The Deference-Based Conception of Rational Belief Updating

Grant Reaber

Northern Institute of Philosophy

University of Aberdeen

draft of 14 August 2010

ABSTRACT

What I call deference is a technical notion best explained by examples: David Lewis's Principal Principle is an attempt to formally express the deference we owe the objective chances, and Bas van Fraassen's Reflection Principle is an attempt to formalize the (false) claim that rational agents always defer to their future selves. On the orthodox Bayesian conception of rational belief updating, agents update by conditionalization. It can be shown that if you are *sure* you will update by conditionalization then you defer to your post-update self. The deference-based conception of rational belief updating reverses the usual order of explanation and explains the appeal of conditionalization by reference to the idea that rational agents defer to the results of rational updating. I argue that the deference-based conception is superior to and more general than the conditionalization conception.

The difference between the two conceptions particularly dramatic if we suppose that a rational agent need not be certain what his or her credence functions is. In this case, it is not clear how to formalize deference, and a subsidiary aim of the paper is to make progress on this problem.

In an appendix, I consider whether there is a viable deference-based conception of rational utility function updating. My results here are negative: if there is such a conception, I have not found it.

Table of contents

1 Advertisement

I like papers that get right to the point, and I try to write them. But this paper requires a lot of setup. Although the setup is interesting in its own right, I won't be able to say what the deference-based conception of rational belief updating is until halfway through the paper. Nonetheless, I can advertise straightaway one of its virtues, and doing so will help with the setup. So that's what I'll do.

Orthodox Bayesians hold that the beliefs of a rational[1] agent can

[1] By 'rational', I mean *ideally* rational—logically omniscient, never forgetful, and so on. Orthodox Bayesianism is a highly idealized theory, and there is no obvious way to derive what rationality requires of us non-ideal agents from what it requires of the ideally rational. But studying the ideally rational can help us learn about the nature of rationality.

Although I will often write as if ideal rationality is a totally univocal notion, I am open to the idea that there are just different intellectual superpowers that agents could have and no interesting fact of the matter about precisely which set of superpowers constitutes ideal rationality. In particular, there may be no any interesting question about whether ideally rational agents must be certain what their own credences are, or about whether ideally rational agents must know in advance how they should respond to getting evidence E. It will be interesting to have a conception of rational belief updating that makes sense even for agents who don't have these particular superpowers even if we ultimately decide to include these superpowers in the "ideal rationality" package.

be represented by a probability function[2] (sometimes called the agent's *credence function*) and that rational agents always and only update their beliefs by conditionalizing their probability functions on propositions that fully characterize the empirical import of the experiences they are having.[3] (The second half of orthodox Bayesianism, which I call the conditionalization conception of rational belief updating, is the main target of the paper.) Richard Jeffrey (1965) criticized orthodox Bayesianism on the grounds that rational agents may fail to have such propositions at their disposal.[4] Imagine that you observe a cloth by candlelight. In response, you may increase your credence that the cloth is green from 1/2 to 2/3, and there may be no proposition A whatsoever that you become certain of such that, conditional on A, your prior credence that the cloth is green is 2/3. Jeffrey suggested replacing the requirement that rational agents update by conditionalization with the requirement that they update by what has come to be called Jeffrey conditionalization. While conditionalization takes as input a proposition—the proposition that fully characterizes the empirical import of the agent's occurrent experience—Jeffrey

[2] The thesis that the beliefs of a rational agent can be represented by a probability function is extremely controversial. Many philosophers think we should use *conditional* probability functions (also known as Popper functions) so that we can represent conditional probabilities given propositions assigned probability zero. Others think that the beliefs of a rational agent can be *imprecise* so we should represent them with a set of probability functions called a *representor*. And there are many other proposals (see Weisberg forthcoming for a few references). I tend to think we can get by with probability functions, but I see no fundamental incompatibility between the deference-based conception of rational belief updating and alternative views about how to represent beliefs. Still, some details of the deference-based conception will no doubt need to be revisited if the representation of belief is changed; for instance the formal notion of deference it employs will probably need revision.

[3] "Fully characterize the empirical import of" is a gloss that I will be using again and again. In an earlier version of this paper, I used the simpler gloss "fully characterize," but it was objected to me that this gloss is incompatible with the materialist doctrine that it is an intrinsic but hidden property of experiences that they are identical to certain brain states! Maybe the new gloss is also flawed; I will leave it to advocates of the conditionalization conception to say what the right gloss is.

[4] A feature of our notion of proposition is that what propositions there are is agent-relative, for a doxastically possible world is doxastically possible for an agent. Given this understanding of propositions, there is nothing mysterious about what it is for a proposition to be at an agent's disposal: it is just for the proposition to exist, to be in the domain of the agent's probability function.

conditionalization takes as input a finite (or perhaps countably infinite) partition,[5] all of whose elements the agent assigns non-zero probability to, together with an assignment of new probabilities to the elements of the partition. If v_0 is your old probability function, your new probability function after Jeffrey conditionalizing on the partition $A_1, \ldots, A_n$ and probability assignment $A_1 \mapsto p_1, \ldots, A_n \mapsto p_n$ is that probability function v such that, for any proposition A,

$$v(A) = \sum_i p_i v_0(A \mid A_i).$$

Unfortunately, Jeffrey had little to say about how experiences determine partitions and assignments, so in a sense there is no such thing as the Jeffrey conditionalization conception of rational belief updating.[6]

I will introduce the deference-based conception of rational belief

[5] A partition is a set of pairwise incompatible propositions whose disjunction is the trivial proposition. I will identify propositions with sets of possible worlds, so two propositions are incompatible if their intersection is empty, and the trivial proposition is the set of all possible worlds. Now, it is widely agreed that the objects of belief cannot be sets of *metaphysically* possible worlds, for you can rationally doubt necessary a posteriori truths even if you are logically omniscient. I am convinced by arguments of David Chalmers (forthcoming a and b) that, for the purposes of doing Bayesian epistemology, we should take the objects of belief to be sets of *doxastically* possible worlds. Scott Soames (1987; 2008) and others reject this view. Presumably, they are either wrong or else their position can be reconciled with orthodox Bayesianism, in which case it can presumably be reconciled with my brand of Bayesianism.

[6] Notice that if you stripped the conditionalization conception of the claim that what you conditionalize on is the proposition that fully characterizes the empirical import of your experience, it would still give a non-trivial constraint on how your probability function can evolve: when you update, you must zero your credence in some possible worlds and then renormalize. If there are only finitely many possible worlds, the only constraint that saying you must Jeffrey conditionalize on a partition and an assignment puts on how you can update is that you can never raise your credence in a proposition you have credence 0 in. When there are uncountably many possible worlds, Jeffrey conditionalization is more constraining, but that just seems to be a technical weakness of Jeffrey conditionalization. See the final section of Diaconis and Zabell 1982 for a way to generalize Jeffrey conditionalization to be less constraining in this case.

My criticism of Jeffrey conditionalization has been extremely quick and breezy. What I really think is that everything Jonathan Weisberg (2009) says about Jeffrey conditionalization is true. Weisberg's conclusion is cautious: ways of spelling out Jeffrey conditionalization that make it non-trivial and not obviously false render it troublingly anti-holistic.

updating in section 4, and in section 6, I will argue that it is superior to the conditionalization conception of rational belief updating that is a component of orthodox Bayesianism. Most of my arguments will have nothing to do with Jeffrey's criticism of orthodox Bayesianism, but there is one virtue of the deference-based conception that I can advertise now: unlike Jeffrey conditionalization, it has the resources to respond to Jeffrey's criticism.

2 The no constraints conception of rational belief updating

I do not say: the deference-based conception of rational belief updating is the right conception of rational belief updating. This is partly because I am aware of certain difficulties that the deference-based conception faces, but it is also because there is another conception of rational belief updating that in a way seems just as valid as the deference-based conception. I suspect there is a version of this conception that is compatible with Jeffrey's criticism, but the version I will present is not, so let's suppose for the rest of this section that there are propositions fully characterizing the empirical import of your experiences.

Let E be the conjunction of all the propositions fully characterizing the empirical import of all the experiences you have ever had. Call E your *total evidence*. Let S_E be the set of probability functions that a rational agent with total evidence E could have. Notice that, on the conditionalization conception, if it is possible to have no evidence at all then

$$S_E = \{\mu(\cdot \mid E) : \mu \in S_\emptyset\}.$$

(Here $S_\emptyset$ is the set of what David Lewis (1980) called reasonable initial credence functions.) The function that takes E to S_E characterizes the *static*, or *synchronic*, constraints of theoretical rationality. Saying what these static constraints are is a very hard problem unless, as de Finetti and Savage held, S_E is simply the set of all probability functions that assign probability 1 to E. My concern is with the *dynamic*, or *diachronic*, constraints of theoretical rationality, and that problem promises to be easier. The no constraints conception of rational belief updating says that there are no distinctively diachronic constraints of theoretical rationality on rational agents.[7] Theoretical rationality just requires that if your total evidence is E then your probability function is a member of S_E.[8]

Roger White (2005) and others have produced arguments for a thesis that White calls Uniqueness. Uniqueness says that the static constraints of theoretical rationality nail down exactly what your probability function has to be: S_E is always a singleton.[9] If Uniqueness is

[7] The restriction to *rational* agents is not idle. One could easily get the impression reading papers like Kolodny 2005 that diachronic rationality is primarily concerned with requirements on *irrational* agents: its function is to help you get back on track when you stray from the true path of rationality. In a probabilistic setting, it would be natural to ask of such a theory that it tell you how to change your credence function if it does not satisfy the axioms of probability. Presumably you should change it so that it does satisfy the axioms of probability, but it is implausible that how you should change it will depend only on its formal properties. This is the probabilistic analog of the problem of what rationality requires you to do if you have contradictory beliefs. Obviously, you should give up some of your beliefs so that your beliefs are no longer contradictory. Some ways of doing so will undoubtedly be more rational than others, but I don't know any substantive general principles concerning which ones you should give up. Similarly, I don't know any substantive general principles concerning how you should change your credence function if it is not a probability function. Therefore, I restrict my attention to constraints of rationality on *rational* agents. The no constraints conception says there are no *distinctively diachronic* ones, by which I mean there are no constraints that rule out patterns of belief change that the static constraints don't. For example, suppose that $S_\emptyset$ is a singleton and the conditionalization conception is true. Then S_E is a singleton for all non-contradictory propositions E, so, while in a sense there are diachronic constraints of rationality they don't rule out anything that the static constraints don't already rule out. The matter may be purely terminological, but I take this possibility to be one in which the conditionalization and no constraints conceptions are both right.

[8] *Objection.* Part of my evidence will bear on what my probability function is. Indeed, if I have perfect introspection, it is plausible that my evidence will *determine* what my probability function is, and this will ruin your way of setting things up. *Reply.* I concede there is a problem here, but I don't know any better way of setting things up. I leave it as a challenge to the defender of the no constraints conception to formulate the view more satisfactorily.

[9] We might also consider the view that S_E never has more than one element but is sometimes the

true then the no constraints conception is trivially true. But even if S_E can sometimes have more than one element, the no constraints conception could still be true if what I will call *free shifting* can be rational. I won't try to define free shifting; I will just give one clear-cut example of it.

> (**Example of Free Shifting**) Your total evidence is E, and your probability function is now $v_0 \in S_E$. You are certain that at noon tomorrow, you will change your probability function to $v_1 \in S_E$. This shift will not be a response to new evidence since you will be sure all along that you will shift.

First, let me respond to a possible objection to my way of describing the example: I do learn something when my probability function changes from v_0 to v_1, namely that my probability function is *now* v_1. If the objection is right, the example will need to be described in some more subtle way. Also, we may need to be more careful in explaining why free shifting is incompatible with the conditionalization conception. A particularly nice explanation would be that self-locating information should never impact your credence in non-self-locating propositions (but see Titelbaum 2008. The main reason I draw attention to the objection is just that I want to acknowledge that self-locating information raises tricky problems for diachronic rationality, and I want to set those problems aside.

Could free shifting be rational? I don't know, but there are apparently just three possible answers. Answer one: no, because Uniqueness is true. Fair enough. Answer two: no, even though Uniqueness is false. But why? I think the only reasonable answer to this question

empty set. For instance, suppose you think that a mind reader will read your credence in a proposition A at t and ensure that, whatever it is, the objective chance of A is something different. Then you might be in a *theoretical dilemma* at t, a situation in which there is no probability function you can rationally have. See [another manuscript of mine].

is that rational agents *defer* to the results of rational updating, in a sense that I will make more precise in the next section. The idea that rational agents defer to the results of rational updating is intuitive to many people but hard to motivate independently. This paper is devoted to arguing that anyone who accepts Answer Two should accept the deference-based conception of rational belief updating.[10] Answer three: yes, because the no constraints conception is correct. I have no answer to Answer Three. Maybe it is right. Or maybe there is just more than one way of using the word 'rational'. In fact, here are the beginnings of an argument for the no constraints conception:

> (**A Parfit-Inspired Thought**) Perhaps, if you radically changed both your probability and utility functions, but your total evidence did not change, that would be tantamount to ceasing to exist and spawning a new person. Ceasing to exist and spawning a new person, even in this fashion, doesn't seem like the sort of thing that theoretical rationality would forbid. But if theoretical rationality doesn't forbid radical discontinuities like this, why should it forbid smaller discontinuities like those exhibited in typical cases of free shifting?

The thought is not, as it stands, a compelling argument. But perhaps it could be fruitfully developed by advocates of the no constraints conception.

[10] Someone might doubt the rationality of free shifting on the grounds that you should not change your credences without good reason. (The idea is related to Gilbert Harman's (1986) Principle of Conservatism, though that principle says that it's OK to keep believing what you currently believe, not that it's not OK to start believing something else.) The problem with this line is that—so long as we are talking about cases in which you clearly understand that your new credences will be just as rational as your old ones—the only reasons I can see to go in for it are deference related.

3 Deference

Before I can say what the deference-based conception is, I must say what I mean by deference. On the one hand, deference is a technical notion. On the other hand, I don't (yet) have a precise definition of deference; it is an uncompleted project to say what exactly the right technical notion is, a project I'll sometimes call the project of formalizing deference. How do we go about it? Luckily, deference has at least two important applications besides the deference-based conception of rational belief updating, so one thing we can do is look at the formal notions that have been used in these applications and, if they are not adequate, at notions that haven't been used but perhaps should be. *Application One.* David Lewis (1980) formulated a principle he called the Principal Principle that was supposed to express the connection between rational credence and credence concerning what the objective chances are. Abstracting from weird cases where you get news from the future about the outcomes of chancy processes, the connection is: rational agents defer to the chances. Lewis ultimately (1994) rejected the Principal Principle in favor of a different principle that he called the New Principle.[11] Although Lewis saw matters differently, I think the right way to see this development is as a recognition that the Principal Principle was a flawed formalization of the idea that rational agents defer to the chances.[12] *Application Two.* Bas van Fraassen (1984) introduced a principle he called the Reflection Principle. I think we should understand the Reflection Principle as an attempt to formalize the idea that rational agents defer to their future selves. Many philosophers have convincingly argued that rational agents do not in fact always defer to their future selves, but con-

[11] The New Principle was independently introduced by Lewis and Ned Hall (1994) in response to Michael Thau's (1994) observation that the Principal Principle could be almost right in typical cases even if Lewis's theory of chance, according to which there is sometimes a chance that the chances are other than they actually are, was right.

[12] Unfortunately, I don't think anyone in the literature sees things quite this way. Ned Hall (2004) comes closest.

sider a different question: is the Reflection Principle even a correct formalization of the idea that they do? I say it isn't, though it works in important special cases.

Here is a conceivable problem that I think the Principal Principle and the Reflection Principle don't have:

> (**The Daphne Case**) Suppose you defer to God because you think He is omniscient: His probability function is the truth function. One day God tells you that, on a lark, He has set Daphne's probability function to the truth function, just for a few minutes. If the Principal Principle and the Reflection Principle embodied the correct formalizations of deference, you would count as deferring to Daphne, for these principles just state formal conditions on your probability function and cannot distinguish between deferring to God and deferring to Daphne. Intuitively, though, you defer only to God, not to Daphne.

While I grant that there is a fine-grained notion of deference according to which you do not defer to Daphne, this fine-grained notion is not the technical notion of deference I aim to capture. The technical notion that will explain the connection between rational credence and credence concerning the objective chances and do duty in the deference-based conception of rational belief updating is a more coarse-grained notion according to which you do defer to Daphne. This technical notion can be construed as a relation between probability functions and probability function-valued random variables.[13]

[13] You can think of a probability function-valued random variable as an *unknown* probability function that is modeled as a function from the set of possible worlds to the set of probability functions over those worlds. The value that a random variable Q takes world s to is called the value of Q in s. Certain grammatical conventions involving terms for random variables are standard in probability theory. For instance, if Q is a probability function-valued random variable and μ is a probability function then $Q = \mu$ is the set of worlds s such that the value of Q in s is μ, and I will assume that sets formed in this way are always in the domain of the agent's probability function. (In general,

With this in mind, let me introduce two possible formalizations of deference. They are the formalizations of deference that are implicit in the Principal Principle and the Reflection Principle, respectively. When we see their inadequacies, we will see that there is a problem about how to correctly formalize deference.

Let v be your probability function, and let Q be a probability function-valued random variable. (We will use 'v' and 'Q' in this way for the remainder of the paper.)

> You *globally defer* to Q if $v(\cdot \mid Q = \mu) = \mu$ for every probability function μ such that $v(Q = \mu) > 0$.[14]
>
> You *locally defer* to Q if $v(A \mid Q(A) = c) = c$ for every proposition A and number c such that $v(Q(A) = c) > 0$.

Global and local deference face two problems, one rather trivial, the other deeper. The trivial problem is what I call the Indiscreteness

this would require requiring random variables to satisfy a "measurability" criterion, but very shortly we will be supposing that *all* sets of worlds are in the domain of the agent's probability function, so we can forget about this complication.) Comparison with standard probability theory may be eased if you bear in mind that probabilists say 'event' where I say 'proposition', that they are more likely to deal with real-valued random variables than probability function-valued ones, and that they are more likely to write '$X = x$' than '$Q = \mu$'.

Here is why I say that deference *can be construed as* a relation between probability functions and probability function-valued random variables instead of saying that it *is* a relation between probability functions and probability function-valued random variables: deference is a technical notion, so we have some freedom in defining it. Just as propositions can be construed as either sets of possible worlds or functions from the set of possible worlds to the set $\{0, 1\}$, deference can be construed in various ways. For instance, it might be construed as a relation between probability function-valued random variables that happens to be *extensional* in its first argument, in the sense that it only depends on the value of its first argument in the actual world. Alternatively, deference might be construed as a relation between ideal agents (equipped, due to their ideality, with probability functions) and probability function-valued random variables. And there are further options.

[14] Notation: $v(\cdot \mid Q = \mu)$ is the result of conditioning v on the proposition that $Q = \mu$. As explained in the previous footnote, the proposition that $Q = \mu$ is the set of possible worlds s such that Q (thought of as a function from the set of possible worlds to the set of probability functions) takes s to μ. If it helps, think of 'μ' as a *hyperrigid designator* for a probability function, a designator that designates the same thing in every doxastically possible world. And think of 'Q' as non-hyperrigid designator for a probability function.

Problem. If your probability function is indiscrete then it might assign probability 0 to $Q = \mu$ for every μ and probability 0 to $Q(A) = c$ for every A and c.[15] And then you would count as globally and locally deferring to Q. But surely if $\nu(Q(A) \geq 1/2) = 1$ and $\nu(A) < 1/2$ then you shouldn't count as deferring to Q. This problem is rather trivial, and one of the earliest formalizations of deference, given by Haim Gaifman (1986), does not suffer from it. In the case of local deference, the problem can be solved by requiring that $\nu(A \,|\, Q(A) \in [a,b]) \in [a,b]$ for every proposition A and numbers a, b such that $\nu(Q(A) \in [a,b]) > 0$. Call this the *tweaked* version of local deference. It is also possible to formulate a tweaked version of global deference.[16] But just

[15] A probability function ν is *discrete* if the sum over all possible worlds s of $\nu(\{s\})$ is 1. Otherwise, i.e., if this sum is less than 1, ν is *indiscrete*.

[16] Formulating a tweaked version of global deference is not important for the purposes of this paper, but it is of some interest because it requires confronting two issues that are likely to come back again when we try to give a fully general formalization of deference.

A set S of probability functions is *convex* if it is closed under finite mixtures, which means that for any $\mu_1, \mu_2 \in S$ and $\lambda \in [0,1]$,

$$\lambda\mu_1 + (1-\lambda)\mu_2 \in S.$$

A first pass at a tweaked version of global deference is: ν tweakedly globally defers to Q if $\nu(\cdot \,|\, Q \in S) \in S$ for every convex set S of probability functions such that $\nu(Q \in S) > 0$. Unfortunately, closure under *finite* mixtures is not enough. Suppose there are propositions $A_1, A_2, \ldots$ and probability functions $\mu_1, \mu_2, \ldots$ such that, for all i and j, $\mu_i(A_j)$ is 1 if $i = j$ and 0 if $i \neq j$. Suppose further that $\nu(Q = \mu_i) = 2^{-i}$, for all i and that ν globally defers to Q. Then ν does not lie in the convex closure of the μ_i because every element of the convex closure of the μ_i assigns probability 0 to all but finitely many A_i but ν assigns non-zero probability to each A_i. What has gone wrong is that the convex closure of the μ_i is not closed under countable mixtures. The example shows that the tweaked version of global deference should only invoke sets that are closed at least under countable mixtures. Perhaps it should only invoke sets that are closed under yet more general mixtures.

If closure under mixtures were the only issue then we could allow open and half-open intervals in the tweaked version of local deference along with closed intervals, for these sets are closed under arbitrary "mixtures" (i.e., weighted averages). However, if we admit merely finitely additive probability functions then we should not admit open and half-open intervals in the tweaked version of local deference, and we must require some sort of topological closure condition—that I won't attempt to state—on the sets invoked by the tweaked version of global deference. For suppose that ν assigns probability 1 to $Q(A) < 1/n$ for all n but probability 0 to $Q(A) = 0$, as can happen if ν is merely finitely additive. Then, intuitively, ν can defer to Q, and deference to Q will require $\nu(A) = 0$. If we allowed open or half-open intervals in the tweaked version of local deference, however, tweaked local deference to Q would require

$$\nu(A) = \nu(A \,|\, Q(A) \in (0,1]) \in (0,1].$$

to simplify the discussion, let's suppose from now on that there are only finitely many possible worlds, so that the Indiscreteness Problem cannot even arise, and let's also suppose that probability functions are defined on every set of worlds.[17] The deeper problem is what I call the Imperfect Introspection problem. Let's say that a probability function-valued random variable Q *perfectly introspects* if $\mu(Q = \mu) = 1$, where μ is the value of Q in the actual world. Your own probability function might be, for you, an unknown probability function, that is, a probability function-valued random variable. *You* perfectly introspect if you are certain what your own probability function is and you are right. Perhaps, if you perfectly introspect, you don't defer to any random variables that don't perfectly introspect. But if you don't perfectly introspect, it is hard to see why you couldn't defer to a random variable that you think might not perfectly introspect. For instance, even if you don't perfectly introspect, perhaps you defer, in a degenerate sense, to your present self (that is, to your present probability function, the random variable that takes each possible world s to your present probability function in s). If you don't perfectly introspect but are certain that you are about to rationally update your beliefs then, if rational agents defer to the results of rational updating, you will defer, in a non-degenerate sense, to your post-update self, and surely there is no more reason for your post-update self to perfectly introspect than there is for you to. Granted, there may be reasons for thinking that rational agents must perfectly introspect, so that this case can never arise. Informal polling shows that philosophers have widely divergent views about whether a rational agent might fail to perfectly introspect, yet I know of no discussions of the issue worth citing. The way forward is to try to develop formal epistemology without the perfect introspection assumption and see what happens, and that is how

That is, tweaked local deference would require $\nu(A) > 0$. Since the original version of tweaked local deference already requires $\nu(A) < 1/n$ for all n, ν could not tweakedly locally defer to Q.

[17] I think everything in the paper would still be true if we made no restriction on the number of worlds but required all probability functions to be discrete and defined on every set of worlds.

I shall proceed.

Here is why I say that global and local deference are incorrect formalizations of deference except in the special case that v is certain Q perfectly introspects. As Lewis (1980) essentially understood,[18] unless v is certain Q perfectly introspects, it is not even possible for v to globally defer to Q.[19] For if v is not certain that Q perfectly introspects then there must be some μ such that $v(Q = \mu) > 0$ but $\mu(Q = \mu) < 1$. But then

$$v(Q = \mu \,|\, Q = \mu) = 1 \neq \mu(Q = \mu).$$

The case of local deference is subtler but equally compelling. Intuitively, the problem with global deference is that, in conditioning on the proposition that $Q = \mu$, you condition on something that Q might not know, and it is unfair to say that you don't count as deferring to Q just because, conditional on some proposition A that neither of you is certain of, you are certain of A, and Q (still) isn't. Of course, *conditional on* A, you are certain of A. But that is no achievement! Similarly, conditional on $Q(A) = c$, there may be respects in which you have an edge an epistemic edge on Q, and you might on that account have a conditional credence in A that is different from c, but intuitively that doesn't mean that you don't defer to Q, for Q didn't get the benefit of conditioning on $Q(A) = c$.[20]

I think: if you are certain Q perfectly introspects then global and local deference are both adequate formalizations of deference. I don't

[18] See the first displayed equation on page 291 of Lewis 1980.

[19] Remember that we are assuming that there are only finitely many possible worlds. If there are infinitely many possible worlds and v assigns probability 0 to $Q = \mu$ for all μ then, as already discussed, v will globally defer to Q.

[20] A less gestural argument that local deference is not necessary for deference is given in footnote 43.

Is local deference sufficient for deference? I doubt it. Suppose that $v(Q(A \,|\, Q(A) = c) = d) = 1$ for some $d \neq c$. Then, intuitively, you do not defer to Q unless $v(A \,|\, Q(A) = c) = d$, in violation of local deference. If you can nonetheless locally defer to Q then local deference is not sufficient for deference. I don't see why it wouldn't be possible to locally defer to Q in such a case, but I haven't taken the time to construct an explicit model to prove that it is.

really have any novel arguments to offer for this position, but both conceptions seem to function well in all the usual applications—so well that the position has gone all but unquestioned. Moreover, global and local deference are provably equivalent under these circumstances. (This equivalence was the "easy" part of Gaifman's (1986) Theorem 2, but the proof, and even the need for a proof, has been ignored by most authors writing on the Principal Principle and the Reflection Principle.)[21]

So what is the correct formalization of deference? I can point to two plausible conditions that it should satisfy. Together these give an interesting necessary condition for deference, but I don't know if it is a sufficient condition. I also know a plausible sufficient conditions for deference, but it is clearly not necessary. So there is work to be done!

A special case of the first plausible necessary condition is that if you are certain that $Q = \mu$ then you don't defer to Q unless your probability function is μ. Another special case is that if you are certain that $Q(A) = c$ then you don't defer to Q unless your credence in A is c. A set S of probability functions is *convex* if it is closed under finite mixtures, which means that for any $\mu_1, \mu_2 \in S$ and $\lambda \in [0,1]$,

$$\lambda\mu_1 + (1-\lambda)\mu_2 \in S.$$

The fully general condition is: if S is a convex set of probability functions and $\nu(Q \in S) = 1$ then ν doesn't defer to Q unless $\nu \in S$. (Compare van Fraassen's (1995) General Reflection Principle, which has a similar flavor.) The second plausible necessary condition is that if ν defers to Q then $\nu(\cdot \mid A)$ defers to $Q(\cdot \mid A)$ for any proposition A such that $\nu(A) > 0$.[22] I don't have any very substantive arguments that these two

[21] Gaifman does not include the proof in his paper, and I know of no published proofs. Therefore, I have included a proof (actually, two proofs) in Appendix 2.

[22] Notation: $Q(\cdot \mid A)$ is the random variable whose value in world s is the result of conditioning the value of Q in s on A. An obvious problem is that there may be some worlds s such that the value of Q in s assigns probability 0 to A. If ν defers to A and $\nu(A) > 0$ then the result of conditioning ν on A will assign probability 0 to such worlds, so it is irrelevant what $Q(\cdot \mid A)$ is in such worlds. I will

conditions are necessary for deference.[23] They just seem plausible to me, and they give good results so far as I've tested them. Also, global and local deference satisfy them. (I leave this to the reader.)

If we combine the two conditions, we get a notion I call S-deference, which is the largest relation satisfying the two conditions.

> You *S-defer* to Q if $v(\cdot\,|\,B) \in S$ for every proposition B and convex set of probability functions S such that $v(Q(\cdot\,|\,B) \in S\,|\,B) = 1$.

I don't know if S-deference is sufficient for deference, but I have some very weak evidence that it is. Consider NP-deference, which I take to be the notion of deference implicit in the New Principle.

> You *NP-defer* to Q if $v(\cdot\,|\,Q = \mu) = \mu(\cdot\,|\,Q = \mu)$ for all μ such that $v(Q = \mu) > 0$.[24]

ignore this glitch. One fix for it is to construe deference as a relation between probability functions and *partially defined* random variables, that is *partial* functions from the set of possible worlds to the set of probability functions. Then we can say that $Q(\cdot\,|\,A)$ is defined for all A but that it only has a value in those worlds in which Q assigns non-zero probability to A.

[23] Though perhaps this is an argument for the necessity of the second condition: If you defer to someone who you think updates rationally and then it is publically announced that A, you should update by conditionalizing on the proposition that it is publicly announced that A, and you should continue deferring to that person.

[24] The New Principle, in its original formulation, involves quantifying over theories of chance and total histories of the world and is not obviously equivalent to the principle that rational agents NP-defer to the chances. Still, that does seem to be the core idea behind the New Principle. James Joyce (2007, 198) evidently agrees, for he introduces a principle that he calls the New Principle that says you must NP-defer to the chances, and he does not distinguish this principle from the original New Principle. NP-deference is a global notion, in the sense that it involves conditioning v on very specific propositions that specify Q in full detail. Joyce realizes that it would be convenient to have a local notion analogous to local deference, or even a local necessary condition for deference. Joyce produces a local principle that he claims is a "special case" of the New Principle. Abstracting from the special case of deference to the chances and putting things in the notation we have been using, what Joyce essentially claims is that NP-deference implies what I will call J-deference, defined as follows: you *J-defer* to Q if

$$v(A\,|\,Q(A) = c \text{ and } Q(A\,|\,Q(A) = c) = d) = d$$

whenever $v(Q(A) = c \text{ and } Q(A\,|\,Q(A) = c) = d) > 0$. I suspect Joyce's claim is false, for J-deference has the same basic problem as local deference: v gets the benefit of conditioning on something that Q

NP-deference is a great response to the arguments we have seen that global and local deference are not necessary for deference. The problem with global deference was that it required conditioning ν on $Q = \mu$ but didn't give Q the benefit of conditioning on $Q = \mu$. NP-deference remedies this problem in the most straightforward possible way, by giving Q the benefit of conditioning on $Q = \mu$. However, NP-deference is insufficient for deference because you can NP-defer without satisfying either of the plausible necessary conditions for deference.[25] S-deference implies NP-deference,[26] so S-deference implies global and local deference in cases where you are certain Q perfectly introspects. (If you are certain Q perfectly introspects, you NP-defer to Q iff you globally defer to Q.) This is all the evidence I have that S-deference is sufficient for deference.

Later on, we will encounter a case where it is clear that you defer to a random variable Q even though you are not certain that Q perfectly introspects. This is because you satisfy the following very plausible sufficient condition for deference to Q: you are certain that, for every proposition A, Q's probability for A is at least as close to the truth value of A (thought of as either 0 or 1) as your own. This sufficient condition for deference is not necessary, for it is not satisfied in many cases in which you are certain that Q perfectly introspects and globally defer to Q. One possible approach to showing that S-deference is a correct formalization of deference is to find some weaker sufficient condition for deference that can be expressed in terms of your confi-

doesn't (namely, $Q(A\,|\,Q(A) = c) = d$). It might be nice to have a local version of NP-deference, but as far as I can tell, there is none.

[25] It is easy to see that you can NP-defer without satisfying the first plausible necessary condition for deference, but it takes a little work to establish that you can NP-defer without satisfying the second condition. See Appendix 3 for a proof. An important question for reductionists about chance is: is the New Principle an adequate expression of the deference we owe the chances or does the inadequacy of NP-deference mean that one can satisfy the New Principle without giving the chances their due? I don't know the answer.

[26] *Proof.* Suppose that you S-defer to Q, and let μ be such that $\nu(Q = \mu) > 0$. Set B to $Q = \mu$ and S to $\{\mu(\cdot\,|\,Q = \mu)\}$. Then $\nu(B) > 0$ and $\nu(Q(\cdot\,|\,B) \in S\,|\,B) = 1$, so $\nu(\cdot\,|\,B) \in S$, which is to say that $\nu(\cdot\,|\,Q = \mu) = \mu(\cdot\,|\,Q = \mu)$, as desired.

dence that Q is in some sense more accurate than you and to prove that this sufficient condition is equivalent to the condition that you S-defer to Q.[27]

Appendix 2 contains some additional information about formal notions of deference, but there is still much that I do not understand. There may be probably some simple observation I am overlooking that makes everything much simpler than it seems to me.

4 The deference-based conception of rational belief updating

Here are two closely related guiding ideas behind the deference-based conception of rational belief updating. One, if you are certain you are about to rationally update your beliefs then (at least if you fully understand the nature of rational belief updating) you defer to your post-update self. Two, rational agents defer to the results of rational updating. Maybe these guiding ideas are wrong; maybe the no constraints conception is right. But let's suppose that they are right. As we will see in the next section, the conditionalization conception can explain why they are true. The deference-based conception reverses the usual order of explanation of the guiding ideas and defines rational belief updating as updating that you defer to. Here is my official statement of the deference-based conception: to rationally update is to change your probability function to the value of a random variable you defer to.[28] In the next subsection, I will put a little more flesh on the proposal, but first let me make two general remarks. One, even supposing the falsity of the no constraints conception, I only really ad-

[27] See the last sentence of footnote 44 for another possible approach.

[28] Of course, so changing your probability function will normally require having some kind of learning experience (though see the discussion of spontaneous belief updating in subsection 6.1). If you defer to a random variable whose actual value is not your probability function then you must not be certain what its actual value is, by the first necessary condition for deference.

vocate the view that the deference-based conception provides a core necessary condition on rational belief updating. We will see many cases of belief updating that is rational according to the deference-based conception but not obviously rational full stop, and there will always be the possibility of adding extra conditions that rational belief updating must satisfy to the deference-based conception in order to rule some of these forms of belief updating out. I myself am most tempted to do this as a response to the "ignoring experience" problem discussed in subsection 4.2. Others may want to add extra conditions to rule out some of the cases discussed in subsection 6.2. Two, the deference-based conception is interesting even if the no constraints conception is correct. The no constraints conception can be understood as saying that theoretical rationality does not require a certain sort of personal continuity; it allows changes that differ only in degree from something that might amount to ceasing to exist and spawning a new person. If that is right, the deference-based conception can become a conception of personal continuity. It can become a conception of what it means to remain exactly the same person. (At least when supplemented with an account of what it means to remain the same person insofar as one's desires, or utilities, go. See Appendix 1 for a discussion of the prospects for a deference-based conception of rational utility function updating that, if a no constraints conception of rational utility function updating is true, could be reconstrued as a discussion of remaining the same person insofar as one's utilities go.)

Now let's turn to an important lack of detail in the deference-based conception that I call the grain problem.

4.1 The grain problem

The grain problem for the deference-based conception is the problem of saying something substantive about what it takes to count as changing your probability function to a random variable you defer to. An extremely coarse-grained answer to this problem is that you change your probability function to the value of a random variable you defer to any time your probability function changes to a probability function that happens to be the actual value of a random variable you defer to. This answer is obviously too coarse grained, but what can stand in its place? It can't be that you must *intentionally* change your probability function to the value of a random variable you defer to; to start with, rational belief updating is not normally intentional at all. It can't be that your probability function must change *in all nearby possible worlds* to the value of a random variable you defer to, at least if 'nearby' is not meant in a so far unexplicated technical sense, for you can rationally update even if you could easily have failed to.

The grain problem is my name for what is not really a novel problem. The grain problem is an instance of the problem of what it is to follow a rule, and it is closely related to the problem of the basing, or grounding, relation.[29] It is a serious problem for the deference-based conception. But is it a *worse* problem for the deference-based conception than it is for the conditionalization conception? One way in which it seems worse is that there are many different ways that your probability function could change that would be rational according to the extremely coarse-grained version of the deference-based conception because, unless you think you are omniscient, there will be many different random variables that you defer to, whose actual values may differ—for instance, the objective chance function and the truth function. On the other hand, given that you undergo a certain experience,

[29] Cf. James Pryor on the grounding relation:

> We introduced the notion of a ground to distinguish between cases where you believe P *for* good reasons, or on grounds that justify you in believing P, and cases where you believe P on bad grounds, ones that do not justify that belief. What does it take for your belief to be *grounded* on some fact or condition C that you are in? A natural thought is that your belief counts as so grounded iff it is formed (or sustained) in a way that is guided by the epistemic norm "When in C, believe P." If that is right, then the best way to understand the grounding relation is by inquiring into what it takes to be guided by such a norm. (2005, 195)

there is only one way your probability function can change that will be rational according to an extremely coarse-grained version of the conditionalization conception. But, for all that, the extremely coarse-grained version of the conditionalization conception is still obviously too coarse grained. If a gamma ray blast causes your probability function to change to the result of conditioning your current probability function on the proposition that fully characterizes what it's like to take a gamma ray blast to the brain, that is not rational. I will gesture at two possible approaches to the grain problem, but (unsurprisingly) I don't have anything terribly satisfying to offer.

The first approach I call the process proposal. There are versions of the process proposal for both the deference-based and conditionalization conceptions. The process proposal for the deference-based conception says that to rationally update is to update by a correctly-functioning process that is such that, when it functions correctly, it changes your probability function to the value of a random variable you defer to. The process proposal for the conditionalization conception says that to rationally update is to update by a correctly-functioning process that is such that, when it functions correctly, it changes your probability function to the result of conditioning your probability function on the proposition that fully characterizes the empirical import of your new experience. The process proposal is too vague to be satisfying. Saying what the relevant process is looks like a horrible problem akin to the generality problem for reliabilism. Saying what it means for that process to function correctly is yet a further problem. Despite these problems, the process proposal will be my official proposal.

The second approach I call the dispositions proposal. According to the dispositions proposal we should, at least to begin with, forget about trying to classify individual cases of belief updating as rational or irrational. Instead, we should just aim to characterize the belief updating dispositions of rational agents. In every case, we might say, there is some particular random variable that such agents are disposed to change their probability function to. The dispositions proposal is initially less ambitious than the process proposal since it doesn't aim to classify individual cases of belief updating as rational or irrational. Perhaps for that reason it will be easier to make the proposal more substantive. Maybe, in the end, the dispositions proposal can be as ambitious as the process proposal: maybe its advocates can derivatively classify particular cases of belief updating as rational or irrational depending on whether they are expressions of good dispositions, just as virtue theorists in ethics can derivatively classify particular acts as right or wrong.[30]

4.2 The ignoring experience problem

Here is a second pressing problem. According to the deference-based conception, it is rational to *never* change your probability function, not even when you have experiences that you really should be learning from. It is OK to ignore your experiences. This problem shows that the deference-based conception needs supplementation, not that it is altogether mistaken. In fact, the conditionalization conception would suffer from the ignoring experience problem if it didn't presuppose that there is some proposition fully characterizing the empirical import of every experience you ever have. On that supposition, we can adequately supplement the deference-based conception by saying that whenever you have an experience, the proposition that fully characterizes its empirical import has to go to 1. So the conditionalization conception doesn't really have an advantage over the deference-based conception in this department. It is just that the conditionalization conception is limited to cases in which the ignoring experience problem is easy to solve.

So what is the right supplement? Maybe we should add some kind

[30] Thanks to XXX for impressing on me the potential advantages of the dispositions proposal.

of “greediness” requirement that says you must update your probability function whenever you can, and as much as you can. I don’t know how to spell out such a requirement in any detail, so I will leave the ignoring experience problem for another day. There are other fish to fry.

5 The connection between conditionalization and deference[31]

There is a close connection between the deference-based and conditionalization conceptions for agents who are certain they will always perfectly introspect. Suppose, for this section, that you are such an agent. If you are certain that Q is the result of conditioning your probability function on a true proposition then you defer to Q. (That is, you globally defer to Q. Equivalently, you locally defer to Q.)[32] In par-

[31] Much of the substance of this section can be found in Weisberg 2007, though Weisberg’s overall argument is in a sense opposed to mine since he wants to vindicate the conditionalization conception against the charge that it carries with it a commitment to the Reflection Principle, while the deference-based conception competes with the conditionalization conception and is inspired by the Reflection Principle.

[32] *Proof.* Let’s write Q_s for the value of Q in s and G for the set of worlds s such that $v(\{s\}) > 0$. For each $s \in G$, let E_s be some true proposition such that $Q_s = v(\cdot \mid E_s)$. Let’s say that a proposition A *almost entails* a proposition B if $v(A \setminus B) = 0$.

> *Lemma.* For all $s \in G$, E_s almost entails $Q = Q_s$, and $Q = Q_s$ almost entails E_s.
>
> *Proof.* Since v is certain that Q perfectly introspects, conditional on E_s, v is certain that $Q = Q_s$. But that means that E_s almost entails $Q = Q_s$. (Notice that this direction doesn’t depend on E_s being true in s.) For the converse, suppose for the sake of a contradiction that $Q = Q_s$ does not almost entail E_s. Then there is some $s' \in (Q = Q_s \setminus E_s) \cap G$. Since $Q_s = Q_{s'}$,
>
> $$Q_{s'}(\{s'\}) = Q_s(\{s'\}) = 0.$$
>
> But since $E_{s'}$ is true in s',
>
> $$Q_{s'}(\{s'\}) = v(\{s'\} \mid E_{s'}) \geq v(\{s'\}) > 0,$$
>
> contradiction. (Notice that this direction doesn’t depend on v being certain Q perfectly introspects.)

ticular, if Q is the result of conditioning your probability function on the proposition fully characterizing the empirical import of the experience you have at t,[33] you defer to Q. Thus, if you update by conditionalization, you update by changing your probability function to a random variable you defer to—at least if we do not give too fine grained an answer to the grain problem. Conversely, if you globally defer to a random variable Q then changing your probability function to the value of Q can be represented as conditionalization on what Q is. (This follows directly from the definition of global deference: $v(\cdot \mid Q = \mu) = \mu$ for all μ such that $v(Q = \mu) > 0$.) In particular, if you defer to your post-update self, your update can be represented as conditionalization on what your new probability function is (cf. Skyrms 1980, Appendix 2). It follows from all this that if we are to distinguish between the deference-based and conditionalization conceptions, we should either distinguish between updating that can be represented as conditionalization and updating that is really *by* conditionalization or we should drop the assumption that you are sure you will perfectly introspect. In the next section we will explore both of these avenues.

It follows from the lemma that for all $s \in G$, conditioning v on E_s is equivalent to conditioning v on $Q = Q_s$. From this, it easily follows that v globally defers to Q. For let μ be such that $v(Q = \mu) > 0$, and let s be an arbitrary member of $(Q = \mu) \cap G$. Then

$$v(\cdot \mid Q = \mu) = v(\cdot \mid E_s) = \mu.$$

[33] This means: for every world s, the value of Q in s is the result of conditioning your probability function on the proposition fully characterizing the empirical import of the experience you have at t in s.

6 Arguments that the deference-based conception is superior to the conditionalization conception

The arguments will fall into three categories. One, I will argue that the deference-based conception is *conceptually prior* to the conditionalization conception, in the sense that no one who didn't already accept the guiding ideas that lead to it would have any reason to accept the conditionalization conception. Two, I will argue that updating that can be represented as conditionalization on what your new probability function is need not be updating by conditionalization. Three, I will discuss a case in which you do not perfectly introspect and are not certain that you will perfectly introspect after updating. In this case, your updating cannot even be represented as conditionalization.

6.1 The conceptual priority of deference

We will inspect two prominent arguments for conditionalization and see that neither has any force against anyone who doesn't already accept the guiding intuitions that that lead to the deference-based conception. In particular, the arguments have no force against a serious advocate of the no constraints conception.[34] These two arguments are not the only arguments that have ever been given for conditionalization, but as far as I can tell, but I am not aware of any arguments for conditionalization that are immune to the sort of criticism I will offer. I conclude that deference to the results of rational updating is, in a sense, conceptually prior to conditionalization, and I take this conceptual priority to constitute some evidence that, if the guiding intuitions behind the deference-based conception are true, then the deference-based conception is right and the conditionaliza-

[34] Curiously, these arguments have no force against an advocate of the deference-based conception either, as I will discuss in section 7.

tion conception is wrong. Not decisive evidence, for the conditionalization conception adds distinctive claims to the deference-based conception: that when a rational agent has an experience, there is some proposition fully characterizing its empirical import that the agent becomes certain of, and that the agent proceeds by conditionalizing on this proposition. I doubt these claims, but if they are true, the conditionalization conception is right, the conceptual priority of deference notwithstanding. Now to the two prominent arguments for conditionalization.

Argument one: diachronic Dutch books. Suppose you are going to learn which element of a partition is true. A *deterministic updating rule* is a function that assigns to each element E of the partition a probability function v_E such that $v_E(E) = 1$. The diachronic Dutch book argument says: if your probability function is v and you are going to use a deterministic updating rule, you should use the rule that has $v_E = v(\cdot|E)$ because, if you use any other rule, a Dutch bookie can make money off of you for sure by placing bets with you before and after you update.[35] It is much debated whether diachronic Dutch book invulnerability is necessary for rationality. I think it is if rational agents have to defer to the results of rational updating and it isn't if they don't. But all I want to establish here is the second direction, that diachronic Dutch book invulnerability isn't necessary for rationality unless rational agents have to defer to the results of rational updating. There is no need for a fancy argument for this claim. The argument is just that if you think that you are going to rationally update, but you don't defer to the result of that updating, then you should be unmoved by the observation that you are going to lose money for certain. "Of course *that guy* (my future self) is going to make some bets that I wouldn't," you will think, if you know that a Dutch bookie lurks, "and that will cost us money. But I don't think what he thinks, and the

[35] According to Brian Skyrms (1997, 286), philosophers generally attribute the diachronic Dutch book argument to David Lewis (reported in Teller 1973), while statisticians think it is implicit in older work by de Finetti.

bets I am prepared to make will decrease our expected loss (calculated according to *my* probability function).” For a serious advocate of the no constraints conception, the diachronic Dutch book argument is no more troubling than the fact that a bookie can make money for sure by placing bets with two different people.

Argument two: cognitive decision theory. Consider the same setup as before: choosing a deterministic updating rule. Hilary Greaves and David Wallace (2006) suppose that you have an epistemic utility function U that takes a world s and a probability function μ and returns the epistemic utility of having probability function μ in world s. Greaves and Wallace define a probability function μ to be *strongly self-recommending* with respect to an epistemic utility function U if, for all $\mu' \neq \mu$,

$$\sum_s \mu(\{s\})U(s,\mu) > \sum_s \mu(\{s\})U(s,\mu').$$

They show that if, for each E in the partition you are updating on, $\nu(\cdot\,|\,E)$ is strongly self-recommending with respect to your epistemic utility function, then the rule that has $\nu_E = \nu(\cdot\,|\,E)$ uniquely maximizes your expected epistemic utility. This argument, too, would be completely unconvincing to anyone who doubted that rational agents always defer to the results of rational updating. For instance, in a case of free shifting, of course your post-shift probability function has a lower expected epistemic utility than your pre-shift probability function, if the expectation is taken using your pre-shift probability function. Calculate the expectation using your post-shift probability function, and the reverse is true. None of this should faze an advocate of the no constraints conception.

6.2 Updating that can be represented as conditionalization versus updating by conditionalization

Suppose for this subsection that you are sure you will always perfectly introspect. We saw in section 5 that updating by changing your probability function to the actual value of a random variable Q that you defer to can be represented as conditionalization on what Q is. In particular, if you defer to your post-update self then your update can be represented as conditionalization on what your post-update probability function is. Nonetheless, there are cases where you are updating in accord with the deference-based conception, but you are not conditionalizing on any proposition whatsoever. I will give three examples in which you are not updating by conditionalization, but you are arguably updating rationally. What they have in common is that your updating is not a predictable response to evidence. Then I will briefly discuss how we might enrich our formal model so that updating that is not by conditionalization cannot misleadingly be represented as conditionalization.

Example one: spontaneous belief updating. Suppose you are certain that at time t a mind reader will read your mind to determine your credence in a proposition A and then set the objective chance of A to your credence in A at t.[36] Consider the following scenario. Your credence in A is now 1/2. Based on really good evidence, you are rationally certain that your credence in A will stay at 1/2 until t and that at t your credence in A will change to either 1/4 or 3/4. You have credence 1/2 in each possibility, so you satisfy the Principal Principle throughout.[37] Your update at t is *spontaneous*: not a response to the world. Of course, when your probability function changes at

[36] The example is not as far-fetched as it may sound. For instance, suppose that the "mind reader" is Becky, your lab partner in an experimental philosophy class. Becky is going to "read your mind" by *asking* you, after t, what your credence in A at t was. You will tell Becky the truth, and she will input the number you tell her into a quantum mechanical random number generator, which flashes green with chance the number she inputs into it. Let A be the proposition that the random number generator flashes green the first time it is used after t.

[37] Perhaps it could never be rational to have all these certainties, but they can be relaxed without losing the essence of the case. We could say instead that your are rationally *nearly* certain that your credence in A will stay at *nearly* 1/2 until *nearly t* and that at *nearly t* your credence in A will change to either *nearly* 1/4 or *nearly* 3/4. We could say that you have credence *nearly* 1/2 in each possibility, so you *nearly* satisfy the Principal Principle throughout. (Understand 'nearly 1/2' to mean 'a value near 1/2', 'nearly t' to mean 'a time near t', etc.)

t, you learn using your perfect introspection what its new value is, so your update can be represented as conditionalization on what the new value is. Spontaneous belief updating is clearly a case of updating in accordance with the deference-based conception that is not by conditionalization. But is it rational?

A reason to think that it is is that you might be convinced by some of the other arguments for the deference-based conception. On the other hand, maybe you will conclude from those arguments and Example One that the conditionalization and deference-based conceptions are both wrong or that the deference-based conception is basically right but needs some kind of supplement to rule out rational spontaneous updating. The dialectical situation will be similar with our other examples: we will examine a form of belief updating; it will be, perhaps, unclear whether it is rational; and we will conclude that the other arguments for the deference-based conception lend support to the claim that it is rational, though another possible response to the data is to supplement the deference-based conception to rule out the mooted form of updating.

Example two: black box learning.[38] In the next example, your updating is responsive to the world. We might say that it has a mind-to-world direction of fit, whereas the updating in the previous example had a world-to-mind direction of fit. Example Two is formally just like Example One: your credence in A is now $1/2$, you expect it to change to $1/4$ or $3/4$ at t, etc. The difference is that you don't think that your credence in A at t will have any causal influence on whether A is true; rather, you expect to *learn*; you expect to know better at t what to think about A. There are two versions of the example. In the first version, your updating is a response to experience, though not to evidence (conceived propositionally). This is essentially what Jef-

[38] Skyrms (1997, 287–8) writes, "the epistemic agent starts with an initial probability, pr_1, passes through a 'black-box' learning situation, and comes out with a final probability, pr_2. We are not supposed to speculate on what goes on inside the black box." This passage is the origin of the locution 'black box learning', but I conceive of black box learning somewhat differently from Skyrms.

frey had in mind in his discussion of seeing a cloth by candlelight and changing your credence about what color it is.[39] In the second, more radical, version, your updating is not a response to experience at all: you expect your probability function to change, responsively to the world, but you don't expect there to be any experience mediating the change. Of course, due to your perfect introspection, you will have the experience of your probability function changing at t, but your updating will not be a response to this experience. Is the more radical kind of updating rational? I have little intuition for this question, but the dialectical situation is much as it was in Example One.

Example three: confidence in your future self. In our last example, your updating is a response to evidence; it is just not a *predictable* response to evidence. Suppose your conditional probability for A given E is now 1/2 and that you expect to learn at t whether E. Moreover, you expect that if you learn that E at t, you will update your credence in A not to 1/2 but to either 1/4 or 3/4. And you think that you will update in a good way: you defer to your post-update self. Sometimes, it is only when the evidence is before us that we know how to evaluate it, and you think this is one of those times. (Perhaps certain cases where our emotions are involved are often like this. Suppose that you are nearly certain that your girlfriend is faithful and that E is a proposition you have a very low credence in that, if true, would constitute evidence that your girlfriend is cheating on you. You might think to yourself, "I don't know what I would think if I learned E; how can I speculate on such a far-fetched possibility? Nonetheless, if I did learn E, I think I would react appropriately.") Example Three is extreme in that you are *certain* you will have a good response to E, but it is possible to tell more complicated and realistic stories. For instance, maybe

[39] However, in section 1, I claimed that according to Jeffrey, "there may be no proposition A whatsoever that you become certain of such that, conditional on A, your prior credence that the cloth is green is 2/3." Now we see that, given our perfect introspection assumption, there may be such a proposition; for instance, the proposition that your new credence that the cloth is green is 2/3, and that you arrive at it rationally, approximately fits the bill. It is just that you don't update by conditionalizing on this proposition, or on any other proposition.

you will only trust your credence in A after learning E if it is not too extreme, or if it is close to your prior conditional credence in A given E.[40]

Assuming that your language is rich enough to fully characterize the empirical import of any possible experience, updating by conditionalization seems to be updating that can be represented as conditionalization on a proposition that is given in experience *prior* to the update. Now, if you were not an ideal agent, when you updated in response to experience, there would presumably be some time lag between the time at which you had the experience and the time at which you updated, so this notion of priority could be understood in the ordinary temporal sense. But when we idealize, we suppose that there is no time lag in your response to evidence, and we lose the ability to make out this priority relation in the usual way. Since the priority relation is real, it would be nice to have a way of representing it in our idealized formalism. Here are two brief thoughts in this direction. One, we could model the stages of evolution of an agent's probability function using a linear order that is finer than the linear order given by the real numbers. For instance, we could model instants of time as pairs (t, n) of a real number and a natural num-

[40] Bryan Weatherson (2007) claims that orthodox Bayesianism and a version of imprecise probabilism he calls *static Keynesianism* don't allow agents to learn about how evidence bears on hypotheses, and he develops a version of imprecise probabilism he calls *dynamic Keynesianism* that does. Without embarking on a full discussion of Weatherson's paper, I want to make three quick points about it. One, Weatherson is wrong to think that there is an intimate connection between imprecise probabilism and learning about how evidence bears on hypotheses. Example Three can be seen as an example of how agents can learn about how evidence bears on hypotheses even if their credences are precise. Two, imprecise probabilists who want to allow for such learning should not necessarily embrace dynamic Keynesianism, for dynamic Keynesianism has limitations that have not been sufficiently motivated. According to dynamic Keynesianism, all learning about how evidence bears on hypotheses has to happen by means of probability functions dropping out of the agent's representor. Probability functions can never get added back into the representor, and updating still involves conditionalization. Three, the core idea that Weatherson and I share can be separated from the anti-skeptical purposes to which Weatherson wants to put it. I don't think the idea in Example Three has any anti-skeptical bite whatsoever; I advocate the traditional Bayesian view that an anti-skeptical bias should be encoded in your prior.

ber. If you update by conditionalization at t, what happens is that you have an experience at $(t,0)$ and then your probability function changes at $(t,1)$. Two, we could say that if you have an experience at t, your probability function at t is your old probability function, and it is not until after t that you have your new probability function. (Your probability function, viewed as a function of time, will thus be left-continuous at t but not right-continuous instead of right-continuous but not left-continuous.) Either way, some updating that is not really by conditionalization can no longer misleadingly represented as conditionalization. The updating in Examples One, Two, and Three will happen at $(t,0)$, or in a right-continuous fashion, while updating that is genuinely by conditionalization will happen at $(t,1)$, or in a left-continuous fashion. There may be other examples of updating that is not by conditionalization that these models do not sort out correctly. I don't know. I just wanted to point out that we might gain the ability to represent some important distinctions by enriching our model.

6.3 Imperfect introspection cases

We have seen some good reasons to prefer the deference-based conception to the conditionalization conception, but we have also seen that the two conceptions can be hard to distinguish from one other. They come apart dramatically when we drop the assumption that rational perfectly introspect (and are certain that they will). Consider the following toy example. There are only two possible worlds, and your credence is evenly split between them. Let Q be the random variable that in each world gives probability 3/4 to that world and 1/4 to the other world. Then you are certain that Q does not perfectly introspect,[41] so we can't use global deference as a criterion for deference,

[41] *Proof.* Let the possible worlds be named s_1 and s_2. Then the proposition $\{s_1\}$ is the same proposition as $Q(\{s_1\}) = 3/4$, so, if the actual world is s_1 then $Q(\{s_1\}) = 3/4$, but

$$Q(Q(\{s_1\}) = 3/4) = 3/4 \neq 1.$$

but you satisfy the plausible sufficient condition for deference to Q: you are certain that, for every proposition A, Q's probability for A is at least as close to the truth value of A as your own. If you update by changing your probability function to the value of Q, you would seem to be eminently rational.[42] However, such an update is not representable as conditionalization, for there is no proposition that you become certain of.[43]

Let me conclude by pointing out an interesting implication of the deference-based conception of rational belief updating: evidence might be less central to epistemology than we thought.

7 Objections and limitations

I will consider a bad objection to the deference-based conception, an interesting limitation of the conception that I don't see as an objection, and a meritorious objection that deserves further consideration.

Bad objection: the standard arguments for conditionalization show that the deference-based conception is wrong. Surprisingly, the diachronic Dutch book and cognitive decision theory arguments for conditionalization don't really favor the conditionalization conception over the deference-based conception, for two reasons. One,

Similarly, if the actual world is s_2 then $Q(\{s_1\}) = 1/4$, but

$$Q(Q(\{s_1\}) = 1/4) = 3/4 \neq 1.$$

So either way Q doesn't perfectly introspect.

[42] Of course, if you learn what Q is, you don't update to Q, since Q doesn't know what Q is. It might seem mysterious how you could update to Q without learning what Q is. I don't think it should. You might just have some learning experience whose effect is that your probability function changes to the value of Q.

[43] With the example in hand, I can give a less gestural argument that local deference is not necessary for deference than I could in section 3. Let the possible worlds be named s_1 and s_2. You do not locally defer to Q because

$$v(\{s_1\} \mid Q(\{s_1\}) = 3/4) = 1 \neq 3/4.$$

Since you do defer to Q, local deference is not necessary for deference.

these arguments don't distinguish between updating that can be represented as conditionalization and updating by conditionalization. You are invulnerable to diachronic Dutch books and maximize expected epistemic utility in all three of the examples given in subsection 6.2. Two, as I presented them, the standard arguments don't even get off the ground in imperfect introspection cases because they presuppose that you learn for certain what the true member of a partition is and have to update your probability function as a function of what you learn.[44]

Interesting limitation: it is hard to give a forgetting-friendly ver-

[44] There is a generalized version of the diachronic Dutch book argument that does apply to imperfect introspection cases, but the agent in our imperfect introspection example is not susceptible to such a generalized diachronic Dutch book. (Perhaps there is also a generalized version of the cognitive decision theory argument. If there is, I doubt it would convict the agent in our imperfect introspection example of irrationality either.) Here is the generalized diachronic Dutch book argument that I have in mind. Here is a stipulative definition.

> A pair (v, Q) is *diachronically Dutch bookable* if there are packages of bets Γ and Δ such that v judges Γ fair, and, in every possible world, Q judges Δ fair, but, taken together, Γ and Δ result in a sure loss.

If *you* update by changing your probability function to Q then I will say that *you* are diachronically Dutch bookable if (v, Q) is diachronically Dutch bookable.

The way the original diachronic Dutch book argument is usually presented, the package of bets the bookie sells the agent depends on which member of the evidence partition is true, so it might seem possible to be subject to a diachronic Dutch book in the original sense without being subject to a diachronic Dutch book in the generalized sense. Actually, the dependence of the package offered on the evidence is inessential, so the original diachronic Dutch book argument really is a special case of the generalized one, a fact that I will use the remainder of this paragraph to sketch a proof of. First, by adjusting the payoffs of packages of bets, we may assume that the packages the bookie offers all have cost zero (assuming we allow negative payoffs). Second, if the agent learns E, he will judge fair a package of bets Δ_E iff he judges fair a package of bets that has the same payoffs as Δ_E if E is true and payoff 0 otherwise. For each E in the partition $\mathscr{E}$, let Δ'_E be a zero cost package of bets that has the same payoffs as Δ_E if E is true and payoff 0 otherwise. Instead of selling the agent Δ_E if E is learned, the bookie can sell the agent $\bigcup_{E \in \mathscr{E}} \Delta'_E$ whatever is learned.

The agent in the imperfect introspection example is not diachronically Dutch bookable. If he were Dutch bookable, there would have to be some world in which Γ lost money; without loss of generality, suppose it is s_2 and that Γ loses \$1 in s_2. Then Γ must win at least \$1 in s_1 if the agent judges it fair. So Δ must lose more than \$1 in s_1 if Γ and Δ together result in a sure loss. But if Q judges Δ fair in s_1 then Δ must win more than \$1 (in fact, more than \$3) in s_2. But Γ only loses \$1 in s_2, so Γ and Δ do not together result in a sure loss.

Conjecture: v defers to Q if and only if (v, Q) is not diachronically Dutch bookable.

sion of the deference-based conception. The conditionalization conception can easily be generalized to accommodate agents who are ideally rational except that they sometimes forget that they have had certain experiences. Let E be the conjunction of all the propositions fully characterizing the empirical import of all the experiences you remember having had. At least if we can make sense of initial credence functions—the credence function you had prior to having any experiences—the conditionalization conception can be generalized to say that, whatever your initial credence function was, your current probability function should be the result of conditioning it on E. There is no obvious way to generalize the deference-based conception in a similar way. Indeed, on the deference-based conception, it is hard to make sense of what forgetting even amounts to. I take this fact to be an interesting feature of the deference-based conception but not an objection to it.

Meritorious objection: the deference-based conception can't accommodate cases in which I gain misleading evidence that I am irrational. David Christensen (2010) gives the following example. You think that someone might drug you tomorrow morning in an undetectable way with a drug that will make you badly assess how evidence bears on scientific hypotheses. Moreover, the drug will also tamper with your memory so that will think you have always assessed evidence the way you do after taking the drug. (Thus, you will not be able to have accurate beliefs about scientific hypotheses by deferring to your pre-drug judgments.) In such a case, if you do get drugged tomorrow morning, you will be irrational, but if you don't, it seems that you can be rational and that the rational thing to do is to cautiously back off from any strong judgments you are inclined to make about how evidence bears on scientific hypotheses. But before Monday morning, you shouldn't defer to the result of this rational Monday morning update.[45] I take Christensen's example to pose a real prob-

[45] More carefully, let Q be the random variable whose value in s is your Monday morning cre-

lem for the deference-based conception, but I don't take the problem to be damning. For one thing, you know in advance that you are going to update in this way, so your change Monday morning is a response to purely self-locating information. And we know already that cases involving self-locating information pose special problems and that it might be best to leave them for another day. For another thing, as Christensen (2007) argues, cases similar to this one involve conflicting rational ideals and pose general problems for a wide variety of epistemological views.

8 Appendix 1: a deference-based conception of rational utility function updating?

Rationality—or personal continuity, if the no constraints conception is correct—requires some sort of diachronic coherence between your probability functions at different times. The sort of coherence it requires has been the topic of this paper. What does rationality—or personal continuity—require in the way of coherence between your utility functions at different times? The traditional answer, which has a lot going for it, is that it requires your utility function (thought of as a function on worlds, not propositions) to remain constant. Now it is prima facie very plausible that if the deference-based conception of rational belief updating is right, so is a deference-based conception of rational utility function updating, according to which to rationally change your probability function is to change it to the value of a utility-function valued you *practically* defer to. If these the deference-based conception of rational utility function updating coincided with the traditional conception, that would be further evidence for the deference-based conception of rational belief updating,

dence if you rationally update in s and whose value in other worlds is something you defer to, such as your pre-Monday morning credences or the objective chances. You don't defer to Q even though you are sure that Q is the result of rational updating.

for there is no practical analog of the conditionalization conception, and we should prefer a unified explanation of theoretical and practical rationality to a disunified one.[46] In order to see whether the two conceptions coincide, we need an account of practical deference. If practical deference turns out to be as rich and interesting as theoretical deference, the two conceptions are unlikely to coincide. But if it turns out to be trivial and uninteresting then they might. I will now sketch a theory of practical deference according to it is trivial and uninteresting. Unfortunately, this theory will fail to vindicate the deference-based conception of rational utility function updating. That conception will turn out not to coincide with the traditional theory but rather to be a crazy conception that no one should accept.

A Theory of Practical Deference. Suppose you practically defer to a utility function-valued random variable U (however such deference should be cashed out). Then you may be ignorant of the actual value of U, but you do know, for each possible world s, the number that the value of U in s takes s to. It strikes me that all of the action-guiding information in U is contained in this "diagonal" utility function. If you defer to U, what you ought to do is to act in such a way that it is likely that the value of U in the actual world takes the actual world to a high number. In other words, to practically defer to U is just for your utility function to be equal to the diagonal of U.

The theory of practical deference sketched in the previous paragraph succeeds in making practical deference trivial and uninteresting. Unfortunately, it still fails to make the deference-based conception coincide with the traditional one, for changing your utility function to the actual value of a utility function-valued random variable you defer to can change your probability function. Moreover, it can lead you to give arbitrary utilities to every world except the actual world, so it is crazy. If the sketched theory of practical deference is

[46] Note that hoped-for unification would be in two ways partial. One, it would do nothing to unify the static constraints of theoretical and practical rationality. Two, it would do nothing to unify the constraints of theoretical and practical rationality on irrational agents.

correct, the theoretical and practical components of diachronic rationality cannot be unified the way hoped for.

9 Appendix 2: two proofs that global and local deference are equivalent, an observation about NP-deference, and random variables that cannot be deferred to

I have claimed that global and local deference are both appropriate formalizations of deference to Q if you are certain that Q perfectly introspects. This claim would be hard to sustain if global and local deference were not equivalent under these circumstances. Luckily, they are. The equivalence was first proved by Gaifman (1986), but Gaifman didn't publish his proof; he just claimed to have one (as he undoubtedly did, since the proof is not very difficult). I will give two proofs of the equivalence. The first is simpler and in a sense more constructive, but the second is more enlightening and looks to be better suited to generalization. Moreover, the second proof will lead to an important observation about NP-deference: for any v and Q, so long as there is no μ such that $v(Q = \mu) > 0$ but $\mu(Q = \mu) = 0$, there is exactly one v' such that v' has the same distribution as v on what Q is and v' NP-defers to Q. This is a neat property of NP-deference that any strictly stronger notion of deference, such as S-deference, must lack. Sometimes, it will be impossible to S-defer to a random variable without changing your opinion about what it is. Since S-deference is necessary for deference, sometimes it will be impossible to defer to a random variable without changing your opinion about what it is.[47] Finally, we will strengthen this result slightly: I will give an ex-

[47] I wonder if this fact has implications for the debate about peer disagreement. To my knowledge, no one in that debate has looked at cases where two peers disagree over what one of them believes. If my hunch is right, no simple "split the difference" sort of view will generate a unique

ample of a random variable Q that never assumes a value μ such that $\mu(Q=\mu)=0$ but is not S-deferred to by any probability function whatsoever.

9.1 First proof

First, note that global deference implies local deference since, assuming that ν globally defers to Q and writing S for $\{\mu : \nu(Q=\mu) > 0 \text{ and } \mu(A)=c\}$,

$$\begin{aligned}
\nu(A \,|\, Q(A)=c) &= \sum_{\mu\in S} \nu(A \,|\, Q=\mu)\, \nu(Q=\mu \,|\, Q(A)=c) \\
&= \sum_{\mu\in S} c\, \nu(Q=\mu \,|\, Q(A)=c) \\
&= c.
\end{aligned}$$

For the converse implication, suppose that ν locally defers to Q and is certain Q perfectly introspects, and let μ be such that $\nu(Q=\mu)>0$. Let A be an arbitrary proposition. We will prove that $\nu(A \,|\, Q=\mu)=\mu(A)$. Since A is arbitrary, it follows that ν globally defers to Q.

If $Q=\mu$ and Q perfectly introspects then $Q(A\cap Q=\mu)=\mu(A)$, so, since ν is certain Q perfectly introspects,

$$\nu(Q(A\cap Q=\mu)=\mu(A)) \geq \nu(Q=\mu) > 0,$$

so, since ν locally defers to Q,

$$\nu(A\cap Q=\mu \,|\, Q(A\cap Q=\mu)=\mu(A)) = \mu(A).$$

If $\mu(A)=0$ then $\nu(A \,|\, Q=\mu)$ must equal zero, for otherwise $\nu(A \,|\, Q(A)=0)$ would have to be greater than zero, in violation of local deference. So we may assume that $\mu(A)>0$. In that case, using again ν's certainty

recommendation in such cases.

that Q perfectly introspects,

$$v(Q=\mu\,|\,Q(A\cap Q=\mu)=\mu(A))=1.$$

It follows from the last two displayed equations that

$$v(A\,|\,Q(A\cap Q=\mu)=\mu(A))=\mu(A).$$

Since $\mu(A)>0$ and v is certain Q perfectly introspects,

$$v(Q(A\cap Q=\mu)=\mu(A) \text{ iff } Q=\mu)=1.$$

It follows from the last two displayed equations that $v(A\,|\,Q=\mu)=\mu(A)$, as desired.

9.2 Second proof

The second proof is new in only one direction: we take the proof that global deference implies local deference from the previous subsection. For the converse, first note that if v locally defers to Q then v is completely determined by its distribution on what Q is since, for any proposition A,

$$v(A)=\sum_c v(Q(A)=c)v(A\,|\,Q(A)=c)=\sum_c v(Q(A)=c)c.^{48} \tag{1}$$

Therefore, if $v\neq v'$ but v and v' have the same distribution on what Q is, v and v' can't both locally defer to Q. For future reference, let's name this property the At Most One property.

> A relation R between probability functions and probability function-valued random variables has the *At Most One* property if, for all v and Q, there is at most one probability

[48] I write '$\sum_c$' for a sum over all numbers c, and below I will write '$\sum_\mu$' for a sum over all probability functions μ. These sums will always have only finitely many non-zero terms.

> function ν' that has the same distribution as ν on what Q is and bears R to Q.

A kind of opposite of the At Most One property is the At Least One property.

> A relation R between probability functions and probability function-valued random variables has the *At Least One* property if, for all ν and Q, there is at least one probability function ν' that has the same distribution as ν on what Q is and bears R to Q.

The unqualified At Least One property is not very interesting since none of the notions of deference we have been looking at have it. In particular, global deference lacks the At Least One property since ν cannot globally defer to Q unless ν is certain Q perfectly introspects. Nonetheless, global deference does enjoy the restriction of the At Least One property to those ν and Q such that ν is certain Q perfectly introspects, as I will now show. Suppose that ν is certain Q perfectly introspects and set $\nu' = \sum_\mu \nu(Q = \mu)\mu$. I will show that ν' has the same distribution as ν on what Q is and that ν' globally defers to Q. For the first claim, note that for any probability function μ_0,

$$\begin{aligned} \nu'(Q = \mu_0) &= \sum_\mu \nu(Q = \mu)\, \mu(Q = \mu_0) \\ &= \nu(Q = \mu_0)\, \mu_0(Q = \mu_0) \\ &= \nu(Q = \mu_0), \end{aligned}$$

where the first equation is by the definition of ν' and the second and third are true because ν is certain Q perfectly introspects. To see that ν' globally defers to Q, let μ_0 be such that $\nu'(Q = \mu_0) > 0$, and let A be arbitrary. Then

$$\begin{aligned}
\nu'(A \mid Q=\mu_0) &= \frac{\nu'(A, Q=\mu_0)}{\nu'(Q=\mu_0)} \\
&= \frac{\sum_\mu \nu(Q=\mu)\, \mu(A, Q=\mu_0)}{\sum_\mu \nu(Q=\mu)\, \mu(Q=\mu_0)} \\
&= \frac{\nu(Q=\mu_0)\mu_0(A)}{\nu(Q=\mu_0)} \\
&= \mu_0(A),
\end{aligned}$$

so ν' globally defers to Q. (The first equation is true by the definition of conditional probability, the second by the definition of ν', the third by the fact that ν is certain Q perfectly introspects, and the fourth by algebra.)

Now suppose that ν is certain Q perfectly introspects and that ν locally defers to Q. By the restricted At Least One property of global deference, there must be some ν' that has the same distribution as ν on what Q is and globally defers to Q. Since global deference implies local deference, ν' locally defers to Q. By the At Most One property of local deference, $\nu' = \nu$, so ν globally defers to Q.[49]

[49] The method of proof used in this subsection has wider application. For instance, say that ν *defers to Q in expectation* if $\nu(A)$ is the expectation under ν of $Q(A)$, for all A. Local deference implies deference in expectation, but it has occasionally been suggested that the reverse implication may fail. However, deference in expectation has the At Most One property—indeed, equation (1), which was our proof that local deference has the At Most One property, only makes use of deference in expectation. Therefore, deference in expectation is equivalent to local and global deference under certainty of perfect introspection. Generalizations to cases in which the discreteness assumption is dropped will be more difficult, but this proof still looks like a better place to start than the previous one. It will just be harder to prove that the relevant notions of deference have the At Most One and At Least One properties.

9.3 An observation about NP-deference, and random variables that cannot be deferred to

Since global and local deference are equivalent for all v and Q such that v is certain Q perfectly introspects, they can't differ as to whether they have the restrictions of the At Most One and At Least One properties to these v and Q: they both have both of these properties. Let K be the class of pairs (v, Q) such that there is no μ such that $v(Q = \mu) > 0$ and $\mu(Q = \mu) = 0$. Intuitively, K is a very large class. NP-deference has the At Most One and At Least One properties restricted to K (i.e., restricted to those v and Q such that $(v, Q) \in K$). (I leave this to the reader.) Any notion of deference that is strictly stronger than NP-deference for members of K must lack the restriction of the At Least One property to K. Since S-deference is strictly stronger than NP-deference for members of K, and S-deference is necessary for deference, deference itself lacks the restriction of the At Least One property to K: sometimes, you can't come to defer to someone without changing your opinion about what they believe (and this can happen even when the two of you are in K). When you think about it, that isn't so surprising. If an agent is quite unsure what her probability function is, and you are certain what her probability function is, you don't defer to her about what her probability function is, so you don't defer to her *tout court*, and you can't come to do so unless you change your opinion about what her probability function is. It turns out that something even a bit stronger is true. Let's say that a probability function is *regular* if it assigns non-zero probability to every world. If Q is regular in every world then it never assumes a value μ such that $\mu(Q = \mu) = 0$. I will construct a probability function-valued random variable that is regular in every world and that no probability function whatsoever S-defers to.

Let the possible worlds be s_1,s_2, and s_3, and let the value of Q in s_i be μ_i for each i, where the μ_i are any regular probability functions satisfying the following inequalities:

$$\mu_1(\{s_3\}) > \mu_1(\{s_1\}) > \mu_1(\{s_2\}),$$
$$\mu_2(\{s_1\}) > \mu_2(\{s_2\}) > \mu_2(\{s_3\}),$$
$$\text{and } \mu_3(\{s_2\}) > \mu_3(\{s_3\}) > \mu_3(\{s_1\}).$$

Suppose for a contradiction that v S-defers to Q. It is clear that v cannot assign probability 1 to any singleton $\{s_i\}$, so $v(\{s_i, s_j\}) > 0$ for all i and j such that $i \neq j$. Setting B to $\{s_1, s_2\}$ and S to $\{\mu : \mu(s_1) > \mu(s_2)\}$, we have

$$v(Q(\cdot \mid B) \in S \mid B) = 1,$$

so $v(\{s_1\} \mid B) > v(\{s_2\} \mid B)$, so $v(\{s_1\}) > v(\{s_2\})$. Similarly, $v(\{s_2\}) > v(\{s_3\})$, and $v(\{s_3\}) > v(\{s_1\})$. But this is a contradiction.

10 Appendix 3: a proof that NP-deference is not closed under conditioning

Recall our second plausible necessary condition for deference: if v defers to Q then $v(\cdot \mid A)$ defers to $Q(\cdot \mid A)$ for any proposition A such that $v(A) > 0$. Let's say that a notion of deference that satisfies this condition is *closed under conditioning*. In this appendix, I will construct an example to show that NP-deference is not closed under conditioning. I will first present the example in a schematic form and then give an explicit model to demonstrate that the various stipulations in the example are jointly consistent.

Suppose you NP-defer to Susie the Mystic Pundit. Susie has no clue how confident she is that Obama will get a second term (in fact her credence in this proposition is 1/2), but one thing she is almost sure of is that she is an infallible expert at predicting the outcomes of presidential elections. Letting B be the proposition that Obama gets a second term, Q be the random variable whose value in each world s is

Susie's probability function in s, and μ be the value of Q in the actual world,

$$\mu(Q(B)=1\,|\,B)=.99$$

and

$$\mu(Q(B)=0\,|\,\text{not-}B)=.99.$$

As a Susie-head, you know exactly what Susie's probability function is (i.e., $\nu(Q=\mu)=1$). So

$$\begin{aligned}\nu(\cdot\,|\,Q(\cdot\,|\,B)=\mu(\cdot\,|\,B),B) &= \nu(\cdot\,|\,B)\\ &= \nu(\cdot,B)/\nu(B)\\ &= \mu(\cdot,B\,|\,Q=\mu)/\mu(B\,|\,Q=\mu)\\ &= \mu(\cdot\,|\,Q=\mu,B).\end{aligned}$$

Suppose Susie knows exactly how she thinks the world must be if Obama will get a second term, in the sense that

$$\mu(Q(\cdot\,|\,B)=\mu(\cdot\,|\,B)\,|\,B)=1.$$

Let's say that you *CNP-defer* to Susie if your probability function is related to Q by the largest relation that entails NP-deference and is closed under conditioning.

If you CNP-defer to Susie,

$$\begin{aligned}\nu(\cdot\,|\,Q(\cdot\,|\,B)=\mu(\cdot\,|\,B),B) &= \mu(\cdot\,|\,Q(\cdot\,|\,B)=\mu(\cdot\,|\,B),B)\\ &= \mu(\cdot\,|\,B).\end{aligned}$$

But $\mu(\cdot\,|\,B)$ cannot equal $\mu(\cdot\,|\,Q=\mu,B)$ because

$$\mu(Q(B)=1\,|\,B)=.99$$

but

$$\mu(Q(B)=1\,|\,Q=\mu,B)=0.$$

Therefore, you do not CNP-defer to Susie.

Here is an explicit model to demonstrate that the various stipulations in the example are jointly consistent. There are four possible worlds, s_1, s_2, s_3, and s_4, so Q can assume four possible values, μ_1, μ_2, μ_3, and μ_4. Let $B = \{s_1, s_2\}$. Let μ_2 assign probability .495 to each of s_1 and s_4 and probability .005 to each of s_2 and s_3. Let $\mu_1 = \mu_2(\cdot \mid B)$; let $\mu_3 = \mu_2$; and let $\mu_4 = \mu_2(\cdot \mid \text{not-}B)$. Let $\nu(\{s_2\}) = \nu(\{s_3\}) = 1/2$. Then, one can check, all the stipulations in the example are realized. In particular, ν NP-defers to Q, but ν does not CNP-defer to Q.

11 References

Chalmers, David J. Forthcoming a. The nature of epistemic space. In *Epistemic Modality,* eds. Andy Egan and Brian Weatherson. Oxford University Press.

Chalmers, David J. Forthcoming b. Propositions and attitude ascriptions: a Fregean account. *Noûs.*

Christensen, David. 2007. Does Murphy's law apply in epistemology? Self doubt and rational ideals. *Oxford Studies in Epistemology,* vol. 2, ed. Tamar Szabo Gendler, 3–31.

Christensen, David. 2010. Higher-order evidence. *Philosophy and Phenomenological Research* 81: 185–215.

Diaconis, Persi and Sandy L. Zabell. 1982. Updating subjective probability. *Journal of the American Statistical Association* 77: 822–30.

Gaifman, Haim. 1986. A theory of higher order probabilities. In *Proceedings of the 1986 Conference on Theoretical Aspects of Reasoning about Knowledge,* ed. Joseph Y. Halpern. Morgan Kaufmann.

Greaves, Hilary and David Wallace. 2006. Justifying conditionalization: conditionalization maximizes expected epistemic utility. *Mind* 115: 607–32.

Hall, Ned. 1994. Correcting the guide to objective chance. *Mind* n.s. 103: 505–17.

Hall, Ned. 2004. Two mistakes about credence and chance. *Australasian Journal of Philosophy* 82: 93–111.

Harman, Gilbert. 1986. *Change in View.* MIT Press.

Jeffrey, Richard. 1965. *The Logic of Decision.* McGraw Hill.

Kolodny, Niko. 2005. Why be rational? *Mind* n.s. 114: 509–63.

Lewis, David. 1980. A subjectivist's guide to objective chance. In *Studies in Inductive Logic and Probability*, vol. 2, ed. Richard C. Jeffrey, 263–93. University of California Press.

Lewis, David. 1994. Humean supervenience debugged. *Mind* n.s. 103: 473–90.

Pryor, James. 2005. There is immediate justification. In *Contemporary Debates in Epistemology*, eds. Matthias Steup and Ernest Sosa, 181–202. Blackwell.

Skyrms, Brian. 1980. *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws.* Yale University Press.

Skyrms, Brian. 1997. The structure of radical probabilism. *Erkenntnis* 45: 285–97.

Soames, Scott. 1987. Direct reference, propositional attitudes, and semantic content. *Philosophical Topics* 15: 44–87.

Soames, Scott. 2008. Why propositions can't be sets of truth-supporting circumstances. *Journal of Philosophical Logic* 37: 267–76.

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

Thau, Michael. 1994. Undermining and admissibility. *Mind* n.s. 103: 491–503.

Titelbaum, Michael. 2008. The relevance of self-locating beliefs. *Philosophical Review* 117: 555–605.

van Fraassen, Bas C. 1984. Belief and the will. *Journal of Philosophy* 81: 235–56.

van Fraassen, Bas C. 1995. Belief and the problem of Ulysses and the Sirens. *Philosophical Studies* 77: 7–37.

Weatherson, Brian. 2007. The Bayesian and the dogmatist. *Proceedings of the Aristotelian Society* 107: 169–85.

Weisberg, Jonathan. 2007. Conditionalization, reflection, and self-knowledge. *Philosophical Studies* 135: 179–97.

Weisberg, Jonathan. 2009. Commutativity or holism? A dilemma for conditionalizers. *British Journal for the Philosophy of Science* 60: 793–812.

Weisberg, Jonathan. Forthcoming. Varieties of Bayesianism. In *Handbook of the History of Logic,* vol. 10, eds. Dov Gabbay, Stephan Hartmann, and John Woods. Elsevier.

White, Roger. 2005. Epistemic permissiveness. *Philosophical Perspectives* 19: 445–59.