Distorted Reflection

Rachael Briggs
Massachusetts Institute of Technology

Why Worry about Reflection?

Counterexamples to Bas van Fraassen's Reflection principle are sometimes thought to raise trouble for conditionalization. Reflection states that agents should treat their future selves as experts or, roughly, that an agent's current credence in any proposition A should equal his or her expected future credence in A. Although Reflection is intuitively plausible, it is vulnerable to numerous prima facie counterexamples. Why is this a problem for conditionalization? Both conditionalization and Reflection can be defended by diachronic Dutch book arguments. If Reflection is not really a requirement for partial believers, then diachronic Dutch book arguments—including the one that supports conditionalization—are undermined.

I suggest a way out of the trouble that builds on the work of several earlier authors. Reflection is perfectly defensible, I argue, provided we weaken it with an intuitively plausible escape clause. I formulate a new Qualified Reflection principle, which states that an agent should obey Reflection only if she is certain that she will conditionalize on veridical evidence in the future. I argue that Qualified Reflection follows from the probability calculus together with a few idealizing assumptions.

Unfortunately, perfect confidence in one's future ability to conditionalize is hard to come by. Under all but the most ideal circumstances,

I would like to thank Mark Colyvan, David Etlin, Jason Grossman, Leah Henderson, Alan Hajek, Ned Hall, Vann McGee, Dilip Ninan, Agustin Rayo, Bob Stalnaker, Valentina Urbanek, Jonathan Weisberg, and Seth Yalcin for their helpful comments on earlier drafts of this essay.

Philosophical Review, Vol. 118, No. 1, 2009
DOI 10.1215/00318108-2008-029

agents will have reasons to *suspect* that future failures of conditionalization are in store. Luckily, near enough is good enough: as long as an agent is close to certain that she will conditionalize in the future, she should come close to obeying Reflection. I cash this idea out precisely using a Distorted Reflection principle. Like Bayes's theorem, Distorted Reflection is a consequence of the probability calculus which permits the expression of hard-to-compute probabilities in terms of easier-to-compute components.

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

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). An agent's overall doxastic state can be represented as a credence function which maps propositions to real numbers in the interval [0, 1]. 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 $\sim A$.) An agent's overall doxastic state in the full-belief model can be represented as the set of propositions he or she believes, or equivalently, as that set's characteristic function.

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 his or her credence function at a particular time:

Nonnegativity: $Cr(A) \geq 0$ for all A.

Normalization: $Cr(T) = 1$, where T is a 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 the relations between beliefs at a time, agents are bound by *diachronic norms*, which govern the relationship 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 $(E \rightarrow A)$ for some E compatible with his or her beliefs, and learns E, should on pain of inconsistency, come to believe A. An analogous axiom for Bayesians is:

Conditionalization: $Cr_E(A) = Cr(A \mid E)$

where Cr is the agent's initial credence function, $Cr(A \mid E)$ is defined as $Cr(A \,\&\, E)/Cr(E)$, and Cr_E is the agent's credence function after he or 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 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 his or her credence function is incoherent.

$\therefore$ Any agent who violates N has an incoherent credence function.

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 topic and one which I will set aside for the remainder of the essay. Cashing out the appropriate notion of coherence is likewise difficult and likewise outside the scope of this essay. I will simply assume that some pretheoretic sense can be made of the concept of incoherence in a Bayesian framework.

The synchronic norms can be defended using *synchronic DBAs*, where the Dutch book consists of several simultaneous bets. A defense of conditionalization requires 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 different and less serious types of epistemic defects.

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 on which the agent is betting 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 him or her $\$m$ just in case A is true will be written as shown in table 1.

Table 1

A	$\$m - n$
$\sim A$	$\$ - n$

A conditional bet which costs the buyer $\$n$, pays him or her $\$m$ if A & B, and is called off if $\sim B$, will be written as in table 2.

Table 2

A & B	$\$m - n$
$\sim A$ & B	$\$ - n$
$\sim B$	$\$0$

On the above way of writing bets, the diachronic DBA for conditionalization can be written as follows. 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.

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.

Let $Cr(A \mid E) = n$

Let $Cr_E(A) = r$

Let $Cr_0(E) = d$ for $0 < d < 1$

Bets 1 and 2 are made before the agent learns whether E (table 3). Bet 3 is made at t_1 if and only if the agent learns E (table 4).

Table 3

Bet 1	A & E	$\$1 - n$	Bet 2	E	$\$(d-1)(r-n)$
	$\sim A$ & E	$\$ - n$		$\sim E$	$\$d(r-n)$
	$\sim E$	\$0			

Table 4

Bet 3	A	$\$r-1$
	$\sim A$	$\$r$

No matter what happens next, the buyer's net gain is $\$d(r-n)$. If $\sim E$, she wins $\$d(r-n)$ on bet 2, and no money changes hands on bets 1 or 3. If E, she wins a total of $\$(r-n)$ on bets 1 and 3 and wins $(d-1)(r-n)$ on bet 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.

Reflection

Bas 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 his or her credence function at some later time t_1, and r rigidly designates a real number, Reflection states:

Reflection $\quad Cr_0(A \mid Cr_1(A) = r) = r$

To understand what this means, it will help to imagine agents who satisfy Reflection. Such agents treat their future selves as experts[2] about all propositions. An agent who satisfies Reflection and is certain that his or her future self believes A to degree r, believes A to degree r.

2. The term 'expert' is due to Gaifman 1985.

An agent who satisfies Reflection and is uncertain about his or her future credence in A, has a credence in A which is a mixture of his or her possible future credences in A, where the strength of each ingredient in the mixture depends on how likely it is (by the agent's lights) to reflect his or her future credence. Thus, an agent who obeys Reflection has a credence in A equal to his or her expected future credence in A. Finally, an agent who satisfies both Reflection and conditionalization, upon learning that his or her later degree of belief in A will be r (and learns nothing stronger), will come to place credence r in A.

Anyone who violates Reflection is vulnerable to a Dutch book—provided that the proposition A for which $Cr_0(A \mid Cr_1(A) = r) \neq r$ satisfies three other assumptions. First, $Cr_0(A \mid Cr_1(A) = r)$ 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 $Cr_0(A \mid (Cr_1(A) = r) = n$

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

Bets 1 and 2 are made at t_0 (table 5); bet 3, at t_1 if and only if $Cr_1(A) = r$ (table 6).

Table 5

Bet 1	A & $Cr_1(A) = r$	$\$1 - n$	Bet 2	$Cr_1(A) = r$	$\$(d-1)(r-n)$
	$\sim A$ & $Cr_1(A) = r$	$\$ - n$		$Cr_1(A) \neq r$	$\$d(r-n)$
	$Cr_1(A) \neq r$	$\$0$			

Table 6

Bet 3	A	$\$r-1$
	$\sim A$	$\$r$

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

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.

Future irrationality: David 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.

Memory loss: William 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 percent 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.

Apparent memory loss: Frank 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 $Cr_0(M \mid Cr_1(M) = .5)$ equal to .5. But $Cr_0(M)$ should be 1—again, you should trust your senses at t_0.

Future misleading evidence: Patrick Maher (1992) suggests the following example. You are 90 percent 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 $Cr_0(H \mid Cr_1(H = 1)$ equal to 1. This is bad advice. Right now, you surely know better than to place so much trust in Persi's testimony!

Current misleading evidence: Luc 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, $Cr_0(H \mid Cr_1(H) = .01) = .01$. Accordingly, the higher $Cr_0(Cr_1(H) = .01)$ is, the lower $Cr_0(H)$ should be.[3] 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.

Changes in epistemic standards: I propose the following example. At t_0, you are deciding whether to enroll in the PhD program at William James University, where all the professors are voluntarists about belief. You are agnostic about God's existence, but you believe to degree .90 that if you are immersed in William James University's voluntarist environment, you will become a theist. Where G is the proposition that God exists, Reflection tells you that $Cr_0(G \mid Cr_1(G) = .90) = .90$. But this is the wrong advice; you shouldn't treat your enrollment in William James University as evidence for God's existence.

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 the 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, such an agent has some beliefs at t_0 which are jointly inconsistent—while an agent who is susceptible to a diachronic Dutch book has inconsistent beliefs across time—that is, such an agent has some beliefs at t_0 and some at t_1 which, taken all together, are jointly inconsistent. The first sort of inconsistency is problematic: it is like simultaneously believing A and $\sim A$. The second sort of inconsistency is perfectly acceptable: it is like believing A and then coming to believe $\sim A$, or like one person's believing A and a second person's believing $\sim A$. Sets of beliefs (or pairs of credence functions) held at different times are not the sorts of things that ought to be coherent.

3. 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 $\sim D$ 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.

Although tempting, this line of reasoning is wrong: pairs of credence functions held at different times are indeed the sorts of things that should be coherent. As a logic, Bayesian decision theory ought to be useful for inference and planning. But some sort of intrapersonal coherence is necessary for inference and planning; an agent who conducts his or her epistemic life correctly will have earlier and later selves that cohere better than a pair of strangers. 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 agents, but not strong enough to paralyze them.

Another possible response to the counterexamples is to accept the validity of diachronic DBAs, but claim that Reflection is only a defeasible norm. Mitchell S. Green and Christopher 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 usually are appropriately grounded in his evidence. Occasionally he errs and over- or underestimates the support that his evidence lends a hypothesis. But whenever he finds himself making this kind of 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 proposition. 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, 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 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 require 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 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 you know that your beliefs will improve over time (Evnine 2007, 94), or 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. One might worry whether there is any formal way of cashing the qualification out.

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 captures the intuitive ideas put forth by Hall, Evnine, and Elga. Unlike the original Reflection principle, it 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 the DBA that seems to establish Reflection in its unqualified form.

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 introspecter—in other words, that $Cr_0(Cr_0(A \mid B) = r) = 1$ if and only if $Cr_0(A \mid B) = r$. (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 represent the totality of what the agent learns between t_0 and t_1—form a partition $\{B_1, B_2, \ldots B_n\}$. Third, I will assume that all agents 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 qualified Reflection principle immediately below.

Given these assumptions, the appropriately qualified version of Reflection is:

> *Qualified Reflection*: $Cr_0(A \mid Cr_1(A) = r) = r$, provided that for all $B \in \{B_1, B_2, \ldots B_n\}$,
>
> i. $Cr_0(Cr_0(A \mid B) = Cr_1(A \mid B)) = 1$ and
> ii. $Cr_0(B \mid Cr_1(B) = 1) = 1$

Qualified Reflection follows from the Kolmogorov axioms, together with my three idealizing assumptions.[4]

4. Proof: By the probability calculus,

$$\begin{aligned} Cr_0(A \mid Cr_1(A) = r) &= Cr_0(A \& Cr_1(A) = r)/Cr_0(Cr_1(A) = r) \\ &= \Sigma_{\{B:Cr1(A \mid B)=r\}} Cr_0(A \mid Cr_1(B) = 1)\, Cr_0(Cr_1(B) = 1)/ \\ &\quad \Sigma_{\{B:Cr1(A \mid B)=r\}} Cr_0(Cr_1(B) = 1) \end{aligned}$$

By (ii),

$$Cr_0(A \mid Cr_1(A) = r) = \Sigma_{\{B:Cr1(A \mid B)=r\}} Cr_0(A \mid B)\, Cr_0(B)/\Sigma_{\{B:Cr1(A \mid B)=r\}} Cr_0(B)$$

And by (i),

$$\begin{aligned} Cr_0(A \mid Cr_1(A) = r) &= \Sigma_{\{B:Cr0(A \mid B)=r\}} Cr_0(A \mid B)\, Cr_0(B)/\Sigma_{\{B:Cr0(A \mid B)=r\}} Cr_0(B) \\ &= \Sigma_{\{B:Cr0(A \mid B)=r\}} r\, Cr_0(B)/\Sigma_{\{B:Cr0(A \mid B)=r\}} Cr_0(B) \\ &= r\Sigma_{\{B:Cr0(A \mid B)=r\}} Cr_0(B)/\Sigma_{\{B:Cr0(A \mid B)=r\}} Cr_0(B) \\ &= r. \end{aligned}$$

Clauses (i) and (ii) capture the intuitive ideas suggested by Hall, Elga, and Evnine.[5] 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—together, these two clauses compose 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—Evnine's criterion.

Distorted Reflection

Qualified Reflection is a helpful piece of advice. But the advice comes with a catch—Qualified Reflection applies *only* in cases where the agent satisfies (i) and (ii). Even in cases where $t_0 = t_1$, 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 results of conditionalizing on veridical evidence. If tomorrow you believe some hard-to-verify scientific hypothesis to degree .45, this might be the result of veridical observation and reasonable updating; if you believe the hypothesis to degree 1, this is more likely to be the result of overconfidence 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

5. Weisberg 2007 proves a very similar result: conditionalization follows from my first two assumptions, the assumption that the agent is certain he or 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 actually might arrive at in the future and the opinions he or she thinks he or she might arrive at in the future.

result of conditionalization on veridical evidence, so that for all $B \in \{B_1, B_2, \ldots B_n\}$,

iii. $Cr_0(Cr_1(A \mid B) = r \mid Cr_0(A \mid B) = r) = 1$ (where defined) and
iv. $Cr_0(B \mid Cr_1(B) = 1 \;\&\; Cr_1(A) = r) = 1$

(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).)

The second axis measures the agent's expected departure from conditionalization on veridical evidence, on the hypothesis that $Cr_1(A) = r$. Expected departure 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 the *magnitude* of the failure she expects. We might cash this out as follows. Let $\{W_1, W_2, \ldots W_m\}$ be a set of doxastic alternatives such that each $W \in \{W_1, W_2, \ldots W_m\}$ is the conjunction of some $B \in \{B_1, B_2, \ldots B_n\}$ 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 follows:[6]

$$D_r = \Sigma_{W \in \{W_1, W_2, \ldots Wm\}} [r - Cr_0(A \mid W)]\, Cr_0(W) / Cr_0(Cr_1(A) = r)$$

D_r measures the degree to which an agent expects r to be an overly optimistic credence in A, relative to the support his or her future evidence lends A. (Where the agent 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 agents to set their credences in A equal to their expected later credences in A, then the following rule requires them to set their credences in A equal to their expectations of the later credences they *would* have in A, *were* they to conditionalize on veridical evidence:

Distorted Reflection: $Cr_0(A \mid Cr_1(A) = r) = r - D_r$

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

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 $Cr_0(A \mid Cr_1(A) = r)$ 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 he or she will come close to conditionalizing, it may approximate Reflection even when the agent is confident (or certain) that he or 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 of the agent's confidence in his or her future judgment—only because of the agent's lack of useful information about *how* he or she will fail to conditionalize. In this case, the fact that the agent satisfies Reflection is something of a lucky evidential coincidence. Green and Hitchcock seem to 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.[7] Even so, it is useful in roughly the way Bayes's theorem is useful: it expresses a hard-to-calculate quantity in terms of easier-to-calculate parts. Typically, given any evidence proposition $B \in \{B_1, B_2, \ldots B_n\}$, the agent's t_1 beliefs will be irrelevant to A—in other words, $Cr_0(A \mid W)$ will equal $Cr_0(A \mid B)$ for the B that serves as a

7. Proof:

$$\begin{aligned} D_r &= \Sigma_{W \in \{W_1, W_2, \ldots W_m\}} [r - Cr_0(A \mid W)] Cr_0(W) / Cr_0(Cr_1(A) = r) \\ &= (\Sigma_{W \in \{W_1, W_2, \ldots W_m\}} [r\, Cr_0(W)]) / Cr_0(Cr_1(A) = r) \\ &\quad - ([\Sigma_{W \in \{W_1, W_2, \ldots W_m\}} Cr_0(A \mid W)] Cr_0(W)) / Cr_0(Cr_1(A) = r) \\ &= ([r\, Cr_0(Cr_1(A) = r)] / Cr_0(Cr_1(A) = r)) \\ &\quad - [Cr_0(A \& Cr_1(A) = r)] / Cr_0(Cr_1(A) = r) \\ &= r - Cr_0(A \mid Cr_1(A) = r) \end{aligned}$$

Distorted Reflection states:

$$Cr_0(A \mid Cr_1(A) = r) = r - D_r$$

By the above argument, this is equivalent to the claim that

$$Cr_0(A \mid Cr_1(A) = r) = r - [r - Cr_0(A \mid Cr_1(A) = r)]$$

conjunct in W. Calculating $Cr_0(W)$ is generally straightforward, and the value of r is simply stipulated.

Back to the Counterexamples

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

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 $B \in \{B_1, B_2, \ldots B_n\}$, $(Cr_0(F \mid B) < Cr_1(F)$. 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 $Cr_0(F \mid Cr_1(F) = .99) = .99$.

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—there is no $B \in \{B_1, B_2, \ldots B_n\}$ such that $Cr_0(F \mid B) = .99$. Furthermore, suppose you think 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—$Cr_0(F \mid Cr_1(F) = .99 \;\&\; B) = Cr_0(F \mid Cr_1(B) = 1)$ for each $B \in \{B_1, B_2, \ldots B_n\}$. By the definition of expected departure from conditionalization on veridical evidence,

$$D_r = \Sigma_{W \in \{W_1, W_2, \ldots Wm\}} [.99 - Cr_0(F \mid W)]\, Cr_0(W) / Cr_0(Cr_1(A) = .99)$$

Since the proposition that $Cr_1(F) = .99$ screens F off from each B,

$$\begin{aligned} D_r &= \Sigma_{W \in \{W_1, W_2, \ldots Wm\}} [.99 - Cr_0(F)]\, Cr_0(Cr_1(A) = .99) / \\ &\qquad Cr_0(Cr_1(A) = .99) \\ &= [.99 - Cr_0(F)] \end{aligned}$$

By Distorted Reflection,

$$Cr_0(F \mid Cr_1(F) = r) = .99 - [.99 - Cr_0(F)] = Cr_0(F)$$

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 it is true or not. I will elaborate the example as follows (where Q is the proposition that you take LSQ):

$$Cr_0(F \mid B_x) = .99$$

$$Cr_0(F \mid B) = .001 \text{ for all } B \in \{B_1, B_2, \ldots B_n\} \neq B_x$$

$$Cr_0(Cr_1(F) = .99) \mid \sim Q) = .001$$

$$Cr_0(Cr_1(F) = .99) \mid Q) = 1$$

$$Cr_0(\sim B_x \;\&\; Cr_1(F) = .99) = Cr_0(\sim B_x \;\&\; Q)$$

$$Cr_0(B_x \;\&\; Cr_1(F) = .99) = Cr_0(B_x)$$

By the definition of D_r,

$$\begin{aligned} D_r &= [(.99 - .99)\ Cr_0(B_x) + (.99 - .001)\, Cr_0(\sim B_x \;\&\; Cr_1(F) = .99)]/ \\ &\qquad Cr_0(Cr_1(F) = .99) \\ &= .989\ Cr_0(\sim B_x \;\&\; Cr_1(F) = .99)/Cr_0(Cr_1(F) = .99) \\ &= .989\ Cr_0(\sim B_x \;\&\; Q)/Cr_0(Cr_1(F) = .99) \\ &= .989\ Cr_0(\sim B_x \;\&\; Q)/[Cr_0(\sim B_x \;\&\; Q) + Cr_0(B_x)] \end{aligned}$$

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 your confidence in F upon learning that $Cr_0(F) = .99$. Again, both results seem right.

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 $B \in \{B_1, B_2, \ldots B_n\}$, $Cr_0(Cr_0(S \mid B) = Cr_1(S)) \neq 1$, and you violate (i). Thus, you needn't set $Cr_0(S \mid Cr_1(S) = .10)$ 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 $B \in \{B_1, B_2, \ldots B_n\}$ from S. Then once again, Distorted Reflection will tell you to ignore the information that $Cr_1(S) = .10$—it is irrelevant to S.

Apparent memory loss (Shangri-la): The Shangri-la case appears superficially different from the spaghetti case, but their underlying structures are 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, $Cr_0(Cr_0(S \mid B) = Cr_1(S)) \neq 1$ for every $B \in \{B_1, B_2, \ldots B_n\}$, and you violate (i).

Future misleading evidence (Persi): You may learn between t_0 and t_1 that Persi has reported a heads outcome, in which case, there is some $B \in \{B_1, B_2, \ldots B_n\}$ such that $Cr_0(H \mid B) > Cr_1(H \mid B)$. Since you place nonzero credence in the proposition that you will speak to Persi, $Cr_0(Cr_0(H \mid B) = Cr_1(H \mid B) \neq 1$, and you violate (i). Qualified Reflection does not require you to set $Cr_0(H \mid Cr_1(H) = 1)$ 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 Persi might quickly report a heads outcome, there is a $B \in \{B_1, B_2, \ldots B_n\}$ which entails that you won't conditionalize as you should: $Cr_0(Cr_0(A \mid B) = Cr_1(A \mid B) \neq 1$.

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:

> $Cr_0(H \mid R \,\&\, Cr_1(H) = 1) = Cr_0(H \mid R) = .95$
>
> (If Persi reports H, he is 90 percent likely to be reporting correctly, 5 percent likely to be guessing correctly, and 5 percent likely to be guessing incorrectly.)

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

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

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

$D_r = [1 - .95]\ .5/\ .5 = .05$

By Distorted Reflection,

$Cr_0(H \mid Cr_1(H) = 1) = .95$

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 a tails outcome.

TS: Persi slowly reports a tails outcome.

All of the following should hold:

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

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

$Cr_0(H \mid RS \,\&\, Cr_1(H) = 1) = Cr_0(H \mid RS) = 1$

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

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

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

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

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

$$D_r = [(1-.5)(.05) + (1-1)(.45)] \,/\, .5 = .05$$

Once again, by Distorted Reflection,

$$Cr_0(H \mid Cr_1(H) = 1) = .95$$

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.

Current misleading evidence (the small-town scientist): This case is markedly different from the others. You presumably believe that you will conditionalize on any $B \in \{B_1, B_2, \ldots B_n\}$ you might learn between t_0 and t_1. Furthermore you don't expect to receive any nonveridical evidence. Thus, it seems that you satisfy (i) and (ii)—Qualified Reflection demands that $Cr_0(H \mid Cr_1(H) = .01) = .01$.

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 nonveridical 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, nonmisleading evidence.[8]

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—something 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,

8. There may be pragmatic reasons for small-town scientists to believe their hypotheses more strongly than their evidence suggests. 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 norm.

so Qualified Reflection does not require that you set $Cr_0(G \mid Cr_1(G) = .90)$ equal to .90.

Second, you may see attending William James University as a wholly educational experience—a way to correct your subpar epistemic practices. My account cannot accommodate *all* versions of this reading—a William James education had better not cause you to justifiably question 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 $(G \mid Cr_1(G) = .90)$ 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 provide you with a way of gathering 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 is 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, where Persi talks faster when he's guessing than when he's reporting. In this case, you believe that attending William James University (or talking to Persi) will constitute either 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 $Cr_0(G \mid Cr_1(G) = .90)$ equal to .90.

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: problems of incoherence and

problems of what I will call *self-doubt*. Agents are self-doubting whenever they suspect themselves of having incoherent beliefs.

We can contrast incoherence and self-doubt using full-belief examples. Those who believe a proposition A and simultaneously believe its negation $\sim A$ are guilty of incoherence, and we might criticize them on the grounds that there is no possible world where all their beliefs are true. Those who believe that they believe both A and $\sim A$, however, are self-doubting. They are not incoherent—there really are possible worlds where the self-doubting individuals believe both A and $\sim A$. Some self-doubting individuals may even have good evidence of their own incoherence. Still, there is something problematic about their overall belief states: there is a sense in which self-doubting agents can't be right about everything. Either they are mistaken about their own beliefs regarding A, or they are mistaken about whether A is the case (since someone who believes both A and $\sim A$ 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 he or she does not believe A.[9] Just like a self-doubting agent, an agent with Moore's paradoxical beliefs is guaranteed to be wrong about something, even though his or 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 between incoherent beliefs and what we might call Moore-paradoxical beliefs. If

9. My exposition of Moore's paradox differs from Moore's more explicit presentations (1959) and (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 he or she implies; I intend to make a point about higher-order belief. Despite these significant differences, Moore's insights are capable of shedding light on the case at hand.

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.[10] But if it is merely Moore-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-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.

Why does Moore's suppositional test distinguish incoherence from other types of problems? An agent's belief function fixes the truth values of her beliefs in two ways. First—almost tautologically—it fixes what she believes. Second and less obviously, it fixes the truth values of some of her higher-order beliefs. In cases of incoherence, the agent is guaranteed to be wrong solely because of how her beliefs operate in the first, belief-fixing role. In Moore's paradox cases and cases of self-doubt, the agent is guaranteed to be wrong because of some faulty interaction between the two roles. Moore's test pulls the roles apart. The agent's actual suppositions occupy the belief-fixing role, while her counterfactual beliefs occupy the truth-value-fixing role.

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 "loses" a bet on A at w just in case A is false at w—whether or not any bets are made at w. And 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 "loses" a conditional bet on A given B at w just in case 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 of those bets loses more than he or 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 of those bets loses more than he or 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 letting $Cr(A \vee B) = Cr(A) + Cr(B) + x$ for some

10. 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 show that everything follows from a contradiction. All I need is the assumption that the distinction between coherent and incoherent suppositions is clear.

disjoint propositions A and B and some nonzero 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 nonzero real numbers x and y. Both problems render agents susceptible to Dutch books, but the suppositional test shows 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:

Table 7

Bet 1	$A \vee B$	$Cr(A) - 1$
	$\sim(A \vee B)$	$Cr(A)$
Bet 2	A	$Cr(A) - 1$
	$\sim A$	$Cr(A)$
Bet 3	B	$Cr(B) - 1$
	$\sim B$	$Cr(B)$

No matter what happens next, the buyer of bets 1–3 (table 7) 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 nonzero 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:

Table 8

Bet 4	I	$(1 - y)x$
	$\sim I$	$-yx$

Suppose, as might happen, that $Cr(A \mid I) = Cr(A)$, $Cr(B \mid I) = Cr(B)$, and $Cr(A \vee B \mid I) = Cr(A \vee B)$. Then the self-doubting agent's credences also condone buying or selling *conditional* versions of bets 1–3 which take place only on the condition that I. But together with bet 4, the conditional versions of bets 1–3 constitute a Dutch book. If I is false, then the buyer loses $\$yx$ on bet 4 and wins back nothing on the

other bets; otherwise, the buyer loses $\$x$ on bets 1–3 and wins back only $\$(1-y)\,x$ on bet 4.

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 4, together withthe conditional versions of bets 1–3, will suffer a net loss. But at counterfactual worlds where I is true, the buyer will win $\$(1-y)\,x$ on bet 4 (table 8), and the buyer's wins and losses on the conditional versions of bets 1–3 (table 7) 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 adopt Lewis's assumption that the agent stands no chance of mistaking her evidence, so that if the agent learns E, then all the suppositional worlds must be ones where E is true. Under this assumption, 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 $Cr(A \mid E)$).

On the other hand, the Dutch book against agents who violate Reflection reveals diachronic self-doubt. At a world where the agent makes bets 1 and 2 in table 6, he or she is already guaranteed to suffer a net loss. But as long as the agent doesn't make bet 3, there are counterfactual worlds where he or she enjoys a net gain. At those counterfactual worlds, of course, the agent's beliefs would have condoned different betting behavior. But the bets we consider at counterfactual worlds are fixed by the agent's actual (not counterfactual) credence function.

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 Patrick Maher (1992, 132–33) 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 nonzero 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 a few 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.

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—agents should obey Reflection. Under less-than-ideal circumstances—where the agent suspects he or she will adopt some less rational policy—agents 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.

References

Arntzenius, Frank. 2003. "Some Problems for Conditionalization and Reflection." *Journal of Philosophy* 100: 356–70.

Bovens, Luc. 1995. "'P and I Will Believe that Not-P': Diachronic Constraints on Rational Belief." *Mind* 104: 737–60.

Christensen, David. 1991. "Clever Bookies and Coherent Beliefs." *Philosophical Review* 100: 229–47.

Elga, Adam. 2007. "Reflection and Disagreement." *Nous* 41: 479–502.

Evnine, Simon. 2007. "Personhood and Future Belief: Two Arguments for Something like Reflection." *Erkenntnis* 67: 91–100.

Gaifman, Haim. 1985. "A Theory of Higher Order Probabilities." In *Causation, Chance, and Credence*, ed. Brian Skyrms and William Harper, 191–219. Boston: Kluwer Academic.

Green, Mitchell S., and Christopher Hitchcock. 1994. "Reflections on Reflection." *Synthese* 98: 297–324.

Hall, Ned. 1999. "How to Set a Surprise Exam." *Mind* 108: 647–703.

Levi, Isaac. 1987. "The Demons of Decision." *Monist* 70: 193–211.

Lewis, D. 1999. "Why Conditionalize?" In *Papers in Metaphysics and Epistemology*, 403–7. Cambridge: Cambridge University Press.

Maher, Patrick. 1992. "Diachronic Rationality." *Philosophy of Science* 59: 120–41.

Moore, George Edward. 1902. *Principia Ethica*. Amherst, NY: Prometheus Books. Republished in 1988.

———. 1944. "Russell's Theory of Definite Descriptions." In *The Philosophy of Bertrand Russell*, ed. Paul Arthur Schlipp, 151–95. Evanston, IL: Northwestern University Press. Reprinted in Moore 1959.

———. 1952. "A Reply to My Critics." In *The Philosophy of G. E. Moore*, ed. Paul Arthur Schlipp, 533–677. New York: Tudor.

———. 1959. *Philosophical Papers*. London: George Allen and Unwin Ltd.

Skyrms, Brian. 1993. "A Mistake in Diachronic Coherence Arguments?" *Philosophy of Science* 60: 320–28.

Talbott, William. 1991. "Two Principles of Bayesian Epistemology." *Philosophical Studies* 62: 135–50.

Teller, Paul. 1973. "Conditionalization and Observation." *Synthese* 26: 218–58.

Van Fraassen, Bas. 1984. "Belief and the Will." *Journal of Philosophy* 81: 235–56.

———. 1995. "Belief and the Problem of Ulysses and the Sirens." *Philosophical Studies* 77: 7–37.

Weisberg, Jonathan. 2007. "Conditionalization, Reflection, and Self-Knowledge." *Philosophical Studies* 135: 179–97.