Rational Feedback

Grant Reaber

(Northern Institute of Philosophy, University of Aberdeen)

ABSTRACT. Suppose you think that whether you believe some proposition A at some future time t might have a causal influence on whether A is true. For instance, maybe you think a woman can read your mind, and either (1) you think she will snap her fingers shortly after t if and only if you believe at t that she will, or (2) you think she will snap her fingers shortly after t if and only if you don't believe at t that she will. Let A be the proposition that she snaps her fingers shortly after t.

In case (1), theoretical rationality seems to leave it open whether you should believe A or not. Perhaps, for all it has to say, you could just directly choose whether to believe A. David Velleman seems to be committed to something close to that, but his view has been unpopular.

In case (2), you seem to be in a *theoretical dilemma*, a situation where any attitude you adopt toward A will be self-undermining in a way that makes you irrational. Such theoretical dilemmas ought to be impossible, just as genuine moral dilemmas ought to be impossible, but it is surprisingly hard to show that they are (perhaps because they aren't).

I study cases analogous to (1) and (2) in a probabilistic framework where degrees of belief rather than all-or-nothing beliefs are taken as basic. My principal conclusions are that Velleman's view is closer to the truth than it is generally thought to be, that case (2) type theoretical dilemmas only arise for hyperidealized agents unlike ourselves, and that there are related cases that *can* arise for agents like us that are very disturbing but might not quite amount to theoretical dilemmas.

1 Introduction

1.1 Non-probabilistic rational feedback

Suppose you have decided that tomorrow you will ask your classmate Laila to go on a date with you, and let 't' be a name for the time at which you will ask her out. You may think that Laila will be able to tell from your demeanor at t whether you think she will say yes, even if you try to hide what you think. Or you might expect to *tell* Laila whether you think she will say yes, perhaps because you think she will ask you whether you do, and either you want to be honest and forthcoming or you think she'll be able to tell if you lie and will punish you for lying or not being forthcoming, perhaps by saying no. In all these cases, you credit Laila with an ability that I will call, as others have, a mind reading ability. If Laila can read your mind then whether you believe she will say yes might have a causal influence on whether she does.

Possibility One. Suppose you think that Laila will say yes if and only if you think at t that she will. Does theoretical rationality say to believe that she will say yes or not to? Could both attitudes be rational? Let's say that you *directly choose* what to believe if you make yourself believe something (or fail to believe it) without performing any mediating action, such as an action that makes the content of your belief true (or might make it false). Does theoretical rationality prohibit directly choosing what to believe?

Possibility Two. Suppose you think that Laila will say yes if and only if you don't think at t that she will. Then it seems that you are bound to be theoretically irrational at t: if you believe Laila will say yes, you will believe it while believing that this will make her say no, and if you don't believe Laila will say yes, you will withhold belief while believing that this will make her say yes. I will call cases in which there is no rational thing to think theoretical dilemmas. I find it intuitively obvious that there are not really any theoretical dilemmas, just as I find it

2

intuitively obvious that there are not really any moral dilemmas. But paradoxes, like good theories, can force us to reject what seems intuitively obvious. Can the paradox threatened by Possibility Two be blocked?

1.2 Roadmap

Questions of theoretical rationality are best studied using a notion of graded belief, so, in the next subsection, I introduce a probabilistic framework that we will use for the rest of the paper. The remainder of the section situates our topic by discussing some of its neighbors and generalizations. Section 2 discusses probabilistic analogs of Possibility One, while sections 3 and 4 discuss probabilistic analogs of Possibility Two. Section 3 surveys arguments that theoretical dilemmas are strictly impossible. Unfortunately, these arguments all fail. Section 4 surveys arguments that theoretical dilemmas can only arise for idealized agents unlike ourselves. These arguments fare better. Finally, section 5 discusses intriguing puzzle cases that can arise even without the idealizations necessary to construct theoretical dilemmas along the lines of Possibility Two.

1.3 The probabilistic framework

I will assume that your beliefs can be represented by a *credence function*, a function that maps propositions to numbers in the unit interval $[0,1]$. I will also assume that, if you are rational, your credence function will satisfy the axioms of probability theory, so we can also call it your *probability function*.

Suppose you are certain that a mind reader will read your mind at a time t to determine your credence in a proposition A and then see to it that the chance of A at some later time u is

$$f(\text{your credence in } A \text{ at } t),$$

where f is some function from the unit interval to itself whose identity you know.[1] If f is non-constant, there is what, for lack of a better name, I will call *rational feedback*. You think your credence in A could have a causal influence on whether A is true, and that introduces a kind of feedback term into the equations of theoretical rationality. (This "feedback" is actually what I think is the central phenomenon; objective chance is a foreign intrusion into the subject. In a previous version of this paper I tried to avoid invoking chance by working in the framework of causal decision theory, but certain difficulties in that approach led me to the present one.)

To complete the framework, we will need a couple of further strong assumptions. (Later in the paper we will look at weakening them.) Let P_t be your probability function at t. Let's say *you know at t that it's t* if P_t assigns probability 1 to the proposition that it's t. And let's say *you know at t what your probability function is* or *you perfectly introspect at t* if P_t assigns probability 1 to the proposition that your current probability function is what P_t actually is. The strong assumptions are that you know at t that it's t and that you perfectly introspect at t.

My main claim about the setup is that, if you are rational, $P_t(A)$ is a fixed point of f—that is, $P_t(A) = f(P_t(A))$.[2] For if $P_t(A) \neq f(P_t(A))$, you have one credence in A at t but think that having that credence in A at t will make the chance of A something else. Since you know that it is t, and you know what credence you have in A, you are irrational. I think this informal argument is enough to establish the main claim, but since it is so important (given that I do work in a chance-based framework), I will now present a more formal argument invoking David Lewis's (1980) Principal Principle.

The Principal Principle is cumbersome to state exactly, and maybe

---

[1] The reason I make u later than t instead of equal to t is just that it makes the mind reader's job a little easier and thus avoids a potential quibble about whether the setup is possible.

[2] Thus, there will be a problem corresponding to Possibility Two precisely if f doesn't have any fixed points.

f(c)   Figure 1A   f(c)   Figure 1B   f(c)   Figure 1C
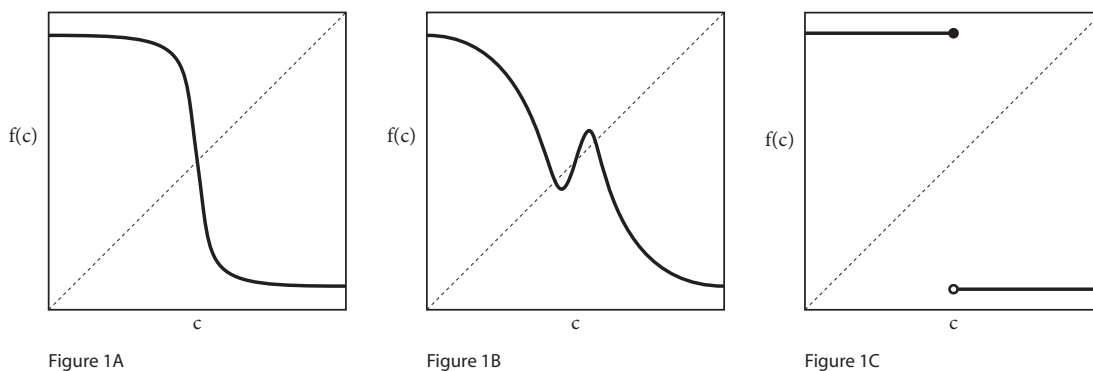
Lewis didn't even get it exactly right, but for our purposes the following approximation is close enough: your conditional credence in A at t, given any proposition wholly about history up to time u (counting what the chances are at u as part of history up to u) that entails that the chance of A at u is c, should be c. Suppose that you satisfy the Principal Principle, and let $P_t(A) = c$. Then

$$
\begin{aligned}
P_t(A) &= P_t(A \mid P_t(A) = c) \\
&= P_t(A \mid \text{chance}_u(A) = f(c), P_t(A) = c) \\
&= f(c) \\
&= f(P_t(A)).
\end{aligned}
$$

The first equation holds because $P_t(P_t(A) = c) = 1$, which is true because you know at t that it's t, you perfectly introspect at t, and $P_t(A) = c$; the second equation holds because you are certain that if $P_t(A) = c$ then $\text{chance}_u(A) = f(c)$; the third equation holds because you satisfy the Principal Principle and the propositions that $\text{chance}_u(A) = f(c)$ and $P_t(A) = c$ are wholly about history up to u; and the fourth equation holds because $P_t(A) = c$.

Figure 1 graphs three possible values for f. The fixed points of these functions are the places where their graphs touch the main diagonal, drawn with a dashed line. If f has a unique fixed point, as in Figure 1A, what rationality requires is clear: your credence in A at t must equal that fixed point (though there may be questions about the path

5

it will take to get there and the rational forces that will push it along that path). If f has multiple fixed points, as in Figure 1B, we have a probabilistic analog of Possibility One. And if f has no fixed points, as in Figure 1C, we have a probabilistic analog of Possibility Two.

The remainder of section 1 deals with special topics that are for the most part independent of one another and of the rest of the paper.

1.4 Newcombian rational feedback

My focus is on cases where there is rational feedback because you think that there might be causal influence running from your credence in A to A itself. We might term this sort of rational feedback *causal* feedback. I think there might be another sort of rational feedback that I will call *Newcombian* feedback. I have in mind cases like the following. Suppose that you are rationally certain that some utterly reliable predictor has predicted $P_t(A)$ and set the chance of A to $f(P_t(A))$, where f is some function from the unit interval to itself whose identity you know. Then (it seems to me) rationality mandates that $P_t(A)$ be a fixed point of f even though A may be a proposition about the past that couldn't possibly be influenced by $P_t(A)$. And presumably, if there are realistic versions of Newcomb's paradox, there will be realistic cases of Newcombian feedback, cases that do not depend on such strong assumptions. I'll have nothing more to say specifically about Newcombian feedback. If it does exist, it might be important to study how it interacts with causal feedback.

1.5 The value of evidence

One of the standard nice things that can be proven in the subjective Bayesian framework is that your expected utility for looking at cost-free evidence is always non-negative (and usually positive) if you

6

think you will act so as to maximize your expected utility.[3] This result threatens to fail if you think your beliefs may have effects in the world that are not mediated by your actions (whether or not you think they may have effects on their *own* truth). For suppose a mind reader vows to kill you if you come to have a credence in A near 0 or 1 and then offers to tell you whether A is true. And suppose your credence in A is middling, but you think that if the mind reader told you what she thought about A, your credence in A would go near 0 or 1, and then the mind reader would make good on her threat. Then you should decline if you value your life. So if the result is to hold good, the evidence the mind reader offers you about A must count as not being cost-free. But it is unclear how to define 'cost-free' in such a way that it doesn't so count without trivializing the result.[4]

1.6 Honest public predictions

Emile Grunberg and Franco Modigliani (1954) asked whether it is possible to make a non-self-falsifying public prediction of the outcome of an election or other social event if you think that your prediction might influence the outcome of the event. A prediction is *non-self-falsifying* if it is a fixed point of the function f that takes a possible prediction x to your prediction on the supposition that you publicly make prediction x. For instance, your prediction may take the form of an estimate of the share of the vote that candidate K will win. If you are an influential blogger, it might be reasonable for you to think

---

[3]Ramsey was the first to prove this result, but it was rediscovered by others before being found in his Nachlass and published as Ramsey 1990.

[4]However, see Kadane, Schervish, and Seidenfeld 2008 for a definition of 'cost-free' that tries to thread this needle. It is worth pointing out that the result also threatens to fail if you attach utility to the wrong things. We have already considered the unfortunate case of honest agents—suppose you are a dyed-in-the-wool Kantian and the mind reader vows to kill you unless you *tell* her, after she's told you what she thinks about A, that your credence in A is middling. Another problem is cases in which you attach value to having mental states rather than just to things in the world. For example, Kadane et al. consider a case in which you are willing to pay a taxi driver not to tell you the ending of the play you are on your way to see.

your prediction might influence the election. For instance, maybe you think a low estimate of K's share would help K by creating an "underdog effect." Or maybe you think a high estimate would be more likely to help K, by creating a "bandwagon effect." If f has multiple fixed points, there are multiple non-self-falsifying predictions that you can make. If you like K, but you want make an honest prediction, you should presumably predict the biggest fixed point (unless you think you can help K even more by keeping your mouth shut). On the other hand, if f has no fixed points then it is impossible to make a non-self-falsifying prediction, so if you want to be honest, you'd better keep your mouth shut.

Herbert Simon (1954) pointed out that if f is continuous then, by Brouwer's fixed point theorem, it must have at least one fixed point,[5] and he argued that there is good reason to think that f will normally be continuous so that non-self-falsifying public prediction will normally be possible.

1.7 Two dimensions of generalization

We have been supposing that you think a mind reader will measure your credence in a single proposition at a single time. More generally, you might think your credence in several different propositions will be measured at several different times. I focus on the special case of a single proposition and time for simplicity and also because generalizing is fairly straightforward. But I want to flag two issues.

First, the idea that continuous functions have fixed points generalizes to the case where there are finitely many pairs (A, t) such that

---

[5]Brouwer's fixed point theorem says that, for any n, any continuous function from a compact, convex subset of n-dimensional Euclidean space to itself has a fixed point. (A subset of n-dimensional Euclidean space is compact and convex if and only if its intersection with any line is empty, a single point, or a line segment including both endpoints.) For scalar predictions, Simon's observation actually follows from the special case where $n = 1$, which is a consequence of the intermediate value theorem applied to the function that takes x to $f(x) - x$.

you think your credence in A at t will be measured,[6] but it does not generalize to the case of infinitely many such pairs without extra restrictions. So if that idea proves to be important in allowing us to resist the possibility of theoretical dilemmas, we will have to be careful if we want to treat infinite cases.

Second, suppose you think a mind reader will measure your credence in A at two different times t_1 and t_2 and set the chance of A to $f(P_{t_1}(A), P_{t_2}(A))$. Call any number x such that $f(x, x) = x$ a "fixed point" of f. If f is continuous then it will have a fixed point in this sense by Brouwer's theorem applied to the function that takes x to $f(x, x)$. However, if x_0 is a fixed point of f, there can be some y_0 that is not a fixed point of f such that $f(x_0, y_0) = y_0$. In this case, let's say that x_0 is a *potentially self-undermining* fixed point of f. If $P_{t_1}(A) = x_0$ then, after t_1 but before t_2, if you know what $P_{t_1}(A)$ is, you will effectively be faced with the function g that takes y to $f(x_0, y)$, and if the conclusions of section 2 are not radically mistaken then you may be such that it will be uniquely rational for $P_{t_2}(A)$ to be y_0 if $P_{t_1}(A)$ is x_0. If, moreover, at t_1 you know all this and think that you will be rational at t_2, the fixed point x_0 of f will be (actually) self-undermining in that you cannot rationally have credence x_0 in A at t_1 if you think that you will have credence y_0 in A at t_2 so that the chance of A will be y_0.

A continuous function f whose fixed points were *all* potentially self-undermining could be used to generate a kind of theoretical dilemma that would be in a way more troubling than theoretical dilemmas that rely on discontinuous functions. It turns out not to be too hard to construct such a function. One example is the function defined by the equation $f(x, y) = |x - y|$. The unique fixed point of f is 0, but every member of the unit interval is a fixed point of the function that takes y to $|0 - y|$. This function has the potentially suspect feature that its unique fixed point lies on the boundary of the unit interval,

---

[6]The generalization requires the full strength of Brouwer's fixed point theorem, not just the special case where $n = 1$ and the compact, convex set is the unit interval.

but there are other examples—I don't have space to give one—whose fixed points all lie in the interior of the unit interval.

I won't discuss these sorts of theoretical dilemmas directly, but in subsection 5.2, I will discuss an example called 'A Guessing Game' that has deep similarities to them.

2 Multiple fixed point cases

The main question posed by multiple fixed point cases is the *fixed point selection problem*: which fixed point or fixed points of f does rationality permit (or require) your credence in A at t to equal? As I am a lover of Bayesianism, I will begin by asking what it has to say about this question. That will lead us to a somewhat more general discussion of rational change in belief. Finally, I will address the connections between multiple fixed point cases, intention, and the issue of whether it is ever rational to directly choose what to believe. This last part of the discussion will produce more questions than answers.

2.1 Bayesian considerations on the fixed point selection problem

Rational belief updating, according to the standard Bayesian account of it, is always a deterministic response to experience, and—at least if we may idealize by supposing that the rational import of an experience is always representable as a special proposition (called your new evidence) that you ought to become certain of in having the experience—if you are ideally rational, your credences always change by means of a process known as conditionalization: evidence comes in, and you react by changing your probability function to the result of conditioning it on your new evidence. The conditionalization model provides an answer, though a strangely uninformative one, to the fixed point selection problem: you should end up at the fixed point

you get to by conditionalizing. (Will you always get to a fixed point if you conditionalize? Yes, if you satisfy the Principal Principle.)

This answer, however, is subtly wrong and in so being exposes a flaw in the standard account. For suppose that you think you might receive evidence E entailing that you are in the situation depicted in Figure 2B, but you aren't certain whether you will react rationally to it. Then your conditional credence in A given that you receive E shouldn't, in general, be a fixed point of f. If, for instance, you have credence .99 that you will respond to receiving E by dropping your credence in A to 0 and leaving it there until t then your conditional credence in A given that you receive E should be close to $f(0)$ and well above all the fixed points of f. Now maybe it is illegitimate in general to assume that E is equivalent to the proposition that you receive E, but suppose that in the case at hand you are almost certain of the biconditional 'E iff I receive E'. Then your conditional credence in A given E itself should also be far above any fixed point of f. But suppose you are also almost certain that if you do respond rationally to receiving E then you will be rational straight through to time t, so that, in particular, your credence in A at t will be a fixed point of f. Then the rational response to receiving E can't be to conditionalize on it, for if you do that, you will violate the Principal Principle by having a high credence in A while being certain that the chance of A will be one of the fixed points of f, which are all close to 1/2. *Objection on behalf of the standard account.* You, apostate, fail to recognize that not just any proposition could be the totality of what a rational agent learns at a time. In fact, only very special propositions satisfying demanding closure conditions are *apt for conditionalization*. You already stumbled toward one such condition: if E is apt for conditionalization then E must be equivalent (up to subjective probability 0) to the proposition that E is received. Another such condition is that E must entail (up to subjective probability 0) that E is conditionalized on. But the E of your example fails to satisfy that closure condition. *Reply.* If you are

right then the raw data of experience, as it were, are not apt for conditionalization. What is apt for conditionalization is only some closed up strengthening of the raw data of experience that already specifies how you will react to that data. But then conditionalization is circular as an updating rule because it requires as input a function of what is supposed to be its output. It might function as a constraint: rational updating must be *representable as* conditionalization (on what the result of such updating is, perhaps). But there can be no requirement that rational updating be *by* conditionalization.

2.2 Informative answers to the fixed point selection problem

The standard account gives a frustratingly uninformative answer to the fixed point selection problem, but we can imagine more informative answers. There could be special fixed point selection constraints of rationality such as—never mind, for the moment, how plausible these constraints appear—'$P_t(A)$ must be one of the closest fixed points of f to 1/2' or '$P_t(A)$ must be one of the farthest fixed points of f from 1/2'.[7] One worry about such constraints is that it seems they will invariably be incomplete. For instance, the two rules just cited, though they come close to completeness (at least ignoring generalizations of the setup like those discussed in subsection 1.7), break down in the case where there two fixed points are tied for being closest to or farthest from 1/2. A further worry is that such rules will conflict with intuitions about learning from experience. Consider for example a case in which f has a unique fixed point closest to 1/2 and a (different) unique fixed point farthest from 1/2. Suppose that, again and again, you have been faced with such cases, and you have always

---

[7]Note that if f is continuous, as we may as well assume for now in case it helps, it will at least have fixed points closest to and farthest from 1/2 because the set of fixed points of a continuous function is closed. On the other hand, there is no obvious way to break all ties if there can be a case where there is no fact of the matter about whether the mind reader is using f on your credence in A or the function that takes x to $1 - f(1 - x)$ on your credence in not-A.

found yourself ending up at t at the fixed point closest to 1/2. Then, in the new case, surely it would be irrational before t for your credence in A to be much different from the fixed point closest to 1/2. If it were, you would be failing to appropriately learn from experience. But then couldn't it be rational for you to keep that same credence at t itself? It is tempting to say yes, but by a parallel argument, if you have a long history of irrationally responding to Figure 1B style cases by not ending up at a fixed point at all, you could eventually rationally be off a fixed point at t, which seems wrong. I am inclined to think that theoretical rationality does not have any informative answer to the fixed point selection problem up its sleeve, but the present arguments are weak.

2.3 Accuracy

One of the most interesting arguments for both updating by conditionalization and the Principal Principle is that you thereby maximize the expected accuracy of your credences.[8] There are two kinds of expected accuracy arguments, those that depend on the idea that you should maximize the expected global accuracy of your credence function and those that depend on the idea that you should maximize the expected accuracy of your credence in A, for each proposition A. I will only address myself to the local arguments, which I find more compelling. It turns out that expected local accuracy considerations give an unambiguous (though incomplete) answer to the fixed point selection problem, if expected local accuracy maximization is even compatible with the Principal Principle. That answer is: you should be at one of the farthest fixed points from 1/2. An example will be more perspicuous than a formal argument.

Example. If your credence in A is 1/2 then it is not terribly accurate

---

[8]See Leitgeb and Pettigrew 2010 and Pettigrew 2012. These arguments turn out to be somewhat less exciting than they at first appear because they depend on contentious principles about how accuracy should be scored.

whether *A* is true or false. If your credence in *A* is .99, and you know that the chance of *A* is .99, then you know that there is a chance of .99 that your credence in *A* is extremely accurate and a chance of .01 that it is extremely inaccurate. But whether the *expected* accuracy of having credence .99 in *A* is higher than the expected accuracy of having credence 1/2 in *A* apparently depends on how accuracy is scored: a hypercautious kind of accuracy score might assign a huge penalty to having credence .99 in *A* when *A* is false that would swamp everything else in the expected accuracy computation. However, if accuracy is scored hypercautiously, then (at least assuming that if your credence in a proposition is 1/2 then it is equally accurate whether the proposition is true or false) having credence 1/2 in a proposition *A* has a higher expected accuracy than having credence .99 in *A* even if you are sure that the chance of *A* will be .99 whatever your credence in *A* is. But if you are sure the chance of *A* is .99 then you should have credence .99 in *A*, not some more cautious credence like 1/2. If accuracy is scored hypercautiously, expected accuracy maximization is incompatible with the Principal Principle.

While the general idea of somehow grounding theoretical rationality in accuracy considerations (though perhaps not the idea of reducing accuracy to a number) is attractive, it is at least prima facie implausible that rationality requires the selection of extreme fixed points.

2.4 Further objections to conditionalization

Arguments for conditionalization as a norm for ideally rational agents tend to assume (1) that rational updating is always a response to experience, (2) that the rational import of experience can be encapsulated in an evidence proposition that takes probability one after the update, and (3) that if you are rational you know in advance, for any evidence proposition you might receive, what the rational response to receiving it would be. All of these assumptions are dubious, but it is mainly

just (2) that has attracted attention.[9]

Against (3), for instance, maybe the right response to receiving evidence E is to become nearly certain that your partner is unfaithful, but your conditional credence in unfaithfulness given E is not all that high because you're not sure if that would be the right response. That could happen even if you are certain that you will respond rightly to whatever evidence comes your way. In that case you, it is natural to think that rationality would require you to satisfy an instance of Bas van Fraassen's Reflection Principle, according to which your conditional credence in a proposition, given that your credence in it at some given future time is c, ought to be c (for all c such that the conditional credence is defined) (van Fraassen 1984). It is a (defeasible) indicator that a proposed form of credence updating is rational that you would satisfy (an instance of) the Reflection Principle if you were certain that you would update that way. There is no obstacle to your satisfying the Reflection Principle in the case at hand.

More interesting for our purposes is the possibility that (1) might fail. I think that might happen even in cases not involving rational feedback. Isn't the idea of *direct receptivity*, whereby your credences are attuned to the world in a way that is not mediated by experience, at least a theoretical possibility? Maybe it is even the right way to understand belief change that comes about by means of reasoning and/or some forms of "intuition." Still, perhaps the most compelling cases do involve rational feedback. They are what I will call cases of *spontaneous belief change.*[10]

The simplest type of case is what we might call *at-t* spontaneous belief change. Suppose you are in a Figure 1B case, and your credence

---

[9]Richard Jeffrey has done most to raise awareness, beginning with Jeffrey 1965.

[10]In the philosophy of action, *spontaneous knowledge* is sometimes identified with *knowledge without observation.* This way of drawing the line would count belief change through direct receptivity as spontaneous, but such belief change is more naturally classed as non-spontaneous, a verdict that is delivered by the more obscure alternative conception of spontaneous knowledge as *knowledge originated by the agent.*

in *A* changes *exactly at t* to some fixed point of *f*. (There are delicate synchronization issues here that we will discuss in section 4, but for now keep in mind that *t* need not be defined as a wall-clock time; *t* could be defined, for instance, as the first time after some given wall-clock time that your credence in *A* leaves some given interval.) If you are to be rational before *t* in a case like this, you had better be uncertain what $P_t(A)$ will be on pain of violating the Principal Principle, but if you do conform to the Principal Principle, I see no reason why you couldn't be rational both before *t* and at *t*. If you are certain before *t* that $P_t(A)$ will be a fixed point of *f* then you can also satisfy the Reflection Principle, and your belief change at *t* can be represented as conditionalization on what $P_t(A)$ is (though, for the reason given in reply to the objection at the end of subsection 2.1, it would nonetheless not be updating *by* conditionalization.)

I have no decisive argument that you could be rational at *t*, which is unfortunate because that claim is in extreme tension with opinions many philosophers have expressed about analogous non-probabilistic cases.[11] The main target of many of these philosophers has been David Velleman, who has expressed the opposite opinion (Velleman 1989; n.d.). There has been a paucity of arguments on both sides of this debate. Velleman's critics insist that rationally formed beliefs must be formed in response to and on the basis of evidence, but Velleman insists that that claim is a prejudice based on an overgeneralization from the case where (to use my jargon) there is no rational feedback. Let me offer two non-decisive considerations in favor of my claim and, by extension, Velleman's side in the non-probabilistic debate. First, suppose that you are off a fixed point up until *t*. Then you will be irrational at *t* *unless* you end up at a fixed point at *t*, so one might hope that rationality would at least permit you to be rational *at t* by moving to a fixed point then. (This is especially compelling if you can be rational up to *t* while being off a fixed point, but if that is not

---

[11]E.g., Dorr 2002; Langton 2004; Setiya 2008; Paul 2009.

possible, what exactly goes wrong?) Second—and Velleman himself makes somewhat similar complaints—it is hard to see what epistemic value would be promoted by the constraint of rationality that Velleman's opponents posit. It can't be truth or accuracy or reliability or calibration. But aren't these the sorts of virtues that would figure in any plausible explanation of why the norms of theoretical rationality are what they are?

Another type of case is what we might call *before-t* spontaneous belief change. Suppose that t is a wall-clock time, and an hour before t your credence in A changes to a fixed point of f and remains there until t. I think that can be rational if it is reasonable for you to think that after the change that your credence in A will stay where it is until t. It could be reasonable for you to think that because, say, you have strong inductive evidence that that is how your credences always evolve when you are in Figure 1B cases. (Of course, even in that case, you are not necessarily irrational if, contrary to your inductive expectation, your credence changes to a new fixed point of f exactly at t.) The case I have just described is contrived and extreme, but it might be interesting to think about cases of more subtle change or cases in which a change is triggered by the reception of evidence but still partially spontaneous. Could such cases even be commonplace?

I don't know of anyone who has explicitly defended the rationality of spontaneous belief change, but some followers of G. E. M. Anscombe might be committed to it. Anscombe (1957) claims that when we act intentionally we have non-observational knowledge of what we are doing. If *spontaneous* action—action where you don't know with certainty ahead of time what you are going to do—can be intentional, and if Anscombe is right, then presumably the belief change by which you come to know what you are doing when you act spontaneously is itself spontaneous.

17

2.5 The link with intention and directly choosing what to believe

We are approaching important questions about intention, which is outside the scope of our discussion but one of the most important reasons to be interested in rational feedback. According to *cognitivism* about intention (Bratman 1991), intentions are beliefs, and practical reasoning is a form of theoretical reasoning. It has mostly been cognitivists about intention, and especially Velleman (1989), who have appealed to phenomena like spontaneous belief change in theories of intention, but the potential appeal is much more general. It is not even restricted to those Sarah Paul (2009) calls "weak cognitivists"—people who hold that intending to φ necessarily involves having some φ-related belief such as that one is at least trying to φ. For everyone has to explain the belief change by which you typically (even if there are many exceptions) rationally come to believe that you will φ when you decide to φ and thereby form an intention to φ, and it is awkward to analyze this belief change as a response to evidence because it is simultaneous with the decision. Maybe, then, the true theory of intention should appeal to spontaneous belief change. But we should also keep in mind that the very fact that spontaneous belief change is unpredicted by the agent can make it seem unchosen and not a suitable accompaniment to decision. I cannot hope to explain intention here, but I would like to close the section by looking at a case that certainly is one of rational feedback and considering how close it might approach to ordinary decision making.

> (**King's Advisor**) The king of Mars has taken you captive and made you his "advisor" on whether to execute or release prisoners who are daily brought before him. Each morning he sends you the case history of the prisoner whose fate is to be decided that day, and, after you have studied it, he determines your credence that the prisoner

18

will be executed (by means of the court mind reader, or perhaps by asking you, if he knows you are honest or has a lie detector). He arranges that the prisoner will be executed with chance equal to his estimate of your credence.

Probably at first you wouldn't be able to bring your credences in line with your preferences about prisoners' fates (though if you pulled it off, perhaps you could count as rational). But after many days have gone by, you might eventually get into a way of thinking where you just think of yourself as deciding the prisoner's fate each day as you look at the dossier. Would that self-conception be accurate? Would you also be directly choosing what to believe about the prisoner's fate? For comparison, consider how you can get yourself to believe that you will be watching channel seven at t by deciding to press '7' on the remote just before t (if it is a while before t) or by pressing '7' (if it is already just before t). It seems like this could count as a case of choosing what to believe, though one that would be universally accounted rational because indirect.[12] What important epistemological difference is there between this case and some iteration of King's Advisor? There is the fact that in the television-watching case, there is a kind of intention or action that mediates the transition, but (Velleman would be quick to point out) why should that make an epistemological difference? And isn't even that difference gone if we use the version of King's Advisor in which you are asked your credence instead of having it determined by a mind reader? Unfortunately, I must leave these questions unanswered.

3 No fixed point cases

This section surveys arguments that theoretical dilemmas are strictly impossible. All the ones I can think of fail.

---

[12]But maybe the lesson of Gregory Kavka's (1983) Toxin Puzzle is that it couldn't, at least not in the case where it is a while before t.

3.1 Argument one: belief is vague

Here is an objection you might make to the claim that the case illustrated in Figure 1C is possible.

> The predicate 'has credence less than or equal to 1/2 in A at t' is vague. There can be borderline cases of it, and if I am in such a borderline case then, for all you have said, I may be rational. The mind reader will do something or other after she reads my mind, but you have not yet said what.

I agree that real-life credence ascriptions are vague. We can respond to the objection in two ways. First, up to now we have only been considering idealized cases anyway. So long as we are idealizing, isn't it just one more idealization to suppose that you have perfectly well-defined credences? Assuming that creatures with perfectly well-defined credences are possible, theoretical dilemmas are possible, and possible is bad enough. But here is a second response: even if we haven't yet said what the mind reader will do in all cases, we can always say more. In principle, we can say, for each particular borderline case, whether the mind reader will set the chance of A high or low (to .9 or .1, say). So long as every borderline case is a case of you determinately having credence less than .9 and greater than .1, and so long as you can always tell whether you are in a case where the mind reader will set the chance of A to .9 or one where she will set it to .1, you will be in a theoretical dilemma. This second response will be developed further in the next section.

3.2 Argument two: belief can be imprecise

Some philosophers hold that the credences of rational agents are sometimes imprecise. There are many approaches to imprecise

probability,[13] but it will suffice to consider a simple one on which credences can be arbitrary subintervals of the unit interval (including degenerate intervals of the form $[c, c]$, which we will identify with precise credences). If credences can be intervals then f will have to be a function that takes intervals to real numbers. (To keep things simple, I will assume that chances are always precise and that you have a precise credence of $[1, 1]$ about what the precise chance of A at u will be if $P_t(A) = I$, for every interval I.) If theoretical rationality demands that your credence in A at t be a degenerate interval $[c, c]$ such that $f([c, c]) = c$ then allowing imprecise credences changes nothing. One reason to think that it does demand that is that if your credence in A at t is a non-degenerate interval or a degenerate interval $[c, c]$ such that $f([c, c]) \neq c$ then you will violate the Principal Principle, for you will know the chance of A at u and yet your credence in A will not be the chance of A at u but rather some non-degenerate interval.

If one is willing to give up the Principal Principle, the suggestion that your credence in A at t must be a degenerate interval $[c, c]$ such that $f([c, c]) = c$ can be resisted. Maybe theoretical rationality allows your credence in A at t to be any interval I such that $f(I) \in I$. If it allows that then you are saved, for while $f(I)$ can fail to be a member of I for every interval except $[0, 1]$, $f([0, 1])$ must lie in $[0, 1]$. However, if it is *ever* irrational to have a maximally imprecise credence in a proposition A, surely it is irrational in cases like ours where you are certain what the exact chance of A is. Imprecise probabilists motivate their view with cases in which you lack knowledge of the objective chances,

---

[13]The term 'imprecise probability' is used in an influential book by the statistician Peter Walley (1991), but, at least among philosophers, it is not standard. Roger White (2010, 173) gives the following list of alternative terms philosophers have used for imprecise probability, which he says is incomplete: 'indeterminate probability' (Levi 1974), 'vague probability' (van Fraassen 1990), 'indefinite credence' (Joyce 2005), 'thick confidence' (Sturgeon 2008). White himself uses 'mushy credence', which he says he owes to Adam Elga, though Elga now prefers 'unsharp degrees of belief' (2010). 'Imprecise probability' seems to be the best established term outside of philosophy, and it is just as suggestive and free of misleading connotations as 'unsharp degrees of belief', which is the best of the alternatives. It even figures in the name of a society, the Society for Imprecise Probability: Theories and Applications.

but here you have it.

3.3 Argument three: ought implies can

If an otherwise promising moral theory M permits genuine moral dilemmas, one might reasonably prefer some theory M^* that requires everything M requires in cases where M does not require the impossible but is more permissive than M in other cases. Similarly, since our otherwise promising theory of theoretical rationality—call it R—sometimes lands you in theoretical dilemmas, maybe R should be replaced by a theory R^* that is similar to R but more permissive so that theoretical dilemmas come out impossible. However, R^* will have to be revisionary—it will have to allow violating the Principal Principle, or violating the axioms of probability theory, or ignoring evidence that the chance of A will depend on your credence in A. So the response doesn't really do away with the weirdness of Figure 1C style cases; at best it establishes that we are misclassifying them as cases in which you are irrational. More likely, it doesn't establish anything but leads to a barren dispute over the best way to use the word 'rational'.

3.4 Does closeness matter?

A suggestion that has been repeatedly put to me is that the *closer* $P_t(A)$ is to $f(P_t(A))$ the *more rational* you are, so that if f has no fixed points, you can't be completely rational, but $P_t(A)$ should nonetheless be such as to minimize the distance between $P_t(A)$ and $f(P_t(A))$.

The suggestion faces two technical problems and, more fundamentally, it misunderstands the nature of theoretical rationality. The first technical problem is that there may be no $c \in [0, 1]$ that minimizes the distance between c and $f(c)$. For instance, if $f(0) = 1/2$ and $f(x) = x/2$ for all $x > 0$ then the distance between c and $f(c)$ can be any number greater than 0 and less than or equal to 1/2, but it can't be

0. The suggestion is thus only a half-measure, and there seems to be no way to make it a full-measure. The second technical problem is that it is not clear why the distance between credences should be measured using the Euclidean metric. Mightn't it be better to have credence .5 in A while believing the chance of A to be .61 than to have credence 0 in A while believing the chance of A to be .1? Moreover, even if an argument can be given for using the Euclidean metric (or some other particular metric), the problem deepens when we consider cases in which many of your beliefs are measured at once: the suggestion turns out to require specifying a metric on probability functions, and it is even less plausible that any such metric has a special connection with theoretical rationality. The more fundamental problem is that it is *simply nuts* for your expectation of the chance of A to differ from your credence in A if you know what both of these quantities are, and it doesn't make it any less nuts if you think the difference is small. (Arguably, practical rationality is different in this respect. If you do something that almost maximizes your expected utility, maybe that's more rational than doing something whose expected utility is far from maximal.)

4 Theoretical dilemmas in real-life cases?

This section surveys arguments that there can be no theoretical dilemmas in real-life cases, or, in other words, that no theoretical dilemmas affect human-like agents in nearby possible worlds. None of the arguments is conclusive, but together they make a passable case.

4.1 Simon's idea

Taking a cue from Simon (1954), we might look for arguments that f will be continuous in real-life cases. (If f is continuous, it will have a fixed point by Brouwer's fixed point theorem.) One way to argue this

is to note that real-life measurements of continuous quantities are all inexact in the sense that, if c and c' are sufficiently close together, the measurer cannot reliably distinguish the case where the quantity has value c from the case where it has value c'. I don't have space to develop this line of argument properly, but it is possible to give an independently plausible definition of inexact measurement that has the property that if you are sure that the mind reader is only capable of inexact measurement then f must be continuous.[14] A weakness of this line of argument is that it only works if you are sure that the mind reader is only capable of inexact measurement, and maybe you could be rational without having such certainty. Here is a slightly different line: just as real-life mind readers are only capable of inexactly reading the minds of others, real-life agents are only capable of inexactly reading their own minds—only capable of imperfect introspection. (Indeed, consider that if you could perfectly introspect at t then you could, in effect, endow a mind reader with the power of exactly measuring $P_t(A)$ by telling the mind reader exact information about the value of $P_t(A)$.) A weakness of this line of argument is that, even if you fail to perfectly introspect, you might still be able to get into a theoretical dilemma. Even if you lack perfect introspection, rationality still demands that your *expectation* of $f(P_t(A))$ be equal to $P_t(A)$. (An argument could be given using the Principal Principle.) Now, it could happen that f has no fixed points but you satisfy this requirement. For instance, in the situation graphed in Figure 1C, you will satisfy the requirement if $P_t(A) = 1/2$ and

$$P_t(P_t(A) \leq 1/2) = P_t(P_t(A) > 1/2) = 1/2.$$

---

[14]Actually, given how things were set up in subsection 1.3, f must be *constant* since otherwise there must be points c and c' arbitrarily close together such that $f(c) \neq f(c')$, but the mind reader cannot reliably set the chance of A to $f(P_t(A))$ unless she can distinguish the case where $P_t(A) = c$ from the case where $P_t(A) = c'$, for all c and c' such that $f(c) \neq f(c')$. This difficulty can be got round by allowing that $f(x)$ is only your *expectation* of the chance of A at u on the supposition that $P_t(A) = x$, not what you are sure the chance will exactly be. (But here is a worry: what sort of supposition is involved?)

However, since you presumably can't control exactly how the imperfection in your introspection works, it seems unlikely that you can always manage to get yourself into such a favorable situation.

4.2 Taking seriously the vagueness of credence ascriptions

Neither argument from the previous subsection takes to heart the lesson of subsection 3.1: credence ascriptions are vague, and credences are not truly continuous quantities. (Indeed, many physicists suspect that there are no truly continuous quantities in our universe.) Let's take that lesson to heart and try to further develop the "second response" from subsection 3.1. Suppose you are certain that a condition C is

1. *Credentially luminous*: if C obtains, your credence that it does is high, and if it doesn't, your credence that it does is low.[15]

2. *A-chance contraindicating at t*: if C obtains at t, the chance of A is low, and if it doesn't, the chance of A is high.

3. *A-credence loosely covariant*: if your credence in A is high, C obtains, and if your credence in A is low, it doesn't, but if your credence in A is middling, C may or may not obtain.

Then you will be in a theoretical dilemma at t if you are sure at t that it is t. For if your credence in A at t is high, you will think that C obtains and thus that the chance of A is low; if your credence in A at t is low, you will think that C doesn't obtain and thus that the chance of A is

---

[15]Timothy Williamson (2000) famously argued that there are no non-trivial "luminous" conditions at all. Even if his argument succeeds (and there has been plenty of criticism), credential luminosity differs from Williamson's luminosity in that it is about credence not knowledge. Knowledge, according to Williamson, requires modal safety, a requirement that is key to his anti-luminosity argument. But credence uncontroversially does not require modal safety. (Certainty of credential luminosity is nonetheless a very strong assumption, and one might try to construct an argument that rational agents will satisfy it only in trivial, or at least very peculiar, cases.)

high; and if your credence in A at t is middling, either you will think the chance of A is high or you think it is low, depending on whether you think C obtains. Moreover, nothing in the setup relies on a sharp distinction between credence levels. Of course, it is a good question whether there are or could easily be any conditions meeting the three conditions or, if not, whether the conditions can be weakened so that there are without being weakened so much that they no longer lead to a theoretical dilemma. Some weakenings to consider are relaxing the condition that you are certain the conditions obtain, allowing time lag for introspection and/or C-perception, and generally weakening the connections between C, the chances, and your credences. But perhaps most interesting is whether the condition that you are sure at t that it is t can be weakened. At least if you are sure that time is continuous, it might be reasonable for you to always be sure that it is not t, in which case this whole approach to generating a realistic theoretical dilemma looks happily doomed.

5 Weird stuff in real-life cases

Even in realistic cases, theoretical rationality can apparently require your credence in A to fluctuate in surprising ways.

Note: in the following examples, I stipulate that certain things are the case or (more carefully) that you "think" they are the case. Either way, the important thing is that you have a high credence in the proposition that they are the case.

5.1 Example one: the sticky inexact brain scanner

(**The Sticky Inexact Brain Scanner**) You are in a room with a *sticky inexact brain scanner* set to measure your credence in A. This is a device that functions as follows. At all times it is either glowing or not. Before t, it is sensitive to your credence in A. Whenever your credence in A

is above .8, it is glowing, and whenever your credence in
A is below .2, it is not glowing (a kind of A-credence loose
covariance). The scanner is *sticky*: if it is glowing, it will
not stop glowing unless your credence in A falls below .3,
and if it is not glowing, it won't start glowing unless your
credence in A rises above .7. Thus, if your credence in A
rises from 0 to 1, the scanner will start glowing when it
is between .7 and .8, and if your credence in A falls from
1 to 0, the scanner will stop glowing when it is between
.3 and .2. The scanner need be capable of only *inexact*
mind reading since there is no particular level between
.7 and .8 or .3 and .2 at which it has to transition. (And
thus its description is compatible with the vagueness of
credence ascriptions.) At t, the scanner will emit a loud
pop and cease being sensitive to your credence in A. If
it is glowing, the chance of A will be about .1; if it is not,
the chance of A will be about .9 (a kind of A-chance con-
traindicatingness at t). You have no clock (though I don't
think it fundamentally changes the nature of the example
if you do have a clock).

The Sticky Inexact Brain Scanner neatly avoids the arguments of the
previous section. Not only is merely inexact mind reading required of
the scanner, but no special assumptions about your powers of intro-
spection are needed, nor do we have to suppose that you are exactly
synchronized with the scanner. Let's see how it goes.

Suppose that the scanner is not glowing at the beginning of the ex-
periment. Then your credence in A must be below .8. You must be-
lieve that there is a significant chance that the scanner will be glow-
ing at t since otherwise you will think that the chance of A will be
about .9, and it is irrational to have a credence in A below .8 while
thinking the chance of A will be about .9. Suppose that your credence
in A does eventually rise high enough that the scanner starts glow-

ing, and suppose you notice that your credence is about x when this happens. How should your credence in A change when you observe that the scanner has begun to glow though there has been no pop? I'm not sure, but here is a tentative analysis. If, just a moment ago, you thought that a non-glowing scanner was reason to have credence about x in A then, by symmetry, now you should think that a glowing scanner is reason to have credence about $1 - x$ in A. So your credence in A will fall to about $1 - x$. This might cause the scanner to stop glowing, in which case your credence should go high again. If that causes the scanner to start glowing then your credence will go low again. And so on. Your credence in A will rapidly oscillate between values between .7 and .8 and values between .2 and .3. Since the scanner is unreliable in these regions, your credence might come to rest, either at a high value with the scanner glowing or at a low value with the scanner not glowing. Either way, you must think that there is a good chance your credence will go through another oscillation cycle before t since otherwise you either think the chance of A is likely to be around .9 while having credence in A below .8 or you think the chance of A is likely to be around .1 while having credence in A above .2. One thing that is unclear to me is *why* your credence would change so as to make scanner change its state. What new information do you expect to get that will make your credence change?

5.2 Example two: a guessing game

In the following example, there is no futuristic brain scanner at all. Although there is someone I call the mind reader she need have no special ability to read minds. Again, none of the arguments of the previous section apply.

> (**A Guessing Game**) The mind reader is going to ask you at some time t' between t and u whether you think it's more likely that $P_t(A) \leq 1/2$ or that $P_t(A) > 1/2$. She will

set the chance of A at u to about .9 if you say you think $P_t(A) \leq 1/2$ is more likely and to about .1 if say you think $P_t(A) > 1/2$ is more likely. If you say 'equally likely' or 'it's semantically indeterminate' or in any other way fail to give a straight answer to the question then she will set the chance of A to about .1. You plan to try to tell the truth (either out of honesty or because you think the mind reader can tell if you lie and will punish you for it). As before, you have no clock, but t' is sufficiently far after t that you will be certain that t has passed before t' arrives.

If you are to be rational in this case, you cannot settle on some credence in A before t and maintain roughly that same credence past t'. For if you settle on a low credence, you will eventually realize that your credence at t was low and that you will report that fact to the mind reader, which will cause the chance of A to be high, while if you settle on a high credence, you will eventually realize that your credence at t was high and that you will report that. In either case you will be irrational because you can foresee that you will report your credence and thereby cause the chance of A to disagree with it. Your only hope for being rational is to adopt a credence in A close to $1/2$ as t approaches. But even if you do this, at some point you will know what you say to the mind reader, and then your credence in A must swing low or high. (Maybe you don't know what you are going to say until you say it; still, *after* you speak you will know what you said. In this case your credence will not swing until after t', but it will still swing.) Your only hope for being rational is for your credence in A to swing from about $1/2$ to about .9 or .1, but it is not obvious *why* your credence in A would change like that. Maybe you will learn by introspection after t something about what $P_t(A)$ was—but maybe you won't and know you won't. Maybe if you answer in a funny voice so that you are not sure whether the mind reader will count you as giving a straight answer or not you can rationally have credence $1/2$ in A at t. But even

29

if you can manage this, it could hardly be a constraint of theoretical rationality that you answer in a funny voice, for theoretical rationality constrains belief not action. I worry that, for ourselves or agents not too different from ourselves, A Guessing Game might amount to a kind of theoretical dilemma.[16]

References

Anscombe, G. E. M. 1957. *Intention.* Basil Blackwell.

Bratman, Michael E. 1991. "Cognitivism about Practical Reason: *Practical Reason* by J. David Velleman." *Ethics* 102:117–28.

Dorr, Cian. 2002. "Non-cognitivism and Wishful Thinking." *Noûs* 36:97–103.

Elga, Adam. 2010. "Subjective Probabilities Should Be Sharp." *Philosophers' Imprint* 10:1–11.

Grunberg, Emile, and Franco Modigliani. 1954. "The Predictability of Social Events." *Journal of Political Economy* 62:465–78.

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

Joyce, James. 2005. "How Probabilities Reflect Evidence." *Philosophical Perspectives* 19:153–78.

Kadane, Joseph B., Mark Schervish, and Teddy Seidenfeld. 2008. "Is Ignorance Bliss?" *Journal of Philosophy* 105:5–36.

Kavka, Gregory. 1983. "The Toxin Puzzle." *Analysis* 63:242–4.

Langton, Rae. 2004. "Intention as Faith." In *Agency and Action,* edited by John Hyman and Helen Steward, 243–58. Cambridge University Press.

Leitgeb, Hannes, and Richard Pettigrew. 2010. "An Objective Justification of Bayesianism (I and II)." *Philosophy of Science* 77:201–72.

Levi, Isaac. 1974. "On Indeterminate Probabilites." *Journal of Philosophy* 71:391–418.

Lewis, David. 1980. "A Subjectivist's Guide to Objective Chance." In *Studies in Inductive Logic and Probability,* edited by Richard C. Jeffrey, 263–93. Vol. 2. Berkeley: University of California Press.

Paul, Sarah K. 2009. "Intention, belief, and wishful thinking: Setiya on 'Practical Knowledge'." *Ethics* 119:546–57.

Pettigrew, Richard. 2012. "Accuracy, Chance, and the Principal Principle." *Philosophical Review* 121:241–75.

Ramsey, Frank P. 1990. "Weight or the value of knowledge." *British Journal for the Philosophy of Science* 41:1–4.

Setiya, Kieran. 2008. "Practical Knowledge." *Ethics* 118:388–409.

Simon, Herbert A. 1954. "Bandwagon and Underdog Effects and the Possibility of Election Predictions." *The Public Opinion Quarterly* 18:245–53.

Sturgeon, Scott. 2008. "Reason and the Grain of Belief." *Noûs* 42:139–65.

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

———. 1990. "Figures in a Probability Landscape." In *Truth or Consequences: Essays in honor of Nuel Belnap,* edited by J. Michael Dunn and Anil Gupta, 345–56. Kluwer.

Velleman, J. David. n.d. "A Note on Practical Knowledge." Available from `http://nyu.academia.edu/DavidVelleman/Papers`. Cited with permission.

———. 1989. *Practical Reflection.* Princeton University Press.

Walley, Peter. 1991. *Statistical Reasoning with Imprecise Probabilities.* Chapman / Hall.

White, Roger. 2010. "Evidential Symmetry and Mushy Credence." In *Oxford Studies in Epistemology,* edited by Tamar Szabó Gendler and John Hawthorne, 161–86. Vol. 3. Oxford University Press.

Williamson, Timothy. 2000. *Knowledge and its Limits.* Oxford University Press.