(Cr_0(S|B) = Cr_1(S)) \neq 1$, and you violate (i). Thus, you needn't set $Cr_0(S|Cr_1(S) = .10)$ equal to .10.

What should you do? Suppose that, just as in the LSQ example, there is no future evidence that could justify you in believing S to degree .10, and the proposition that $Cr_1(S) = .10$ screens off each $B \in \{B_1, B_2, \dots, B_n\}$ from S . Then once again, Distorted Reflection will tell you to ignore the information that $Cr_1(S) = .10$ —that information is irrelevant to S .

2.8.3 Apparent memory loss (Shangri La)

The Shangri La case appears superficially different from the spaghetti case, but their underlying structure is remarkably similar. Although you expect to remember your mountain travels at t_1 , you expect to doubt the veridicality of your memories. Just as in the spaghetti case, $Cr_0(Cr_0(S|B) = Cr_1(S)) \neq 1$ for every $B \in \{B_1, B_2, \dots, B_n\}$, and so you violate (i).

2.8.4 Future misleading evidence (Persi)

You may learn between t_0 and t_1 that Persi has reported a heads outcome, in which case there will be some $B \in \{B_1, B_2, \dots, B_n\}$ such that $Cr_0(H|B) > Cr_1(H|B)$. Since you place nonzero credence in the proposition that you will speak to Persi, $Cr_0(Cr_0(H|B) = Cr_1(H|B)) \neq 1$, and you violate (i). Qualified Reflection does not require you to set $Cr_0(H|Cr_1(H) = 1)$ equal to 1.

Notice that this result holds even if there is evidence that could in principle help you distinguish the case where Persi is fully informed about the coin toss from the case where he's just guessing. Suppose Persi speaks more quickly when guessing than he does when reporting something he knows. You may learn either that Persi slowly reports a heads outcome, in which case you will respond appropriately to your evidence, or that Persi quickly reports a heads outcome, in which case you will believe him even though you shouldn't. Even in this version of the example, you violate (i). Since your t_1 evidence B may entail that Persi quickly reports a heads outcome, you can't be sure that you will conditionalize as you should: again, there is some $B \in \{B_1, B_2, \dots, B_n\}$ for which $Cr_0(Cr_0(H|B) = Cr_1(H|B)) \neq 1$.

No matter which way we read the example, Distorted Reflection gives the same advice. Consider the first version, in which there is no observable difference between sincere and insincere Persi. Let your possible evidence propositions be

- R Persi reports a heads outcome
- T Persi reports a tails outcome

All of the following claims should hold:

$$Cr_0(H|R \ \& \ Cr_1(H) = 1) = Cr_0(H|R) = .95$$

(If Persi reports H , he is 90% likely to reporting correctly, 5% likely to be guessing correctly, and 5% likely to be guessing incorrectly.)

$$Cr_0(H \ \& \ Cr_1(H) = 1) = .5$$

(You are 50% certain that Persi will report a heads outcome and you will become convinced that the coin landed heads.)

$$Cr_0(T \ \& \ Cr_1(H) = 1) = 0$$

(You are 100% certain that if Persi reports a tails outcome, you will not become convinced that the coin landed heads.)

$$Cr_0(Cr_1(H) = 1) = .5$$

(You are 50% certain that you will come to believe H to degree 1.)

Thus, by the definition of D_r for $r = 1$,

$$D_r = \frac{(1 - .95).5}{.5} = .05$$

By Distorted Reflection,

$$Cr_0(H|Cr_1(H) = 1) = .95$$

Now consider the second version of the example, where Persi talks faster if he does not know the outcome of the coin toss. Let your possible evidence propositions be

- RQ Persi quickly reports H
- RS Persi slowly reports H
- TQ Persi quickly reports $\neg H$
- TS Persi slowly reports $\neg H$

All of the following should hold:

$$Cr_0(H|RQ \ \& \ Cr_1(H) = 1) = Cr_0(H|R) = .5$$

$$Cr_0(RQ \ \& \ Cr_1(H) = 1) = .05$$

$$Cr_0(H|RS \ \& \ Cr_1(H) = 1) = Cr_0(H|RS) = 1$$

$$Cr_0(RS \ \& \ Cr_1(H) = 1) = .45$$

$$Cr_0(TQ \ \& \ Cr_1(H) = 1) = 0$$

$$Cr_0(TS \ \& \ Cr_1(H) = 1) = .0$$

$$Cr_0(Cr_1(H) = 1) = .5$$

By the definition of expected departure from conditionalization on veridical evidence,

$$D_r = \frac{(1 - .5)(.05) + (1 - 1)(.45)}{.5} = .05$$

Once again, by Distorted Reflection,

$$Cr_0(H|Cr_1(H) = 1) = .95$$

On either version of the example, you come close to satisfying (i). Your expected deviation from conditionalization on veridical evidence is low in the first version, because you don't expect to depart very far from conditionalization, and in the second version, because you don't think you're very likely to depart from conditionalization. It makes no difference how finely we divide the evidence propositions.

2.8.5 Current misleading evidence (the small-town scientist)

This case is significantly different from the others. You presumably believe that you will conditionalize on any $B \in \{B_1, B_2, \dots, B_n\}$ you might learn between t_0 and t_1 . Furthermore, you don't expect to receive any non-veridical evidence. Thus, it seems that you satisfy (i) and (ii)—Qualified Reflection demands that $Cr_0(H|Cr_1(H) = .01) = .01$.

Some readers may think I have bitten a bullet in responding to this case, but I claim that the bullet is not as unpalatable as it seems. I have construed the case as one of current misleading evidence, where your future judgment is unimpugned. If I had construed the case as one of future misleading evidence, then it would collapse back into the Persi example. If I had construed it as a case of future non-veridical evidence, it would collapse back into the second version of the LSQ example. You should obey Reflection only in the version of the example where you expect veridical, non-misleading evidence against H .⁷

2.8.6 Changes in epistemic standards (William James University)

There are three ways of cashing out the William James University example. First, you may see attending William James University as an epistemic pitfall—a disaster which will ruin your future epistemic standards. On this reading, the William James University example is easily assimilated to the first three examples: you expect not to conditionalize on veridical evidence, so Qualified Reflection does not require that you set $Cr_0(G|Cr_1(G) = .90)$ equal to .90.

Second, you may see attending William James University as a wholly educational experience—a way to correct your current subpar epistemic practices. My

⁷ There may be pragmatic reasons for small-town scientists to believe their hypotheses more strongly than their evidence warrants. Perhaps confidence is crucial to the success of one's scientific career, or perhaps scientific disciplines are best advanced when individual scientists believe hypotheses even in the absence of sufficient evidence. This is compatible with my claim that Qualified Reflection is a *prima facie* epistemic norm.

account cannot accommodate all versions of this reading—a William James education had better not cause you to justifiably doubt the adequacy of conditionalization!—but it can accommodate some versions. Perhaps you are bad at gathering evidence for the existence of God, and attending William James will attune you to a new, more sentimental type of evidence. On this reading, the William James University example can be assimilated to the small town scientist example: you satisfy both (i) and (ii), and Qualified Reflection requires you to set $Cr_0(G|Cr_1(G) = .90)$ equal to .90.

Third, you may see attending William James University as a mixture of educational experience and epistemic pitfall. Perhaps you think that a William James education will improve your ability to gather sentimental evidence, but will lead you to slightly overvalue that evidence. Or perhaps you are unsure whether your education will be salutary or detrimental, and you grant some credence to each possibility. In either case, the William James University example can be assimilated to the Persi example. The first case corresponds to the original version of the Persi example, where there is no observable difference between Persi when he's being reliable and Persi when he's being unreliable. In this case, you are certain that if you attend William James University (or talk to Persi), you will acquire evidence which is useful but misleading. The second case corresponds to the modified version of the Persi example, where Persi talks faster when he's guessing than when he's reporting the truth. In this case, you believe that attending William James University (or talking to Persi) will either constitute an unadulterated epistemic gain or an unadulterated epistemic loss, though you're not sure which. In both cases, you violate clause (i), so you are not required to set $Cr_0(G|Cr_1(G) = .90)$ equal to .90.

2.9 Back to the Dutch Book

I claim that Qualified Reflection is a norm of coherence, but Reflection is not. How can I reconcile this with van Fraassen's DBA, which seems to establish Reflection in its unqualified form? I will argue that the DBA conflates two types of epistemic problems: incoherence and what I will call *self-doubt*. An agent is self-doubting just in case she believes or suspects that she has incoherent beliefs.

We can contrast incoherence and self-doubt using full belief examples. Someone who believes a proposition A and simultaneously believes its negation $\neg A$ is guilty of incoherence, and we might criticize her on the grounds that there is no possible world where all her beliefs are true. Someone who believes that she believes both A and $\neg A$, however, is self-doubting. She is not incoherent—there are possible worlds where she believes both A and $\neg A$. She may even have good evidence that she inhabits one of these worlds. Still, there is something problematic about her overall belief state: there is a sense in which she can't be right about everything. Either she is mistaken about whether she believes both A and $\neg A$, or she is mistaken about whether A is the case (since she believes both A and $\neg A$, and must be mistaken in one of these beliefs).

Self-doubt bears a close resemblance to Moore's paradox, in which an agent

believes a proposition A while simultaneously believing that she does not believe A .⁸ Just like a self-doubting agent, an agent with Moore's-paradoxical beliefs is guaranteed to be wrong about something, even though her beliefs are perfectly coherent. The analogy will prove useful: one of the lessons Moore draws from his paradox can be adapted to the case of self-doubt.

Moore [1902, 132] writes,

It is often pointed out that I cannot at any given moment distinguish what is true from what I think so: and this is true. But though I cannot distinguish what is true from what I think so, I can always distinguish what I mean by saying that it is true from what I mean by saying that I think so. For I understand the meaning of the supposition that what I think true may nevertheless be false.

Moore can be understood as proposing a test for distinguishing incoherent beliefs from what we might call Moore's-paradoxical beliefs. If it is incoherent to believe both A and B , then it is equally incoherent to suppose both A and B at the same time and in the same context.⁹ But if it is merely Moore's-paradoxical to believe both A and B , then it is perfectly coherent to believe both A and B at the same time and in the same context. Self-doubting beliefs are like Moore's-paradoxical beliefs: there is nothing odd or contradictory in supposing that I hold contradictory beliefs. After all, there is some possible world in which I do.

When formulating Dutch books for partial believers, we can run a version of Moore's suppositional test. Say that someone 'wins' a bet on A at a possible world w just in case A is true at w , and say that she 'loses' a bet on A at w just in case A is false at w —whether or not any bets are made at w . Say that someone 'wins' a conditional bet on A given B at w just in case A and B are both true at w , and say that she 'loses' a conditional bet on A given B at w just in case $\neg A$ and B are both true at w —again, whether or not any bets are made at w . A set of bets reveals incoherence just in case at every possible world, the buyer (or the seller) of those bets loses more than she wins. But a set of bets counts as a Dutch book just in case at every possible world where the agent's beliefs condone the bets, the buyer (or the seller) of those bets loses more than she wins. So every set of bets that reveals incoherence counts as a Dutch book, but not every Dutch book reveals incoherence.

We can illustrate the difference between the two types of Dutch book using a pair of synchronic examples. It is incoherent to violate Finite Additivity by

⁸ My exposition of Moore's paradox differs from Moore's more explicit presentations [Moore, 1944, 1952] both in form and in function. Moore presents the paradox as involving an agent who asserts a conjunction of the form, ' A and I do not believe that A '. I am interested in agents who believe pairs of propositions of the form ' A ' and 'I do not believe that A '. Moore intends to make a point about the distinction between what a speaker says and what she implies; I intend to make a point about higher-order belief. Despite these significant differences, Moore's insights shed considerable light on the issues at hand.

⁹ There may be special cases in which it is a good idea to make incoherent suppositions—for instance, cases where one intends to perform a *reductio* or argue that everything follows from a contradiction. All I need is the distinction between coherent and incoherent suppositions.

letting $Cr(A \vee B) = Cr(A) + Cr(B) + x$ for some disjoint propositions A and B and some positive real number x . It is self-doubting to suspect that one violates Finite Additivity by letting $Cr(Cr(A \vee B) = Cr(A) + Cr(B) + x) = y$ for some disjoint propositions A and B and some positive real numbers x and y . Both problems render agents susceptible to Dutch books, but the suppositional test reveals that the Dutch books are of different types.

First, consider an agent with an incoherent credence function Cr such that $Cr(A \vee B) = Cr(A) + Cr(B) + x$ for some disjoint propositions A and B and some positive real number x . Cr condones buying or selling each of the following bets:

$$\begin{array}{ll} A \vee B & \$1 - Cr(A \vee B) \\ \neg(A \vee B) & \$ - Cr(A \vee B) \end{array} \quad (2.7)$$

$$\begin{array}{ll} A & \$Cr(A) - 1 \\ \neg(A) & \$Cr(A) \end{array} \quad (2.8)$$

$$\begin{array}{ll} B & \$Cr(B) - 1 \\ \neg(B) & \$Cr(B) \end{array} \quad (2.9)$$

No matter what happens next, the buyer of bets 2.7-2.9 is guaranteed a net loss. The buyer pays a total of $\$(Cr(A \vee B) - (Cr(A) - 1 + Cr(B) - 1)) = \$(1 + x)$, and wins exactly $\$1$, for a net loss of $\$x$. Furthermore, the buyer suffers this net loss at every possible world, since at every possible world, exactly one of $\{A, B, A \vee B\}$ is true.

Contrast the self-doubting agent whose credence function Cr is such that $Cr(Cr(A \vee B) = Cr(A) + Cr(B) + x) = y$ for some disjoint A and B and some positive real numbers x and y . Let I be the proposition that $(Cr(A \vee B) = Cr(A) + Cr(B) + x)$. Then the self-doubting agent's credences condone buying or selling the following bet:

$$\begin{array}{ll} I & \$(1 - y)x \\ \neg I & \$(-y)x \end{array} \quad (2.10)$$

Suppose, as might happen, that $Cr(A|I) = Cr(A)$, $Cr(B|I) = Cr(B)$, and $Cr(A \vee B|I) = Cr(A \vee B)$. Then the self-doubting agent's credences also condone buying or selling *conditional* versions of bets 2.7-2.9 which take place only on the condition that I . But together with bet 2.10, the conditional versions of bets 2.7-2.9 constitute a Dutch book. If I is false, then the buyer loses $\$yx$ on bet 2.10 and wins back nothing on the other bets; otherwise, she loses $\$x$ on bets 2.7-2.9 and wins back only $\$(1 - y)x$ on bet 2.10.

In the Dutch book against the self-doubting agent, however, there are possible worlds where the buyer does not suffer a net loss. Suppose $Cr(A \vee B) = Cr(A) + Cr(B)$. Then at the actual world, a buyer of bet 2.10, together with the conditional versions of bets 2.7-2.9, will suffer a net loss. But at counterfactual worlds where I is true, the buyer will win $\$(1 - y)x$ on bet 2.10, while her wins and losses on the conditional versions of bets 2.7-2.9 will cancel each other out. So this Dutch book reveals not incoherence, but self-doubt.

The Dutch book against agents who violate conditionalization reveals diachronic incoherence. I'll accept Lewis's assumption that the agent stands no chance of mistaking her evidence, so that if the agent learns E , the suppositional worlds must be ones where E is true. Under this constraint, every suppositional world ensures a net loss for either the buyer or the seller of the bets (depending on whether $Cr_E(A)$ is greater than or less than $Cr(A|E)$).

On the other hand, the Dutch book against agents who violate Reflection reveals diachronic self-doubt. At a world where the agent makes bets 2.4 and 2.2, she is guaranteed to suffer a net loss. But as long as she doesn't make bet 2.3, there are counterfactual worlds where she wins more than she loses. At those counterfactual worlds, her beliefs would have condoned different bets—in particular, they would have condoned bet 2.3. So there is no possible world where she makes bets 2.1 and 2.2 and still wins more than she loses.

According to the suppositional test, then, violating conditionalization is a type of incoherence, while violating Reflection is a type of self-doubt. This result makes sense. As Maher [1992, 132-133] points out, an agent who implements a shift that violates Reflection (such as taking LSQ or attending William James University) thereby violates conditionalization. But to violate Reflection is to afford positive credence to the proposition that one will implement a (specific and predictable) Reflection-violating shift—whether or not one actually does. Thus, to violate Reflection is to suspect one will fail to conditionalize—that is, to suspect oneself of diachronic incoherence.

Self-doubt needn't be objectionable—in fact, it needn't even be *prima facie* wrong. Whenever you suspect that incoherence is either advisable or inevitable, self-doubt is perfectly in order. If you decide that the mind-expanding potential of LSQ outweighs its epistemic side effects, or that Persi's testimony is worth listening to despite its tendency to mislead you, then you should choose incoherence over coherence, and you should expect to choose incoherence over coherence. If you grant some credence to worlds in which you are force-fed LSQ, threatened with undetectable memory-wiping in Shangri La, or beset with memory loss, then you should be self-doubting. There may be something epistemically problematic about getting into such situations, but there is nothing problematic about doubting yourself once you are in them.

In some cases, you know that diachronic incoherence is neither inexorable nor advisable. In these cases, Qualified Reflection tells you to set your credence in A equal to your expectation of A 's conditional credence, given your future evidence. Therefore, if you're certain that you will remain a conditionalizer, your credence in A should equal your expectation of your future credence in A . In other words, if you expect to remain diachronically coherent, you should obey Reflection.

2.10 Conclusion

Christensen was right: the DBA for Reflection is not enough to establish Reflection as a norm of coherence. Since Reflection involves higher-order beliefs, the

DBA for Reflection reveals not incoherence, but self-doubt. The DBA for conditionalization, on the other hand, involves no higher-order beliefs. Therefore, we have grounds for rejecting Reflection even if we accept conditionalization. Not all diachronic DBAs are on equal footing.

Green and Hitchcock were right. Under ideal circumstances where the agent believes that it is both possible and uniquely rational to obey a policy of conditionalization the agent should obey Reflection. Under less-than-ideal circumstances where the agent suspects she will adopt some less rational policy she need not obey Reflection. Thus, when Green and Hitchcock claim that Rational Joe is justly confident in his future abilities, they should include the ability to conditionalize on current evidence. Since we are often unsure about whether we'll be able to conditionalize on our current evidence, we should often avoid emulating Rational Joe.

Hall, Evnine, and Elga were right. Reflection applies only in cases where, roughly speaking, you wholeheartedly expect your later epistemic state to be an improvement on your earlier epistemic state, where you expect to respond to veridical evidence in a rational manner, and where you trust both your future memory and your future judgment. Qualified Reflection is a way of cashing out these intuitions.

So Reflection (properly qualified) is not as bad as it seems, and susceptibility to Dutch books does not always reveal inconsistency. As long as defenders of DBAs carefully distinguish questions about coherence from questions about self-doubt, they can safely accept the diachronic DBA for conditionalization alongside the synchronic DBAs for the Kolmogorov axioms.

Appendix on Probability Kinematics

I have assumed that evidence comes in the form of a proposition which the agent believes to degree 1. But there may be evidence that comes in the form of a shift in credence across a partition $\{B_1, B_2, \dots, B_n\}$. Jeffrey [1983] claims that in the face of such evidence, agents should update their credences in accord with the following rule:

Probability kinematics $Cr_0(A|B_i) = Cr_1(A|B_i)$ for all $B_i \in \{B_1, B_2, \dots, B_n\}$.

Later work suggests that Jeffrey was right—Skyrms [1987] provides a DBA for probability kinematics. Conditionalization is just a special case of probability kinematics where $Cr_1(B_i) = 1$ for some $B_i \in \{B_1, B_2, \dots, B_n\}$.

Suppose that an agent's evidence comes, not in the form of a proposition learned, but in the form of a *probability gradient*—a set of real numbers $\{g_1, g_2, \dots, g_n\}$, where g_i is her credence in B_i after she takes in new information. For her, we can reformulate Qualified Reflection as follows. Let $\{G_1, G_2, \dots, G_m\}$ be a set of propositions such that each G_j is a proposition of the form $g_1 \in [x_{1j}, y_{1j}] \& g_2 \in [x_{2j}, y_{2j}] \& \dots \& g_n \in [x_{nj}, y_{nj}]$, where $\{g_1, g_2, \dots, g_n\}$ is the agent's probability gradient as of t_1 . Let the G_j s be chosen so that so

that for each value of j , either $\sum_k y_{ij} Cr_0(A|B_i) < r$, $\sum_k x_{ij} Cr_0(A|B_i) = r$ and $\sum_k y_{ij} Cr_0(A|B_i) = s$, or $\sum_k x_{ij} Cr_0(A|B_i) > s$.

General Qualified Reflection $Cr_0(A|Cr_1(A) \in [r, s]) \in [r, s]$, provided that for all $B_i \in \{B_1, B_2, \dots, B_n\}$ and all $G_j \in \{G_1, G_2, \dots, G_m\}$,

$$(i) Cr_0(Cr_0(A|B_i) = Cr_1(A|B_i)) = 1 \text{ and}$$

$$(vi) Cr_0(B_i|G_j) \in [x_{ij}, y_{ij}]$$

Qualified Reflection is a special case of General Qualified Reflection (where $r = s$ and x is constrained to equal 1). General Qualified Reflection is motivated by the same considerations as Qualified Reflection: clause (i) states that the agent updates by probability kinematics, while clause (vi) states that her evidence is veridical.

General Qualified Reflection is a consequence of the probability calculus. We can prove this by assuming (i) and (v) and showing that $Cr_0(A|Cr_1(A) \in [r, s]) \in [r, s]$. Let Γ be the set of values of j such that $\sum_k x_{ij} Cr_0(A|B_i) = r$ and $\sum_k y_{ij} Cr_0(A|B_i) = s$. Then by the probability calculus,

$$\begin{aligned} Cr_0(A|Cr_1(A) \in [r, s]) &= \sum_j Cr_0(A|G_j) \\ &= \sum_j \sum_i Cr_0(A|B_i) Cr(B_i|G_j) \end{aligned}$$

By (vi),

$$\begin{aligned} \min Cr_0(A|Cr_1(A) \in [r, s]) &= \sum_{j \in \Gamma} \sum_i Cr_0(A|B_i) x_{ij} \\ \max Cr_0(A|Cr_1(A) \in [r, s]) &= \sum_{j \in \Gamma} \sum_i Cr_0(A|B_i) y_{ij} \end{aligned}$$

By (i),

$$\begin{aligned} \min Cr_0(A|Cr_1(A) \in [r, s]) &= \sum_{j \in \Gamma} \sum_i Cr_1(A|B_i) x_{ij} \\ \max Cr_0(A|Cr_1(A) \in [r, s]) &= \sum_{j \in \Gamma} \sum_i Cr_1(A|B_i) y_{ij} \end{aligned}$$

By the definition of Γ ,

$$\begin{aligned} \min Cr_0(A|Cr_1(A) \in [r, s]) &= r \\ \max Cr_0(A|Cr_1(A) \in [r, s]) &= s \end{aligned}$$

In other words,

$$Cr_0(A|Cr_1(A) \in [r, s]) \in [r, s]$$

QED.

As for the value that the agent should assign to $Cr_0(A|Cr_1(A) \in [r, s])$, we don't need the concept of departure from rationality to calculate it. It's simply:¹⁰

$$Cr_0(A|Cr_1(A) \in [r, s]) = \sum_j Cr_0(A|G_j)$$

¹⁰ I assume that $Cr_0(A|G_i)$ is well-defined for each G_i , where the G_i are selected using the interval $[r, s]$.

We can, however, still give a measure of expected departure from rationality with respect to A , on the supposition that $Cr_1(A) \in [r, s]$.

$$\mathbf{Overshoot} = \max[0, \sum_{j \in \Gamma} (\sum_i x_{ij} - Cr_0(A|G_j))]$$

$$\mathbf{Undershoot} = \max[0, \sum_{i \in \Gamma} (Cr_0(A|G_j) - \sum_i y_{ij})]$$

Intuitively, overshoot measures the degree to which the agent expects a t_1 credence in A that falls within the interval $[r, s]$ to overshoot the of appropriate t_1 credence in A . Likewise, overshoot measures the degree to which the agent expects a t_1 credence in A that falls within the interval $[r, s]$ to undershoot the appropriate t_1 credence in A . At least one of the two quantities must be zero— one can't expect to overshoot the appropriate credence and undershoot the appropriate credence at the same time. $Cr_0(A|Cr_1(A) \in [r, s]) \in [r, s]$ just in case the undershoot and overshoot are both zero.

Where the agent certain that she will update by conditionalizing and $r = s$, D_r is equal either to the overshoot (where the overshoot is positive), the negative undershoot (where the undershoot is positive), or both (where both are zero). In footnote 7, I proved that

$$D_r = r - Cr_0(A|Cr_1(A) = r)$$

Where the overshoot is greater than or equal to zero,

$$\begin{aligned} \mathbf{Overshoot} &= \sum_{j \in \Gamma} (\sum_i x_{ij} - Cr_0(A|G_j)) \\ &= \sum_{j \in \Gamma} \sum_i x_{ij} - \sum_{j \in \Gamma} Cr_0(A|G_j) \\ &= r - Cr_0(Cr_1(A) = r) \\ &= D_r \end{aligned}$$

Likewise, where the undershoot is greater than or equal to zero,

$$\begin{aligned} \mathbf{Undershoot} &= \sum_{j \in \Gamma} (Cr_0(A|G_j) - \sum_i y_{ij}) \\ &= \sum_{j \in \Gamma} Cr_0(A|G_j) - \sum_{j \in \Gamma} \sum_i y_{ij} \\ &= Cr_0(Cr_1(A) = s) - s \\ &= Cr_0(Cr_1(A) = r) - r \\ &= -D_r \end{aligned}$$

QED.

3. PUTTING A VALUE ON BEAUTY

3.1 *Stage Setting*

Adam Elga's Sleeping Beauty problem is an instance of a more general puzzle: what is the relevance of purely *de se* information to *de dicto* beliefs? In part 3.1, I explain the distinction between *de dicto* and *de se* beliefs (section 3.1.1), remind readers of the standard Bayesian story about belief updating (section 3.1.2), and summarize the debate between halfers and thirderers about how to extend the standard Bayesian story to the Sleeping Beauty case (section 3.1.3).

In part 3.2, I explain why the trouble posed by the Sleeping Beauty problem—and by *de se* information in general—is much deeper than it first appears. I build a framework that gets around the trouble by dividing agents' belief functions into a *de dicto* component and a *de se* component. Using my framework, I formulate a 'Halfer Rule' and a 'Thirder Rule'—extensions of the halfer and thirder solutions to the Sleeping Beauty problem.

The Halfer Rule and the Thirder Rule both seem appealing, but unfortunately, they are incompatible. Which is preferable? I consider two ways of resolving the question: Dutch books in part 3.3, and scoring rules in part 3.4. I argue that on both ways of resolving the conflict, whether one should prefer the Halfer Rule or the Thirder Rule depends on one's views about decision theory. Evidential decision theorists should prefer the Halfer Rule, while causal decision theorists should prefer the Thirder Rule.

In part 3.5, I argue that there is a consideration that favors the Thirder Rule over the Halfer Rule independently of any scoring considerations. The Thirder Rule is stable in a way that the Halfer Rule is not. Enriching an agent's belief worlds with irrelevant information will dramatically change the advice of the Halfer Rule. Since the Thirder Rule is closely connected to causal decision theory and the Halfer Rule is closely connected to evidential decision theory, the stability of the Thirder Rule constitutes a reason for preferring evidential decision theory to causal decision theory.

3.1.1 *The de dicto-de se distinction*

In the terminology of Lewis [1979], *de dicto* beliefs concern only what the world is like. A person's *de dicto* beliefs might include the belief that the meek will inherit the Earth, the belief that flossing prevents tooth decay, or the belief

that a Republican won the 1992 presidential election.¹ If two inhabitants of the same possible world have the same *de dicto* beliefs, they are either both wrong or both right. *De se* beliefs, by contrast, concern the believer's location (in addition to what the world is like).² A person's *de se* beliefs might include the belief that today is Tuesday, the belief that her shopping cart contains a leaky bag of sugar [Perry, 1971], or the belief that she will be devoured by ravenous dogs [Quine, 1969]. It's possible for two inhabitants of the same world to have the same *de se* beliefs, but for one to be right and the other to be wrong; Napoleon might truly believe, and I might falsely believe, the *de se* content expressed by the sentence "I am Napoleon". A *de dicto* belief can be represented using the proposition that its content is true—i.e., the set of possible worlds where its content is true. A *de se* belief cannot.

Why not? Consider the example of Rudolph Lingens, an amnesic lost in the Main Library at Stanford [Perry, 1974]. Lingens happens upon a biography of himself, and discovers all sorts of interesting *de dicto* facts about Lingens. We might suppose (embellishing Perry's story) that the book was written by an oracle, and contains the sentence "Lingens is in aisle five, floor six, Main Library, Stanford, on October 1, 1977". But no matter how much *de dicto* information Lingens acquires by reading the book, he is missing the crucial *de se* information that would allow him to find his way out of the library. Until he learns that *he* is Rudolph Lingens, and that *today* is October 1, 1977, he is lost.

The contents of *de se* beliefs can be represented using *centered propositions*, or sets of centered worlds—where a centered world is an ordered pair consisting of a world *W* and a *center*, or the spatiotemporal location of an individual, in *W*.³ (In the remainder of the paper, I will refer to the worlds and propositions

¹ Whether these beliefs are truly *de dicto* is open to debate. Perhaps 'the Earth' really means something like 'the Earth-like planet inhabited by *me*'. If so, then the belief that the meek will inherit the Earth is irreducibly *de se*. In a universe with two very similar planets, both called Earth, I might believe truly that the meek will inherit the Earth (because the meek will inherit the Earth-like planet that I inhabit), while my twin on the other planet might believe falsely that the meek will inherit the Earth (because the meek will not inherit the Earth-like planet that she inhabits). It is hard to formulate an English sentence that uncontroversially expresses a *de dicto* claim. But if there are no purely *de dicto* beliefs, then this is all the more reason to attend to the epistemic role of *de se* evidence.

² Whether these beliefs are really *de se* is open to debate. Millikan [1990] claims that there are no self-locating contents, only different roles that a belief might play in the believer's mental life. Millikan objects to one reason for positing self-locating beliefs—the theory that an agent's self-locating beliefs explain her ability to act on her other beliefs. Millikan's argument is persuasive, but it does not establish that there are no self-locating beliefs. Rather, it establishes that self-locating beliefs are not sufficient to explain an agent's ability to act on her beliefs. It may be that irreducibly *de se* beliefs explain some (but not all) of the behavior that *de dicto* beliefs cannot account for alone.

³ Are spatiotemporal locations really enough to capture the intuitive concept of a center? To pre-empt any mischief on the part of my readers, I'll go ahead and formulate the counterexample from hell: Suppose some sorcerer turns me into an extended ghost that can superimpose itself on ordinary matter. The sorcerer then packs me into her convenient time machine and sends me back to my own past. There, I superimpose my ghostly body on my older, more substantial body. My 'later', ghostly self experiences different thoughts and qualia than my 'earlier', physical self. There seem to be two centers here, although there is only one individual (me) who occupies one spatiotemporal location. One might refine my rough-and-ready

used to represent *de dicto* belief as *uncentered*. Every uncentered proposition A is equivalent to some centered proposition—roughly, the set of centers located in worlds where A is true—but not every centered proposition is equivalent to an uncentered proposition.)

We can use the terms ‘*de dicto*’ and ‘*de se*’ to describe properties of overall belief states, as well as properties individual beliefs. Say that a person suffers from irreducibly *de se* ignorance just in case some of her doxastically possible worlds contain more than one doxastically possible center. When someone suffers from irreducibly *de se* ignorance, uncentered worlds are too coarse-grained to distinguish among her doxastic alternatives. Her epistemic state is best represented using centered worlds. Say that someone’s ignorance is purely *de dicto* just in case she is ignorant, but her ignorance is not irreducibly *de se*.

One might read Perry’s Lingens story as an example of someone suffering irreducibly *de se* ignorance. For all Lingens knows, there are two people lost in the library who satisfy all of his *de se* beliefs. For all he knows, these people are facing indistinguishable shelves of books, thinking indistinguishable thoughts, and wearing shoes of indistinguishable colors. Described this way, Lingens’ ignorance is irreducibly *de se*: some of his doxastically possible worlds contain two doxastically possible centers.

Note that not all ignorance about one’s identity or location is irreducibly *de se*. In order for Lingens’ ignorance to be irreducibly *de se*, it is not sufficient for him to be unsure whether he is Rudolf Lingens or Bernard J. Ortcutt. If Lingens is unsure whether he is Rudolf Lingens or Bernard J. Ortcutt, believes that he is lost in a library, and believes that exactly one person is lost in the library, then his ignorance can be characterized in purely *de dicto* terms. He has some doxastic alternatives where Lingens is lost in the library (and Ortcutt is not), and some where Ortcutt is lost in the library (and Lingens is not). These alternatives can easily be represented by uncentered worlds.

3.1.2 De dicto Confirmation Theory

Standard Bayesian confirmation theory is built to cope with *de dicto* ignorance, rather than irreducibly *de se* ignorance. It posits credences that attach to uncentered propositions. Updating is bound by the following norm (where Cr is the agent’s original credence function, E is her new evidence, Cr_E is her credence function updated on E , and A is any proposition):

$$\text{Conditionalization: } Cr_E(A) = Cr(A|E) = \frac{Cr(A \& E)}{Cr(E)}$$

A and E are assumed to be uncentered propositions.

Conditionalization is a useful rule in most situations—but not in all. Sometimes, conditionalization is too restrictive, as in cases of memory loss. If an

definition either by adding in a clause about the unity of qualia, or by quibbling about personal timelines in the definition of ‘spatiotemporal location’. For the purposes of this chapter, characterizing centers as spatiotemporal locations will work well enough.

agent forgets a proposition A , she cannot possibly update by conditionalization. Her old credence in A conditional absolutely any proposition was 1, but if she forgets whether A , then her new credence in A is less than one.

At other times, conditionalization is not restrictive enough, as in cases where the agent learns something incompatible with her old beliefs. Where $Cr(E)$ is 0, the ratio $\frac{Cr(A \& E)}{Cr(E)}$ is undefined. Mathematically speaking, it's easy to pick out a conditional credence function $Cr(\cdot|E)$ which is equal to $\frac{Cr(A \& E)}{Cr(E)}$ where $E > 0$ and well-defined even where $Cr(E) = 0$. The trouble is that there *many* possible ways to do so. If $Cr(E) = 0$, then as long as A doesn't entail E , the agent's real-valued credences allow $Cr(A|E)$ to be anything. Unless we have some ways of narrowing down the range of acceptable conditional credence functions, conditionalization doesn't amount to a real constraint.

One more note on the decision-theoretic apparatus: Throughout this essay, I will assume that the domains of agents' credence functions can be represented using finitely many possible worlds. I'll be treating possible worlds not as maximally specific (descriptions of) states of affairs, but as (descriptions of) states of affairs that are sufficiently rich for the theoretical purposes at hand.

3.1.3 Sleeping Beauty

Elga [2000] has argued that irreducibly *de se* ignorance generates exceptions to *de dicto* conditionalization. He supports this claim with the following *Sleeping Beauty* example [2000, 143], addressing the protagonist, whom I will call 'Beauty', in the second person:

Some researchers are going to put you to sleep. During the two days that your sleep will last, they will briefly wake you up either once or twice, depending on the toss of a fair coin (Heads: once; Tails: twice). After each waking, they will put you back to sleep with a drug that makes you forget that waking.

When you are first awakened, to what degree ought you believe that the outcome of the coin toss is heads?

The first waking will occur on Monday, and the second, if it occurs at all, will occur on Tuesday. We can assume that the coin flip occurs after the Monday waking and before the Tuesday one. When Beauty wakes, she will know neither her temporal location in the world (i.e., what day it is) nor what the world is like (i.e., whether the coin lands heads or tails). Her overall state of opinion on awakening will thus be a credence function defined over a set of centered propositions.

Elga claims that Beauty should place credence 1/3 in the proposition that the coin lands heads. More specifically, her credence function on waking up (which I will call Cr^{up}) should assign the following values to centered propositions:

	<i>Heads</i>	<i>Tails</i>
<i>Monday</i>	1/3	1/3
<i>Tuesday</i>	0	1/3

Each column in the table represents an uncentered proposition, while each entry represents a centered proposition. I will follow Elga's scheme for naming these centered propositions: H_1 = Monday, heads; T_1 = Monday, tails; and T_2 = Tuesday, tails. *Heads* is the uncentered proposition that the coin lands heads; *Tails* is the uncentered proposition that the coin lands tails. Likewise, *Monday* is the centered proposition that it is Monday; *Tuesday* is the centered proposition that it is Tuesday. Thus, according to Elga $Cr^{up}(H_1) = Cr^{up}(T_1) = Cr^{up}(T_2) = 1/3$. Call the adoption of this credence function *thirring* and those who support it *thirders*.

Elga's argument for thirring is as follows. First, he claims that

$$Cr^{up}(T_1) = Cr^{up}(T_2) \quad (3.1)$$

3.1 is a consequence of the following highly restricted indifference principle:

Weak Indifference If two centered worlds W_1 and W_2 are subjectively indistinguishable, and the same propositions are true in both, then W_1 and W_2 should receive equal credence.

Next, let Cr^+ be the credence function that it would be rational for Beauty to adopt after waking up and coming to believe Monday. Then

$$Cr^+(H_1) = Cr^+(T_1) = 1/2 \quad (3.2)$$

Why so? We can imagine that Beauty knows the coin will not be flipped until after she forms Cr^+ . When she forms Cr^+ , she learns that the coin (which is fair), has not yet been flipped, so she should assign equal credence to H_1 and T_1 . Since Beauty should conditionalize on her total evidence, says Elga,

$$Cr^{up}(H_1|Monday) = Cr^{up}(T_1|Monday) = 1/2 \quad (3.3)$$

Together, 3.1 and 3.3 fix the probabilities of all the cells in the table.

If Elga is right, then the *de dicto* version of conditionalization is wrong. According to Elga, Beauty's Sunday credence in *Heads* should be 1/2, while her Monday credence in *Heads* should be 1/3. But Beauty gains no uncentered evidence between Sunday and Monday. According to conditionalization, her Monday credence in *Heads* should be 1/2 rather than 1/3.

Contra Elga, Lewis [2001] claims that the correct value of Cr^{up} is:

	<i>Heads</i>	<i>Tails</i>
<i>Monday</i>	1/2	1/4
<i>Tuesday</i>	0	1/4

Call the adoption of this credence function *halving* and those who support it *halfers*.

Lewis argues for halving on the grounds that Beauty acquires no new information between Sunday and Monday. Arntzenius [2002] points out that Lewis should invoke a second premise—that Beauty loses no information between Sunday and Monday. A halfer might support both premises by pointing out that Beauty neither loses nor gains *de dicto* information between Sunday and Monday. If any change in her *de se* evidence is irrelevant to her beliefs about the coin toss, then she ought to half.

Oddly enough, Lewis grants that *de se* evidence is sometimes relevant to *de dicto* beliefs. He claims that

$$Cr^{up}(H_1|Monday) = 2/3 \quad (3.4)$$

Like Elga, Lewis believes that Beauty should conditionalize on her total evidence. So once she learns that it's Monday,

$$Cr^+(H_1) = 2/3 \quad (3.5)$$

As we will see in part 3.2, halfers need not agree to 3.5, or grant that *de se* evidence is ever relevant to *de dicto* beliefs. In fact, I will argue in section 3.2.1 that they should do neither of these things.

3.2 Rules for Halfers and Thirders

Elga and Lewis disagree about how Beauty should update her beliefs when she receives new, irreducibly *de se* evidence, but they agree that she should conditionalize on her total evidence. Bostrom [2007] and Meacham [forthcoming] have pointed out that this shared assumption is misguided.

I argued in section 3.1.2 that conditionalization is useful only when the agent's new evidence is compatible with her old beliefs. But even routine *de se* updating involves new *de se* evidence that is incompatible with the agent's old *de se* beliefs. To borrow an example from Meacham [forthcoming], suppose that I start out believing that it is noon, then glance at my clock and realize that it is 12:01. My new belief that it is 12:01 is incompatible with my old belief that it is noon.

Several authors suggest sophisticated ways around this problem. Meacham [forthcoming] uses a 'hypothetical prior' which assigns probabilities to centered worlds outside the space of the agent's doxastic alternatives. Titelbaum [forthcoming] proposes a system where sentences, rather than propositions, receive credences. I suggest a simpler solution: we can just divide the agent's belief state into one part that represents her uncentered information, and a second part that represents her additional centered information.

The first, uncentered part of the agent's belief state is a credence function Cr_u that ranges over uncentered propositions. We can think of Cr_u as the function the agent would endorse as the correct 'view from nowhere', given her priors and her evidence. The second, centered part of the agent's belief state is a function that takes each doxastically possible uncentered world W to a natural number N_W equal to the number of doxastically possible centers in W .

Like the agent's belief state, any new evidence that the agent acquires can be divided into centered and uncentered parts. (Throughout this chapter, I'll assume that evidence comes in the form of a centered proposition the agent learns with certainty.) The uncentered part of a centered evidence proposition E is the strongest uncentered proposition entailed by E —the set of all and only the uncentered worlds that contain centers in E . (I'll call this proposition $u(E)$) Cr_u can then be updated by conditionalizing on $u(E)$ when the agent learns E . The centered part of E is a new function taking each world W to a natural number N'_W —the number of centers in W after the agent has updated.

Let $Cr_{\textcircled{a}}$ be the agent's actual credence function. What is the appropriate relationship between Cr_u and $Cr_{\textcircled{a}}$? Two answers suggest themselves.

3.2.1 The Halfer Rule

We might claim $Cr_{\textcircled{a}}$ should simply coincide with Cr_u for all uncentered propositions. In other words, we might espouse the

$$\textbf{Halfer Rule: } Cr_{\textcircled{a}}(A) = Cr_u(A)$$

The Halfer Rule has been endorsed (albeit in different notation) by Halpern [2005] and Meacham [forthcoming].

As its name suggests, the Halfer Rule entails that Beauty should half. But unlike Lewis's version of halving, the Halfer rule is incompatible with Lewis's equation 3.5. Lewis espoused 3.5 because he believed that Beauty should update by conditionalizing on her total evidence. But as we've seen, conditionalization is useless for *de se* updating. The Halfer Rule recommends conditionalizing the *uncentered portion* of one's credence function on the *uncentered portion* of one's total evidence, and then within each world, dividing one's credence among the doxastically possible centers. Since Beauty gains no relevant uncentered evidence when she learns that it's Monday, the Halfer Rule entails Elga's equation 3.2.

It's just as well that the Halfer Rule leads us to reject 3.5. Draper and Pust [2008] formulate a Dutch book against agents who satisfy 3.5. (For reasons that will become clear in section 3.3.2, their Dutch book does *not* automatically carry over to agents who satisfy 3.4.)

3.2.2 The Thirder Rule

We might embrace the following alternative to the Halfer Rule, where A is any uncentered proposition, W and W^* are variables ranging over doxastically possible uncentered worlds, and N_W is the number of centers in W :

$$\textbf{Thirder Rule: } Cr_{\textcircled{a}}(A) = \frac{\sum_{W \in A} Cr_u(W)N_W}{\sum_{W^*} Cr_u(W^*)N_{W^*}}$$

Elga [2007] endorses the Thirder Rule (albeit in different notation).

You might think of the Thirder Rule (a bit fancifully) as recommending that you follow this procedure: First, use the *de dicto* portion of your evidence to

calculate Cr_u , the credence which it would be appropriate to place in each doxastically possible uncentered world W if you occupied the ‘view from nowhere’. Next, use the *de se* portion of your evidence to increase your credence in worlds with multiple centers. If a world contains n centers, then n -tuple your credence in it. Finally, re-normalize your credences so that they sum to 1.

Of course, the Thirder Rule doesn’t *really* recommend following the above procedure—it merely recommends adopting the credences you would arrive at if you followed the procedure, but it doesn’t say how you should arrive at them. As its name suggests, the Thirder Rule entails that Beauty should third.

Notice that although the Thirder Rule disagrees with the Halfer Rule about what Beauty should do upon waking up, it agrees with the Halfer Rule about what she should do upon waking up and learning that it’s Monday. Once Beauty learns that it’s Monday, she has only one doxastically possible center per doxastically possible uncentered world. Like the Halfer Rule, the Thirder Rule entails Elga’s equation 3.2.

3.2.3 Weak Indifference

Neither the Halfer Rule nor the Thirder Rule tells agents how to divide their credences among multiple centers within uncentered worlds. Halfers and thirders might enrich their accounts by espousing Elga’s Weak Indifference principle. Weak Indifference seems to correctly capture an important symmetry in cases of *de se ignorance*: when two centered worlds are subjectively indistinguishable, and when the same propositions are true in both, it’s hard to see any reason for granting one greater credence than the other.

On the other hand, Weak Indifference has worrisome consequences. Elga [2004] points out that someone who satisfies Weak Indifference, upon learning that the world contains subjective duplicates of her, should become very confident that she herself is one of the duplicates. This consequence of Weak Indifference has an important practical upshot—if you are convinced that the world contains subjective duplicates of you, you should be highly concerned for those duplicates’ welfare. For all you know, any of the duplicates may be you.

Weatherson [2005] points out that when an agent has doxastically possible worlds containing a countable infinity of subjective duplicates, Weak Indifference is incompatible with countable additivity. There is no additive way to distribute credence evenly over countably many individuals. One might say, ‘So much the worse for countable additivity!’, but this would be, at the least, a substantial commitment. (Weatherson spends part of his article detailing some of its counterintuitive consequences.)

Finally, Weatherson [2005] suggests that Weak Indifference conflates uncertainty with risk. An agent who is unsure of her spatiotemporal location may not be in a position to assign any precise credence to propositions about her spatiotemporal location. Perhaps she should treat all centers in the same uncentered world equally by being equally vague about the credences she assigns them, rather than by assigning them equal credence.

Halfers and thirders who find Weak Indifference intuitively appealing, but

are worried by these problems, might get around them by adopting a weaker version of Weak Indifference.

Weaker Indifference If two centered worlds W_1 and W_2 are subjectively indistinguishable, the same propositions are true in both, and W_1 and W_2 both receive precise nonzero credence, then W_1 and W_2 should receive equal credence.

Weaker Indifference does not entail Elga's counterintuitive corollary, because it says nothing at all about when to assign nonzero credence to a centered world. For all that Weaker Indifference says, in a world that contains duplicates of you, you should assign zero credence to the proposition that you are one of the duplicates. It avoids Weatherson's conflict with countable additivity, because in worlds that contain a countable infinity of subjective duplicates of you, you can assign credence one to the proposition that you belong to some (particular) countable subset of the set of duplicates. And it avoids Weatherson's third worry by permitting you to have vague credences.

3.3 Dutch Books

Two authors [Hitchcock, 2004, Halpern, 2005] suggest that we adjudicate between the Halfer Rule and the Thirder Rule using Dutch books. Oddly enough, they come up with diametrically opposed results. Hitchcock claims that Dutch book considerations favor a thirder approach, while Halpern claims that they favor a halfer approach.

Who's right? The answer depends on which decision theory we adopt. I'll argue that causal decision theorists should be thirders, while evidential decision theorists should be halfers. I'll begin in section 3.3.1 by articulating an important and sometimes-overlooked constraint on what counts as a Dutch book, and illustrating the importance of the constraint by showing that Halpern's putative Dutch book [2005] violates it.

In section 3.3.2, I'll explain Hitchcock's 2004 Dutch book against halfers. I'll argue that Hitchcock's Dutch only works if Beauty is a causal decision theorist. In section 3.3.3, I argue that if Beauty is an evidential decision theorist who thirds, she is vulnerable to a Dutch book (though not the one suggested by Halpern). At this point, I will have given Dutch book arguments against halving (if you're a causal decision theorist) and thirding (if you're an evidential decision theorist).

In the remainder of part 3.3, I will prove a pair of converse Dutch book results: causal decision theorists who obey the Thirder Rule and evidential decision theorists who obey the Halfer Rule are immune to Dutch books. In section 3.3.4, I'll establish a claim that serves as an important part of both converse Dutch books: anyone who bets at what I call *thirder odds* is immune to a Dutch book. In section 3.3.5, I'll argue that causal decision theorists who obey the Thirder Rule, and evidential decision theorists who obey the Halfer Rule, should bet at thirder odds. Putting these two results together yields

the conclusion that causal decision theorists who obey the Thirder Rule, and evidential decision theorists who obey the Halfer Rule, are immune to Dutch books.

3.3.1 A Constraint on Dutch Books

A Dutch book is a set of bets that an agent is willing to accept individually (where an agent's willingness to accept a bet cashed out in terms of the bet's having positive expected value), but that jointly result in a sure loss. If a set of bets is to count as a Dutch book, however, the agent must be willing to accept the bets even when she is fully informed about their nature. For instance, if you believe to degree $1/2$ that it will rain tomorrow, it is no objection to your credence function that a bookie could exploit you by selling you the following 'bet'.

$$\begin{array}{ll} \text{a ticket that costs \$5 and pays \$10 in the case of rain} & \text{if it will not rain} \\ \text{a ticket that costs \$5 pays and \$10 in the case of no rain} & \text{if it will rain} \end{array} \quad (3.6)$$

Bet 3.6 isn't a Dutch book because your willingness to accept it depends on your ignorance of the betting conditions. If you knew that you would be offered the no-rain side of the bet in case of rain, and the rain side in case of no rain, you wouldn't take the bet. It's not your incoherence that bet 3.6 is punishing, but your ignorance.

Authors sometimes phrase the no-deception requirement in terms of whether the bookie knows anything that the buyer does not, but this is misleading. The bookie's epistemic state is beside the point—after all, a bookie could deceive an agent unwittingly. What matters is whether the bookie's behavior matches up with the agent's expectations.

In cases of irreducibly *de se* ignorance, it's easy to run afoul of the no-deception requirement. Halpern [2005] claims that if Beauty thirds, she is vulnerable to Lewis's 1999 Dutch book against agents who violate *de dicto* conditionalization. On closer examination, Halpern's extension of Lewis's Dutch book turns out to require deception.

On Sunday, a bookie might sell Beauty the following bet.

$$\begin{array}{ll} \text{Heads} & \$(15 + \epsilon) \\ \text{Tails} & \$(-15 + \epsilon) \end{array} \quad (3.7)$$

When she wakes up on Monday, the bookie might sell her

$$\begin{array}{ll} \text{Heads} & \$(-20 + \epsilon) \\ \text{Tails} & \$(10 + \epsilon) \end{array} \quad (3.8)$$

Together, bets 3.7 and 3.8 seem to constitute a Dutch book. If the coin lands heads, Beauty wins $\$(15 + \epsilon)$ on bet 3.7, but loses $\$(20 - \epsilon)$ on bet 3.8, and if the coin lands tails, she loses $\$(15 - \epsilon)$ on bet 3.7 and wins back only $\$(10 + \epsilon)$ on bet 3.8. Either way, she loses a total of $\$(5 - 2\epsilon)$.

But Hitchcock [2004] points out that this setup isn't a Dutch book at all. What if the coin lands tails? Does the bookie offer bet 3.8 only once, on Monday, or does he offer it twice, once on Monday and once on Tuesday? If the bookie offers bet 3.8 only on Monday, then Beauty should refuse to bet. Once she learns that the bookie has approached her, she'll know that it's Monday, and her credence in *Heads* will be $1/2$ —meaning that bet 3.8 has negative expected value. On the other hand, if the bookie offers 3.8 on Tuesday as well, then Beauty will end up $\$(5 + 2\epsilon)$ richer when the coin lands tails.

The only way the bookie can trap Beauty into a sure loss is by deceiving her. If Beauty is convinced that she'll make bet 3.8 on Monday and Tuesday, is offered bet 3.8 only on Monday, then she's guaranteed to lose money. But since this scenario involves deception about the nature of the bets, it's not a Dutch book.

3.3.2 Hitchcock's Dutch book

Hitchcock suggests a Dutch book against halfers that seems to avoid the deception pitfall described above. After phrasing the no-deception requirement in terms of the bookie's epistemic state, he writes:

There is one way in which the bookie can ensure that he has no information that is unavailable to Beauty: he can sleep with her. That is, he can place his first bet, go into a deep sleep when Beauty does, arrange to have himself awakened under the same protocol as Beauty, and sell a follow-up bet to Beauty whenever they wake up together. The bookie, like Beauty, will awaken having no idea whether it is the first or second awakening, having no idea whether an initial follow-up bet has already been placed. Thus he must sell the same bet to Beauty whenever they both wake up.

On Hitchcock's way of individuating bets, the same bet is offered at the same stakes in all subjectively indistinguishable centers. We can consider this a stipulation about how to use the word 'bet'. (It's not the only acceptable way to individuate bets, but it's an acceptable way. The important thing is that Beauty know what the setup is.)

Hitchcock then argues that Beauty is vulnerable to the following Dutch book if she halves. On Sunday, the bookie offers Beauty:

$$\begin{array}{ll} \textit{Heads} & \$(-15 + 2\epsilon) \\ \textit{Tails} & \$(15 + \epsilon) \end{array} \quad (3.9)$$

On Monday (and on Tuesday, if she wakes up), he offers her:

$$\begin{array}{ll} \textit{Heads} & \$(10 + \epsilon) \\ \textit{Tails} & \$(-10 + \epsilon) \end{array} \quad (3.10)$$

Together, bets 3.9 and 3.10 result in a sure loss of $\$(5 - 3\epsilon)$ for Beauty. If the coin lands heads, she loses $\$(15 - 2\epsilon)$ on bet 3.9, and wins back only $\$(10 +$

ϵ) on bet 3.10, which is made only once (on Monday). If the coin lands tails, she wins $\$(15 + \epsilon)$ on bet 3.9, but loses $\$(20 - 2\epsilon)$ on bet 3.10, which is made twice (on Monday and on Tuesday).

But will Beauty really accept bet 3.10? Adopting a thirder's credences doesn't necessarily translate to betting at a thirder's. Arntzenius [2002], who anticipates a version of Hitchcock's Dutch book, points out that that Beauty will only accept bets like 3.10 if she is a causal decision theorist as well as a halfer. If Beauty is an evidential decision theorist, she won't think bet 3.10 is worth accepting.

Arntzenius supports his point by drawing an analogy between *Sleeping Beauty* and a two-person prisoner's dilemma game. But no such analogy is necessary; we can see that Arntzenius is right by carefully going through the expected utility calculations. Suppose that Beauty has just awoken, that she is unsure whether it's Monday or Tuesday, and that she is deciding whether to accept bet 3.10. (Call the centered proposition that she accepts bet 3.10 right now *Accept*, and call the centered proposition that she accepts the same bet on a different day *D*.) The evidential expected values of *Accept* and \neg *Accept* are

$$\begin{aligned} V_E(\textit{Accept}) &= Cr(\textit{Heads} \& D | \textit{Accept})V(\textit{Heads} \& D \& \textit{Accept}) \\ &\quad + Cr(\textit{Heads} \& \neg D | \textit{Accept})V(\textit{Heads} \& \neg D \& \textit{Accept}) \\ &\quad + Cr(\textit{Tails} \& D | \textit{Accept})V(\textit{Tails} \& D \& \textit{Accept}) \\ &\quad + Cr(\textit{Tails} \& \neg D | \textit{Accept})V(\textit{Tails} \& \neg D \& \textit{Accept}) \end{aligned}$$

$$\begin{aligned} V_E(\neg \textit{Accept}) &= Cr(\textit{Heads} \& D | \neg \textit{Accept})V(\textit{Heads} \& D \& \neg \textit{Accept}) \\ &\quad + Cr(\textit{Heads} \& \neg D | \neg \textit{Accept})V(\textit{Heads} \& \neg D \& \neg \textit{Accept}) \\ &\quad + Cr(\textit{Tails} \& D | \neg \textit{Accept})V(\textit{Tails} \& D \& \neg \textit{Accept}) \\ &\quad + Cr(\textit{Tails} \& \neg D | \neg \textit{Accept})V(\textit{Tails} \& \neg D \& \neg \textit{Accept}) \end{aligned}$$

I'll assume Beauty is certain that if she wakes twice, she will make the same bets both times. This assumption is highly plausible. Beauty's epistemic situation on when she is offered a bet Monday is exactly like her situation when she is offered a bet on Tuesday. Presumably the same dispositions that influence her Monday decision also influence her Tuesday decision. Using the assumption, we can simplify the above expected value calculations:

$$\begin{aligned} V_E(\textit{Accept}) &= Cr(\textit{Heads} | \textit{Accept})V(\textit{Heads} \& \neg D \& \textit{Accept}) \\ &\quad + Cr(\textit{Tails} | \textit{Accept})V(\textit{Tails} \& D \& \textit{Accept}) \\ &= Cr(\textit{Heads} | \textit{Accept})(10 + \epsilon) \\ &\quad + Cr(\textit{Tails} | \textit{Accept})(-20 + 2\epsilon) \end{aligned}$$

$$\begin{aligned} V_E(\neg \textit{Accept}) &= Cr(\textit{Heads} | \neg \textit{Accept})V(\textit{Heads} \& \neg D \& \neg \textit{Accept}) \\ &\quad + Cr(\textit{Tails} | \neg \textit{Accept})V(\textit{Tails} \& \neg D \& \neg \textit{Accept}) \\ &= 0 \end{aligned}$$

Furthermore, it seems reasonable to assume that Beauty's acceptance or rejection of a bet gives her no evidence as to the outcome of the coin toss—in other words, that $Cr(\textit{Heads} | \textit{Accept}) = Cr(\textit{Heads})$. Thus,

$$V_E(\textit{Accept}) = Cr(\textit{Heads})(10 + \epsilon) + Cr(\textit{Tails})(-20 + 2\epsilon)$$

If Beauty is a halfer (that is, if $Cr(\text{Heads}) = 1/2$), it turns out that $V_E(\text{Accept}) = -10 + 3\epsilon$. Since ϵ is minute, $-10 + 3\epsilon < 0$, and so $V_E(\text{Accept}) < V_E(\neg\text{Accept})$. So if Beauty is both a halfer and an evidential decision theorist, she should reject bet 3.10, and Hitchcock's Dutch book argument fails.

On the other hand, if Beauty is a halfer and a causal decision theorist, she should accept bet 3.10. *Heads*, *Tails & D*, and *Tails & ¬D* are paradigmatic dependency hypotheses. Each of these propositions determines which outcome Beauty's choice will cause, but Beauty's choice has no causal impact on which of them is true. So the causal expected values of *Accept* and $\neg\text{Accept}$ are

$$\begin{aligned} V_C(\text{Accept}) &= Cr(\text{Heads})V(\text{Heads} \& \text{Accept}) \\ &\quad + Cr(\text{Tails} \& D)V(\text{Tails} \& D \& \text{Accept}) \\ &\quad + Cr(\text{Tails} \& \neg D)V(\text{Tails} \& \neg D \& \text{Accept}) \\ &= 1/2(10 + \epsilon) \\ &\quad + Cr(\text{Tails} \& D)(-20 + \epsilon) \\ &\quad + Cr(\text{Tails} \& \neg D)(-10 + \epsilon) \\ \\ V_C(\neg\text{Accept}) &= Cr(\text{Heads})V(\text{Heads} \& \neg\text{Accept}) \\ &\quad + Cr(\text{Tails} \& D)V(\text{Tails} \& D \& \neg\text{Accept}) \\ &\quad + Cr(\text{Tails} \& \neg D)V(\text{Tails} \& \neg D \& \neg\text{Accept}) \\ &= Cr(\text{Tails} \& D)(-10 + \epsilon) \end{aligned}$$

No matter what the value of $Cr(\text{Tails} \& D)$ is,

$$\begin{aligned} V_C(\text{Accept}) - V_C(\neg\text{Accept}) &= 1/2(10 + \epsilon) \\ &\quad + Cr(\text{Tails} \& D)(-10 + \epsilon) \\ &\quad + Cr(\text{Tails} \& \neg D)(-10 + \epsilon) \\ &= 3\epsilon \end{aligned}$$

In other words, accepting the bets has a slightly higher causal expected value than rejecting them. If Beauty is a causal decision theorist as well as a halfer, she should accept bet 3.10, and she is vulnerable to Hitchcock's Dutch book.

The moral of this section is that in cases of irreducibly *de se* ignorance, an agent's credences do not always match her betting odds—at least if she is an evidential decision theorist. In examples like *Sleeping Beauty*, accepting a bet at one center is *correlated* with gains or losses at other centers, although it does not *cause* gains or losses at other centers. (When the agent does not suffer from irreducibly *de se* ignorance, we can expect her betting odds to match her credences whether or not she is an evidential decision theorist, since she bets at only one center per world.)

The divergence between credences and betting odds answers an outstanding question from section 3.2.1. Why is Beauty vulnerable to a Dutch book if she satisfies equation 3.5, but not necessarily if she satisfies equation 3.4? In formulating their Dutch book, Draper and Pust [2008] assume that Beauty's betting odds match her credences. This assumption is reasonable after Beauty learns that it's Monday, since has only one center per doxastically possible uncentered world. It is not reasonable when she wakes up, since she has more than one center per doxastically possible uncentered world.

3.3.3 A New Dutch Book

In the previous section, I showed that that causal decision theorists who obey the Halfer rule fall victim to Hitchcock's Dutch book. Evidential decision theorists who obey the Thirder Rule are vulnerable to a Dutch book of their own. Consider the following set of bets, the first of which takes place on Sunday:

$$\begin{array}{ll} \text{Heads} & \$(15 + 2\epsilon) \\ \text{Tails} & \$(-15 + \epsilon) \end{array} \quad (3.11)$$

and the second of which takes place on Monday and (and on Tuesday, if Beauty is awake):

$$\begin{array}{ll} \text{Heads} & \$(-20 + \epsilon) \\ \text{Tails} & \$(5 + \epsilon) \end{array} \quad (3.12)$$

If Beauty accepts bets 3.11 and 3.12, she is bound to lose $\$(5 - 3\epsilon)$. If the coin lands heads, she wins $\$(15 + 2\epsilon)$ on bet 3.11, but loses $\$(20 - \epsilon)$ on bet 3.12. If it lands tails she loses $\$(15 - \epsilon)$ on bet 3.11, and wins back only $\$(10 + 2\epsilon)$ on bet 3.12.

If Beauty is a thirder and an evidential decision theorist, then she should accept bets 3.11 and 3.12.

$$\begin{aligned} V_E(\text{Accept}) &= Cr(\text{Heads}|\text{Accept})V(\text{Heads} \& \neg D \& \text{Accept}) \\ &\quad + Cr(\text{Tails}|\text{Accept})V(\text{Tails} \& D \& \text{Accept}) \\ &= (1/3)(-20 + \epsilon) + 2/3(10 + 2\epsilon) \\ &= 3\epsilon \\ V_E(\neg \text{Accept}) &= Cr(\text{Heads}|\neg \text{Accept})V(\text{Heads} \& \neg D \& \neg \text{Accept}) \\ &\quad + Cr(\text{Tails}|\neg \text{Accept})V(\text{Tails} \& \neg D \& \neg \text{Accept}) \\ &= 0 \end{aligned}$$

In other words, for evidential decision theorists who obey the Thirder Rule, bets 3.11 and 3.12 constitute a Dutch book.

So far, I've proved two Dutch book results: causal decision theorists who obey the Halfer Rule are vulnerable to Hitchcock's Dutch book, and evidential decision theorists who obey the Thirder Rule are vulnerable my Dutch book. In the next two sections, I will prove two converse Dutch book results: causal decision theorists who obey the Thirder Rule and evidential decision theorists who obey the Halfer Rule are immune to Dutch books. Section 3.3.4 proves a lemma which is crucial to both results: agents who bet at what I call *thirder odds* are immune to Dutch books. Section 3.3.5 proves that (given some reasonable background assumptions) causal decision theorists who obey the Thirder Rule and evidential decision theorists who obey the Halfer Rule bet at thirder odds.

3.3.4 Why Everyone Should Bet at Thirder Odds

In order to define thirder odds, I must first stipulate what counts as a betting arrangement. Let a betting arrangement be a partition \mathcal{A} that divides uncentered worlds into equivalence classes such that the agent receives the same payoff at

any two centers belonging to uncentered worlds in the same equivalence class. I'll switch back and forth between writing agent's payoff at world W as X_W and writing the agent's payoff in an equivalence class A as X_A .

An agent bets at *thirder odds* for an uncentered credence function Cr_u just in case she is willing to accept all and only betting arrangements such that, where W is a variable ranging over uncentered worlds,

$$\sum_W N_W Cr_u(W) X_W > 0$$

Halfer odds will play an important role in my argument. I'll that an agent bets at halfer odds for Cr_u just in case for every uncentered world W , she is willing to accept all and only betting arrangements such that

$$\sum_W Cr_u(W) X_W > 0$$

(Intuitively, thirder odds are odds that 'correspond' to thirder credences, and halfer odds are odds that 'correspond' to halfer credences.) I will argue that agents who bet at thirder odds are immune to Dutch books (provided Cr_u is coherent and updated by conditionalization).

The bare bones of my argument are as follows. Suppose an agent (called 'the original agent' for reasons that will soon become apparent) has a coherent uncentered credence function Cr_u , and suppose she bets at thirder odds for Cr_u at all of her doxastically possible centers. We can describe a second, imaginary agent who bets at halfer odds for Cr_u , and who bets only once at each uncentered world. Teller [1973] has proved that (as long as Cr_u is updated by conditionalizing) such an imaginary agent is immune to Dutch books. But for any betting arrangement that the original agent is willing to accept, we can generate a corresponding bet β' such that

- (a) the imaginary agent is willing to accept β' ,
- (b) for every $A \in \mathcal{A}$ such that $X_A > 0$, there is some $W \in A$ such that at W , the original agent wins at least as much as the imaginary agent, and
- (c) for every $A \in \mathcal{A}$ such that $X_A < 0$, there is some $W \in A$ such that at W , the original agent loses at least as little as the imaginary agent.

If we could subject the original agent to a Dutch book, then we could subject the imaginary agent to a Dutch book by selling her the corresponding bets before and after she updated. But by Teller's result, the imaginary agent is immune to Dutch books. Therefore, the original agent must be immune to Dutch books too.

The bare bones need fleshing out in a few places. First, I must specify a procedure for generating the imaginary agent's betting arrangement β' from the original agent's betting arrangement β . Second, I must show that (a)-(c) hold true for β' . I'll take these tasks in turn.

Let the expected number of centers in $A \in \mathcal{A}$ be defined as

$$C_A = \frac{\sum_{W \in A} Cr_u(W)N_W}{Cr_u(A)}$$

Let β' be a betting arrangement that pays, for any $A \in \mathcal{A}$,

$$\$X'_A =_{df} \$C_A X_A$$

Now that I've defined β' , it only remains to show (a)-(c).

I'll begin with (a). Suppose the original agent is willing to accept β . Since she bets at thirder odds,

$$\begin{aligned} \sum_W N_W Cr_u(W) X_W &> 0 \\ \sum_{A \in \mathcal{A}} \sum_{W \in A} N_W Cr_u(W) X_A &> 0 \\ \sum_{A \in \mathcal{A}} \frac{\sum_{W \in A} Cr_u(W) N_W}{Cr_u(A)} X_A Cr_u(A) &> 0 \\ \sum_{A \in \mathcal{A}} C_A X_A Cr_u(A) &> 0 \\ \sum_W X'_W Cr_u(W) &> 0 \end{aligned}$$

Since the imaginary agent bets at halfer odds, it follows that she'll be willing to accept β . β' satisfies (a), and it only remains to show that β and β' satisfy (b) and (c).

Suppose $X_A > 0$, and let α_A be a world in $A \in \mathcal{A}$ with at least as many centers as any other world in A . The maximum number of centers in A is at least as large as the expected number of centers in A , so $X_{\alpha_A} \geq X'_{\alpha_A}$. β' satisfies (b).

Suppose $X_A > 0$, and let ω_A be a world in A with at least as few centers as any other world in A . The minimum number of centers in A is at least as large as the expected number of centers in A , so $X_{\omega_A} \leq X'_{\omega_A}$. β' satisfies (c).

3.3.5 Who Bets At Thirder Odds?

I've shown that agents who bet at thirder odds are immune to Dutch books. In this section, I'll show causal decision theorists who satisfy the Thirder Rule and evidential decision theorists who satisfy the Halfer Rule are immune to Dutch books. My argument will require a few assumptions.

First, I'll assume that in cases of irreducibly *de se ignorance*, agents are certain that within each uncentered world, they will bet the same way at all subjectively indistinguishable centers. I gave a reason for accepting this assumption in section 3.3.2: at any two subjectively indistinguishable centers in

an uncentered world, an agent's beliefs and desires are the same (else she would be able to distinguish between the two centers by taking note of her beliefs and desires). The agents I'll be discussing are either consistent causal decision theorists or consistent evidential decision theorists, so at any two subjectively indistinguishable centers in an uncentered world, they use the same decision procedure. If an agent's choices are determined by her beliefs, desires, and decision procedure, then she will have to bet the same way at any subjectively indistinguishable centers in an uncentered world.

Second, I'll assume that an agent's accepting or refusing bets gives her no evidence about the propositions she is betting on. Cases that violate this assumption are problematic for any proponent of Dutch books, not just for me, and a full discussion of such cases would be a distraction from the topic of the paper. So for now, I will set them aside.

Third, I'll assume that accepting bets does not cause the agent to gain or lose anything, aside from what she wins or loses as the bet's explicit payoff. I take this to be a stipulation about what counts as a bet.

When the above three assumptions are satisfied, causal decision theorists who obey the Thirder Rule are invulnerable to Dutch books. For suppose there is a causal decision theorist who obeys the Thirder Rule, and suppose she is considering a betting setup β .

We know that for each $A \in \mathcal{A}$ if A is the case, then no matter what else is true, the agent will be $\$X_A$ richer if she accepts β than if she rejects it. Therefore, the causal expected value of accepting β is

$$V_C(\text{Accept}) = \sum_{A \in \mathcal{A}} X_A Cr_{\textcircled{0}}(A)$$

By the Thirder Rule,

$$V_C(\text{Accept}) = \sum_{A \in \mathcal{A}} X_A \frac{\sum_{W \in A} N_W Cr_u(W)}{\sum_{W^*} N_W Cr_u(W^*)}$$

The agent, being a causal decision theorist, will accept the bet only when $V_C(\text{Accept}) > 0$. This will happen just in case

$$\sum_{A \in \mathcal{A}} X_A \frac{\sum_{W \in A} N_W Cr_u(W)}{\sum_{W^*} N_W Cr_u(W^*)} > 0$$

Or, since $\sum_{W^*} N_W Cr_u(W^*) > 0$, just in case

$$\sum_{A \in \mathcal{A}} X_A \sum_{W \in A} N_W Cr_u(W) > 0$$

$$\sum_{W \in \mathcal{A}} N_W Cr_u(W) X_W > 0$$

I've proved that the agent bets at thirder odds.

Likewise, an evidential decision theorist who obeys the Halfer Rule will bet at thirder odds. For suppose there is an evidential decision theorist obeys the

Halfer Rule, and suppose she is considering the a bet of form β . If the agent occupies an uncentered world with N_W centers, then a centered world where she accepts β will have value $\$N_W X_W$ (since if she accepts the bet on this occasion, she is guaranteed to accept it again on $N_W - 1$ other occasions). The expected value of accepting β is:

$$V_E(\text{Accept}) = \$ \sum_W Cr_{@}(W) N_W X_W$$

By the Halfer Rule,

$$V_E(\text{Accept}) = \$ \sum_W Cr_u(W) N_W X_W$$

Since the agent is an evidential decision theorist, she will accept β if and only if $V_E(\text{Accept}) > 0$. But to accept a betting setup if and only if $\$ \sum_W Cr_u(W) N_W X_W > 0$ is just to bet at thirder odds.

3.4 Scoring Rules

Kierland and Monton [2005] suggest that we use scoring rules to adjudicate the debate between halfers and thirders. I will follow Kierland and Monton in focusing on the Brier score. It's not obvious that the Brier score the only acceptable method of measuring inaccuracy, but there are reasons to expect that what holds for the Brier score holds for scoring rules generally.

In section 3.4.1, I'll explain how the Brier score works as a measure of inaccuracy, discuss Kierland and Monton's suggestion that agents ought to minimize expected inaccuracy, and criticize Kierland and Monton's definition of expected inaccuracy. In section 3.4.2, I'll suggest two ways of revising the definition of expected inaccuracy. In section 3.4.3, I'll argue causal decision theorists should favor one definition of expected inaccuracy, while evidential decision theorists should favor the other. According to the definition that's best for causal decision theorists, the Thirder Rule minimizes expected inaccuracy. According to the definition that's best for evidential decision theorists, the Halfer Rule minimizes expected inaccuracy. Once again, it turns out that causal decision theorists should obey the Thirder Rule, while evidential decision theorists should obey the Halfer Rule.

3.4.1 Measuring Inaccuracy

The idea behind scoring rules is something like this: Just as full belief aims at truth, partial belief aims at accuracy. If A is true, it's good to fully believe A and bad to fully believe $\neg A$, while if A is false, it's good to fully believe $\neg A$ and bad to fully believe A . Likewise, if A is true, it's is good to believe A to a high degree (the higher the better) and bad to believe A to a low degree (the lower the worse), while if A is false, then it's good to believe A to a low degree (the lower the better) and bad to believe A to a high degree (the higher the worse).

If it turns out that halving (or thirding) conduces to inaccuracy, this will give us a reason to reject halving (or thirding).

In addition to *ranking* partial beliefs in terms of inaccuracy, we can *measure* their inaccuracy. Among the many possible measures, the *Brier score* (developed by Brier [1950]) is the most popular. Where X is a *de dicto* proposition, $Cr_{\textcircled{A}}(X)$ is an agent's credence in X , and $W(X)$ is X 's truth value, the agent's Brier score for X is

$$S_W(X) = (W(X) - Cr_{\textcircled{A}}(X))^2 \quad (3.13)$$

The lower the Brier score—that is, the closer the agent's degree of belief in A to A 's truth value—the better.

In addition to measuring inaccuracy for individual partial beliefs, we can use the Brier score to measure inaccuracy for sets of partial beliefs. Where $\{X_1, X_2, \dots, X_n\}$ are (*de dicto*) propositions in the domain of the agent's credence function, $W(X_i)$ is X_i 's truth value, and $Cr_{\textcircled{A}}(X_i)$ is the agent's credence in X_i , the agent's expected inaccuracy for $\{X_1, X_2, \dots, X_n\}$, as calculated using the Brier score, is:

$$S_W(X_1, X_2, \dots, X_n) = \sum_i S_W(X_i) \quad (3.14)$$

An important advantage of the Brier score is that it's a *proper* scoring rule, meaning that an agent whose goal is to minimize her Brier score will have no incentive to 'cheat' by adjusting her credences in ways that are epistemically unwarranted. Savage [1971] shows that if value is defined in terms of accuracy, so that the value of adopting a set of partial beliefs $\{X_1, X_2, \dots, X_n\}$ is $-S_W(X_1, X_2, \dots, X_n)$, then an agent will always assign higher expected value to keeping her current beliefs about $\{X_1, X_2, \dots, X_n\}$ than to adopting any other set of beliefs about $\{X_1, X_2, \dots, X_n\}$. Savage uses the evidential definition rather than the causal definition of expected value, but in ordinary cases, we shouldn't expect this to make much difference. As long as the agent's adopting a set of beliefs is uncorrelated with her accumulating inaccuracy by some other means, the two definitions will coincide. The Brier score is not the only proper scoring rule, but what I have to say about the Brier score should apply to any other proper scoring rule.

Advising a partial believer to minimize her inaccuracy isn't very helpful. Someone who is unsure what the world is like won't know how inaccurate her partial beliefs are. But if she knows the objective chances of the world's turning out various ways, she can calculate her expected inaccuracy. Let $\{W_1, W_2, \dots, W_m\}$ be a set of 'possible worlds', or logically consistent valuation functions assigning truth values to each of $\{X_1, X_2, \dots, X_n\}$. Then where $\{X_1, X_2, \dots, X_n\}$ are uncentered propositions and $P(W_j)$ is the objective chance of W_j , we can define the agent's expected inaccuracy for $\{X_1, X_2, \dots, X_n\}$ as follows:

$$S_E(X_1, X_2, \dots, X_n) = \sum_j P(W_j) \sum_i S_{W_j}(X_i)$$

Kierland and Monton suggest that when an agent knows the relevant objective chances, she should minimize her expected inaccuracy.

This suggestion is on the right track, but there are three potential problems with it. An agent may have what Lewis [1986b] calls ‘inadmissible information’-information that tells her more about whether an event will occur than the event’s objective chance alone. To take a somewhat farfetched example, suppose an agent has a single atom of carbon 14, which a trustworthy oracle tells her will decay within three months. On the basis of the oracle’s testimony, it may be reasonable for her to place a high credence in the proposition that the atom decays within three months, even though the objective chance of its decaying so quickly is low. When the agent has inadmissible information, it looks like minimizing expected inaccuracy is the wrong thing to do.

Second, the advice to minimize expected inaccuracy is not invariably useful (as Kierland and Monton point out). Agent don’t always know the objective chances. If an agent doesn’t know the value of $P(W_j)$ for some W_j generated by $\{X_1, X_2, \dots, X_n\}$, then she won’t know how to minimize $S_E(X_1, X_2, \dots, X_n)$.

The third problem is a bit more subtle than the first two. So far, my talk of ‘propositions’ and ‘worlds’ has been ambiguous. Objective chances attach to uncentered worlds, but the agent’s credences attach to centered worlds. So the W_j in $P(W_j)$ can’t be the same sort of thing as the W_j in $S(W_j)$. The ambiguity is harmless in cases of purely *de dicto* ignorance, where the agent’s epistemic state can be represented in terms of uncentered worlds. But in cases of irreducibly *de se* ignorance, the ambiguity is ineliminable.

On Kierland and Monton’s definition of expected inaccuracy, the advice to minimize expected inaccuracy is sometimes wrong and sometimes useless. Furthermore, in cases of irreducibly *de se* ignorance, there’s no univocal way of interpreting the definition. These are the problems—how best to remedy them?

3.4.2 Revising the Concept of Expected Inaccuracy

For my current purposes (adjudicating between the Halfer Rule and the Thirder Rule), there’s a convenient fix for the first two problems. We can define expected inaccuracy not in terms of the objective chance function, but in terms of the agent’s uncentered credence function Cr_u . Just as Cr_u ought to satisfy conditionalization, it ought to satisfy the Principal Principle— $Cr_u(A|P(A) = x)$ should equal x , provided the agent has no admissible information. And if an agent satisfies the Principal Principle, then if she knows A ’s objective chance and has no inadmissible information, $Cr_u(A)$ will equal $P(A)$.

So when a definition of expected inaccuracy in terms of the objective chance function gets the right answers, so does a definition in terms of Cr_u . When a definition of expected inaccuracy in terms of the objective chance function gets the wrong answers (because the agent has inadmissible information) a definition in terms of Cr_u can do better, because Cr_u takes the inadmissible information into account. And when definition of expected inaccuracy in terms of the objective chance function yields no useful answers (because the agent has too little information about the objective chances), a definition in terms of Cr_u can do

better, because the agent presumably has better epistemic access to Cr_u than to the objective chance function.

The third problem is not yet solved—like P , Cr_u assigns probabilities only to uncentered propositions. We can solve the third problem by letting the W_j s be centered worlds, and letting $u(W_j)$ designate the uncentered world associated with W_j . Instead of calculating expected inaccuracy using $Cr_u(W_j)$, which is undefined, we can calculate it using $Cr_u(u(W_j))$.

We're almost ready to introduce a new, improved definition of expected inaccuracy. There's just one problem: the correspondence between centered worlds and uncentered worlds is not one-one, but many-one. There are two natural ways extending the definition of expected inaccuracy. (These ideas are faintly suggested by the calculations in [Kierland and Monton, 2005], though the authors don't comment very extensively on the general principles behind the calculations. Nonetheless, I will follow Kierland and Monton's terminology.)

We might measure expected inaccuracy as expected total inaccuracy.

$$S_{ET}(X_1, X_2, \dots, X_n) = \sum_j Cr_u(u(W_j)) \sum_i S_{W_j}(X_i)$$

On the other hand, we might measure expected inaccuracy as expected average inaccuracy.

$$S_{EA}(X_1, X_2, \dots, X_n) = \sum_j Cr_u(u(W_j)) \sum_i \frac{S_{W_j}(X_i)}{N_{W_j}}$$

Both measures look plausible. Expected total inaccuracy is uniform across centered worlds—each centered world makes the same contribution to S_{EA} , proportional to the probability assigned to the corresponding centered world by Cr_u . Expected average inaccuracy is uniform across uncentered worlds—each uncentered world makes the same contribution to S_{ET} , proportional to the probability assigned to it by Cr_u . Which measure is best? And what does the best measure have to tell us about irreducibly *de se* ignorance?

3.4.3 Average or Total? Halfer or Thirder?

I pointed out in section 3.4.1 that the Brier score had the advantage of being a proper scoring rule. A good and important thing about proper scoring rules is that they don't generate pragmatic rewards for epistemically unwarranted credences.

One might worry that choosing the wrong measure of expected inaccuracy will generate pragmatic rewards for epistemically unwarranted credences, in something like the way an improper scoring rule does. In particular, it would be a problematic if a measure of expected inaccuracy weighted an agent's inaccuracy more heavily at some centered worlds than at others.

Does either S_{EA} or S_{ET} have a problem with weighting? This depends on what the 'penalty' for inaccuracy is. In one sense—a causal sense—the agent's inaccuracy at a center is independent of the number of subjectively indistinguishable centers in the same uncentered world. Her having accurate or inac-

curate beliefs at one center does not cause her to have accurate or inaccurate beliefs at subjectively indistinguishable centers in the same world. If we understand ‘reward’ causally, the agent’s inaccuracy at each center should count for the same amount. Otherwise, she will be rewarded for reporting credences that are too low for worlds with multiple centers (since her accuracy at each of the centers will count for less). So on the causal understanding of ‘reward’, S_{ET} seems to be the better scoring rule.

In an evidential sense, on the other hand, the agent’s inaccuracy at a center depends not only on what she believes at that center, but also on the number of subjectively indistinguishable centers in the same world. Her having accurate or inaccurate beliefs at one center is decisive evidence that she has equally accurate or inaccurate beliefs at other, subjectively indistinguishable centers.

If we understand ‘reward’ in this evidential sense, the agent’s inaccuracy at each uncentered world should count for the same amount, no matter how many centers the uncentered world contains. Otherwise, she will be rewarded for reporting credences that are too high for worlds with multiple centers (since her accuracy at each of the centers will count, evidentially, for more). On the evidential understanding of ‘reward’, S_{EA} seems to be the better scoring rule.

So causal decision theorists should minimize expected total inaccuracy, while evidential decision theorists should minimize expected average inaccuracy. It will turn out that agents who aim to minimize expected total inaccuracy should adhere to the Thirder Rule, while agents who want to minimize expected average inaccuracy should adhere to the Halfer Rule.

Suppose that an agent wants to minimize her expected total inaccuracy for a particular world W . Then she’ll want to minimize

$$\begin{aligned} S_{ET}(A) &= Cr_u(W)N_W(1 - Cr_{\textcircled{u}}(W))^2 + \sum_{W^* \neq W} Cr_u(W^*)N_{W^*} Cr_{\textcircled{u}}(W)^2 \\ &= Cr_u(W)N_W(1 - 2Cr_{\textcircled{u}}(W) + (Cr_{\textcircled{u}}(W))^2) + \sum_{W^* \neq W} Cr_u(W^*)N_{W^*} Cr_{\textcircled{u}}(W)^2 \end{aligned}$$

We can find the minimum of this equation by setting its derivative with respect to $Cr_{\textcircled{u}}(W)$ equal to zero.

$$Cr_u(W)N_W(-2 + 2Cr_{\textcircled{u}}(W)) + 2 \sum_{W^* \neq W} Cr_u(W^*)N_{W^*} Cr_{\textcircled{u}}(W) = 0$$

$$-2Cr_u(W)N_W + 2Cr_{\textcircled{u}}(W)(Cr_u(W)N_W + \sum_{W^* \neq W} Cr_u(W^*)N_{W^*}) = 0$$

$$-2Cr_u(W)N_W + 2Cr_{\textcircled{u}}(W)(\sum_{W^*} Cr_u(W^*)) = 0$$

$$Cr_u(W)N_W = Cr_{\textcircled{u}}(W)(\sum_{W^*} Cr_u(W^*))$$

$$Cr_{\textcircled{u}}(W)(\sum_{W^*} Cr_u(W^*)) = Cr_u(W)N_W$$

$$Cr_{\textcircled{u}}(W) = \frac{Cr_u(W)N_W}{\sum_{W^*} Cr_u(W^*)}$$

This equation has its minimum where $Cr_{\textcircled{u}}(W) = \frac{\sum_{W \in W} Cr_u(W)N_W}{\sum_{W^*} Cr_u(W^*)N_{W^*}}$. So the agent’s total inaccuracy for an uncentered world is minimized when she satisfies

the Thirder Rule.

Suppose that an agent wants to minimize her expected average inaccuracy for a particular uncentered world W . Then she'll want to minimize

$$\begin{aligned} S_{EA}(W) &= Cr_u(W)N_W \frac{(1-Cr_{\textcircled{a}}(W))^2}{N_W} + \sum_{W^* \neq W} Cr_u(W^*)N_{W^*} \frac{Cr_{\textcircled{a}}(W)^2}{N_{W^*}} \\ &= Cr_u(W)(1 - Cr_{\textcircled{a}}(W))^2 + \sum_{W^* \neq W} Cr_u(W^*)Cr_{\textcircled{a}}(W)^2 \\ &= Cr_u(W)(1 - 2Cr_{\textcircled{a}}(W) + (Cr_{\textcircled{a}}(W))^2) + \sum_{W^* \neq W} Cr_u(W^*)Cr_{\textcircled{a}}(W)^2 \end{aligned}$$

We can find the minimum of this equation by setting its derivative with respect to $Cr_{\textcircled{a}}(W)$ equal to zero.

$$\begin{aligned} Cr_u(W)(-2 + 2Cr_{\textcircled{a}}(W)) + \sum_{W^* \neq W} Cr_u(W^*)2Cr_{\textcircled{a}}(W) &= 0 \\ -2Cr_u(W) + 2Cr_{\textcircled{a}}(W)(Cr_u(W) + \sum_{W^* \neq W} Cr_u(W^*)) &= 0 \\ -2Cr_u(W) + 2Cr_{\textcircled{a}}(W) &= 0 \\ Cr_u(W) &= Cr_{\textcircled{a}}(W) \end{aligned}$$

So an agent's expected average inaccuracy for an uncentered world has its minimum where $Cr_{\textcircled{a}} = Cr_u$ -in other words, when she satisfies the Halfer Rule.

The Thirder Rule furthers the goal of minimizing expected total inaccuracy (the criterion that causal decision theorists should favor), while the Halfer Rule furthers the goal of minimizing expected average inaccuracy (the criterion that evidential decision theorists should). Again, causal decision theory goes with the Thirder Rule, and evidential decision theory goes with the Halfer Rule.

3.5 Stability

There is one criterion that favors the Thirder Rule over the Halfer Rule independently of any decision-theoretic considerations. An adequate *de se* confirmation theory should be insensitive to relatively trivial changes in the way we represent an agent's credence function. In particular, an adequate *de se* confirmation theory should be stable in the following sense: if we care about the change in an agent's credence regarding A , then once the agent's doxastic worlds are represented richly enough to include all information she considers relevant to A , it should be possible to enrich them without changing the theory's advice.

We can illustrate the concept of stability using conditionalization. Suppose that we are interested in the effects of conditionalizing on an agent's beliefs regarding some proposition A . And suppose that, according to our way of representing the agent's credence function, E is the strongest proposition of which the agent becomes certain between between t_0 and t_1 . There might be some irrelevant proposition B such that the agent becomes certain of B between t_0 and t_1 , but the belief worlds in our representation need not settle whether B is true or false. We can enrich our representation, dividing each of the agent's

belief worlds into one world where B is true, and another where B is false. Since B is irrelevant to A , we should ensure that E screens A off from B at t_0 (i.e., $Cr_0(A|E \& P) = Cr_0(A|E)$). Enriching our representation this way won't affect the conditionalization's advice regarding the proposition A .

The following simple example shows how conditionalizing is a stable rule: Linda is about to witness a fair coin toss. Before the toss, she grants credence $1/2$ to the proposition that the coin is fair and credence $1/2$ to the proposition that it is biased with $P(\text{Heads}) = \frac{1}{4}$. She then learns that the coin is fair, and updates accordingly. Of course, while learning that the coin is fair, she gains plenty of information which is irrelevant to the outcome of the coin toss—that the coin's owner told her the coin was fair, that he told her in a pleasant baritone voice, and so forth. We might represent Linda's initial credence function Cr_0 as follows, where *Fair* is the proposition that the coin is fair, and *Biased* is the proposition that it is biased:

	<i>Heads</i>	<i>Tails</i>	
<i>Fair</i>	1/4	1/4	(3.15)
<i>Biased</i>	1/8	3/8	

On this way of representing things, Linda conditionalizes on *Fair*, so that $Cr_1(\text{Heads}) = 1/2$.

On the other hand, we might represent Cr_0 in a somewhat richer way, where B is the proposition that Linda learns about the coin's bias from someone with a nice baritone voice. (Since B is irrelevant information, it must be the case that $P(\text{Heads}|B \& \text{Fair}) = P(\text{Heads}|\text{Fair})$.)

	<i>Heads</i>	<i>Tails</i>	
<i>Fair</i> & B	1/16	1/16	(3.16)
<i>Fair</i> & $\neg B$	3/16	3/16	
<i>Biased</i> & B	1/32	3/32	
<i>Biased</i> & $\neg B$	3/32	9/32	

On this richer way of representing things, Linda conditionalizes on *Fair* & B , and once again, $Cr_1(\text{Heads}) = 1/2$.

There are numerous other ways of representing Linda's credence function, corresponding to different ways of representing irrelevant information B . But as long as we make sure that E screens A off from B , all of them, taken together with conditionalization, yield the verdict that $Cr_1(\text{Heads}) = 1/2$.

We might enrich our representation of SB's belief worlds, just as we enriched our representation of Linda's belief worlds. Titelbaum [forthcoming] illustrates this idea using the following variant on the Sleeping Beauty story, which he calls *Technicolor Beauty*:

Everything is exactly as in the original Sleeping Beauty Problem, with one addition: Beauty has a friend on the experimental team, and before she falls asleep Sunday night he agrees to do her a favor. While the other experimenters flip their fateful coin, Beauty's friend will go into another room and roll a fair die. (The outcome of the

die roll is independent of the outcome of the coin flip.) If the die roll comes out odd, Beauty's friend will place a piece of red paper where Beauty is sure to see it when she awakens Monday morning, then replace it Tuesday morning with a blue paper she is sure to see if she awakens on Tuesday. If the die roll comes out even, the process will be the same, but Beauty will see the blue paper on Monday and the red paper if she awakens on Tuesday.

Certain that her friend will carry out these instructions, Beauty falls asleep Sunday night. Some time later she finds herself awake, uncertain whether it is Monday or Tuesday, but staring at a colored piece of paper. What does ideal rationality require at that moment of Beauty's degree of belief that the coin came up heads?

Technicolor Beauty is just an enriched version of *Sleeping Beauty*. Seeing the piece of paper tells Beauty nothing about the outcome of the coin toss—she is equally likely to see a red piece of paper whether the coin lands heads or tails. Likewise, seeing the piece of paper tells Beauty nothing about what day it is—she is equally likely to see a red piece of paper whether it is Monday or Tuesday. Whatever Beauty ought to do in *Sleeping Beauty*, she ought to do in *Technicolor Beauty*. The only difference between the two scenarios is that in *Technicolor Beauty*, Beauty gains irrelevant *de dicto* information upon waking up.

Although the Halfer Rule and the Thirder Rule disagree in *Sleeping Beauty*, they agree in *Technicolor Beauty*. I'll assume that on Sunday night in the *Technicolor Beauty* example, Beauty assigns credence $1/4$ to each of the *de dicto* propositions *Heads & Even*, *Heads & Odd*, *Tails & Even*, and *Tails & Odd*, where *Even* and *Odd* designate the relevant outcomes of the die roll. Upon waking up and seeing the paper, Beauty eliminates either *Heads & Even* (if she sees red paper), or *Heads & Odd* (if she sees blue paper). In either case, if Beauty updates Cr_u by conditionalizing on the *de dicto* portion of her total evidence, then

$$Cr_u(Heads) = 1/3 \quad (3.17)$$

But once Beauty has seen the paper, there is only one center in each of her doxastically possible uncentered worlds. Knowing the outcome of the coin toss and the die toss would be enough to tell her what day it is. Therefore, the Halfer Rule and the Thirder Rule agree that $Cr_{\textcircled{u}}(Heads) = Cr_u(Heads) = 1/3$.

Enriching Beauty's doxastically possible centered worlds with irrelevant information has changed the Halfer Rule's advice. In *Sleeping Beauty*, the Halfer Rule recommended that $Cr^{up}(Heads) = 1/2$, but in *Technicolor Beauty*, it recommends that $Cr^{up}(Heads) = 1/3$. The Halfer Rule is unstable. On the other hand, enriching Beauty's doxastically possible worlds with irrelevant information has not effected the Thirder Rule's advice. In both *Sleeping Beauty* and *Technicolor Beauty*, it recommends that that $Cr^{up}(Heads) = 1/3$. So the Thirder Rule is stable—at least in this particular case.

Notice that Beauty's friend is not essential to the point Titelbaum is making in his example. What's crucial is that Beauty be convinced that she will have a different experience every time she wakes up. Beauty can convince herself of this point by putting a coin in her pocket on Sunday night, flipping it ten times whenever she wakes up, and observing the string of outcomes.⁴ If she follows this coin-flipping procedure, her chance of getting the same string of outcomes on both days is tiny—less than .0005. The Halfer Rule will instruct her to set her credence in *Heads* close to 1/3 upon waking up and tossing the coin.

The Thirder Rule is stable in general. Imagine a representation of a *de se* ignorant agent's credence function at time t_1 . Let B be an irrelevant centered proposition which is both true and directly observable at t_1 . (I have in mind the sort of centered proposition expressed by sentences like, "I am looking at a red piece of paper now", "The birds are singing loudly now", or "I am now having a sui-generis subjective experience that cannot be duplicated".) Suppose that on our original representation of the agent's credence function, none of the agent's doxastically possible centered worlds settles whether B .

Next, suppose we enrich each of the agent's doxastically possible centered worlds so that it does settle whether B . We'll have to correspondingly enrich the agent's doxastically possible uncentered worlds. Where the original representation had a single uncentered world W , the the new representation will have a set of uncentered worlds, each of which settles, for each center in W , whether B is true in W .

We can consider an earlier time t_0 when the agent didn't know whether B would be true in any of the centers that are doxastically possible at t_1 . Let B^1 be the proposition that B is true at at least one of the centers. It seems reasonable to require that there be some real number δ such that, for any one of the agent's original belief worlds W at t_1 , $Cr_u(B^1|W) = \delta N_W$ in the enriched representation. (This requirement can be justified on the grounds that in a world with N_W centers, B has N_W chances to be true at one of those centers, and is thus N_W times as likely to be true as at a world with one center.) Although B is irrelevant information, B^1 is not—the more centers a world contains, the more likely B^1 is to be true in that world, all things considered.

At t_1 , when the agent observes that B , she learns an uncentered proposition—namely B^1 . So where Cr_u is her uncentered credence function at t_0 , Cr_u^+ is her uncentered credence function at t_1 , E is the uncentered information she amasses between t_0 and t_1 according to the original representation, W is a variable ranging over propositions that counted as worlds in the original representation, and A is any proposition,

⁴ I thank Robert Stalnaker for pointing this out to me.

$$\begin{aligned}
Cr_u^+(A) &= Cr_u(A|E \& B^1) \\
&= \frac{Cr_u(A \& E \& B^1)}{Cr_u(E \& B^1)} \\
&= \frac{\sum_{W \in (A \& E)} Cr_u(W \& B^1)}{\sum_{W^* \in E} Cr_u(W^* \& B^1)} \\
&= \frac{\sum_{W \in (A \& E)} Cr_u(W) \delta N_W}{\sum_{W^* \in E} Cr_u(W^*) \delta N_{W^*}} \\
&= \frac{\sum_{W \in (A \& E)} Cr_u(W) N_W}{\sum_{W^* \in E} Cr_u(W^*) N_{W^*}}
\end{aligned}$$

At t_1 , E is true at all the agent's doxastically possible uncentered worlds, and the agent has only one center per doxastically possible uncentered world. Therefore, at t_1 ,

$$Cr_{@}(A) = \frac{\sum_{W \in (A)} Cr_u(W) N_W}{\sum_{W^*} Cr_u(W^*) N_{W^*}}$$

So if the agent satisfied the Thirder Rule in the old representation, switching to the new representation won't make any difference.

Thus, the Thirder Rule boasts an advantage over the Halfer Rule. Enriching the representation of an agent's belief state spells trouble for halfers, but not thirders. Adding irrelevant *de se* information to a belief world seems to generate relevant *de dicto* information, since the more centers a world contains, the likelier it is to contain any given experience. Halfers have trouble accounting for this, since they hold that purely *de se* information is never relevant to *de dicto* matters. But thirders who treated *de se* information as relevant to *de dicto* matters all along all along, will have no trouble. For thirders, the relevant *de dicto* information in the new representation is not really new information; it's just relevant *de se* information redescribed.

3.6 Conclusion

The *Sleeping Beauty* example shows that there is more to belief updating than conditionalizing on one's total evidence—at least if one's total evidence is taken to include *de se* information. Agents' belief states can be divided into centered and uncentered parts, and conditionalization can be understood as a constraint on the uncentered part. Halfers and thirders disagree about how the uncentered part of an agent's belief state should interact with the centered part to generate her credence function.

The answer to the debate between halfers and thirders turns on the answer to the debate between causal and evidential decision theorists: causal decision theorists should favor the Thirder Rule, while evidential decision theorists should favor the Halfer Rule. At first glance, this claim might seem bizarre. Why should one's commitments about *how to choose a course of action* influence

one's beliefs about *the result of a coin toss*? Is this just a blatant conflation of pragmatic rationality with epistemic rationality?

It is indeed a conflation of pragmatic rationality with epistemic rationality, but I'm in good company. Justifications for adopting partial beliefs which correspond to the probability calculus almost invariably have a pragmatic component. Dutch books turn on the assumption that one should not have partial beliefs that expose one to a foreseeable sure loss—where 'sure loss' is cashed out in pragmatic terms. Representation theorems turn on qualitative requirements regarding agents' preferences, and requirements regarding preferences are surely pragmatic rather than epistemic. Scoring rules, while they may treat the goal of partial belief as non-pragmatic [Joyce, 1998, e.g.,] still require commitments about the best means to attaining those goals. I haven't introduced new pragmatic commitments to the concept of partial belief—the pragmatic commitments were there all along.

When it comes to stability, the Thirder Rule outperforms the Halfer Rule. Since evidential decision theory entails a commitment to the Halfer Rule (at least for evidential decision theorists who countenance Dutch books and scoring rules), this is a point against evidential decision theory. Cases of irreducibly *de se* ignorance are rare, so perhaps the point is weak one, which might be outweighed by evidential decision theory's good points. Still the stability of the Thirder Rule is a hitherto unappreciated reason—even if a weak reason—for preferring causal decision theory over evidential.

4. CONCLUSION

I'm now in a position to say more about the two questions I began with: When and how should deference to experts or other information sources be qualified? Second, how closely is epistemology related to other philosophical fields, such as metaphysics, ethics, and decision theory?

The first and second chapter provide one sort of answer to the first question: PP and Reflection—principles that advise deference to different sorts of experts—must be qualified with structurally similar escape clauses. The parallels between the two escape clauses may not have been obvious from the discussion in chapters 1 and 2. Qualified Reflection can be understood as the Reflection principle with a PP-style admissibility constraint. I will reformulate PP so that the similarities are more obvious. In the process, I will show how to make PP more realistic, thereby circumventing the worries raised in section 1.1.2.

Recall that where A is any proposition, C is any reasonable initial credence function, t is any time, P_t is the objective chance function at t , x rigidly designates a real number in the unit interval, and E is any proposition admissible with respect to A , PP states:

PP $C(A|P_t(A) = x \ \& \ E) = x$ (Lewis, 1986)

According to Lewis, C assigns nonzero probability to every proposition, and carries no information about what the world is like. In section 1.1.2, I remarked that these features are somewhat worrisome. Fortunately, PP can be rephrased to eschew mention of C , and the paraphrase bears an obvious structural resemblance to Reflection.

Where Cr_0 is any reasonable credence function held at t_0 , and P_1 is the objective chance function at some time t_1 (which may be past, present, or future relative to t_0) we can formulate a new Qualified Principal Principle.:

QPP $Cr_0(A|P_1(A) = x) = x$, provided the agent has no inadmissible information (relative to A and P_1) as of t_0 .

QPP follows from PP together with the claim that the agent arrives at Cr_0 by conditionalizing her initial credence function Cr_{init} on her total evidence.

Qualified Reflection states

Qualified Reflection $Cr_0(A|Cr_1(A) = r) = r$, provided that for all $B \in B_1, B_2, \dots, B_n$,

(i) $Cr_0(Cr_0(A|B) = Cr_1(A|B)) = 1$ and

$$(ii) Cr_0(B|Cr_1(B) = 1) = 1$$

We can think of (i) and (ii) as two halves of an admissibility clause.

When does an agent have inadmissible information? In the case of PP, she has admissible information about A just in case she knows something relevant to A which the chance function does not or might not ‘know’ (i.e., assign probability 1). In the case of Reflection, an agent has inadmissible information just in case

- (a) she knows something relevant to A which her future self does not or might not know (clause (i)),
- (b) she knows something relevant to A which her future self does not or might not weigh properly as evidence (clause (i)), or
- (c) she suspects her future self will be misinformed (clause (ii)).

Clauses (b) and (c) are unnecessary when the expert in question is the chance function, since the chance function always weighs its evidence properly and never assigns probability 1 to falsehoods.

Chapter 3, deals with qualified deference to another sort of expert—the ideal agent with a view from nowhere. The qualifications in chapter 3 are structurally different from the admissibility clauses of PP and Reflection. Barring cognitive mishaps, an agent’s new *de dicto* evidence is typically compatible with her old beliefs, even if both the old beliefs and the new evidence are inadmissible. On the other hand, an agent’s new evidence is incompatible with her old beliefs when the old beliefs and the new evidence are *de se*. One can conditionalize on inadmissible *de dicto* information, but not on *de se* information.

Chapter 3 ends up generating a new set of qualifications to PP and Reflection. If the agent has no inadmissible information, then the uncentered portion of her credence function, Cr_u , should satisfy PP and Reflection. But in some situations where she suffers from irreducibly *de se* ignorance, her actual credence function $Cr_{\textcircled{u}}$ may violate PP and Reflection. (The Sleeping Beauty case is one of these situations.) Lewis thought that in the Sleeping Beauty case, Beauty had inadmissible *de se* information. It would be more accurate to say that Beauty has relevant *de se* information that plays a different role from ordinary *de dicto* inadmissible information.

On to the second question: how closely is epistemology related to other philosophical fields? In chapter 1, I suggest that there are good reasons to suspect that a metaphysical conclusion (Humean supervenience is wrong) follows from epistemic considerations (the Principal Principle and the claim that $Cr(B|A)$ should be 0 for A and B that are known to be incompatible).

At first glance, this conclusion seems surprising; it suggests that metaphysics and epistemology are more interdependent than they might first appear. At second glance, however, it’s not so shocking that epistemology should constrain metaphysics. What the world is like at a deep metaphysical level constrains which epistemic norms we can obey. Therefore, claims about which epistemic norms we can obey must sometimes have deep metaphysical consequences.

It might still seem odd that epistemological considerations should lead to the conclusion that there are more things in heaven and earth than are dreamed of in Humean philosophy. Epistemological considerations typically favor ontological parsimony—usually, the more unknowable entities or properties, the worse. But sometimes, positing entities or properties (or chances) is methodologically useful, even if those entities or properties (or chances) are not directly observable. Russell pointed out that since we value simple theories, external world realism might provide the best explanation of our sense data—even if the only things we directly observe are sense data. Similarly, since we subscribe both to the the Principal Principle and to the rule that where A and B are incompatible $Cr(B|A)$ should be zero, irreducible objective chances might provide the best explanation of the statistical patterns we observe. This is so even if we cannot directly observe the objective chances.

Chapter 2 illustrates a way in which epistemology is more independent from value theory than it appears. van Fraassen [1984, 255] claims that Reflection illustrates an important feature of the ethical relationship between an agent and later self. After presenting his arguments for Reflection, he writes,

I conclude that my integrity, qua judging agent, requires that, if I am presently asked to express my opinion about whether A will come true, on the supposition that I will think it likely tomorrow morning, I must stand by my own cognitive *engagement* as much as I must stand by my own expressions of commitment of any sort.

My discussion shows that van Fraassen is wrong to see Reflection as an ethical matter. Its plausibility has nothing to do with integrity can be explained by appeal to purely epistemic considerations. (Indeed, nothing crucial in Reflection turns on the assumption that the credence function I defer to is *mine*; chapter 2 could be just as easily adapted to the literature on disagreement.)

Chapter 3 suggests that decision-theoretic disagreements have an epistemological upshot. As with chapter 1, the moral looks surprising at first glance, but looks less surprising reasonable on closer examination. Almost all the reasons for having probabilistic beliefs in the first place are pragmatic. It's no wonder that pragmatic disagreements should sometimes lead to epistemic ones.

So the three papers yield three morals about the non-epistemological implications of epistemological principles. First, epistemological considerations do not always favor a sparse ontology—sometimes a richer ontology is better on methodological or explanatory grounds. Second, Reflection and Reflection-like principles have less ethical importance than they might have appeared to. Third, because talk of partial belief is pragmatically loaded, sometimes different pragmatic commitments lead to different commitments about what to believe in light of one's evidence.

BIBLIOGRAPHY

- Frank Arntzenius. Some problems for conditionalization and Reflection. *The Journal of Philosophy*, 100:356–370, 2003.
- Frank Arntzenius. Reflections on Sleeping Beauty. *Analysis*, 62(2):53–62, January 2002.
- Frank Arntzenius and Ned Hall. On what we know about chance. *The British Journal for the Philosophy of Science*, 54:171–179, 2003.
- Nick Bostrom. Sleeping Beauty and self-location: A hybrid model. *Synthese*, 157:59–78, 2007.
- Luc Bovens. P and I will believe that not-p: Diachronic constraints on rational belief. *Mind*, 104:737–760, 1995.
- Glenn Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.
- David Christensen. Clever bookies and coherent beliefs. *The Philosophical Review*, 100:229–247, 1991.
- Kai Draper and Joel Pust. Diachronic Dutch books and Sleeping Beauty. *Synthese*, 164(2):281–287, 2008.
- Andy Egan. Some counterexamples to causal decision theory. *Philosophical Review*, 116(1):93–114, 2007.
- Adam Elga. Self-locating belief and the Sleeping Beauty problem. *Analysis*, 60(2):143–147, April 2000.
- Adam Elga. Reflection and disagreement. *Noûs*, 41:479–502, 2007.
- Adam Elga. Defeating Dr. Evil with self-locating belief. *Philosophy and Phenomenological Research*, 69(2):383–396, 2004.
- Simon Evnine. Personhood and future belief: two arguments for something like Reflection. *Erkenntnis*, 67:91–100, 2007.
- Haim Gaifman. A theory of higher order probabilities. In Brian Skyrms and William Harper, editors, *Causation, Chance, and Credence*, pages 191–219. Kluwer Academic, 1985.

- Mitchell Green and Christopher Hitchcock. Reflections on Reflection. *Synthese*, 98:297–324, 1994.
- Ned Hall. Correcting the guide to objective chance. *Mind*, 103:505–517, 1994.
- Ned Hall. Two mistakes about credence and chance. *Australasian Journal of Philosophy*, 82:93–111, 2004.
- Ned Hall. How to set a surprise exam. *Mind*, 108:647–703, 1999.
- Joseph Halpern. Sleeping Beauty reconsidered: Conditionalization and reflection in asynchronous systems. In Tamar Szabo Gendler, editor, *Oxford Studies in Epistemology*, volume 1, pages 111–142. Oxford: Clarendon Press, 2005.
- Chris Hitchcock. Beauty and the bets. *Synthese*, 139:405–420, 2004.
- Carl Hoefer. On Lewis’s objective chance: ‘Humean supervenience debugged’. *Mind*, 1997.
- Carl Hoefer. The third way on objective probability. *Mind*, 116:549–596, 2007.
- Jenann Ismael. Raid! The big, bad bug dissolved. Forthcoming in *Noûs*.
- Richard Jeffrey. *The Logic of Decision*. University of Chicago Press, Chicago, 2nd edition, 1983.
- James Joyce. A non-pragmatic vindication of probabilism. *Philosophy of Science*, 65(4):575–603, 1998.
- Brian Kierland and Bradley Monton. Minimizing inaccuracy for self-locating beliefs. *PPR*, 70(2):384–395, 2005.
- Isaac Levi. The demons of decision. *The Monist*, 70:193–211, 1987.
- David Lewis. Sleeping Beauty: A reply to Elga. *Analysis*, 61(3):July, 171–176 2001.
- David Lewis. Why conditionalize? In *Papers in Metaphysics and Epistemology*, pages 403–407. Cambridge University Press, 1999.
- David Lewis. Humean supervenience debugged. *Mind*, 103:473–490, 1994.
- David Lewis. Attitudes *de dicto* and *de se*. *The Philosophical Review*, 88(4): 513–543, October 1979.
- David Lewis. Introduction. In *Philosophical Papers: Volume II*. Oxford University Press, New York, 1986a.
- David Lewis. A subjectivist’s guide to objective chance. In *Philosophical Papers: Volume II*. Oxford University Press, New York, 1986b.
- Barry Loewer. David Lewis’s Humean theory of objective chance. *Philosophy of Science*, 71:1115–1125, 2004.

- Patrick Maher. Diachronic rationality. *Philosophy of Science*, 59:120–141, 1992.
- Tim Maudlin. Why be Humean? In *The Metaphysics Within Physics*, pages 50–77. Oxford University Press, 2007.
- Chris Meacham. Sleeping Beauty and the dynamics of *de se* beliefs. Forthcoming in *Philosophical Studies*.
- John Stuart Mill. *A System of Logic*. Longmans, Green, and Co., London, 1911.
- Ruth Millikan. The myth of the essential indexical. *Noûs*, 24(5):723–734, 1990.
- George Edward Moore. *Philosophical Papers*. George Allen and Unwin Ltd., 1959.
- George Edward Moore. *Principia Ethica*. Prometheus Books, 1902. Republished in 1988.
- George Edward Moore. A reply to my critics. In Paul Arthur Schlipp, editor, *The Philosophy of G. E. Moore*, pages 533–677. Tudor Publishing Company, 1952.
- George Edward Moore. Russell's theory of definite descriptions. In Paul Arthur Schlipp, editor, *The Philosophy of Bertrand Russell*. Northwestern University Press, 1944. Reprinted in Moore 1959, 151–195.
- John Perry. Frege on demonstratives. *The Philosophical Review*, 86(4):474–497, 1974.
- John Perry. The problem of the essential indexical. *Noûs*, 13(1):3–21, 1971.
- W.V. Quine. Propositional objects. In *Ontological Relativity and Other Essays*. Columbia University Press, 1969.
- Frank Ramsey. General propositions and causality. In D.H. Mellor, editor, *Foundations*. Routledge & Kegan Paul, London, 1978.
- John Roberts. Undermining undermined: Why Humean supervenience never needed to be debugged (even if it's a necessary truth). *Philosophy of Science*, 68:S98–S108, 2001.
- Denis Robinson. Matter, motion, and Humean supervenience. *Australasian Journal of Philosophy*, 67:394–409, 1989.
- Leonard Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.
- Jonathan Schaffer. Principled chances. *The British Journal for the Philosophy of Science*, 54:27–41, 2003.
- Brian Skyrms. Dynamic coherence and probability kinematics. *Philosophy of Science*, 54:1–20, 1987.

- Brian Skyrms. A mistake in diachronic coherence arguments? *Philosophy of Science*, 60:320–328, 1993.
- William Talbott. Two principles of Bayesian epistemology. *Philosophical Studies*, 62:135–150, 1991.
- Paul Teller. Conditionalization and observation. *Synthese*, 26:218–258, 1973.
- Michael Thau. Undermining and accessibility. *Mind*, 103:491–503, 1994.
- Mike Titelbaum. The relevance of self-locating beliefs. Forthcoming in *The Philosophical Review*.
- Bas van Fraassen. Belief and the problem of Ulysses and the sirens. *Philosophical Studies*, 77:7–37, 1995.
- Bas van Fraassen. Belief and the will. *The Journal of Philosophy*, 81:235–256, 1984.
- Peter Vranas. Have your cake and eat it too: The old Principal Principle reconciled with the new. *Philosophy and Phenomenological Research*, 69:368–382, 2005.
- Peter Vranas. Whos afraid of undermining: Why the Principal Principle might not contradict Humean supervenience. *Erkenntnis*, 57:151–174, 2002.
- Brian Weatherson. Should we respond to Evil with indifference? *Philosophy and Phenomenological Research*, 70:613–635, 2005.
- Jonathan Weisberg. Conditionalization, Reflection, and self-knowledge. *Philosophical Studies*, 135:179–197, 2007.