ACCURACY AS EPISTEMIC UTILITY

by

BENJAMIN ANDERS LEVINSTEIN

A dissertation submitted to the
Graduate School-New Brunswick
Rutgers, The State University of New Jersey
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
Graduate Program in Philosophy
written under the direction of
Branden Fitelson
and approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey
October 2013

ABSTRACT OF THE DISSERTATION

Accuracy as Epistemic Utility

by BENJAMIN ANDERS LEVINSTEIN

Dissertation Director:

Branden Fitelson

As an epistemic agent, I have the single ultimate goal of matching my doxastic attitudes to the world. Matching isn't an all-or-nothing affair. Instead I face a gradational criterion of success: the closer I come to fitting my beliefs to the world, the better. I then have reason to follow a given epistemic constraint on my credences only insofar as I have reason to think it will help me in my quest for accuracy. Truth, in other words, is the highest epistemic good, and closeness to truth is *epistemic utility*.

In matters both pragmatic and epistemic, an agent ought to maximize her utility, and I exploit the standard apparatus of decision theory throughout the dissertation. However, while rational agents may have vastly different practical utility functions, they mostly agree about which doxastic states are preferable to others. Even if Hume is right that it is not *pragmatically* irrational to prefer the destruction of the world to the scratching of his little finger, it's surely *epistemically* irrational to prefer a credence of $1/2$ to a credence of $3/4$ in a true proposition. Because the space of reasonable epistemic utility functions is so limited, cognitive decision theory is a powerful formal tool for answering normative questions in epistemology.

This dissertation develops the decision-theoretic approach. Chapter 1 explores how the quest for accuracy ought to determine derivative norms on doxastic states. I argue that if a kind of epistemic attitude can be taken as primitive—i.e., not derived from a

more fundamental kind of attitude—then it should come packaged with a sufficiently robust notion of accuracy. Unlike full-belief and credence, comparative confidence has no prospects for its own measure of accuracy. Chapter 2 (of which Levinstein (2012) is an earlier version) examines which epistemic utility functions are rationally permissible. I attack the most popular measures of inaccuracy—quadratic scoring rules—and provide considerations in favor of logarithmic rules. Chs. 3 and 4 apply cognitive decision theory to the problem of peer disagreement.

Dedication

To my parents, Janna and Irwin, and to the memory of my grandparents, Mary and Mo.

Acknowledgments

First and foremost, I owe thanks to my parents, Janna and Irwin Levinstein. The standard they set in our household for clear thought and rigorous argumentation was always high, and I hope that I have lived up to it here. Any success I owe to them.

Outside of my immediate family, many others have played prominent roles in my intellectual growth throughout my life. Jackie and Larry White were always up for a conversation about anything I was interested in and have provided great emotional support over the years. Phil McKenzie continues to afford an MBA's perspective while enjoying the abstraction of philosophical discussion. The late Mo Levinstein encouraged all the right interests, especially math and physics. He was ultimately unsuccessful in persuading me to study engineering proper, but if one takes Quine's perspective endorsed at the beginning of the dissertation, he did achieve a pyrrhic victory. I am additionally indebted to David and Sue Levinstein, Jennifer Swain, Ellen and the late Stanley Wolfson, Mary Levinstein (now deceased), Marjorie Levinstein and all my cousins. Personal support from my friends Alex Anthony, Sarah Birgé, L. Sulin Carling, Amanda Faraone, Brian Gittis, Adi Habbu, and Dan Schnitzer has been of tremendous value during the time that the dissertation was written.

Having now served as the instructor of a number of college classes, I have great appreciation for the tremendous good primary and secondary school teachers can impart upon their students. Four such educators stand out in my life. Kathryn Morton has been an intellectual and ethical inspiration since my childhood. Mary Anna "Domina"

White, an exceptionally dedicated teacher, often met with me outside of school to read Latin poetry and teach me basic ancient Greek. I gained a great love for language because of her influence and eventually even learned the meaning of *arx*, *arcis*. Her husband Richard White, though employed at a different high school, was similarly committed to my classical education. In addition to thanks, I owe Rich a great number of books, some of which I will one day return. With luck, he may even get his copy of DJ Mastronarde's *Introduction to Attic Greek* before I get tenure. Christopher Fischer led my first formal philosophy course, *Theory of Knowledge*, and supervised my senior essay. After working with him, I was set on pursuing philosophy in college.

At the University of Chicago, I had too many great professors to mention. I am especially grateful to Tim Bays and Michael Kremer who exposed me early on to logic, set theory, and philosophy of math and to Josef Stern for introducing me to Frege.

As for the dissertation itself, I'll start by thanking all my fellow graduate students at Rutgers. I've benefitted especially from conversations with Alex Anthony, Saba Bazargan, Nick Beckstead, Robert Beddor, Adam Crager, Pavel Davydov, Marco Dees, Heather Demarest, Tom Donaldson, Gabriel Greenberg, Preston Greene, Michael Hicks, Erik Hoversten, Michael Johnson, Katy Meadows, Ricardo Mena, Zak Miller, Amanda Montgomery, Carlotta Pavese, Mary Salvaggio, Andrew Sepielli, Derek Shiller, William Starr, Meghan Sullivan, and Jack Woods. Nick Beckstead, David Christensen, Pavel Davydov, Tom Donaldson, Andy Egan, Aditya Habbu, Jim Joyce, Stephanie Leopold, Irwin Levinstein, and Richard Pettigrew each read over chapter drafts and met with me or sent me written comments. I'm also very grateful to two anonymous referees for *Philosophy of Science* who were of great help improving Ch. 2 and to audiences at the Groningen/Munich Summer School on Formal Methods in Philosophy, the 2012 Rationality and Decision Conference in Groningen, the Rutgers Graduate Colloquium, the Rutgers Dissertation Workshop, and the 2012 Formal Epistemology Workshop in Munich and especially to Christopher Meacham who served as commentator on a version of Ch. 2.

My committee has been fantastic in guiding me to the completed version of this dissertation. First, thanks to the unofficial member: Mercedes Díaz, without whom I had no hope of finishing a dissertation on anything. Barry Loewer, Thony Gillies, Brian Weatherson, and Branden Fitelson are all responsible for great improvements in content and presentation. Brian is especially patient for reading and commenting on very early drafts and notes on accuracy while my ideas were egregiously inchoate and disorganized. Branden is the model of a committee chairperson. He is profoundly dedicated to his students and the profession, and anybody would be lucky to study with him. Thank you, Branden, for everything.

Finally, thanks to Amy Hilty for her love, support, and companionship. In the words of Frank Sinatra:

Once in love with Amy,
Always in love with Amy.

Contents

Introduction

> *Normative epistemology is a branch of engineering. It is a matter of efficacy for an ulterior end, truth. The normative here, as elsewhere in engineering, becomes descriptive when the terminal parameter is expressed.*
>
> — W.V.O. Quine

Vera, in her capacity as an epistemic agent, strives to fit her doxastic state to the world. Perfect fit is best: if she had things her way, she'd be certain of all truths and completely reject all falsehoods. Unfortunately, in matters both practical and epistemic, she can't always get everything she wants. Her ultimate goal of alethic perfection—i.e., perfect accuracy—is nearly always out of reach regardless of how rational she is. So, she'll usually be stuck in some less-than-optimal state or other. Not all such predicaments, however, are equally desirable.

The more accurate she is—the closer Vera's doxastic state comes to matching the way the world really is—the greater her epistemic well-being, or, to use a more technical term, the greater her *epistemic utility*. If she's epistemically rational, her goal for accuracy will drive her to gather evidence, form judgments, and revise her beliefs.

Indeed, from a purely epistemic point of view, there can be no tradeoffs between the quest for accuracy and the desire for any other kind of epistemic good, such as responding well to our evidence, maintaining coherence, or obeying the Principal Principle. We wouldn't opt for any policy of belief formation, no matter its other features, if we thought it would hinder our pursuit of accuracy. While these other virtues are important, they aren't worth sacrificing accuracy for. Indeed, in the attitude taken throughout this dissertation, these virtues are best thought of as good *strategies* for acquiring accuracy,

not as ends in themselves.

Rationality and Cognitive Decision Theory

Rationality is about finding good ways to obtain what you want. Practical decision theory helps us to determine how to *act* rationally. We first come up with a measure of how well an agent actually ends up. Formally, this step requires specifying a utility function that encodes her preferences over outcomes and then checking how much utility she got after she completed her action. Given her utility function, we can also provide an assessment of whether a given action is choice-worthy *before* discovering which outcome really will result. For instance, we can provide an *ex ante* evaluation of the act based on how much utility we (or she) expect it to produce.

The cognitive case works by analogy. We first come up with a measure of how well an agent actually ends up epistemically. Again, we need to specify a utility function. This time, we identify utility with accuracy and check how accurate her doxastic state really is. Given a measure of accuracy, we can also provide an assessment of whether a given doxastic state is choice-worthy *before* discovering which world we're really in. For instance, we can provide an *ex ante* evaluation of the act based on how much accuracy we (or she) expect it to produce.

In both cases, your actual level of well-being is not what determines whether you're rational. Suppose a fair coin will be flipped. Alice has credence .5 that it will land heads, but Bob has credence 1. Bob may end up more accurate, but he's still less rational than Alice. While his epistemic well-being is greater, he didn't take appropriate *means* to obtain accuracy. This method of evaluation does not conflict in any way with the ultimate end of fitting one's doxastic state to the world. Compare: Alice and Bob later play heads-up Texas Hold 'Em. Alice picks up on Bob's tells and realizes she can lure him into going

all-in before the flop with a much weaker hand.[1] Bob lucks out: after the five communal cards are revealed he ends up with the stronger hand and wins. We may criticize Bob for his bone-headed play and praise Alice despite her loss. Nonetheless, both players had one and only one goal: to win the hand.

It's better to be lucky than good in both the practical and epistemic cases, but hoping for luck isn't a rational way to get what you want. You tend to do better by playing a different strategy. Alice places no value in rationality *per se* and wishes she had acted differently, but she won't change her strategy in future cases. As an agent, all she cares about is getting what she wants. Decision theorists—who are interested in rationality as such—are the kibitzers who are less focused on actual outcomes and instead attempt to guide and evaluate actions and beliefs.

Epistemic Utility

So far, we've been emphasizing the strong analogies between practical and cognitive decision theory. There is, however, one striking difference. In practical decision theory, the utility function is almost entirely up for grabs. As Hume famously put it:

> 'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. 'Tis not contrary to reason for me to chuse my total ruin, to prevent the least uneasiness of an Indian or person wholly unknown to me. (2000: Bk II, Pt III, S. III)

Less dramatically, we note that aside from a few structural constraints like transitivity, what you prefer and to what you degree you prefer it is not up for rational assessment from the point of view of standard decision theory. Some people like chocolate ice cream,

---

[1] In Texas Hold 'Em, each player is dealt two cards face down (called hole cards). After a round of betting, three communal cards are revealed (the flop), followed by a second round of betting. Next, a communal fourth card is revealed (the turn), followed by a third round of betting. Finally, the fifth communal card is revealed (the river), with a final round of betting. A player's hand—supposing she hasn't folded before the river—comprises the best five-card hand she can make out of her two hole cards and the five communal cards.

while others like vanilla. There's no accounting for taste. Because of this permissiveness, it's hard to derive many substantive constraints on action that apply to all agents regardless of their desires. Nearly anything is rational under *some* set of desires and beliefs.

There may likewise be some variance in rationally permissible epistemic utility functions. William James long ago argued that the goal of accuracy is insufficient to determine exactly what our epistemic values are:

> Believe truth! Shun error!—these, we see, are two materially different laws; and by choosing between them we may end by coloring differently our whole intellectual life. We may regard the chase for truth as paramount, and the avoidance of error as secondary; or we may, on the other hand, treat the avoidance of error as more imperative, and let truth take its chance. James (1979: S. VII)

These two directives—to believe truths and disbelieve falsehoods—are in tension, and every agent must strike some balance between them. It's easy to believe all the truths there are: simply believe every single proposition. However, that strategy is obviously flawed because you'll also believe all the falsehoods there are. If, on the other hand, you avoid falsity by refusing to believe anything, you'll be stuck in a paralyzing state of skepticism. Agents must find a way to balance these values against one another, and it may be permissible for different agents to settle for different tradeoffs. If so, their epistemic utility functions won't be identical.

Nevertheless, the space of rational epistemic utility functions is much more constrained than the space of rational practical utility functions. While deciding which functions exactly are permitted is a subtle matter, we can all agree that higher credence in a true proposition is preferable to a lower one. Because of the commonality of goals that we find in the epistemic case, cognitive decision theory delivers verdicts on epistemic rationality for which we find no analog in practical case. There is simply more we can say that rationally binds all doxastic agents.

Outline of the Dissertation

This dissertation comprises four (nearly) self-standing papers that develop the notion of epistemic utility and apply cognitive decision theory to various issues in contemporary epistemology. Because it is not meant to be a monograph, I've varied the notation and definitions across chapters and repeated some material to suit my local rhetorical purposes.

Big themes addressed include:

- Exploring how the quest for accuracy ought to determine substantive derivative norms on doxastic states (Ch. 1).
- Determining which epistemic utility functions are rationally permissible (Ch. 2).
- Understanding how our assessment of ourselves and others in terms of accuracy ought to affect how we update in cases of disagreement (Chs. 3 and 4).

In more detail, the individual chapters cover the following:

Comparative Confidence and Coherence

Here, I explore the notion of accuracy for three different kinds of doxastic attitudes: full-belief, credence, and comparative confidence. I argue that if a kind of epistemic attitude can be taken as primitive—i.e., not derived from a more fundamental kind of attitude—then it should come packaged with a sufficiently robust notion of accuracy. Unlike full-belief and credence, comparative confidence has no prospects for its own measure of accuracy. In particular, no such notion is strong enough—as far as I can see—to ground the standard coherence requirements needed for comparative confidence to stand on its own feet.

Against Quadratic Scoring Rules

In this chapter, I look at various alternative epistemic utility functions and ask which—if any—is the "true" measure of inaccuracy. I argue that logarithmic measures are significantly more attractive in general than their more popular cousins: quadratic scoring rules. Building on work from Levinstein (2012), this chapter highlights problems for updating rules that quadratic measures generate. The logarithmic rule, on the other hand, leads to the standard and superior method of Jeffrey Conditionalization. I also show that logarithmic rules mesh much better for the purposes of decision theory than any other kind of scoring rule. Therefore, we should tentatively endorse it as the correct measure.

With All Due Respect: The Macro-Epistemology of Disagreement

For the last few years, the debate over how to handle disagreement with a peer has been raging in the philosophical literature. One common position is the Equal Weight View, which is usually characterized as follows: If you take somebody to be your epistemic peer with regard to some proposition p, you ought to give that person's own credence in p equal weight after learning what she thinks. The notion of equal weight is then understood as splitting the difference, *viz.*, taking the average of your two levels of confidence. I show here how accuracy-theoretic considerations prove this version of the view is misguided. First, I demonstrate that equal expected accuracy (under any reasonable measure) is necessary and sufficient for equal *expected* weight. However, I go on to show that this relationship only exists in the large. That is, an agent will often have to deviate from her expectations after she learns what her advisor thinks about p, but over the course of many disagreements, her response should average out to equal weight.

Next, I obtain a second result to show that the notion of symmetry that motivates the equal weight view has been importantly misunderstood. Straight averaging is what I call a *spineless* policy. If you go around splitting the difference with your peers, you'll generally end up with levels of confidence that are too timid, i.e., too close to 50/50

and not close enough to 100% or 0% confidence. In general, if agents are taken to be relatively reliable and unbiased, the respect given to one party should tend to *grow* with her confidence. In other words, if two equally *good* agents disagree, but one of them has a much stronger opinion than the other, then the one with the stronger opinion should usually be favored.

What Your Credence Tells Me about Whether p

This chapter examines standard conciliatory views of disagreement in the literature, tries to show where they go wrong, and develops the variable weight view that was motivated in the previous chapter. I claim that other theorists of disagreement haven't sufficiently appreciated that a doxastic agent's primary goal is to be accurate, not to be rational. This error has led the literature astray.

I focus on conciliationists Adam Elga and David Christensen, who have claimed that evidence from disagreement is somehow epistemically special and requires agents to handle it differently from normal evidence. I argue that by thinking of ourselves and our advisors as truth-measurement devices, we can treat evidence from disagreement as normal evidence. I then discuss in more detail how evidence from disagreement works, how it interacts with other evidence, and how one's *ex ante* assessment of an advisor's expected inaccuracy should affect updating. While the last chapter shows there are broad constraints on updating over the course of many disagreements, this chapter shows that the right response in any particular case is highly case-specific. Nevertheless, there are some general rules of thumb that will often be useful for responding to disagreement, and I conclude with a discussion of how to incorporate Christensen's and Elga's insights for handling evidence from disagreement without treating it as fundamentally different.

Chapter 1

Accuracy and Comparative Confidence

1.0 Introduction: Three Grades of Doxastic Involvement

Epistemologists these days focus primarily on two different kinds of models of doxastic attitudes rational agents might have. The first is what I'll call *full-belief models*. In these models, agents can have one of either two or three doxastic attitudes toward propositions. In the simplest case, we can represent agents as either fully believing or fully disbelieving any proposition p. Sometimes we expand this model to allow for agents to have an intermediate attitude of suspension of judgment as well. Even with this expansion, however, full-belief models are structurally impoverished. Clearly, rational agents may have all sorts of intermediate attitudes—not just one—between full belief and full disbelief. Furthermore, even some full beliefs are stronger than others. I fully believe that I'll teach on Wednesday, but I believe even more strongly that $1 + 1 = 2$. Thus, even for unidealized agents like us, such models have a hard time capturing all the different sorts of doxastic attitudes we might have.

The other kind of popular model I'll call *credence models*. The most popular example requires rational agents to have fixed numerical degrees of belief in every proposition in

some algebra under consideration. Here there are continuum-many distinct attitudes an agent can have. Other kinds of credence models allow for intervals or sets of probability functions to represent agents. One problem with all of these—at least for thinking about human agents—is the opposite of what we have in the full-belief models. Here there's an embarrassment of riches. Idealizing agents to have fixed numerical credences to infinitely many decimal places seems to require much more structure than any real agent could hope for.

It's surprising, then, that intermediate models of doxastic attitudes have fallen out of vogue in recent decades. One natural kind of model goes by various names: *comparative probability*, *comparative confidence*, or *comparative likelihood models*.[1] Though there are a number of different species of these models, the basic idea is the following: For any two propositions p and q, an agent S judges p to be more likely than q ($p \succ_S q$), less likely than q ($p \prec_S q$), or as likely as q ($p \approx_S q$).[2]

Historically, comparative probability had a more central role in formal epistemology. Some, such as de Finetti (1964), took it as the fundamental doxastic attitude from which numerical credence should be derived. There are a few reasons, I think, to consider assigning comparative confidence this bedrock role. First, it doesn't take nearly as much idealization as the credence case does. Though I don't always have comparative judgments between propositions, it's much easier for me by and large to figure out which of two propositions I'm more confident in than to assign them both precise numerical credences. For example, I'm more confident that Hillary Clinton will be elected in 2016 than I am that Bernie Sanders will, but I have no idea what my credence in either proposi-

[1]The former name is slightly inapt because I won't be assuming that rational agents necessarily obey important analogs of the probability axioms, but I'll slip into it on occasion.

[2]I just want to give the basic idea here and am not yet worried about particular differences between the models. For instance, some will, for reasons of theoretical elegance, take $\succeq$ to be the fundamental relation. Some models also allow agents to abstain from judging the relative likelihood between propositions, but I'll put those to the side as well. I'll also drop the subscripted S when the context is clear.

tion should be. Second, there's no suspicious notion of 'distance' built into comparative confidence. That is, there's no non-derivative notion of how much more confident I am in one proposition over another. The metric structure that comes with numerical credence is justly suspect, since there are all sorts of mathematical structures other than the real numbers with the usual topology that could be used to do the same work. Without additional justification, we should worry that our notion of 'distance between credences' or 'distance between credence functions' is overly dependent on an arbitrary choice of representational device. Facts about which propositions an agent takes to be more likely, on the other hand, seem representation-independent.

Due to these features, one might reasonably think comparative probability should be assigned a more central or even a fundamental role in epistemology.[3] However, as we'll see below, comparative probability is lacking certain features common to both full-belief and credence that I think make it much less attractive than it may at first appear. Most importantly, there cannot—as far as I can tell—be an adequate notion of fitness or accuracy for comparative probability.

The fundamental aim of a doxastic agent is to fit her attitudes to the world. For non-omniscient agents like us, we're stuck in states of alethic imperfection. That is, we're not perfectly accurate. We have to use our capabilities and resources to try to improve our overall accuracy. After all, the reason we change our beliefs when new evidence comes in is that we expect that our updated state better matches the way things really are. If we can't tell, aside from special cases, whether comparative ranking of propositions is more accurate than another at a world, then comparative confidence lacks the resources for serving as a non-derivative kind of doxastic attitude.

---

[3]By 'fundamental role', I mean at the least comparative confidence is not taken as a derivative kind of attitude that is parasitic on full-belief or credence. Some epistemologists, such as de Finetti, want comparative confidence not only to be self-standing but to be the kind of attitude that grounds full-belief and credence. That is, they take comparative confidence to be the only fundamental attitude, with all others derivative from it.

What counts as an adequate notion of accuracy? Some of the details will emerge below, but we can lay down some criteria here. First, any good measure of accuracy should have some intuitive motivation. That is, it shouldn't be blatantly *ad hoc*. Second, it should be sufficiently powerful to do necessary theoretical work. Since epistemic agents strive for accuracy, we should usually be able to compare how well two different agents are doing at a world. Otherwise, our measure of accuracy won't be able to do any serious theoretical leg-work.

To make matters vivid, I'll focus on the theoretical connection between accuracy and coherence. Coherence norms—such as logical consistency, probabilism, and the usual axioms for comparative confidence—limit the set of rationally permissible global doxastic states. I'll argue below that such norms should be derived from accuracy considerations. Unfortunately, unlike in the successful cases of full-belief and credence, we have strong reasons to doubt that the notion of accuracy for comparative confidence is robust enough to carry out the project.[4]

Because comparative confidence models seem to fit somewhere in between full-belief and credence models, it's surprising that the latter two have adequate notions of accuracy but the former doesn't. The fundamental difference is that full-belief and credence can be broken down into local attitudes, whereas comparative confidence is perniciously global. It makes sense to talk about the accuracy of a belief or a credence in a proposition p without any reference to other propositions. By contrast, an agent's comparative judgments toward a set of propositions are entangled. To specify her attitude toward p, she must compare p to every other proposition in the set. It's this distinctive feature at the heart of comparative confidence that leads to its undoing.

Here's the plan. §1.1 accomplishes two goals. First, it explores the connection be-

[4]Even if the reader is unconvinced that coherence requirements should come from accuracy considerations alone, I hope the considerations advanced will be enough to demonstrate that it's unlikely any adequate notion of accuracy for comparative confidence is forthcoming.

tween accuracy and coherence norms. Second, for contrast, it shows off the success of accuracy measures for full-belief and credence. §1.2 introduces comparative confidence models in more detail and discusses the conceptual reasons for doubting that any adequate measure of accuracy is forthcoming. §1.3 examines the desired coherence constraints on comparative confidence, and §1.4 shows why none of various attempts will work. §1.5 concludes.

1.1 Coherence Norms

To see where comparative confidence falls flat, we need to examine the justification of coherence norms on doxastic attitudes. Coherence norms are essentially non-local. They don't say what my attitude toward any particular proposition should be, but instead say how my attitudes toward various propositions should relate. Consider, for instance, the standard norm that my beliefs be logically consistent. Consistency is a global property of my entire doxastic state. I can't simply check each belief for consistency in isolation from the others.

One naturally wonders what ought to ground coherence norms. As we've stressed, a rational agent strives to match her doxastic state to the world. Coherence norms don't appear to have anything to do with the relationship of your doxastic state to the world; they're simply concerned with the relationship of your individual doxastic attitudes to one another.

If there's no way to connect alethic and coherence norms, then it's hard to see why rational agents are bound to follow them. Suppose, for instance, that I claim you ought to obey a given coherence norm C. I concede, however, that there's no important connection between obeying C and fitting your doxastic state to the world. In other words, whether you obey C or not gives no information about your accuracy. It's hard to see

how you could have any compelling rational obligation to obey C.[5]

For this reason, a number of epistemologists have sought to connect coherence to the goal of accuracy.[6] There's a basic schema behind all these arguments, which goes as follows:

1. Specify what the perfectly accurate doxastic state at a world is.

2. Come up with some appropriate way of measuring how far from the mark less than perfectly accurate doxastic states are.

3. Show that if an agent violates a given norm, then she's culpably less accurate than she could be.

While I hope the discussion below should clarify and refine each of these steps, we'll go through them in broad strokes now. Suppose we're interested in a given type of doxastic attitude like full-belief, credence, or comparative confidence. Step 1 requires that we identify what the alethically ideal state for that type looks like. A perfectly accurate agent matches her attitudes exactly to the world she's in. In other words, she obeys the truth-norm for the relevant type of attitude. Step 2 requires coming up with an appropriate *evaluative criterion* that tells us how far away a given agent is from perfect accuracy at a world. That is, we need a way to tell how well agents who aren't in full compliance with the truth norm are doing. Step 3 demands an argument that an agent who doesn't obey the relevant norm is leaving accuracy on the table. That is, she is less accurate than she could be in a way that makes her epistemically blameworthy.

---

[5]It's tempting to say that while C may well be orthogonal to accuracy, it could still be important evidentially. That is, perhaps C has nothing to do with accuracy, but violating it shows you're not obeying evidential norms. It seems to me that this just pushes the issue back one step. If following an alleged evidential norm isn't thought to promote accuracy one way or another, then it's hard to see what reason there could be to follow it.

[6]E.g., Joyce (1998, 2009); Leitgeb and Pettigrew (2010a); Lindley (1982); Rosenkrantz (1981) for credence, and Briggs et al. (ms) for full-belief.

Each of these steps is non-trivial and will require further treatment. To see how the strategy works, I'll first sketch how to develop it for full-belief and credence. Next, I'll try to glean some general lessons about Steps 2 and 3: general constraints on measuring how far a doxastic state is from perfect accuracy and how to show that violating of a proposed coherence requirement indicates genuine irrationality. I'll then argue that no such argument can be made for the standard coherence requirements on comparative confidence.

1.1.1 Full-Belief

For simplicity, we begin with a binary full-belief model.[7] Suppose that for any proposition p in a given algebra Ω over a set of worlds W, an agent S can form one of only two *judgments* toward any proposition: either S believes p (denoted $B(p)$) or S disbelieves p (denoted $D(p)$), where $B(p) \equiv \neg D(p)$. An agent's *doxastic state* in this model is then the set of all her judgments.

Say that a judgment $B(p)$ ($D(p)$) is *accurate at* w if p is true (false) at w. An agent with perfect accuracy completely complies with the *Truth Norm for Full Belief*:

TN-FB: Believe a proposition p iff it's true and disbelieve it iff it's false.

By specifying the relevant truth-norm and what full compliance amounts to, we've completed Step 1.

Of course, non-omniscient agents can't in general obey TN-FB. Step 2 is what allows TN-FB to play an important regulatory role in the epistemic lives of such agents. For the case of full-belief, the most natural way to measuring how "far" an agent is from the alethically ideal state is simply to count the number of disagreements between the agent's

[7] This section is heavily indebted to and largely is a summary of some of the work in Briggs et al. (ms).

doxastic state and the perfectly accurate doxastic state. That is, we simply count how many inaccurate judgments the agent has.

Now that Steps 1 and 2 are complete, let's turn to Step 3. We begin with the most popular coherence norm on full-belief, *viz*., the standard consistency norm:

CON-FB: An agent's beliefs ought to be logically consistent.

CON-FB is often justified as follows: If your belief-set is logically inconsistent, then it's *a priori* true that you have at least one false belief. So, disobeying CON-FB leads to a kind of alethic defect.

The problem with CON-FB, however, is that disobeying it doesn't necessarily indicate that an agent is foregoing an alternative that she could tell or should have been able to tell was superior. Rational agents recognize that they're fallible and cannot avoid violating TN-FB at least sometimes. Giving up the hope of perfect accuracy isn't necessarily irrational, so long as there is no clear path to a more accurate belief set.

Imagine, for instance, that an agent is in a preface case: She has very strong evidence for and believes the propositions $p_1,\ldots,p_{1,000}$ but she also thinks that q: Some p_i is false. For brevity, we can represent (the salient part of) her doxastic state as follows: $\mathscr{B}_1 = \{B(p_1),\ldots,B(p_{1,000}),B(q)\}$.

$\mathscr{B}_1$ violates CON-FB, since there's no world where each p_i is true but some p_i is false. What keeps her from irrationality, however, is that there's no clear route for her to improve her overall level of compliance with TN-FB. Suppose, for instance, that she retained her belief in $p_1,\ldots,p_n$ but disbelieved q. Her new doxastic state would then be $\mathscr{B}_2 = \{B(p_1),\ldots,B(p_{1,000}),D(q)\}$. $\mathscr{B}_2$ is logically consistent, but it may well be less accurate than the one she started out with. It's possible that each p_i is highly likely while their conjunction $\bigwedge p_i$ is highly unlikely. So, she'd be irrational to opt for $\mathscr{B}_2$ over $\mathscr{B}_1$, since she'd have good reason to think her compliance with TN-FB would go down. Similarly, if she dropped her belief in any of the p_i's, her accuracy would more than likely decrease, since each of them is individually highly probable. Therefore, while

she may well recognize she's given up on having a possibly perfectly accurate belief state, she doesn't have any appealing options.

So, we're in need of a less demanding type of coherence norm. Suppose we've fixed a set of propositions Ω and $\mathscr{B} \subset \Omega$ is the set of propositions the agent believes. Instead of demanding that agents are possibly perfectly accurate, we demand that they obey *Weak Accuracy Dominance Avoidance*:

WADA-FB: An agent's doxastic state should not be weakly dominated. That is, if $\mathscr{B}$ is her set of beliefs, there should be no set $\mathscr{B}' \subset \Omega$ such that:

1. At every world, $\mathscr{B}'$ is *at least as accurate* as $\mathscr{B}$.
2. At some world, $\mathscr{B}'$ is *strictly more accurate* than $\mathscr{B}$.

Violating WADA-FB is a sign of irrationality. If an agent S is weakly dominated, then we can identify some alternative doxastic state that is guaranteed to do at least as well as her own and will perhaps do better. In that case, she has a clear route to potentially improving her doxastic state with no risk, but doesn't.

To get a feel for how WADA-FB works, we'll give an equivalent characterization. Call a set $\mathscr{C}$ of judgments *mostly bad* if (1) at every world, at least half of the judgments in $\mathscr{C}$ are inaccurate, and (2) at some world, more than half the judgments in $\mathscr{C}$ are inaccurate. An agent violates WADA-FB *if and only if* some subset of her doxastic state is mostly bad. Call such an agent *seriously incoherent*.

We then get:

COH-FB: An agent's doxastic state ought not be seriously incoherent.

Since any seriously incoherent agent doesn't obey WADA-FB, we've thus established a coherence norm on full-belief. Any agent in violation is culpably alethically defective.

To recap: For Step 1, we stated TN-FB and noted that the alethically ideal agent believes all the truths and the disbelieves all the falsehoods. For Step 2, we measured how

far an agent was from perfect accuracy merely by counting the number of false beliefs she had. For Step 3, we used WADA-FB to establish the coherence norm COH-FB.

1.1.2 Credence

Step 1 isn't as easy when it comes to credence, since it's not as obvious what the truth-norm is. Ramsey (1931) argued long ago that there were important disanalogies between the domains of full belief and credence on this front. Following Hájek (ms), we might set the problem up thus: The fact that p is true is to the belief that p as ___ is to credence x that p. A belief that p is in some sense vindicated by the fact that p, but it seems far less clear what could vindicate a non-extremal credence in p. In other words, a belief that p is correct iff p is true, but there seems to be no analog notion of correctness for credence.

Some authors have suggested supposed analogs in terms of gradational accuracy, such as the following two:

NGA1: An epistemically rational agent must evaluate partial beliefs on the basis of their gradational accuracy, and she must strive to hold a system of partial beliefs that, in her best judgment, is likely to have an overall level of gradational accuracy at least as high as that of any alternative system she might adopt. (Joyce (1998: p. 579))

NGA2: An epistemic agent ought to approximate the truth. In other words: she ought to minimize her inaccuracy. (Leitgeb and Pettigrew 2010a: p. 202)

Now, NGA1 isn't a good analog of TN-FB, since the former is clearly internalist, while the latter is externalist. NGA2 is the correct analog, but it's stated in an overly complicated way. For any proposition, there's a uniquely minimally inaccurate credence to have in it. So, at least as stated, NGA2 is equivalent to the following command:

TN-C: Have a credence in p that matches the truth-value of p. In other words, have credence 1 in p iff p is true, and have credence 0 in p iff p is false.

Note that an agent can't follow TN-C and have a non-extremal credence in any proposition. One may worry, then, that either there's something wrong with the Truth Norm or there's something wrong with credence.

There isn't. Just as it is sometimes rational to have a set of full-beliefs that are mutually inconsistent (as in preface cases), it's also sometimes rational to have non-extremal credences. Agents who recognize that they're non-omniscient sometimes rationally give up on the hope of being completely accurate.

We simply are in need of a good *evaluative criterion* that tells us how far an agent is from perfect accuracy. Not all alethically imperfect states are equally bad, so we need some way to evaluate how far off the mark we are. In the case of full-belief, counting discrepancies between the agent's doxastic state and the perfectly accurate doxastic state stood out as obviously the most natural. In the case of credence, matters are much trickier.

Suppose we have two credence functions $\mathfrak{b}$ and $\mathfrak{c}$ that assign credences as follows:

Table 1.1:

	p	$\neg p$	q	$\neg q$
$\mathfrak{b}$	.7	.3	.7	.3
$\mathfrak{c}$	1	0	.5	.5

Suppose p and q really are true. Which credence function is more accurate overall? The answer isn't obvious, and different standard methods of evaluation lead to different results in this case.[8]

We'll return to this issue shortly, but we can side-step it for a moment and move on to Step 3. Suppose we wish to argue that a rational agent's credence function ought to be a probability function. In other words, suppose we want to argue that:

[8]For instance, $\mathfrak{b}$ does better according to the Brier Score, but worse according to Logarithmic Score.

PROB: Agents with credences should be probabilistically coherent.

It turns out that under *any standard measure of accuracy*, we get this result through the analog of WADA-FB.[9]

Let $\mathfrak{C}$ be the set of credence functions (over an algebra), and let W be the set of worlds.

WADA-C: Let $I : \mathfrak{C} \times W \to \mathbb{R}$ be a reasonable measure of accuracy. An agent's doxastic state should not be weakly dominated. That is, if $\mathfrak{c}$ is her credence function, there should be no $\mathfrak{c}' \in \mathfrak{C}$ such that:

1. At every world, $\mathfrak{c}'$ is *at least as accurate* according to I as $\mathfrak{c}$.
2. At some world, $\mathfrak{c}'$ is *strictly more accurate* according to I than $\mathfrak{c}$.

For any reasonable I, a credence function $\mathfrak{c}$ is dominated just in case it is probabilistically incoherent.[10] So, even though we haven't pinned down the right measure of accuracy for credences, we get PROB from WADA-C.

1.1.3 Takeaways

1.1.3.1 Dominance

Let's start with Step 3. In both the full-belief case and the credence case, we sketched an argument for coherence norms based on dominance considerations. If an agent could choose between two different doxastic states and opts for one that never does better and sometimes does worse, then she's epistemically culpable. So, if violating a coherence constraint results in accuracy dominance, then an agent is irrational if she violates it.

If a coherence constraint is a rational requirement, must we be able to show that violating it results in accuracy dominance? I think so. Suppose that your doxastic state

[9]The relevant class is the set of strictly proper scoring rules. I don't want to get into the details of why this class is standard, so I refer to the reader to Joyce (1998, 2009); Pettigrew (2011); Selten (1998).

[10]For a demonstration, see Predd et al. (2009).

d is not dominated. So, for any alternative doxastic state d', there's some w where d is more accurate than d'. In that case, we can't categorically recommend d' over d, since doing so might lead the agent to be less accurate than she currently is.

This mere possibility should be enough to demonstrate that d isn't always irrational. After all, coherence constraints prohibit agents from ever—under any circumstances—having certain doxastic states. Furthermore, coherence constraints aren't dependent on the agent's background evidence. So, if we can't offer the agent an option that is guaranteed to do at least as well as her own and sometimes better, it's hard to see how she is necessarily rationally culpable.

We were also able to show in the cases above that no doxastic state that obeyed PROB or COH-FB was dominated. Therefore, since all and only the violating states are always irrational, we know we've arrived at the strongest correct coherence norms.

So, to establish coherence constraints on comparative confidence, we'll need a dominance argument. To establish that we've identified the strongest such norms, we'll need to show that no state that complied with those norms is itself dominated.

1.1.3.2 Evaluative Criteria

Step 2 requires that we find a decent way of measuring how far from perfect accuracy a given doxastic state is at a world. This step will usually be the most controversial, since it's a tall order to establish that a given measure or even a given class of measures is correct. While in the full-belief case, counting discrepancies was the most natural, it's not obviously the only reasonable way to go. In the credence case, there is a large class of reasonable measures that disagree on which doxastic states are overall more accurate. Our dominance argument is convincing only because each member of the class leads to the same result, *viz.*, all and only probabilistically incoherent credence functions are dominated.

In spite of these difficulties, evaluation is obvious at the local level. $B(p)$ is better

than $D(p)$ when p is true. A credence of .8 is better than a credence of .7 in a true proposition according to every reasonable measure. It may not be clear just how much better .8 is than .7 nor how to aggregate the local scores to arrive at a global score, but for any world and any proposition, it's self-evident how to order potential judgments in terms of accuracy.

Call an measure of accuracy *local* if gives a score to a doxastic attitude toward a single proposition at a world. A *local* measure of accuracy is a function of the set of possible doxastic attitudes (of a given type) toward a proposition and truth-values.

For illustration, we'll examine accuracy measures for credences. Consider the Local Brier Score:

$$\mathrm{LBS}(x, p, w) = \begin{cases} (1-x)^2 & p \text{ is true at } w \\ x^2 & p \text{ is false at } w \end{cases}$$

where $x \in [0,1]$, p is a proposition, and w is a world. Such a score is local because it doesn't need any input other than p's truth-value and the agent's attitude toward p alone.

By contrast, a measure of accuracy is *global* if it gives a single score to an entire doxastic state at a world. If $\mathfrak{b}$ is a credence function defined over an algebra, a global measure of inaccuracy will give a score to $\mathfrak{b}$, not to individual credences such as $\mathfrak{b}(p)$. For example, consider the Global Brier Score:

$$\mathrm{GBS}(\mathfrak{b}, w) = \frac{1}{|\Omega|} \sum_{p \in \Omega} (v_w(p) - \mathfrak{b}(p))^2$$

where $\mathfrak{b}$ is a credence function and $v_w(p)$ is p's truth-value at w.

Notice that the value of $\mathrm{GBS}(\mathfrak{b}, w)$ is determined by aggregating $\mathrm{LBS}(\mathfrak{b}(p), p, w)$. Furthermore, GBS works by aggregating the local scores LBS gives without regard to how they interact with one another. GBS scores $\mathfrak{b}$ by looking at how accurate $\mathfrak{b}(p)$ is for each proposition p. $\mathfrak{b}(p)$'s contribution to $\mathfrak{b}$'s overall accuracy is independent of how accurate $\mathfrak{b}(q)$ is for any other proposition q.

To formalize this idea, we note that GBS is separable in the following sense:

Separability Suppose $\Gamma \subseteq \Omega$, let $\mathfrak{b}, \mathfrak{c}, \mathfrak{b}', \mathfrak{c}'$ be credence functions over Ω such that:

1. $\mathfrak{b}(p) = \mathfrak{b}'(p)$ and $\mathfrak{c}(p) = \mathfrak{c}'(p)$ for any $p \in \Gamma$
2. $\mathfrak{b}(p) = \mathfrak{c}(p)$ and $\mathfrak{b}'(p) = \mathfrak{c}'(p)$ for any $p \in \Omega - \Gamma$

A global measure of accuracy G is *separable* if $G(\mathfrak{b}, w) \geq G(\mathfrak{c}, w)$ iff $G(\mathfrak{b}', w) \geq G(\mathfrak{c}', w)$.[11]

Separable measures of accuracy base their scores only on the accuracy of the individual judgments in isolation. Thus, there's no essentially global component of the score.[12]

Reasonable measures of global accuracy for both full-belief and credence should be separable. Accuracy, after all, is simply a matter of matching one's doxastic state to the world. Coherence, on the other hand, has to do with how individual doxastic attitudes mesh with one another. If an alleged measure of accuracy for full-belief or credence were non-separable, it would depend on purely global properties of an agent's doxastic state. Because it wouldn't simply look at how doxastic attitudes match up against the world but on the internal relationships between the judgments, it would fail to measure accuracy and accuracy alone.

Below, I'll argue that because of the nature of comparative confidence, we can't come up with any good measure of accuracy that results in the desired coherence norms. The underlying cause is that any comparative attitude toward a proposition isn't isolatable. My attitude toward p is bound together with my attitude toward q, which prevents us from coming up with attractive, local measures of accuracy. It's now time to look more closely at how how comparative confidence works.

---

[11] I take this definition and terminology from Joyce (2009: pp. 271-2).

[12] Notice that separable measures obey the epistemic analog of the Sure-Thing Principle in decision theory.

1.2 Comparative Confidence

1.2.1 Global Nature of Doxastic Attitudes

Before looking directly at the notion of accuracy or fit for comparative judgment, we can already see a fundamental difference between it and full-belief and credence. If you ask an agent with full beliefs and disbeliefs what her attitude toward p is, she'll either say she believes it or she disbelieves it. Likewise, if you ask an agent with credences what her attitude toward p is, she'll report a precise number to you. Agents with full-beliefs and credences can tell you their doxastic attitudes toward a given proposition without having to mention any other propositions at all. That is, they can usefully isolate their attitudes toward any given proposition from their attitude toward other propositions. However, if you ask an agent who only has comparative confidence judgments what her attitude is, there's no obvious answer. The best she can do, it seems, is to give you a list of all the propositions she's more confident in than p, all those she's less confident in than p, and all those that she's as confident in as p. Though such a report would presumably count as a full disclosure of her doxastic attitude, it's troubling that she'd have to survey her whole doxastic state just to tell you how she feels about whether p. As we'll see, the fact that her attitude toward p is entangled with her attitude toward other propositions makes the prospects dim for understanding how comparative judgments are supposed to fit the world.

1.2.1.1 Digression on Possible Variations

We can partially alleviate this problem in two ways. The first is to move to something like an ordinal scale. The idea here is that an agent could quickly report her attitude toward p merely by naming where p was ranked among some algebra of propositions. If, for instance, there are 4 possible states and 16 propositions, an agent could report that some given proposition p was ranked 3$^{\text{rd}}$ out of 16. This helps to some extent in the report

of doxastic attitudes. Obviously, some reference is still made toward other propositions, but an agent could report an attitude toward a given proposition without reporting her attitude toward *every* proposition. There is a slight disadvantage, however: Moving to an ordinal scale would force agents with comparative credences to have a well-ordered ranking, whether they're rational or not. Transitivity would simply be a structural pre-condition for having ordinally ranked beliefs, just as determinate point-wise credences are a structural pre-condition in many models. Ideally, we could generate a dominance argument for transitivity instead of building it into the framework.

Another way to help is to introduce what I'll call the pseudo-proposition Mid. Here's how Mid works: We say that $p \succ_S \text{Mid}$ just in case S thinks that p is more likely than not. Now, we can go either of two ways here. The first way is just through the following abbreviation: $p \succ_S \text{Mid} \equiv_{df} p \succ_S \neg p$. That is, if S thinks p is more likely than its negation, then we say that she thinks p is more likely than not. The advantage is that we don't need any additions to our system, but there's the disadvantage that we rule out thinking that both p and $\neg p$ are more likely than not. So, the second way is to give Mid a bit more autonomy. When asking agents what their comparative judgments are, we can also ask them which propositions they think are more likely than not. If we take this second option, we can allow agents to think both $p \succ_S \text{Mid}$ and $\neg p \succ_S \text{Mid}$.

Both the addition of ordinal structure and the addition of Mid I take to be well within the spirit of comparative probability. The whole point is to give yourself more structure than full belief without the more wild idealizations necessary for credences. In particular, we don't want built in notions of distance nor absolute levels of confidence, aside from some small few. Ordinal structure and Mid avoid both of these.

1.2.2 Evaluative Criteria for Doxastic Attitudes

As I've argued, the very notion of a doxastic attitude is dependent on the goal of truth. However, no agents of interest can perfectly follow any relevant truth-norm. As epis-

temologists, we're in the business of telling believers what to do and how well they're doing. Therefore, if we're going to take a certain kind of doxastic attitude as fundamental, we must be able to provide sufficiently robust evaluative criteria. At the least, we'll have to be able to make non-trivial claims either about which of two agents is doing better overall or which of two doxastic attitudes toward a given proposition is superior. *It is on this front that comparative probability fails*. That is, there simply is no good way of generating strong enough criteria of evaluation for comparative probability that could justify giving it a fundamental role in epistemology.

The primary aim of the rest of the paper will be devoted to showing that there can be no adequate measures of fit for comparative belief. I don't have any knock-down argument along these lines, but I think that, given the failures so far, no such measure is forthcoming. I'll try nonetheless to highlight some loose-ends and unexplored routes in the hope that future progress might still be made.

Here's the basic strategy: We'll look first at how we could develop good evaluative criteria for a given kind of doxastic attitude vis-à-vis the truth-norm. In particular, we'll focus on how our evaluative criteria might interact with coherence norms. Then, we'll look at a few different approaches to specifying how to measure the fitness of a bunch of comparative beliefs and see that they fail. Finally, we'll throw up our hands and conclude that it can't be done.

1.2.2.1 Conceptual Reasons for Failure

Before going into all this depth, I want to look at what I think are the basic conceptual reasons to doubt the possibility of a successful measure of fit for comparative probability. Let's first look at the relevant truth-norm:

TN-CB: Rank all truths above all falsehoods, be equally confident in two truths, and be equally confident in two falsehoods. Restated, we have: If p and q are both true or

both false, be such that $p \approx q$; if p is true and q is false, be such that $p \succ q$.[13]

TN-CB may be unfortunately phrased, but any agent who follows it will end up in what is clearly the best possible state from a purely veritistic point of view. What we don't have, however, is much of a start on any sort of evaluative criterion for states that don't fully meet this norm. I.e., we don't have a good way of completing Step 2.

There are two primary reasons why I think no evaluative criterion is forthcoming, which I'll refer to as *non-locality* and *incomparability*. First, let's look at locality. In the case of full-belief and credence, local measures are primary. We start with the accuracy of a given belief or credence. The only factors that matter when measuring accuracy are what the world is like and what the doxastic attitude toward p is. The accuracy of a set of beliefs or credences is then just a function of the accuracy of the doxastic attitudes toward the individual propositions.

When thinking about accuracy, we want to check how close our doxastic attitudes are to the world. However, because of the interdependence of the doxastic attitudes of comparative believers, it's hard to see how any sense could be made of checking how well my doxastic attitude *toward p alone* could match up against the world. The attitude toward p is impossible to specify *without at least implicitly referencing every other proposition in the algebra*. Unlike full-belief and credence, with comparative probability, what my attitude toward p is only makes sense when other propositions are brought into the fold. A comparative believer cannot in general change her attitude toward p without changing her attitude toward at least one other proposition as well. If previously she ranked p above q and now thinks that they're equally likely, her attitude has changed both toward p and toward q.[14] In turn, any good measure of fitness for comparative belief has to be differ-

[13]If we add Mid as a non-derived entity—i.e., if we don't define $p \succ \mathsf{Mid}$ simply as $p \succ \neg p$—then right way to state the norm gets a bit more complicated, since we'll have to specify in addition that truths should be ranked above Mid and falsehoods below it.

[14]One might think that it's indeterminate whether my attitude toward q has changed or whether my attitude toward p has changed or both. In this way of thinking, it still comes out that doxastic attitudes are

ent in kind from our measures of accuracy for full-belief and credence. Our measure of fitness must be *essentially non-local*.

This non-locality wouldn't be a problem *per se* if it were usually easy to tell which orderings (or sub-orderings) were more accurate than others. Unfortunately, it doesn't seem that there's any non-arbitrary way to do so. We have no easy way of deciding whether two distinct attitudes toward p that both rank p strictly between $\top$ and Mid are better, even when holding everything else fixed.

For full-belief and credence, accuracy was a matter of having more confidence in true propositions and less confidence in false propositions. However, there's simply no fact of the matter as to whether an agent in doxastic state S_1:$\top \succ p \succ q \succ \mathsf{Mid}$ is more confident in p than an agent who's in doxastic state S_2:$\top \succ p \approx q \succ \mathsf{Mid}$.[15] When we try to assign some sort of measure of accuracy to S_1 or S_2, then, it's unclear which is more accurate in a world where p and q.

The issue becomes more perspicuous if we wish to evaluate an agents who have credences both for the accuracy of their credences and for the accuracy of their comparative confidence judgments. Without loss of generality, suppose that if p and q are both true, S_2 is more accurate than S_1. Suppose further that we can read off comparative judgments from an agent's credences. That is, an agent is more confident in p than in q just in case she has a higher credence in p than in q. Given these assumptions, it's clear that more accurate credences will sometimes lead to less comparative accuracy and vice versa. Consider: $1 > \mathfrak{c}(p) > \mathfrak{c}(q) > \mathfrak{b}(p) = \mathfrak{b}(q) > .5$. $\mathfrak{c}$ is more accurate than $\mathfrak{b}$, but an agent with credence function $\mathfrak{b}$ has more accurate comparative judgments (per our assumptions) than an agent with credence function $\mathfrak{c}$.[16]

---

inter-dependent.

[15]Strictly speaking, both agents have ordinal rankings over an algebra. For brevity, I only include the salient features of their respective doxastic states.

[16]While I take this argument to elucidate the problems with the notion of accuracy for comparative confidence, I don't think it is as damning as it might appear. The defender of comparative confidence can

In sum, here's the problem: To come up with a measure of fit, we can't have a bottom up approach that builds a global measure of accuracy from a local measure, we can't make many judgments about the different strengths of attitudes between different agents, and we also don't have any intuitive means for deciding which global doxastic states are more accurate than others. So, any global measure seems like it will either be arbitrary or it will be too anemic to do the work we need it to. Still, it's worth taking a look in a bit more depth and seeing whether anything can be done. A potential source of help comes from the grasp we already have on coherence norms, which we'll take a look at immediately below.

1.3 Desiderata

We've committed to the idea that a dominance argument should vindicate coherence norms on comparative confidence. I take it, however, that we already have a fairly good idea of what these norms are supposed to look like: something like the standard axioms. While some of these requirements are negotiable, reflective equilibrium demands that an adequate measure of accuracy deliver something in the ballpark. We review those axioms presently, with commentary.[17]

Standard Axioms

(1) $p \nsucc \top$, $p \nprec \bot$, and $\top \succ \bot$

I take it that this axiom is obvious. If a measure of accuracy allows rational agents to be more confident in p than in the necessary proposition, less confident in p than in the

respond that credences are parasitic on comparative confidence.

[17] I ask the reader to forgive the meandering nature of what follows. While my main goal is to show why it's especially difficult to find a measure of accuracy that will deliver the right results, I also hope to help point the way for potentially successful future efforts.

necessarily false proposition, or not more confident in the necessary proposition than in the necessarily false proposition, then that measure is obviously wrong.

(2) $\succeq$ is a total order.

In other words, rational agents should **(2a)** make comparisons between all elements of the algebra, and furthermore, **(2b)** $\succ$ should be transitive and **(2c)** $\approx$ should be an equivalence relation. Now, I take it that of these conditions, only (2b) is clearly mandatory. It's contentious whether all elements need to be compared. It seems like it might be overly demanding to require agents to have thorough opinions about the relative likelihood of any two propositions whatsoever. Condition (2c) is also not obvious. Consider, for instance, cases of perceptual indiscriminability. If I'm a witness to a bank robbery, I may recall that the getaway car was dark blue. Suppose the police want me to help them identify which car the robbers used and show me a line up with thirty sedans, each of which is a slightly different shade of blue. Some will obviously not be the getaway car, while others I judge more likely. I won't generally be able to discriminate which of two is more likely when they're both nearly the same shade of blue. However, since I can discriminate between cars that are very different shades, I can judge some less likely than others. It seems that in a case like this I could be perfectly rational while not having $\approx$ be an equivalence relation. Therefore, we might instead list as desiderata just (2b) and take (2a) or (2c) as not speaking to the adequacy of a measure. Note, however, that we could simply build in (2) in its entirety by moving to an ordinal model instead of a merely comparative one.

(3) Suppose $\langle p, q\rangle$ and $\langle p, r\rangle$ are mutually exclusive pairs of propositions. Then $(p \vee q) \succ (p \vee r)$ iff $q \succ r$ (and *mutatis mutandis* for $\approx$).

This axiom is due to de Finetti (1964) who thought that, when combined with (1) and (2), it was sufficient for probabilistic representability. That is, all agents who obeyed (1), (2), and (3) were such that some numerical probability function preserved their confidence ranking. It turned out he was wrong about this, but the axiom is still the most nat-

ural analog of the additivity axiom for numerical probability: $\Pr(p \vee q) = \Pr(p) + \Pr(q)$ when p and q are mutually exclusive. In the case of finding measures of accuracy that deliver the numerical probability, this axiom is hard to achieve. Finding out which measures get (3) through dominance considerations, then, would help illuminate the underlying assumptions necessary. As far as I can tell, this can't be done. Even more difficult would be:

(3′) Scott's Axiom: Let $\pi = \langle p_1, \ldots, p_n \rangle$ and $\rho = \langle q_1, \ldots, q_n \rangle$ be arbitrary sequences of propositions such that at every world $w \in W$, π and ρ contain the same number of true propositions. Suppose that for $i \in \{2, \ldots, n\}$, $p_i \preceq q_i$. Then, $p_1 \succeq q_1$.

Scott's Axiom, when combined with (1) and (2) is necessary sufficient for probabilistic representability.[18] However, it's not intuitive in any way. Its only motivation, as far as I can tell, is simply to guarantee such representability. (3′) definitely does not pass the analog of the Russell test for set theory. No theorist would have listed it as an axiom were she not already aware that it was needed to get the desired result. If some intuitive measure of fit or accuracy could deliver (3′), it would be a major accomplishment for this approach.

1.4 Possible Approaches

As far as I can tell, there are three natural kinds of approaches to developing measures of fit for comparative probability, which I'll call *intrinsic*, *extrinsic*, and *entropic*.

1.4.1 Intrinsic Approach

On the intrinsic approach, our measure of fit looks directly at the comparative confidence judgments directly to generate a score. In other words, we don't move the discussion from

[18] For the proof, see Scott (1964).

the space of comparative judgments to some other mathematical domain. This approach is akin to synthetic geometry, not analytic geometry. One instantiation of this method is as follows.

1.4.1.1 Extensional Approach

Let's review what the game is. We're trying to come up with a formal notion of fitness for comparative beliefs that is sufficiently intuitively plausible and that delivers the dominance results necessary for the desired constraints on coherence. One natural and simple way to score comparative judgments is by examining each judgment individually based on the truth-values of the two propositions.

Call a scoring scheme *extensional* if the score for a set of comparative judgments C is a function of the truth-values of the propositions in each judgment in C. That is, the scoring scheme simply looks at each judgment pRq, where $R \in \{\succ, \approx\}$, and gives an individual score to it based only on the truth-values of p and q, and then gives an overall score to C as a function of the individual scores. There are a few different ways to do this, but the easiest one is as follows:

- $p \succ q$ is (in)correct at w iff $p \wedge \neg q$ is true (false) at w.
- $p \approx q$ is (in)correct at w iff $p \equiv q$ is true (false) at w.

We say that a set of comparative judgments C *e-dominates* another C' iff C has strictly fewer incorrect judgments at some worlds and has more incorrect judgments at no worlds.

Unfortunately, despite its elegance, this sort of approach delivers some undesired consequences. It's not hard to show that any such scoring scheme that meets a few other basic desiderata has the following problem: Either (1) any C with a judgment of the form $p \approx \neg p$ is dominated, or (2) it faces the problem for lotteries listed a few sentences below. The scoring scheme above faces issue (1). David Christensen pointed out that by its lights, $p \succ \neg p$ will never do worse than $p \approx \neg p$ and will sometimes do better. That problem

can be fixed by, e.g., moving to a multi-valued scoring scheme. (2) is more intractable. Suppose we knew that one of three tickets was going to win the lottery, and they were going to be drawn randomly. Let $L(i)$ be the proposition that ticket i loses. Intuitively, we should have $C1 : \top \succ L(1) \approx L(2) \approx L(3)$. However, this ordering is dominated by $C2 : \top \approx L(1) \approx L(2) \approx L(3)$.[19] Since we don't want our dominance argument to rule out any rational states, the extensional approach is doomed.

I think the fundamental problem with the extensional approach is just that it misunderstands what a comparative confidence judgment is. When I judge $p \succ q$ I don't necessarily think that p and q have different truth-values. However, on the scoring scheme proposed above, I get points only when they in fact do have different truth-values. Likewise, judging $p \approx q$ is not the same as judging $p \equiv q$, but the extensional approach identifies them.

This confusion stems, I think, from TN-CB as it's stated. TN-CB tells us to have the ideal state. However, the way it describes that ideal state misleads us into the wrong evaluative criterion. Recall TN-CB's wording from above: *If p and q are both true or both false, be such that $p \approx q$; if p is true and q is false, be such that $p \succ q$*. The problem is that it's not so clear that having $p \approx q$ when and only when they have the same truth-value is of any worth. It's true that the ideal state does co-rank propositions with the same truth-value, but that's only because it gives maximal belief to truths and minimal belief to falsehoods. Likewise, the ideal state only ranks p above q when p is true and q is false, but that alone doesn't show that there's anything wrong *per se* with ranking p above q when both are true. The mere fact that the ideal state has a certain property D doesn't

[19] To see this, assume without loss of generality, that ticket 1 wins. Then $C1$ is right about: $\top \succ L(1)$ and $L(2) \approx L(3)$. $C1$ is wrong about: $\top \succ L(2), \top \succ L(3), L(1) \approx L(2)$, and $L(1) \approx L(3)$. Overall, that's 2 right judgments and 4 wrong judgments. $C2$, on the other hand, is right about: $\top \approx L(2), \top \approx L(3), L(2) \approx L(3)$ and wrong about $\top \approx L(1), L(1) \approx L(2)$, and $L(1) \approx L(3)$. Overall, that's 3 right and 3 wrong. No matter how we weight right and wrong judgments in an extensional scoring system, this problem will emerge for big enough lotteries.

provide license to use the bearing of D in our evaluation of sub-ideal states.

Here's one way of bringing out the confusion. I could restate the truth-norm for credence in a logically equivalent way thus: Have only a rational number for your credence in any proposition, and make that rational number be 1 is the proposition is true and 0 if the proposition is false. Despite the phrasing, it would be wrong to dock accuracy points from a partial believer merely for having an irrational number as a credence in a given proposition.

1.4.1.2 More Generally

One way to start to fix this problem is to invoke Mid. If I judge $p \succ q \succ \mathsf{Mid}$, then I should overall do best when p and q are both true. After all, I think they're both more likely than not, so regardless of how they compare to each other, my overall doxastic attitude fits the world better when they're both true. Next best for me, in terms of fit, is when p is true and q is false. Third best is when p is false and q is true. Worst is when both p and q are false. The scoring scheme will be different when $p \succ \mathsf{Mid} \succ q$ and when $\mathsf{Mid} \succ p \succ q$.

We can expand this approach further by giving a score to a proposition p based on, e.g., its rank and truth-value and the cardinality of the algebra Ω of propositions. For instance, the higher p is ranked, the better the score when it turns out to be true, and the worse it does when it turns out to be false. With an eye toward the desiderata above, we can try rig up the scoring to get us desired results. For instance, we might avoid the problem of lotteries we found in the extensional method by making the penalty for a false proposition co-ranked with $\top$ especially severe. In addition to p's rank and truth-value, we can also look at specifically comparative facts. For instance, we might think we should additionally penalize an agent for ranking a false proposition above a true one. If we have $\top \succ q \succ p$, we might give p a different score if q is false than if q is true.

1.4.1.3 Problems

One problem is figuring out how we should now treat $p \approx q$ versus $p \succ q$. If I rank them both above Mid and they both turn out to be true, it's not completely obvious whether I should do better with $p \approx q$ or $p \succ q$. Intuitively, the former should get a better score, *ceteris paribus*. However, we could imagine an agent with credences who's asked to report only her comparative judgments. If her credence in p is very slightly higher than her credence in q, we want her to report $p \succ q$. But, if there is some fixed numerical scoring scheme, and the difference between her credence in p and q is less than some sufficiently small ϵ, she'll expect to do better if she lies and reports $p \approx q$. That result seems unwelcome, but we're stuck, since we need for there to be some difference. We want the ideal set of comparative judgments, where all truths are ranked with $\top$ and all falsehoods are ranked with $\bot$, to do strictly better than any other set. In particular we want it to do better than a set in which all truths (falsehoods) are ranked above (below) Mid, but $\top$ is ranked strictly above all other propositions.

More importantly, though, are a bunch of technical problems. First, it's quite hard to generate a scoring system even under these looser constraints that will deliver results of interest, even if we allow the scoring system to be *ad hoc*. Furthermore, some desired results seem impossible to get. Consider, for instance, the following ordering $R : \top \succ p \succ \neg p \approx \bot$. We definitely want R to be dominated, but it's hard to see how it could be. The only relevant alternative orderings that look rationally permissible are:

- R_1: $\top \succ p \succ \neg p \succ \bot$
- R_2: $\top \approx p \succ \neg p \approx \bot$
- R_3: $\top \succ p \approx \neg p \succ \bot$

Now, on any reasonable method of measuring fit, R will beat R_1 and R_2 when p is false, and it should beat R_3 when p is true. Without a bunch of further justification of a measure, it's hard to see anyway around this fact.

Perhaps, then, we should retreat and consider an extrinsic method.

1.4.2 Extrinsic Approach

As we just saw, the most natural way to score comparative probability doesn't work. To check fit against the world, it would be especially helpful to have some notion of distance to work with. One step in this direction is to move the analysis over to a domain with more structure. The most obvious thing to do is to consider homomorphisms from a set of comparative judgments to $\mathbb{R}$. For this to work, we'll have to demand transitivity, since otherwise it will be impossible to represent the set of comparative judgments faithfully. Since we already have scoring rules for credence, we can use some of the tools developed for those models to help us. In more detail, here's how this kind of idea would work. Let $\mathscr{O}$ be an ordinal ranking of the propositions in some algebra Ω. We say that $h : \mathscr{O} \to [0,1]$ is a homomorphism if:

- All propositions p such that $p \succeq q$ for all q are sent to 1.
- Mid gets sent to .5.[20]
- All propositions p such that $p \preceq q$ for all q are sent to 0.
- $h(p) > h(q)$ iff $p \succ q$, and $h(p) = h(q)$ iff $p \approx q$.

Let $\mathsf{Hom}(\mathscr{O}) := \{h : \mathscr{O} \to [0,1] : h \text{ a homomorphism}\}$. The idea here is that everything that matters about the ordering $\mathscr{O}$ is common to every $h \in \mathsf{Hom}(\mathscr{O})$, though as we'll see below, we can refine $\mathsf{Hom}(\mathscr{O})$ further. Thus, as a first pass we can represent an ordering as the set of all credence functions that agree with it.

By sending comparative judgments to sets of credences, we can then use a scoring rule for credences to make dominance arguments. How dominance works here will be

[20] Alternatively, if we don't want to have Mid officially represented, we make h sent p to a number $> .5$ iff $p \succ \neg p$, to a number $< .5$ iff $p \prec \neg p$ and to .5 otherwise.

something of a subtle matter, however, since as we'll see, we can define a number of different notions. For the sake of simplicity, I'll take the Brier score to be our scoring rule.

At this point, it's worth considering a special kind of homomorphism that will be important. Call $g : \mathscr{O} \to [0,1]$ an *interpretation of* $\mathscr{O}$ if g is a homomorphism and it's not the case that there's a $g' \in \mathsf{Hom}(\mathscr{O})$ such that g' Brier-dominates g. Let $\mathsf{Int}(\mathscr{O})$ be the set of all interpretations of $\mathscr{O}$.

Interpretations are charitable homomorphisms. The idea is if one homomorphism interprets you in a way that is necessarily more accurate than another, then we should consider the less accurate one as disqualified. $\mathsf{Int}(\mathscr{O})$ has some nice properties. In particular, as is not hard to show, it's convex (though not necessarily closed), which will make it easier to compare distinct orderings.

We should take a moment to think about how dominance works here. The most natural way to go is to say that $\mathscr{O}_1$ *i-dominates* $\mathscr{O}_2$ just in case every $h \in \mathsf{Int}(\mathscr{O}_1)$ is Brier-dominated by every $g \in \mathsf{Int}(\mathscr{O}_2)$. However, we'll need some way to throw out more credence functions as ineligible if we're going to get dominance results of any significance using this definition. To see why, it's best to look at an example. Let $\mathscr{Q}$ be the ordering: $\top \succ p \succ q \succ \mathsf{Mid} \succ \neg p \succ \neg q \succ \bot$. Though we'd like $\mathscr{Q}$ to be dominated, there's obviously not going to be any particular alternative ordering which is up to the task. Some elements of $\mathsf{Int}(\mathscr{O})$ send p to .999, while others send it to .5001. In general, whenever $\top \succ p \succ \mathsf{Mid}$, there will be an element of Int that sends p to x for any $x \in (.5, 1)$. So, there's no way some alternative ordering $\mathscr{O}'$ could be such that *every* element of $\mathsf{Int}(\mathscr{O}')$ dominated *every* element of $\mathscr{O}$. Though we could perhaps find creative ways to exclude more homomorphisms from contention, it's unlikely that we'll ever be able to use such a strong definition of dominance to get any important results, since orderings (of which there are only finitely many) drastically underdetermine compatible credence functions (of which there are uncountably many).

The best we can do is to weaken the definition of dominance. Instead of requiring every element of $\mathsf{Int}(\mathcal{O}')$ to dominate every element of $\mathsf{Int}(\mathcal{O})$, we could try the following. $\mathcal{O}'$ *w-dominates* $\mathcal{O}$ just in case for every $h \in \mathsf{Int}(\mathcal{O})$ there exists a $g \in \mathsf{Int}(\mathcal{O}')$ such that g Brier-dominates h. Unfortunately, even on this weaker notion, there just aren't even the minimal results we want. It's not hard to check (by computer) that $\mathcal{Q}$ isn't even w-dominated. As far as I can tell, no further weakening is of much interest.

Of course, one could object either to this representation or to the scoring rule used. However, since this representation scheme was relatively natural, it's doubtful that anything else could have much intuitive pull.

1.4.3 Entropic Method

One standard way of measuring distance between numerical probability functions is to look at their relative entropy. We could tell, for instance, how far away a given probability function was from the true state by considering how much of a surprise it would be when starting with the original probability function to learn the actual state of affairs. The less the surprise the better.

Computer scientists have some tools for measuring the entropic distance between sequences and strings, such as Levenshtein distance. Perhaps there's some good way of checking how "far" in an information theoretic sense a given ordering is from another. Unfortunately, nothing I've tried so far has worked, but I take it that this method has the most potential going forward.

1.5 Conclusion

Comparative probability does not seem up for the task de Finetti (1964) and others wanted. We argued above that if comparative probability is to serve a foundational role, there must be a sufficiently robust and attractive measure of fit to the world. Because dox-

astic attitudes in a comparative model are so heavily entangled with one another, no such measure appears possible. So, without further advances, comparative judgments should be taken at best as subsidiary to some more fundamental kind of doxastic attitude.

Chapter 2

Against Quadratic Scoring Rules

2.0 Introduction

In their (2010a; 2010b), Leitgeb and Pettigrew argue for a number of constraints on rational belief through appeal to the following norm:

ACCURACY: An epistemic agent ought to approximate the truth. In other words: she ought to minimize her inaccuracy. (2010a: 202)

Of course, for ACCURACY to be of much use to epistemologists, more must be said, and Leitgeb and Pettigrew are able to make it mathematically precise. In order to show how, I'll first streamline the discussion and assume all agents under consideration obey the following synchronic norm:

PROBABILISM: At any given time, an agent ought to have a probabilistically coherent credence function.

We can treat inaccuracy at a world formally as follows. Let A be a proposition, w a world, and let $v_w(A)$ be 1 if A is true at w and be 0 otherwise.[1] The idea is that v_w—the

[1] I assume throughout that the set of possible worlds W is finite. This restriction may seem implausible, but (1) it's fairly standard, and (2) we can understand W to be the set of the most fine-grained possibilities

characteristic function of w—represents the best possible credence to have at w, since it assigns maximal (minimal) credence to all propositions true (false) there. Leitgeb and Pettigrew argue that we should measure the *inaccuracy* of a probability function $\mathfrak{b}$ at w by seeing how far it is from v_w under the average of the squared Euclidean distance between them.[2] We end up with the *Brier score* as our inaccuracy measure:[3]

$$I(\mathfrak{b}, w) = \frac{1}{|W|} \sum_{w' \in W} \left(v_w\left(\{w'\}\right) - \mathfrak{b}\left(\{w'\}\right)\right)^2$$

The Brier score has a natural generalization to a larger class of inaccuracy measures, known as *quadratic scoring rules*, which are of the following form:

$$I(\mathfrak{b}, w) = \sum_{i=1}^{N} \lambda_i \left(v_w(A_i) - \mathfrak{b}(A_i)\right)^2, \text{ where} \sum_{i=1}^{N} \lambda_i = 1, \text{ and each } \lambda_i > 0.$$

Quadratic scoring rules let us thumb the scale and count some propositions as more important than others. If we really care about whether it will rain tomorrow but not so much about whether there are an even number of stars in the universe, we assign a higher weight to the former proposition when measuring our inaccuracy.[4]

In addition to Leitgeb and Pettigrew (2010a,b), quadratic scoring rules have a number of defenders and admirers; in fact, it's safe to say that they are the clear front-runners in the debate over how best to measure the inaccuracy of probability functions.[5]

---

an agent is concerned with. Since real-world agents can only track finitely many distinguished possibilities, I take it the finite case is of primary interest.

[2]We take the average square of the Euclidean distance instead of just the Euclidean distance to guarantee that all probability functions assign themselves lowest expected inaccuracy. I thank a referee for *Philosophy of Science* for catching earlier sloppiness about this point.

[3]Named after Glenn Brier, who originally used it to measure the inaccuracy of weather forecasts. (See Brier 1950.)

[4]To allow for some propositions to count more than others, quadratic inaccuracy measures may give a score to the agent's credence in each proposition A in the algebra. With the Brier score, since we're counting each proposition equally and all agents under consideration are probabilistically coherent, we only have to consider the agent's credences in propositions of the form $\{w\}$ for w a world.

[5]For a partial list, see: de Finetti (1974); Greaves and Wallace (2006); Joyce (1998, 2009); Leitgeb and Pettigrew (2010a,b); Savage (1971); Selten (1998). In §2.1 I briefly look at considerations in favor of the Brier score. For much more, see Selten (1998); Leitgeb and Pettigrew (2010a); Joyce (2009: ğ12).

Nonetheless, I think that quadratic scoring rules serve as poor measures of inaccuracy because they naturally lead to an extremely unattractive updating procedure, as can be shown through application of a result in Leitgeb and Pettigrew (2010b). Other measures, such as the logarithmic scoring rule, lead to the more attractive and standard procedure of Jeffrey-Conditionalization. Therefore, quadratic rules should be rejected.[6]

2.1 Background on Brier

Because the updating norm we'll be studying is so counter-intuitive, it's worth examining some considerations that lead authors to favor the Brier score. First, we discuss Leitgeb and Pettigrew's argument and then turn to Reinhard Selten's.

Leitgeb & Pettigrew's Motivation

Since the Euclidean notion of distance is familiar to us, Leitgeb and Pettigrew (2010a) naturally incorporate it as a default starting point for the framework they develop. To argue for the Brier Score as the right measure of inaccuracy, they insist on the following norm on global measures of inaccuracy:

Global Normality and Dominance If I is a legitimate inaccuracy measure, then there is a strictly increasing function $f : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ such that, for all worlds w and and probability functions $\mathfrak{b}$,

$$I(w, \mathfrak{b}) = f(||w - \mathfrak{b}||)$$

where $||\cdot||$ is the Euclidean norm on vectors. (p. 219, with minor changes)

This constraint is crucial to their argument for the Brier Score, but it is only a preference for the Euclidean metrics that motivates it, as they readily admit.

[6]For our purposes below, I'll focus my attack on the Brier score, but the basic problem I'll highlight will apply to quadratic scoring rules in general.

> Global Normality and Dominance is a consequence of taking seriously the talk of inaccuracy as 'distance' from the truth, and it endorses the geometrical picture provided by Euclidean n-space as the correct clarification of this notion.... [T]he assumption of this geometrical picture is one of the presuppositions of our account, and do not have much to offer in its defense. (*ibid.*)

Leitgeb and Pettigrew are right to try the Euclidean norm as the starting point. However, many other potential measures of inaccuracy, including the Logarithmic Scoring Rule, violate GND.

Selten's Motivation

Selten (1998) provides an axiomatic characterization of the Brier Score from the assumption of PROBABILISM. The potentially controversial axioms are the following:

PROPRIETY I is a strictly proper scoring rule.

NEUTRALITY For any probability functions $\mathfrak{b}, \mathfrak{c}$ over $\mathscr{P}(W)$,

$$\mathrm{E}_{\mathfrak{b}}(I(\mathfrak{c})) - \mathrm{E}_{\mathfrak{b}}(I(\mathfrak{b})) = \mathrm{E}_{\mathfrak{c}}(I(\mathfrak{b})) - \mathrm{E}_{\mathfrak{c}}(I(\mathfrak{c}))$$

where $\mathscr{P}(W)$ is the power set of W.

PROPRIETY requires that for any probability function $\mathfrak{b}$, $\mathrm{E}_{\mathfrak{b}}(I(\mathfrak{b})) \leq \mathrm{E}_{\mathfrak{b}}(I(\mathfrak{c}))$ with equality only if $\mathfrak{b} = \mathfrak{c}$ (where $\mathrm{E}_b(X)$ is the expected value of X according to $\mathfrak{b}$). The idea here is that every probability function should expect a better score for itself than for any other function. Why? Well, if some probability function $\mathfrak{b}$ expected another to be strictly more accurate under some rule I, then (assuming I is a legitimate measure of inaccuracy) it would be *a priori* irrational for an agent to adopt $\mathfrak{b}$ as her credence function. After all, by her own lights, she thinks she can do better with $\mathfrak{c}$ regardless of any possible evidence she may have. That result seems wrong. Surely, we can't rule out any probability function as automatically irrational in any evidential situation. Now, if a probability function $\mathfrak{b}$ expected (under I) that some other function $\mathfrak{c}$ would do *equally* well, we would be forced

to violate any reasonable notion of distance. In particular, as Schervish (1989) shows, there would be an interval $[\alpha, \beta] \subseteq [0, 1]$ such that for any $\gamma, \gamma' \in [\alpha, \beta]$ I counts γ and γ' as equally far from 0 and 1. Imagine, for instance, if I counted .5 and .6 as equally far from 1. Then an agent who has a credence of .5 in a true proposition would get as good of a score (*ceteris paribus*) as an agent with a credence of .6 in a true proposition. Surely, however, the second agent is more accurate, so I is illegitimate. While the assumption of propriety has been questioned, I take no issue with it here.[7]

NEUTRALITY requires that for any legitimate measure of inaccuracy and any two probability functions $\mathfrak{b}$ and $\mathfrak{c}$, $\mathfrak{b}$ expects $\mathfrak{c}$ to lose as much accuracy as $\mathfrak{c}$ expects $\mathfrak{b}$ to lose. More explicitly: suppose you start with credence function $\mathfrak{b}$ and you're asked how much worse off you think you'd be if you—without any reason—were to switch to credence function $\mathfrak{c}$. The answer, according to Selten, should be the same as in the reverse—when you have credence function $\mathfrak{c}$ and you're asked how much worse off you think you'd be were you to switch to credence function $\mathfrak{b}$. Selten sums up the idea as follows:

> [Consider] the hypothetical case that one and only one of two theories $\mathfrak{b}$ and $\mathfrak{c}$ is right, but it is not known which one. The expected score loss of the wrong theory is a measure of how far it is from the truth. It is only fair to require that this measure is "neutral" in the sense that it treats both theories equally. If $\mathfrak{b}$ is wrong and $\mathfrak{c}$ is right, then $\mathfrak{b}$ should be considered to be as far from the truth as $\mathfrak{c}$ in the opposite case that $\mathfrak{c}$ is wrong and $\mathfrak{b}$ is right.
>
> A scoring rule should not be prejudiced in favor of one of both theories in the contest between $\mathfrak{b}$ and $\mathfrak{c}$. The severity of the deviation between them should not be judged differently depending on which of them is true or false.
>
> A scoring rule which is not neutral is discriminating on the basis of the location of the theories in the space of all probability distributions over the alternatives. Theories in some parts of this space are treated more favorably than those in some other parts without any justification. Therefore, [NEUTRALITY] is a natural requirement to be imposed on a reasonable scoring rule. (p. 54 with minor changes)

[7]For objections to assuming PROPRIETY without argument, see Hájek (2008); Maher (2002); Pettigrew (2011).

The most important reply to this argument that I'll make is simply that the argument against the Brier score below shows that it can't work. We can, however, attack the motivation more directly as follows. Epistemic agents should tend to be risk averse. In general, we want to encourage moderation and discourage extreme opinion. Imagine, then, that we examined two credence functions in the propositions H and $\neg H$: $\mathfrak{b} = \langle .5, .5\rangle$ and $\mathfrak{c} = \langle .99, .01\rangle$. $\mathfrak{b}$ is more moderate than $\mathfrak{c}$. So, it's reasonable for a scoring rule to make it so that an agent with credence function $\mathfrak{b}$ could *lose* much more accuracy by switching to $\mathfrak{c}$ than an agent with credence function $\mathfrak{c}$ could lose by switching to $\mathfrak{b}$. $\mathfrak{b}$'s accuracy level is constant no matter what happens. However, an agent with $\mathfrak{c}$ takes a big risk. In order to discourage extremity without good reason, a reasonable scoring rule may give only a small additional reward to $\mathfrak{c}$ if it turns out that H, while severely punishing $\mathfrak{c}$ should it turn out that $\neg H$. In this case, $\mathfrak{b}$ expects to lose a lot more by switching to $\mathfrak{c}$ than $\mathfrak{c}$ does by switching to $\mathfrak{b}$.

2.2 Inaccuracy and Updating

2.2.1 Conditionalization

Leitgeb and Pettigrew have four different precisifications of ACCURACY. We'll be interested in how to update our *entire* credence function *across time*, so the version we want is:

ACCURACY (DIACHRONIC EXPECTED GLOBAL): Suppose an agent has learned evidence between t and t' that imposes constraints C on her belief function $\mathfrak{b}_{t'}$ at t', or on the set E of worlds that are epistemically possible for her at t', or both. Then, *at time* t', such an agent ought to have a belief function that satisfies constraints C and is minimal amongst belief functions thus constrained with respect to expected global inaccuracy by the lights of her belief function *at time* t, relative to a *legitimate* global inaccuracy measure [i.e., relative to the Brier score for Leitgeb

> and Pettigrew], and over the set of worlds that are epistemically possible for her *at time t'* given the constraints C. (From (2010a: 207) with minor changes)

Here's how ADEG leads to conditionalization: Suppose I start with credence function $\mathfrak{b}$ and learn for certain some new information E. I'm now epistemically obligated to pick some probability function in set C of probability functions that assign 1 to E. But which should I pick? If I follow ADEG, I ought to pick the member of C that minimizes expected inaccuracy by the lights of $\mathfrak{b}$. So, if $I(\cdot,\cdot)$ is my favored measure of inaccuracy, the best new credence function to adopt upon learning E according to $\mathfrak{b}$ is the credence function $\mathfrak{b}^* \in S$ such that:

$$\mathrm{E}_{\mathfrak{b}}(I, \mathfrak{b}', E) := \sum_{w \in E} \mathfrak{b}(\{w\}) I(\mathfrak{b}', w)$$

is minimal.

It turns out that for nearly any measure of inaccuracy one might think is reasonable (including the quadratic scoring rules), the function that minimizes expected inaccuracy under these constraints is simply $\mathfrak{b}(\cdot|E)$. In other words, I minimize expected inaccuracy upon learning new information if I update by conditionalization.[8]

2.2.2 Updating with Uncertain Evidence

I'll assume with Leitgeb and Pettigrew that not all updating should be by conditionalization. Sometimes, I get perceptual input that doesn't raise my credence in any proposition fully to 1. For instance, I might be in a pitch black room with a piece of paper in front of me that I believe to degree .3 is red. A small bit of light comes in under the door which allows me a better but not perfect view of the paper. Even though I didn't learn any new salient proposition for sure, my credence in red changes to .7, and I'm obligated to update

[8]For demonstrations, see Greaves and Wallace (2006); Oddie (1997); Leitgeb and Pettigrew (2010b).

the rest of my beliefs. For ease of reference, let's call a situation where we don't update any of our partial beliefs non-inferentially to 0 or 1 an *Uncertain Evidential Situation* (UES).

To set up the primary issue of the paper, we can consider only a special class of UES's, which I'll refer to as *Jeffrey-UES*'s. Let W be the set of possible worlds (with $|W| < \infty$), and let $\{E_i\}_{i \in I}$ be a partition of W with $0 < \mathfrak{b}(E_i) < 1$. We then get some new uncertain evidence that forces us (non-inferentially) to adopt new credence q_i in E_i (where $0 < q_i < 1$ and $\sum q_i = 1$). In other words, we partition W into a bunch of cells. The new evidence forces changes in some or all of those cells, and our problem is to decide how to update $\mathfrak{b}$ under that constraint.

The classic answer to this updating problem is that our new credence $\mathfrak{b}_{t'}$ should be given by:[9]

Jeffrey-Conditionalization: In Jeffrey-UES's, agents should update their credences through Jeffrey-Conditionalization. I.e., for any proposition A:

$$\mathfrak{b}_{t'}(A) = \sum q_i \mathfrak{b}_t(A|E_i)$$

Jeffrey-Conditionalization is the obvious way to extend standard conditionalization to cases with uncertain evidence, and it has long been more or less the only game in town when it comes to updating in Jeffrey-UES's.

To get the basic idea, first consider the case in which W is partitioned into E and $\neg E$. The new evidence forces your new credence in E to be q, so your new credence in $\neg E$ is $(1-q)$. You need to determine $\mathfrak{b}_{t'}(A)$. In the extreme case where $q = 1$, you would simply conditionalize on E, so $\mathfrak{b}_{t'}(A) = 1 \cdot \mathfrak{b}_t(A|E) + 0 \cdot \mathfrak{b}_t(A|\neg E)$. In non-extremal cases, your new credence in A is $\mathfrak{b}_{t'}(A) = q \cdot \mathfrak{b}_t(A|E) + (1-q) \cdot \mathfrak{b}_t(A|\neg E)$. So, you simply take the weighted average of what you would have thought about A had you learned that E

[9] For extensive discussion, see Jeffrey (1983).

and what you would have thought had you learned that $\neg E$ with the weights determined by your new credences in E and $\neg E$. The general case is analogous.

Because Jeffrey-Conditionalization seems so natural, it's surprising that *quadratic scoring rules don't lead to Jeffrey-Conditionalization*! Leitgeb and Pettigrew (2010b) show that under the Brier score, ADEG requires a different updating procedure entirely, which I'll argue has absurd consequences.[10]

As before, we want to update by minimizing expected inaccuracy under constraints. This time, however, we don't restrict the class of worlds under consideration from W to E. Instead, we just consider the same class of worlds W, but we have new constraints imposed, given by the q_i's. That is, we choose the credence function with minimal expected inaccuracy that assigns credence q_i to E_i for all i.

Here's how this updating procedure works: We first take some element E_j of the partition and then add (!) a constant d_j to $\mathfrak{b}_t(\{w\})$ for each world $w \in E_j$. In order to prevent negative credence, our new credence $\mathfrak{b}_{t'}(\{w\})$ is just $\max(\mathfrak{b}_t(\{w\}) + d_j, 0)$. By requiring that $\mathfrak{b}_{t'}(E_j) = q_j$, we end up with only one possible choice of d_j.[11] In other words, we end up with:[12]

LP-CONDITIONALIZATION: In Jeffrey-UES's, agents should update their credences through Leitgeb-Pettigrew-Conditionalization. That is, let d_i be the unique real number such that

$$\sum_{\{w \in E_i \mid \mathfrak{b}_t(\{w\}) + d_i > 0\}} \mathfrak{b}_t(\{w\}) + d_i = q_i.$$

---

[10]There will be different updating procedures for the different quadratic scoring rules, but each will have a similar form, and similar problems, to the one for the Brier score. To be clear, the objection isn't that quadratic scoring rules don't lead to Jeffrey-Conditionalization *per se*, but that they lead to extremely unattractive alternatives. It's only because of the general success of Jeffrey-Conditionalization that I'll hold it up as the procedure to beat below.

[11]For instance, in the special case where no worlds in E_j end up getting assigned credence 0, we have $d_j = \frac{q_j - \mathfrak{b}_t(E_j)}{|E_j|}$.

[12]From Leitgeb and Pettigrew (2010b: 254).

Then the agent ought to have belief function $\mathfrak{b}_{t'}$ at t' such that for $w \in E_i$:

$$\mathfrak{b}_{t'}(\{w\}) = \begin{cases} \mathfrak{b}_t(\{w\}) + d_i & \mathfrak{b}_t(\{w\}) + d_i > 0 \\ 0 & \text{Otherwise} \end{cases}$$

It's worth emphasizing that choosing this updating procedure is inevitable for adherents of ADEG and the Brier score. Let's review why exactly: An agent starts with some credence function $\mathfrak{b}_t$. She then finds herself in a Jeffrey-UES, and her new evidence compels her to have credence q_i in the propositions E_i. She's still left with uncountably many credence functions to choose from that meet these constraints, but she wants the one she expects at t is the most accurate of the ones available. If she uses the Brier score, she'll pick the one mandated by LP-Conditionalization.[13] After discussing the problematic cases, I'll argue below that the problems with LP-Conditionalization should lead to the rejection of the Brier score even as a measure of synchronic accuracy.

2.3 Problems

2.3.1 Initial Issues

Some apparent problems with LP-conditionalization are discussed and defended in Leitgeb and Pettigrew (2010b). These won't be my primary focus here, but it's worth going over them briefly:

The first is that the probability of some worlds can be lowered from a positive prior all the way to 0 in a UES. For instance, consider the following case:

[13] The proof is rather long, so I omit it here and refer the interested reader to Leitgeb and Pettigrew (2010b). The basic reason is as follows: ADEG combined with the Brier score leads us to pick the Euclidean-closest credence function to the original that meets the constraints. It's then not hard to show that we minimize Euclidean distance by adding the same constant (so long as we can) to each world in a given element of the partition.

Table 2.1:

	w_1	w_2	w_3
$\mathfrak{b}_t$	.3	.2	.5
$\mathfrak{b}^{LP}_{t'}$	.8	0	.2

$\mathfrak{b}^{LP}_{t'}$ is the result of raising the probability of $\{w_1\}$ to .8 and lowering that of $\{w_2, w_3\}$ to .2 under LP-conditionalization. Oddly, this results in assigning w_2 credence 0.[14]

A related issue is that LP-Conditionalization, unlike Jeffrey-Conditionalization, does not have standard conditionalization as a limiting case. In other words: Raising the probability of one element E of the partition to 1 and LP-Conditionalizing isn't the same as simply conditionalizing on E. To see this, consider the following case, in which the posterior of $\{w_1, w_2\}$ is raised to 1:

Table 2.2:

	w_1	w_2	w_3	w_4
$\mathfrak{b}_t$	.3	.2	.25	.25
$\mathfrak{b}_t(\cdot \mid \{w_1, w_2\})$	.6	.4	0	0
$\mathfrak{b}^{LP}_{t'}$	.55	.45	0	0

Unlike Leitgeb and Pettigrew (2010b), I do take both of these facts to be problematic, though I don't think they're necessarily fatal. They argue, perhaps rightly, that there's a stark difference between UES's and situations that call for full conditionalization. The latter involve ruling out worlds from the set of epistemic possibilities. The former in-

[14] It's worth working out this example to get a better sense of LP-Conditionalization. Since $\{w_1\}$ is a single-membered element of the partition, we'll just look at what happens to the elements of $E_2 := \{w_2, w_3\}$. We're given the constraint that $\mathfrak{b}^{LP}_{t'}(\{w_2\}) + \mathfrak{b}^{LP}_{t'}(\{w_3\}) = .2$. LP-conditionalization tells us to meet this constraint by finding some constant d_{E_2} such that $\max(\mathfrak{b}_t(\{w_2\}) + d_{E_2}, 0) + \max(\mathfrak{b}_t(\{w_3\}) + d_{E_2}, 0) = .2$. Now, since probabilities can't be negative, to meet the constraints we need $\mathfrak{b}^{LP}_{t'}(\{w_3\}) \leq .2$, so we have $d_{E_2} \leq -.3$. Since $.2 - .3 < 0$, we know that $\mathfrak{b}^{LP}_{t'}(\{w_2\}) = 0$ and $d_{E_2} = -.3$.

volve the same set of epistemic possibilities, just new constraints on the attitudes toward those possibilities. Thus, credence 0 in a proposition, even when there are only finitely many worlds, can represent two very different doxastic attitudes toward it. One of those attitudes rules the proposition out and precludes it from every being assigned positive probability in the future, whereas the other maintains its epistemic possibility, but assigns it maximally low credence. Though these issues call for further discussion, I'll bracket them here.

2.3.2 Main Problem

Regardless of what one thinks of these results, a more important problem—and a fatal one, in my view—is the potentially dramatic effect LP-Conditionalization can have on the likelihood ratios between different propositions. Since LP-Conditionalization *adds* a constant to the prior credence in a world, important evidential relationships reflected in the prior can then be violated. For illustration, we turn to the following two cases.

Case 1: Ghost Riding

There's a car behind an opaque door, which you're almost sure is blue, but which you know might be red. You're almost certain of materialism, but you admit that there's some minute possibility that ghosts exist. For ease of reference, we introduce the following abbreviations for propositions:

- B: The car is blue
- G: There are ghosts

To make the case precise, we stipulate you have the prior shown below:

Table 2.3:

	w_1 B ∧ G	w_2 B ∧ ¬G	w_3 ¬B ∧ G	w_4 ¬B ∧ ¬G
$\mathfrak{b}_t$	∼ .000476	.95	∼ .000025	.0495

Now the opaque door is opened, and the lighting is fairly good. You're quite surprised at your sensory input: Your new credence that the car is red is .99! Let's look at your posterior credence under the two updating procedures:

Table 2.4:

	w_1	w_2	w_3	w_4
$\mathfrak{b}^{LP}_{t'}$	0	.01	∼ .470262	∼ .519738
$\mathfrak{b}^{J}_{t'}$	∼ .000005	∼ .009995	∼ .000495	∼ .989505

Jeffrey-Conditionalization leads to no change in opinion about ghosts. Under LP-Conditionalization, however, seeing the car makes you about 47% sure there are ghosts. Note that the case was originally set up so that you thought the questions of what color the car is and whether there are ghosts were independent, but somehow, merely acquiring information about car color has drastically changed your opinion about materialism. Had you come to know the car was red and conditionalized on ¬B, you would have ended up with credence .0005 that there were ghosts. So, the difference between the near certainty of credence .99 and full knowledge that ¬B is the difference between credence .47 and credence .0005 that G. I think it's clear that something's gone wrong with LP-Conditionalization here. Becoming more confident in one proposition shouldn't alone raise your credence in another if you initially take them to be independent.

Case 2: Unemployment

Consider the following propositions:

- U: Unemployment will rise before the next election.
- P: The president will be re-elected.

Suppose your prior and posterior credences after LP-conditionalization are given in the chart below:

Table 2.5:

	w_1 U ∧ P	w_2 U ∧ ¬P	w_3 ¬U ∧ P	w_4 ¬U ∧ ¬P
$\mathfrak{b}_t$	.05	.15	.38	.42
$\mathfrak{b}^{LP}_{t'}$	.44	.54	0	.02

$\mathfrak{b}_{t'}$ is the result of becoming .98 sure that unemployment will rise from an initial credence of .2. In this example, we have $\mathfrak{b}_t(\mathsf{P}) = .43$, while $\mathfrak{b}_t(\mathsf{P}|\mathsf{U}) = .25 < .43$. However, after becoming more confident that unemployment will rise, LP-Conditionalization leads to $\mathfrak{b}_{t'}(\mathsf{P}) = .44 > .43$. So, despite the fact that you thought that it was less likely that the president would be re-elected given that unemployment rises, becoming nearly sure that unemployment will rise ends up *raising* your credence that the president will be re-elected.

The general problem in both cases is that LP-Conditionalization does not respect important evidential relationships reflected in the agent's prior. Furthermore, there's nothing particularly odd or unusual about the structure of either case. Therefore, LP-Conditionalizers will often end up with unreasonable posteriors despite perfectly sensible priors. Below, we consider two potential escapes for defenders of the Brier score, which I'll argue don't work.

2.4 Rigidity to the Rescue?

For those who accept both ADEG and the Brier score, one way to avoid the unwelcome consequences above is to build in a requirement that certain structural relationships from the prior credence function be preserved. For instance, one might suggest something like the following as an additional constraint:

RIGIDITY: Suppose $\{E_1,\ldots,E_n\}$ is a partition of W, $0 \leq q_1,\ldots,q_n$ and $\sum q_i = 1$. If the agent acquires uncertain evidence that requires $\mathfrak{b}_{t'}(E_i) = q_i$ for all i, then we have that for all $A \subseteq W$, $\mathfrak{b}_{t'}(A|E_i) = \mathfrak{b}_t(A|E_i)$.

With this added requirement, quadratic scoring rules do lead to standard Jeffrey-Conditionalization and indeed do so trivially, since:

$$\mathfrak{b}_{\text{New}}(A) = \sum \mathfrak{b}_{\text{New}}(E_i) \cdot \mathfrak{b}_{\text{New}}(A|E_i) = \sum \mathfrak{b}_{\text{New}}(E_i) \cdot \mathfrak{b}_{\text{Old}}(A|E_i)$$

However, I don't think this move is attractive. Though preservation of these particular conditional probabilities may generally be desirable, it should result from the choice of accuracy measure, not from any added constraint. Why? First, such a move looks like an *ad hoc* fix unless more motivation can be provided. RIGIDITY should be earned through earnest toil. Second, being accurate is more important than maintaining initial opinions about any conditional relationships. Structural requirements on a credence function should *emerge from* evidential and alethic requirements. Put differently: Our quest as epistemic agents is for the truth, which we pursue by means of obeying evidential requirements. If we expect to sacrifice some accuracy in order to maintain nice structural relationships between propositions, then we're expecting to go against our primary ends as epistemic agents. If I'm faced with a choice between two credence functions, one of which preserves my prior probabilities conditional on elements of the partition, but the other of which I think is more accurate, I should prefer the latter. It's then unreasonable to add the constraint that the new probability function not be the one with the

highest expected accuracy that's compatible with the new information *simpliciter*, but instead be the one with the highest expected accuracy that's *both* compatible with the new information *and* that preserves some other features of the prior.

Here's another way to understand the problem. Take a non-extremal probability functions $\mathfrak{b}$ that makes propositions A and E independent. I.e., $\mathfrak{b}(A|E) = \mathfrak{b}(A)$ and $0 < \mathfrak{b}(A), \mathfrak{b}(E) < 1$. There will always be some other credence function $\mathfrak{b}'$ such that $\mathfrak{b}'(A|E) \neq \mathfrak{b}'(A)$ and $\mathfrak{b}'$ is more accurate than $\mathfrak{b}$ in the actual world. Since $\mathfrak{b}'$ is actually more accurate, an agent with credence function $\mathfrak{b}'$ is better off than an agent with credence function $\mathfrak{b}$. Now, imagine that we adopted a policy that imposed two requirements on an agent when she updates in Jeffrey-UES's: (1) She adopt credence q_i in E_i for all i, and (2) She also maintain RIGIDITY. She's told to minimize her expected inaccuracy over the credence functions that remain. In this case, the addition of Requirement (2) will sometimes eliminate a credence function she would otherwise have chosen. Since we want agents to be as epistemically well-off as possible, we ought only impose RIGIDITY if it is expected to decrease an agent's overall inaccuracy. But why should it? After all, an agent who uses the Brier Score expects—by her own lights—to do better by minimizing expected inaccuracy under only Requirement (1) than she does by minimizing expected inaccuracy under both Requirement (1) and RIGIDITY. Assuming the agent has a reasonable credence function at t, her expectation is justified. So, it's hard to see how a proponent of the Brier Score could argue that the addition of the RIGIDITY requirement will lead agents to do greater accuracy.

2.5 The Brier Score and Synchronic Accuracy

The argument against the Brier score as a measure of inaccuracy comes from its failings in recommending an updating procedure in Jeffrey-UES's. Now, a number of authors endorse ADEG or an equivalent norm and use it to argue for constraints on updating

procedures.[15] The only options they have are rejecting the Brier score, accepting LP-Conditionalization, or endorsing different synchronic and diachronic accuracy measures. The last option appears *ad hoc* at best. It amounts to claiming that the kind of accuracy I care about my future beliefs having is different from the kind of accuracy I care about my current beliefs having. Though I won't rule this possibility out, it would require a lot of further motivation. Since we've seen that LP-conditionalization is unattractive, supporters of ADEG ought to take the first option of rejecting the Brier score.

Many defenders of the Brier score, however, have been silent about any accuracy norms for updating beliefs and have considered only its synchronic features. Indeed, there is something *prima facie* odd about ADEG, since it requires evaluation of potential posterior credence functions through the use of a prior credence function that we know is out-dated. That is, it requires us to use the prior $\mathfrak{b}_t$ that is—*ex hypothesi*—an inappropriate response to the total evidence at t'. To make the case that the considerations above should lead everyone to reject the Brier score, I'll try to strengthen the argument that a norm like ADEG is at least descriptively adequate.

Let's go back to Case 1. Instead of jumping into the diachronic case, this time we have three agents with the following *prior* credence functions.

Table 2.6:

	w_1 $B \wedge G$	w_2 $B \wedge \neg G$	w_3 $\neg B \wedge G$	w_4 $\neg B \wedge \neg G$
Alice	$\sim .000476$	.95	$\sim .000025$	.0495
Leopold	0	.01	$\sim .470262$	$\sim .519738$
Jeff	$\sim .000005$	$\sim .009995$	$\sim .000495$	$\sim .989505$

That is, Alice's prior is $\mathfrak{b}_t$ in the original case, Leopold's is $\mathfrak{b}^{LP}_{t'}$, and Jeff's is $\mathfrak{b}^{J}_{t'}$.

[15] Aside from Leitgeb and Pettigrew (2010a,b), see Greaves and Wallace (2006), Kierland and Monton (2005), and Oddie (1997).

Suppose Alice likes the Brier score as a synchronic measure of accuracy, but she doesn't endorse ADEG and instead updates by Jeffrey-Conditionalization. She's now in an odd situation. She *expects* Leopold to be more accurate than Jeff overall. However, were she to have the same constraints placed on her credence function that we find in Case 1, she'd adopt Jeff's current credence function as her own.

Here's one way to think about this. Using ACCURACY and the Brier score, Alice can create a preference ranking for all possible credence functions in terms of how well she expects them to do.[16] She of course expects her own credence function to have a higher level of accuracy than any other, so it's first on her preference list. Still, she doesn't think all other credence functions are on a par. In particular, she thinks that Jeff's is worse than Leopold's. However, she knows now (or at least can be shown) that were nature to place constraints on her credence function that required her to have credence .99 in $\neg \mathsf{B}$, she would adopt Jeff's credence over Leopold's. We can dramatize the case as follows: Suppose an evil scientist told Alice that he was going to force her to take one of two pills. Pill 1 would change her prior to Leopold's, and Pill 2 would change it to Jeff's. Her memory of this event would be wiped clean either way, and she'll get no new evidence, uncertain or otherwise, that should affect her credence in either B or G. In this situation, she would opt for Pill 1. If, however, instead of an evil scientist, she has a sensory experience that a-rationally raises her credence in $\neg \mathsf{B}$ to .99, she'll opt for Jeff's credence.

I think that this oddity should worry even someone who denies that an agent should literally use her prior probability function to justify her posterior. Such a philosopher might object to ADEG or any diachronic norm as follows: At t, you have total evidence E, which you should use to generate a rational credence function $\mathfrak{b}_t$. At t', you have evidence E'. Throw out your earlier $\mathfrak{b}_t$. It's outdated, and you now shouldn't care about

[16] I assume that all participants in the discussion accept something like the simple ACCURACY norm.

what you thought in the past. Instead, just look at your total evidence at t' and figure out how best to respond to it. It can turn out that $\mathfrak{b}_t$ and $\mathfrak{b}_{t'}$—if they're good—will exhibit certain formal relationships. It might even look as if you've updated by some form of conditionalization, and it might be useful to do so to save computational time. But that's just a spandrel from a normative perspective. Your updated credence function is justified by the evidence you have at t' not by what you thought at t.

Even if we concede that this is the right normative picture, we can maintain an 'as if' picture of updating in accord with ADEG and a good inaccuracy measure. Here's why: Suppose Alice doesn't really update, but instead just follows some rational policy for evaluating evidence, which we'll call P. P is a function from potential sets of total evidence to credence functions. That is, whenever you get some body of total evidence, P tells you what your credence function should be without regard for what you thought in the past. The reply then goes as follows: If P is actually a good policy, then it will generally tend to give you accurate credence functions. Of course, sometimes it will be off, and sometimes you'll be in an uncooperative world that feeds you misleading evidence all the time, but in typical worlds with typical evidence, a good policy should be fairly successful. Now suppose the Brier score is the right way to measure inaccuracy. So, if P is generally good, it will let Alice do a good job at coming up with an accuracy ranking of credence functions in terms of their expected Brier scores. However, if P makes it look as if Alice updates by a rule seriously different from LP-Conditionalization and the Brier score really is the right way to measure inaccuracy, then Alice will either end up with bad posterior credence functions or her prior rankings of expected accuracy will be systematically and predictably off. That is, either P could do better at figuring out a good posterior or P could generate prior credence functions that do better at evaluating other credence functions. Since P could do better either way, it must not be an ideal policy. Of course, this argument is rough-and-ready and falls short of showing that P is dominated, but I think it nonetheless puts an advocate of the Brier score in an uncomfortable

position regardless of her views on updating.

Therefore, a norm like ADEG should be at least descriptively accurate at least in normal circumstances where the agent's prior is reasonable. Consequently, we should reject the Brier score as a measure of synchronic accuracy, so long as better alternatives are available. It turns out other, less popular measures of accuracy do result in a superior updating procedure, so we should opt for one of them. In the Appendix, we show that the logarithmic rule, in particular, leads to Jeffrey-Conditionalization.

2.6 Additional Support for the Logarithmic Rule

The *Logarithmic Score* of a credence function $\mathfrak{b}$ is given as follows:

LOGARITHMIC RULE (LR): $I(\mathfrak{b}, w) = -\ln(\mathfrak{b}(\{w\}))$.

Aside from leading to leading to Jeffrey-Conditionalization, the Logarithmic Rule has some additional features that I find attractive. The considerations below are only suggestive and not intended as compelling. Epistemologists who aren't as drawn to LR should still, however, benefit from the discussion as it reveals important properties that distinguish LR from other rules.

2.6.1 The Logarithmic Rule and Sensitivity

One feature mentioned briefly above is that LR grows rapidly more sensitive to small absolute changes in credence the closer it is to 1 or 0. Suppose an agent starts off with credence $1/10$ in the actual world w at t_0. At t_1, she becomes less confident that w is the actual world and adopts credence $1/100$. At t_2, her credence falls even further to $1/1000$. According to LR, the agent loses the same amount of accuracy between t_0 and t_1 as she does between t_1 and t_2 even though $\mathfrak{b}_0(\{w\}) - \mathfrak{b}_1(\{w\}) = 10 \cdot (\mathfrak{b}_1(\{w\}) - \mathfrak{b}_2(\{w\}))$. Furthermore, if the agent were to rule out w completely, her credence of 0 would make her

maximally inaccurate (i.e., give her an inaccuracy score of ∞) regardless of the rest of her credence function.[17]

While Selten (1998) considers these properties bugs, I take them to be features of LR. In general, we want to encourage epistemic modesty. Entirely ruling out a true proposition is an unforgivable epistemic sin, especially if we endorse conditionalization. Once a proposition is ruled out, it is ruled out forever.[18] LR adopts a flinty and unyielding attitude toward agents who mistakenly and avoidably blind themselves to the true state of the world.

In the less extreme cases, the rapid change in score is similarly justified. While the difference between credence .99 and .999 appears to be small, an agent with the latter credence is *ten times* more confident than the former. Extreme confidence is only called for with extreme evidence, and it should take a lot of evidence to make somebody ten times more confident in a proposition. Intuitively, most of the time, raising one's credence in a proposition A from .5 to .509 requires much less evidence than raising one's credence in A from .99 to .999 does. LR meshes well with this sort of policy toward evidence, since it severely punishes agents who become especially confident in false propositions.

2.6.2 The Logarithmic Rule and Practical Decision Theory

Imagine that Alice has to choose between two gambles:

Gamble A Receive \$1 if Coin 1 lands heads (H_1) and \$0 otherwise.

Gamble B Receive \$2 if Coin 2 lands heads (H_2) and \$0 otherwise.

---

[17] Note that this is not the case with the Brier score. An agent achieves maximum inaccuracy only when she adopts credence 1 in the wrong world. An agent who has credence 0 in the actual world but who doesn't have credence 1 in any of the non-actual worlds won't get the worst score.

[18] As Dennis Lindley (1991) puts it: "Leave a little probability for the moon being made of green cheese; it can be as small as 1 in a million, but have it there since otherwise an army of astronauts returning with samples of the said cheese will leave you unmoved" (p. 104).

Should Alice pick Gamble A or Gamble B? The answer depends on her credences in H_1 and H_2. Letting $\mathfrak{b}$ represent her credence function and assuming Alice's credence function is linear with dollars, we see that: Alice should take Gamble A if $\mathfrak{b}(H_2) < \frac{1}{2}\mathfrak{b}(H_1)$, she should take Gamble B if $\mathfrak{b}(H_2) > \frac{1}{2}\mathfrak{b}(H_1)$, and she is rationally permitted to take either if $\mathfrak{b}(H_2) = \frac{1}{2}\mathfrak{b}(H_1)$. The rational decision, of course, is determined not by the absolute difference $|\mathfrak{b}(H_1) - \mathfrak{b}(H_2)|$ *but by the ratio* $\frac{\mathfrak{b}(H_1)}{\mathfrak{b}(H_2)}$. If Alice's credence in $H_1 = 1$, then she'll choose H_2 if $\mathfrak{b}(H_2) > 1/2$. If her credence in $H_1 = .1$, then she'll choose H_2 if $\mathfrak{b}(H_2) > .05$.

While scoring rules are sometimes taken to serve as epistemic utility functions without regard to practical import, we still want them to fit well with (practical) bayesian decision theory. Assuming you're an expected utility maximizer, how well you do in life depends entirely upon the accuracy of your credences. Since ratios are driving your decisions, we should in turn measure how well off you are epistemically based on ratios and not absolute distances. Whenever you become half as confident in a true proposition, your inaccuracy increases by $\ln 2$, which is a constant. Therefore, LR is an appropriate choice for those concerned with integrating practical and epistemic decision theory.

2.7 Conclusion

The argument of the paper went as follows. First, we should strive to minimize our expected inaccuracy under some reasonable measure, by ACCURACY. If we use a quadratic scoring rule and follow ADEG, we end up with LP-Conditionalization. We should keep ADEG (at least usually), but we shouldn't update by LP-Conditionalization. Therefore, quadratic scoring rules aren't reasonable inaccuracy measures.

2.A The Logarithmic Rule and Jeffrey-Conditionalization

We here sketch a proof of the earlier claim that ADEG combined with the following inaccuracy measure leads to Jeffrey-Conditionalization.

First, it will be useful to have the following redescription of Jeffrey-Conditionalization. Suppose Alice's prior is $\mathfrak{b}_t$ and let $\{E_i\}_{i\in I}$ partition the set of possible worlds W. News comes in, and Alice must update her credence function so that she now has credence q_i in each element E_i of the partition. If she Jeffrey-Conditionalizes, she updates her credence to $\mathfrak{b}^J_{t'}$ as follows: for each element E_i she finds some constant c_i such that for all $w \in E_i$, $\mathfrak{b}^J_{t'}(\{w\}) = c_i \cdot \mathfrak{b}_t(\{w\})$. Since $\sum_{w\in E_i} \mathfrak{b}^J_{t'}(\{w\}) = q_i$, exactly one constant will work. Therefore, it will suffice to show that LR together with ADEG require multiplying the prior credence in each member of a given element of the partition by the same constant.

Suppose now that Alice follows ADEG and measures inaccuracy with LR. For the sake of simplicity, we first consider the case in which $W = \{w_1, w_2, , w_3, w_4\}$ with $E_1 = \{w_1, w_2\}$ and $E_2 = \{w_3, w_4\}$. For readability, we set $\mathfrak{b}_t(\{w_i\}) = \alpha_i$. To follow ADEG, we're now looking for the quadruple $\langle \beta^*_1, \beta^*_2, \beta^*_3, \beta^*_4\rangle$ that minimizes:

$$-\sum_{i=1}^{4} \alpha_i \ln \beta_i \tag{2.1}$$

where $\beta_1 + \beta_2 = q_1$, $\beta_3 + \beta_4 = q_2$, and $0 \leq \beta_i$ for all i. Since E_1 and E_2 are disjoint, minimizing (2.1) under these side-constraints is just minimizing both of the following:

$$f(\beta_1) = -\alpha_1 \ln \beta_1 - \alpha_2 \ln(q_1 - \beta_1)$$
$$g(\beta_3) = -\alpha_3 \ln \beta_3 - \alpha_4 \ln(q_2 - \beta_3)$$

To minimize f, we let:

$$f'(\beta_1) = \frac{\alpha_2}{q_1 - \beta_1} - \frac{\alpha_1}{\beta_1} = 0 \tag{2.2}$$

Note that $\beta_1 = \alpha_1 \cdot c$ and $\beta_2 = \alpha_2 \cdot c^*$ for some constants c and c^*. Substituting these identities in (2.2) gets us: $\frac{1}{c} - \frac{1}{c^*} = 0$. So, $c = c^*$. To minimize g, we follow the same procedure *mutatis mutandis*. We then have that $\langle \beta_1^*, \beta_2^*, \beta_3^*, \beta_4^* \rangle = \langle c\alpha_1, c\alpha_2, c'\alpha_3, c'\alpha_4 \rangle$ for constants c and c'. In other words, each element of E_i gets multiplied by the same constant c_i as desired. Thus, in this special case in which each $|E_i| \leq 2$ for all i, we end up with Jeffrey-Conditionalization.

For the more general case, suppose $|E_1| = n$. We then wish to minimize $-\sum_{i=1}^{n} \alpha_i \ln \beta_i$ under the side-constraints. To do so, pick two distinct β_i's to treat as variable and treat the rest as constant. Without loss of generality, we choose β_1 and β_2, and set $\sum_{i=3}^{n} \beta_i = K \leq q_1$ for some constant K. Now, we just need to minimize: $h(\beta_1) = -\alpha_1 \ln \beta_1 - \alpha_2 \ln(q_1 - K - \beta_1)$. Running essentially the same argument as above, we find that $\beta_1^* = c\alpha_1$ and $\beta_2^* = c\alpha_2$ for some constant c. Since the choice of K and the two β_i's was arbitrary, we again end up with Jeffrey-Conditionalization as desired.

Chapter 3

With All Due Respect: The Macro-Epistemology of Disagreement

3.0 Introduction

Suppose you (A) and I (B) are weather forecasters. Very early each morning, we look at a bunch of data and come up with private predictions for whether it will rain later that day. Sometimes before I go on television, I learn what you think. Based on my estimation of your skill as a weatherperson, how should I revise my own forecast upon hearing yours?

We'll be using this template as a backdrop for the problem of disagreement. As in other cases of philosophical interest, the agents have overlapping—and sometimes even the same—evidence that they've thought through, but they nonetheless might arrive at substantially different verdicts.[1] The goal is to help give an agent advice for what to do based on her antecedent assessment of her advisor's ability when it turns out they have

[1]Indeed, if we believe Silver (2012: Ch. 4), weather forecasting is a surprisingly apt case for modeling the problem of peer disagreement. Many clever weatherpersons use computer models with breath-taking levels of power and sophistication, but there aren't clear or agreed upon algorithms for how to use the computer outputs to arrive at forecasts. The human agents, then, have various degrees of shared background evidence that they're not entirely certain how to handle and often end up disagreeing.

distinct levels of confidence in some proposition p.

Commonly, authors take what I'll call a *micro-approach*. On our weather model, they tell me how my antecedent expectations about you should affect my prediction for that token disagreement on some particular day i after I learn your view. A simple but orthodox version of the equal weight view, for instance, says that if I expect you *ex ante* to be as good as I am at predicting whether it will rain that day, then upon learning your forecast, I should give you equal weight. In particular, because I take us to be peers with respect to that proposition, symmetry demands that I split the difference and take the average of our credences. Thus, on this view, we have actionable advice for what my posterior credence in rain should be *today* as a function of a few simple input parameters.

Instead of looking at the problem from this angle, I'll take a *macro-approach* to the question of how to handle disagreement. Our task will be to characterize the trends that should emerge *in the long-run* or *in expectation* as far as my updating behavior is concerned. Rather than give rules for how to update my forecast on day i, we'll look at what should happen on average. Given an assessment of an advisor's expected level of accuracy, there are broad constraints on how to update once we learn her opinion. In any particular case, these constraints don't amount to much. Nearly anything is compatible with any non-extremal level of antecedent expected accuracy assigned to an advisor. Over many cases, however, the constraints become much stronger. If expected and actual behavior don't converge after enough trials, something's gone wrong somewhere. Either we'll have to rethink our process of *ex ante* evaluation, or we'll have to start updating differently when we learn other people's views, or both.

As we'll see, there are payoffs both in terms of results and methodology. After some initial discussion of the latter, we give necessary and sufficient conditions for expecting to give equal weight, *viz*., you expect to give two parties equal weight just in case you expect them both to be equally accurate. Thus, if we take one common understanding of *peer*—understood as an agent with *ex ante* equal expected reliability—a version of the

equal weight view falls out.[2] This strong relationship between accuracy and weight is, however, invisible at the micro-level and cannot be maintained as anything close to an exceptionless rule. That holds for two reasons: first, the actual amount of weight given often has to depend on the actual credences reported, and second, *ex ante* evaluations are often based upon expectations about what an advisor thinks. The relationship between expected accuracy and actual weight, then, can only be taken as a wide-scope macroscopic norm.[3,4]

Second, we'll show that the notion of symmetry motivating equal weight views has been importantly misconstrued. On orthodox versions of EWV, an agent is advised to split the difference, i.e., take the straight average of her credence x and her advisor's credence y more or less regardless of the values of x and y. Indeed, this understanding of equal weight is so common that it's often *identified* with the view.[5] Given the supposed parity between peers, splitting the difference seems unavoidable. However, both defenders and detractors of EWV worry that such a policy will lead to *spinelessness*.[6] That is, if you go around splitting the difference a lot, you'll end up overly agnostic on too many propositions. Nearly all your credences will be around the middle of the $[0, 1]$ interval, especially in areas rife with disagreement.

It turns out splitting the difference *is* spineless. When both agents are relatively re-

---

[2]Authors like Elga (2007); Enoch (2010); White (2009) *inter alia* adopt (roughly) this understanding of *peer*.

[3]It's macro-norm because the constraints are emergent over many trials. It's wide-scope (in the sense of Kolodny (2005)) because one could either change her evaluative practices or her updating practices or both should she find herself in violation.

[4]If, on the other hand, we adopt a less directly alethic notion of *peer*—understood in terms of some epistemic virtue like intelligence or general rationality—we'll see what's required for expected equal weight, *viz.*, some argument that equality of that virtue should require equal expected accuracy. For this type of understanding of *peer*, see Gutting (1982); Kelly (2005, 2010).

[5]Lehrer and Wagner (1981) defend splitting the difference, and Elga (2007), Kelly (2010), and Enoch (2010) claim EWV generally requires difference-splitting but allow some wiggle-room. Christensen (2009, 2011) is careful to point out that his view is more flexible, but he thinks difference-splitting is likewise often called for. For technical problems with difference-splitting of a different sort than we'll explore below, see Jehle and Fitelson (2009); Loewer and Laddaga (1985).

[6]See Elga (2007); Enoch (2010); Kelly (2005, 2010).

liable, we show the level of weight given to one party will tend to *grow* with boldness. That is, on average (though again not necessarily in any particular case), if you take both yourself and your advisor to be reliable, then if one of you has a credence closer to 0 or 1, you ought to give that person more weight. Because splitting the difference (or any method of linear averaging) keeps the relative weights fixed regardless of the extremity of the actual credences reported, it usually gives undue weight to the more agnostic of the two disputants. As a result, splitting the difference will produce overly timid attitudes and thus falls prey to the spinelessness objection.

Here's the plan. Throughout the body of the paper, we'll proceed in a largely informal manner with more technical discussion relegated to the appendices. §3.1 covers the basics of the macro-approach. §3.2 first explicates the claim that equal expected inaccuracy is equivalent to equal expected weight and explores the consequences. §3.3 informally lays out the import of Theorems (3.A.3), (3.A.5), and (3.A.6) and examines the various flaws of the split-the-difference view. §3.4 looks at the upshots of earlier discussion for updating in areas with rampant disagreement and wraps up. Appendix 3.A states and provides proofs of the background results. Appendix 3.B explores some formal relationships between expert and peer disagreement.

3.1 Expectations and Calibration

As an epistemic agent, your ultimate goal is to be accurate. Given your cognitive powers and resources—such as computational ability, memory, reasoning power, and so forth—you ought to develop the best policies for matching your doxastic state to the world. A big advantage of the macro-approach is that over time, certain patterns should emerge that are invisible at the micro-level. Any given case can diverge more or less arbitrarily far from expectation. Over time, however, your expectations and reality should tend to converge. One particular method of *diagnosing* flaws in your behavior that's only

applicable in the macro comes through the following criterion:

CALIBRATION Good epistemic agents are long-run approximately calibrated.

To get the idea behind CALIBRATION, let's go back to the weather example at the beginning. Look at your history as a weather forecaster over your career and consider how many of the days on which you've forecast, say, a 60% chance of rain it actually did rain. Supposing you've been at this gig for a while, the answer should be right around 60%.

Why is CALIBRATION an evaluative criterion worth caring about? Informally, the answer is that *ceteris paribus* being calibrated improves your overall accuracy. Let $\mathscr{X}$ be the set of propositions you have credence .6 in, and suppose that as a matter of fact only 50% of the propositions in $\mathscr{X}$ are true. From a pragmatic perspective, if I thought I could predict that only 50% of the propositions in $\mathscr{X}$ were true, I could run a kind of stochastic dutch-book. That is, I can, over time, pump you for money by selling you bets for \$1 on each proposition in $\mathscr{X}$ for \$.60 a piece. In more purely epistemic terms, I could exploit you for accuracy. Even if I had no idea how to read a Doppler forecast, if I announce a 50% chance of rain every time you forecast a 60% chance of rain, I'd end up more accurate.

Now, we need to make a big caveat before proceeding. CALIBRATION (along with the more general requirements that expectation and reality eventually converge) is not an epistemic *goal*. Your only goal as an agent is to be accurate, and an agent can use calibration considerations as a means toward greater accuracy. Treating calibration as an epistemic end-in-itself, however, results in obviously bad epistemic behavior. If you adopted credence .5 in every proposition, you'd be guaranteed to be perfectly calibrated, but at the expense of accuracy. After all, you'd do better if you had credence 1 that $2+2=4$. As your credences grow bolder, calibration is harder to maintain. If you don't know much about weather prediction, but you always forecast either a 90% or a 10% chance of rain, you'll eventually end up far from well-calibrated.

Your overall accuracy is determined by how close your credences are, on average, to truth-values, so high average accuracy requires both boldness and near-calibration. If you can increase boldness without losing much calibration, the epistemic risk you take will paid off, and you'll improve your average accuracy.[7]

Good policies result in calibrated credence functions. When you realize you're uncalibrated, you can use that information as a signal that something's gone wrong with your epistemic behavior and search for ways to improve it. That is, you know that if you were behaving correctly, you would likely be calibrated. When you discover you're not close to calibrated after a sufficient amount of time, you then recognize the need for some kind of change in policy: Being bold and uncalibrated is a sign of over-confidence, while being timid and uncalibrated is a sign of under-confidence.

Now that we've cleared up the role of CALIBRATION, let's see how to extend its application. In addition to evaluating credences alone, we can also use CALIBRATION to check whether we're forming estimates well generally. Suppose you're a veteran logic instructor. You try to set your exams so that the average score is always around 70% without any adjustment *post hoc* for a curve. Nothing necessarily is off if your students do worse one time, but you're clearly doing something wrong if they on average get a score of 40% year after year.

When we take a macro-approach to disagreement, our updating behavior should converge in the long run to our expectations. Imagine you expect an advisor B to be a really good forecaster of election outcomes. Each election, you ask for her credence that some candidate will win, and she always give you a sincere reply. Maybe she turns out to be way off in election 1: she had a very high level of confidence that one candidate would win, but she ended up wrong. That's okay. She might still be reliable generally. However,

[7]For further problems with calibration as a goal, see Joyce (1998: pp. 593–6) and Seidenfeld (1985). For more detail on the relationship between accuracy, boldness, and calibration, see the discussion DeGroot and Fienberg (1982, 1983), Schervish (1989), and §4 of Winkler (1996).

if she is consistently way off, but you continue to give her a very high antecedent level of expected accuracy, you're probably doing something wrong yourself. Presumably, you should change your *ex ante* assessment of her and downgrade your expectations for how well she'll do in the future. There are no guarantees, of course. But over the long run, consistent poor performance is a good sign of epistemic error.

The same goes for your own updating behavior. You may well expect to give an advisor equal weight but find out she has an out-of-left-field credence and decide not to pay her much heed. That's fine. Regardless of how much weight you expected to give somebody *ex ante*, you shouldn't listen to people who tell you Obama is secretly a space alien. However, as with estimates generally, CALIBRATION requires long run convergence.

Note that CALIBRATION is evaluatively useless in the micro. Suppose a hapless weather forecaster Jim makes only one prediction—say a 60% chance of rain on March 15, 2013—and then dies. Jim is necessarily uncalibrated, since either 100% or 0% of the days he forecast a 60% chance of rain, it in fact rained. However, he clearly didn't do anything wrong just because he made a non-extremal prediction that day. Our appeal to CALIBRATION only can gain traction on disagreement when the sample-size is sufficiently large. Thus, by taking a macro-approach, we gain a genuinely new angle on the original problem.

Before looking at some additional liberation granted by the macro-perspective, let's quickly discuss two potential worries one may have. First, the world may not cooperate. You could be behaving in a rationally optimal manner, but even so, you don't end up close to calibrated. That's possible, but it's only an in-principle worry. The bigger the sample size becomes, the more likely it is that you're to blame. Just as it's possible that a fair coin can come up heads arbitrarily many times in a row, it becomes more and more certain that the frequency will eventually converge around $1/2$.

You might also worry about what counts as "the long run". After all, as Keynes famously remarked, "In the long run, we are all dead" (1924: p. 80). As with some equilibrium theorems of economics, theorems in macro-epistemology aren't guaranteed to be of much use. Here, however, I think we can point out an empirical fact. Disagreement is really quite common even with people we have a lot of antecedent respect for. You, like me, will probably have no shortage of sample size.

Now for some details of the new approach.

3.1.1 The t_0-Perspective

Because we're free to depart from our expectations in any given case, we can step back and look at updating policies more generally. Instead of only considering disagreements that arise once the agent has already already gotten a bunch of evidence and arrived at a credence, the macro-approach allows us to consider *strategies* or *policies* for handling potential future disagreements.

To see how this idea works schematically, let's return to the weather forecasters. On day 0, I don't have much evidence concerning whether it will rain on day 1, at least not relative to my later self. I know that between now and then, I'll get a bunch of Doppler radar images and other data that will probably help me get a more accurate credence in whether it will rain (r_1). I also know that you will have access to some data (possibly the same) and will also have a credence at t_1 about whether r_1 is true. From my perspective at t_0, I can consider what should happen if my better informed future time-slice disagrees with your future self at t_1.

Because CALIBRATION will be doing a lot of evaluative work, let's spell this out. Suppose it's t_0 and you take B to be a peer of your t_1-self. By the time t_1 rolls around, pretty much anything could have happened. On many occasions, you'll obtain unexpected evidence either about r_1 or about B that will legitimately lead to dissonance between your two time-slices. That's as it should be. When the evidence changes, you should change

your mind. Nonetheless, over the long run, things should average out.

For instance, suppose every morning you and your friend decide to gather evidence—possibly the same, possibly not—for some proposition of your choosing. After deciding on the proposition of the day, you think about how accurate you expect your friend to be. At the end of the day, you make a second evaluation before learning your friend's credence, and then you update one final time based just on what her credence turns out to be.

Suppose your actual behavior and your expected behavior don't converge. We now know something went wrong. You could be forming bad evaluations in the morning or at the end of the day. You could be bad at predicting the sort of evidence you'll end up getting. You might be bad at predicting what your friend will end up thinking. Regardless, some feature(s) of your doxastic habits should change. Since you're not well-calibrated, you'll have some guidance as to where to look for improvement. Compare: If I believe both p and $\neg p$, I know I've messed up somewhere, but I don't yet have enough information to decide which of the two beliefs I should drop. Nonetheless, realizing this fact is helpful for improving my epistemic state.

Note we can let $t_0 = t_1$ and treat present disagreements as a special case. In such circumstances, you've already gathered the salient evidence, and you know your own credence going into the disagreement. In the weather example, that corresponds to the time after having collected data like Doppler images but before finding out what the other forecaster thinks about rain. When you've already reasoned through your evidence and have credence b in p, you're simply a trivial advisor to yourself: one about whom the information is known and to whom you defer.

Indeed, there's also an important second level of generality we gain. Instead of thinking about an advisor's (or your own) expected epistemic performance with respect to some *fixed* proposition p, we can consider some domain $\mathscr{D}$ of propositions if we like. Again, let's take the weather case. You might have some expectations about your and

your advisor's level of accuracy for predicting whether it will rain tomorrow (r_{i+1}), and the macro-approach will have something to say about your expected updating behavior. However, you might instead generalize and consider your expectations about how well you and your advisor will do at predicting the weather on average over the course of the year. In this case, $\mathscr{D}$ is the set of all r_i. Thus, as we'll see, we can look at expected updating behavior given expected accuracy levels both for individual propositions and for larger subject areas. Of course, the more general the subject matter, the more individual cases will vary from expectations. Nonetheless, CALIBRATION requires eventual convergence.

3.1.2 Some Notation and Housekeeping

I'll use capital letters like A and B to refer to agents, and Fraktur lower-case letters $\mathfrak{a}_t$ and $\mathfrak{b}_t$ to represent their credence functions at t. We'll generally be considering disagreement from the t_0-perspective with B as the representative agent, so most of the time the salient function for evaluation is $\mathfrak{b}_0$, which I'll sometimes abbreviate further as just $\mathfrak{b}$ when appropriate.

At t_0, I know neither what I'll think in the future after acquiring more evidence about some proposition p nor what you'll think. So, from the perspective of $\mathfrak{b}_0$, both $\mathfrak{a}_1(p)$ and $\mathfrak{b}_1(p)$ are random variables whose value is generally unknown. Since we're interested in what my *expectations* about updating are given assumptions about the epistemic performance of A and B, we'll often be looking at distributions of the form $\mathfrak{b}_0(p|\mathfrak{a}_1(p)\,\&\,\mathfrak{b}_1(p))$, but that's quite a mouthful. So, since we'll focus mostly on expectations of disagreement over a single proposition p (like it will rain on day i), I'll use $\mathfrak{b}_0^{A_1B_1}$ (or sometimes just $\mathfrak{b}^{AB}$) to represent $\mathfrak{b}_0(p|\mathfrak{a}_1(p)\,\&\,\mathfrak{b}_1(p))$. That is, $\mathfrak{b}_0^{A_1B_1}$ represents B-at-t_0's credence conditional on what A and B think at t_1. Again, $\mathfrak{b}_0$ doesn't know yet what $\mathfrak{a}_1(p)$ and $\mathfrak{b}_1(p)$ are, so $\mathfrak{b}_0$ treats $\mathfrak{b}_0^{A_1B_1}$ as a variable whose value is unknown.

One further simplification: when it's clear what's meant, I'll sometimes shamelessly

abuse notation and write $A_t = x$ for $\mathfrak{a}_t(p) = x$ and also use A_t to refer to agent A at time t.

3.2 Necessary and Sufficient Conditions

First, let's take a moment to examine the notion of epistemic *weight* given to an advisor. Suppose an agent B disagrees with her advisor A. How much relative weight did B give herself and her advisor? The most natural way, I think, to understand the notion is in terms of the relative *distance* (under an appropriate measure) between $\mathfrak{b}_0^{AB}$ and the two credences $\mathfrak{a}_t(p)$ and $\mathfrak{b}_t(p)$.[8] The closer to the former B ends up, the more weight she gave her advisor. The closer to the latter she ends up, the more weight she gave to herself. In the special case where $\mathfrak{b}_0(p|\mathfrak{a}_t(p) = x) = x$ no matter what the value of x, B treats A_t as an *expert*. That is, she'd defer to A_t regardless of her actual credence. When she was going to ignore A_t no matter the value of A_t's credence, she was determined to *stick to her guns* or *stay pat*. She gives A and B equal weight when $\mathfrak{b}_0^{AB}$ is equally far from both $\mathfrak{a}_t(p)$ and $\mathfrak{b}_t(p)$.

Now, there are lots of different ways to measure distance, but in the body of the paper we'll use *squared-divergence*. That is, we'll say that the distance between two credences x and y is $(x-y)^2$.[9] So, you give two advisors A_t and B_t *equal weight* when $(\mathfrak{b}^{A_tB_t} - \mathfrak{a}_t(p))^2 = (\mathfrak{b}^{A_tB_t} - \mathfrak{b}_t(p))^2$.[10] That is, according to this metric, $\mathfrak{b}(p|A = x \,\&\, B = y)$ gives equal weight to A and B just in case $\mathfrak{b}(p|A = x \,\&\, B = y) = \frac{x+y}{2}$, which is not the case on every standard

[8]The word *distance* here is not meant to imply that the relevant measure is actually a metric. Whichever measure we choose should correspond to some scoring rule for measuring inaccuracy (see Thm (3.A.1)), and scoring rules won't in general correspond to metrics, as Nau (1985) demonstrates.

[9]See Appendix 3.A for brief discussion of more general measures. For more thorough treatment, see DeGroot and Eriksson (1985); Gneiting and Raftery (2007); Schervish (1989); Winkler (1996).

[10]This notion works well for expositional reasons since it simply measures the square of the Euclidean distance between credences. It works well for dialectical reasons because it's the measure that best supports the Split-the-Difference view.

measure.

3.2.1 Inaccuracy

Inaccuracy—negative accuracy, which we'll use for technical convenience—measures how far your credence is from p's truth value $v(p)$. The closer to $v(p)$ you end up being, the less inaccurate you are.

Your ultimate goal as an epistemic agent is to be accurate—not to be rational, thoughtful or prudent. You try to collect evidence and evaluate it as best you can because generally that helps you in your quest for accuracy. Therefore, we'll assume—for now, at least—that an agent's primary method of evaluating herself and her advisors is in terms of accuracy.

As with weight, there are all sorts of different measures of inaccuracy. For ease, we'll go with the most intuitive and most widely known scoring rule, which simply tracks Euclidean distance between p and $v(p)$:[11]

Definition 1. *The* Brier Score of credence x in p *is*

$$BS(x) := (v(p) - x)^2 = \begin{cases} (1-x)^2 & p \textit{ is true} \\ x^2 & p \textit{ is false} \end{cases}$$

Here's an example of how it works. Suppose you have credence .3 in the proposition r_1: it will rain tomorrow. We measure your inaccuracy with the square of the distance between .3 and $v(r_1)$, which is just .09 if it doesn't rain and .21 if it does.

[11] This measure was used originally by Brier (1950) to measure the inaccuracy of weather forecasts. I actually don't particularly like the Brier Score as a measure of inaccuracy (see Levinstein (2012) for objections), but it'll do for our purposes. In Appendix (3.A), we'll show that—while a few wrinkles are added—the results of this discussion apply much more generally to nearly any measure of inaccuracy you might choose.

3.2.2 The Role of Expectations

Our task is to look at how antecedent views about ourselves and our advisors relate to how we end up updating upon discovery of everybody's credences in p. In cases of interest, we don't yet *know* whether p, and we don't yet know what our advisors think. Therefore, we can't look to people's *actual* level of inaccuracy as a guide. Determining actual inaccuracy requires knowing both whether p and what an advisor's credence is.[12]

Instead, we'll have to give updating advice based on *expected* inaccuracy and/or *expected* levels of weight. The important question, then, is:

Question How do your *expectations* about the various levels of accuracy of yourself and advisors or about the weight you'll give yourself and advisors relate to your updating behavior after finding out everybody's credence?

Because (1) our primary method of evaluation is in terms of accuracy, and (2) it's expectations that we'll have to use for updating, we'll define *peer* along these lines, which we keep intentionally somewhat rough:

Definition 2. *Two advisors A and B are* peers with respect to p according to C, *which we abbreviate* $\mathsf{Peers}_p(A_t, B_t; \mathfrak{c})$ *just in case* $\mathfrak{c}$ *expects A and B to be equally inaccurate with respect to p at t. More generally, we say that A and B are* peers with respect to a set of propositions $\mathscr{D}$ according to $\mathfrak{c}$, *abbreviated* $\mathsf{Peers}_{\mathscr{D}}(A, B; \mathfrak{c})$, *just in case* $\mathfrak{c}$ *expects A and B to be equally inaccurate on average over the domain* $\mathscr{D}$.

Now, we could make this definition more rigorous, but I hope the idea is clear. When you assign the same level of expected inaccuracy to two people A and B either over some proposition (like whether it will rain tomorrow) or over some general area (like predicting the weather not just tomorrow but in general), you consider them peers (with respect

[12] Of course that's not *always* the case. I know that an omniscient agent will have a constant level of 0 inaccuracy, for instance, even if I don't know what she thinks.

to that proposition or domain).[13] When the salient proposition and/or domain is clear, I'll often not index to them explicitly.

Three further notes: First, we emphasize again that our definition is relativized to a given credence function $\mathfrak{c}$ and is in terms of *expected* accuracy, not actual accuracy. So, $\mathsf{Peers}(A, B; \mathfrak{c})$ if $\mathrm{E}_{\mathfrak{c}}(\mathrm{BS}(A)) = \mathrm{E}_{\mathfrak{c}}(\mathrm{BS}(B))$, where $\mathrm{E}_{\mathfrak{c}}$ denotes the expected value function by the lights of credence function $\mathfrak{c}$. Second, I won't have much to say about when you should take two agents to be peers in this sense. Our task is to help clarify the disagreement debate and understand the relationships between many of the salient concepts that epistemologists studying disagreement have made appeal to. I'll follow Elga (2007) in ducking out of this task:

> How should one judge the epistemic abilities of weather forecasters, dentists, math professors, gossipy neighbors, and so on? This is a question with the same sort of massive scope as the question: "When does a batch of evidence support a given hypothesis?" Fearsome questions both, and worthy of investigation. But leave them for another day. Here I will focus on [another] question. . . . Given [the] rating of the advisor's judgment, how should you take her opinions into account? (483).

Third, if you don't like this understanding of *peer*, as defined in these purely alethic terms, bear with me. Given the result we'll discuss, we can see how alternative notions relate to this one.

3.2.3 Equal Expected Accuracy Iff Equal Expected Weight

Now that we have the notions of weight and inaccuracy in place, we can now state:

---

[13] Generally, the notion of being a peer over some domain $\mathscr{D}$ will be interesting only when the relative skill of the agents is thought to be roughly constant over $\mathscr{D}$. For instance, you might be way better than I am at predicting the weather, while I'm much more talented at determining election outcomes. If $\mathscr{D}$ contains propositions about elections and weather, then we could well end up peers-over-$\mathscr{D}$, and on average we should give each other equal weight. However, since I'll tend to listen to you about the weather, and you'll tend to listen to me about elections, most cases will deviate greatly from the mean.

Theorem 3.2.1. *Suppose A and B are advisors. Then the expected amount of weight given to A equals the expected amount of weight given to B if and only if A and B have equal expected accuracy.*

In other words, you expect to give your peers equal weight. As we show in Appendix 3.A, this relationship is surprisingly robust and holds across all sorts of different measures. The notion of expected inaccuracy of an advisor B simply *factors* into the sum of the expected inaccuracy of $\mathfrak{c}^B$ and the divergence between B's credence and $\mathfrak{c}^B$ *on any reasonable way of measuring inaccuracy*. The measures of inaccuracy and divergence vary, but this relationship stays the same. From here, it's easy to show that Thm. (3.2.1) holds.[14] Thus, on this rather common understand of *peer*, expected equal weight falls out as a mathematical relationship on any standard measure.

Furthermore, as we've indicated, it doesn't matter whether we're talking about a single proposition p or a domain of propositions. If you expect that somebody else is as good of a weather forecaster as you under some set of circumstances, then you expect that in general you'll give her equal weight on average in those circumstances.

Let's now look at some consequences of this result.

3.2.4 Cleaning Up the Equal Weight View

As we've mentioned, a number of authors—such as Elga (2007); Enoch (2010); White (2009)—understand *peer* in essentially accuracy-theoretic terms. On the standard view, once you expect your advisor to be equally accurate, you're supposed give her equal weight in actuality, not just in expectation. That is, according to EWV-orthodoxy, once you judge her to be equally accurate *in expectation* you commit yourself to giving her equal weight *ex post*. Of course, there are a number of caveats and potential out-clauses,

[14] For the details and a more careful statement of this result, see Thms. (3.A.1) and (3.A.2).

but even with those nuances thrown in, such a view is untenable, and we can show where it goes wrong.

Think about the nature of expectation and how we calculate the expected inaccuracy of an advisor A whose credence we don't know.[15] For simplicity, let's assume we know our advisor A has one of some finitely many potential credences $x_1 < \ldots < x_n$.[16] For each such credence x_i, we'd first consider our *conditional* expected inaccuracy of A were she to announce credence x_i. Since it's clear which credence function we're using for calculations, we'll write the *conditional expected Brier Score* of A when A's credence is x_i simply as:

$$\mathrm{EBS}(A|A = x_i)$$

We then multiply $\mathrm{EBS}(A|A = x_i)$ by $\mathfrak{b}(A = x_i)$ and sum over the values. We'll arrive at:

$$\mathrm{EBS}(A) = \sum_{i=1}^{n} \mathfrak{b}(A = x_i)\mathrm{EBS}(A|A = x_i)$$

What's important to note is that how likely you think it is that A has various credences x_i can and often should affect the expected level of inaccuracy you assign her. Let's take an extreme case:

SUNRISE I ask Tom what his credence that the sun will rise tomorrow is.

I consider Tom—and indeed, nearly everybody I ever come in contact with—roughly a peer when it comes to the question of whether the sun will rise tomorrow. The reason I hold such an egalitarian view is almost entirely based on my expectations about what these people think. My distribution $\mathfrak{b}$(Tom's credence is x) over potential credences he

[15] Of course, in practice even when we have credences over events like rain and expectations about how accurate an advisor will be, we usually don't have well-defined distributions over what an advisor might think. However, the idea is that by looking at how such expected inaccuracy is calculated in more idealized circumstances, we'll be able to identify flaws with the orthodox EWV.

[16] The more general case in which A could adopt any credence in $[0, 1]$ is a fairly straightforward generalization, so I omit it in order to avoid technical distractions.

may have in the proposition that the sun will rise tomorrow has nearly all its weight right around 1. That is, $\mathfrak{b}(\text{Tom's credence is in } [1-\epsilon, 1]) \approx 1$ for some very small value of ϵ. And, conditional on him having a credence right near 1, I assign him very low expected inaccuracy. If he surprises me and tells me that his credence is close to .5, I'll legitimately ignore him (almost) entirely, since reality diverged so extremely from my expectations.

Indeed, lots of similar cases have plagued orthodox versions of the equal weight view. To take another extreme one, consider the contrast between:[17]

RESTAURANT Five of us go out to dinner, and at the end of the night, the bill is $187.50. We agree to split the bill evenly and tip 20%. You and I do the math in our heads, and I become very confident that we each owe $45, while you're highly confident that we each owe $43. Since the two of us go way back, we're both aware that we're each equally talented at doing arithmetic in our heads, and neither of us is especially tired, or drunk, or distracted.

And

EXTREME RESTAURANT Just like before, except I think we owe $45 each, while you think we owe $450 each, which is more than the total bill.

Conciliationists have struggled to come up with explanations of principled reasons the latter case is different from the former.[18] From our macro-perspective, the answer is clearly a difference in degree and not of kind. What the micro-view misses is that the antecedent expectations about an advisor's expected inaccuracy can *depend on expectations about what she thinks*. That is, your evaluation of her was based, in part, on expectations of her opinion. In SUNRISE and EXTREME RESTAURANT, you—at least implicitly—

[17] These examples involve disagreement over expected values of a quantity, not credences in a proposition. It turns out we can extend our inaccuracy measures to cover the more general case (see Gneiting and Raftery (2007) and Winkler (1996)), but I won't go into the details of how to do so here.

[18] See, for instance, Christensen (2011).

thought it very unlikely you'd get an answer of credence around .5 or an estimate of $450. Were your advisor to surprise you with such an answer, it's perfectly legitimate not to listen very much to her because your *ex ante* assessment of her accuracy was based on the assumption that she would almost surely have a much different view.[19]

3.2.4.1 The Conceptual Relationship Between Accuracy & Weight

The relationship between expected weight and expected accuracy only exists as a *wide-scope macro-norm*. It's wide-scope because it doesn't alone tell you which people you should expect to be as accurate as you are. If, over the long haul, you discover that you tend to give people way less weight than you expected to, you're doing something wrong. Supposing you keep assigning people low expected inaccuracy, you'll have to start giving them more weight. However, you may be right to change your *ex ante* assessments as well. Imagine, for instance, an egalitarian steadfaster who will never give any weight to anybody else but who also thinks lots of people are her peers even when they have the same evidence. That can only happen if she's sure (or almost sure) they'll always have the same credence in p that she herself does. Our theorem relating expected equal weight and expected inaccuracy doesn't say whether she's doing something wrong. However, if people often do as a matter of fact end up disagreeing with her, then some changes are called for. She can perfectly well retain her steadfast position, but she'll have to start expecting people to be less accurate than she did before.

It's a macro-norm because the relationship between accuracy and weight is virtually *invisible* for any particular case. Coherence is only violated over repeated trials, and the relationship only emerges in the macroscopic limit. However, we still manage to put serious constraints on updating behavior. Any micro-norms on disagreement must, then,

[19]Indeed, as we'll see below, actual relative weight given often *has* to be a function of the actual credence reported.

be compatible with this emergent connection between weight and accuracy. That is, a necessary condition on the adequacy of a micro-norm is that it on average advises giving equal weight to peers.

3.2.5 Other Notions of Peer

We defined the notion of *peer* alethically. However, in the literature, we often find less directly truth-connected notions of *peer*. In the abstract, we think of our peers as people who are our (expected) epistemic equals in some important sense, such as rationality, intelligence, familiarity with the evidence, or some other kind of epistemic virtue. I don't here wish to get into a dispute over which if any of these notions is most appropriate for debates over disagreement. Instead, I'll highlight how Thm. (3.2.1) allows for a general kind of argument-schema for conciliationists.

First, pick some epistemic virtue v that you think is appropriate. Call advisors A and B v-peers according to $\mathbf{c}$ if $\mathbf{c}$ assigns equal expected levels of virtue v to A and B. The conciliationist will have to argue that generally if you expect A and B to be v-peers then you should also expect them to be (accuracy) peers. That is, the salient virtue v should track accuracy well enough that an agent ought to take them to be more or less substitutable. Since being an expected accuracy peer is necessary and sufficient for expected equal weight, this move must be made for any non-alethic understanding of *peer*.[20]

Summing up, we see that if we adopt an alethic notion of *peer*, Thm. (3.2.1) shows that a version of the EWV falls out. To expect equal accuracy just is to expect to give equal weight. However, that can't (as we'll see below) and shouldn't require giving equal

[20]One common notion of *peer* concerns evidential rationality in particular. That is, A and B are taken to be rationality-peers if they're taken to be equally rational in handling of the background evidence. If rationality is permissive and allows for multiply distinct but equally rational credences on the same evidence, then a number of authors (such as Kelly (2005); Enoch (2010); Ballantyne and Coffman (2011)) have thought that that spells trouble for conciliationism. I think that if anything, permissive rationality supports conciliationism. For more, see my (ms).

weight in each case. For non-alethic notions of *peer*, Thm. (3.2.1) shows what additional premises are needed to maintain any version of an equal weight view.

3.3 Bold Conciliationism

As we've remarked, any notion of *peer* has to do with being (expected) epistemic equals in some important sense. Therefore, there's a real or perceived symmetry that holds between you and those you (take to be) peers. Based on this thought, many take equal weight to require *splitting the difference* (at least in normal circumstances). As we noted above, however, both defenders and detractors of EWV worry that such a policy will lead to spinelessness. That is, by giving your peers equal weight, you'll end up overly timid: too close to the middle of the unit interval and too far from 0 or 1.

We want to get a sense of Split the Difference as a *pure* updating policy with no confounding, disagreement-independent factors. You can think of that as getting an idea of what Split the Difference looks like *ceteris paribus*. So, we'll first consider the case when I take us both to be reliable, or to use a technical term, I take us to be *experts* for me with respect to the proposition in question. Here's how that works. Suppose I find out at t_0 what my credence at t_1 is, and I don't learn anything else. Then, in the cases we'll look at, I'll defer entirely to my future self. That is, $\mathfrak{b}(p|\mathfrak{b}_1(p) = x) = x$.[21] Note that that doesn't say that I'll defer entirely if I learn my future belief *and some other evidence too*. It just says that conditional *only* on my future belief in rain being x, I have credence x.

Likewise, let's say I just learn what your credence at t_1 is. In the scenarios we'll look at for now, I treat you as an expert too, so $\mathfrak{b}(p|\mathfrak{a}_1(p) = x) = x$. That would happen, for instance, if I generally trusted you to assess the evidence. Imagine, say, that one day you got to look at the data but I didn't. Treating you as an expert means that I'd defer to you

[21] This requirement is just the (in)famous Reflection Principle, from van Fraassen (1984).

in such a situation. The question, then, is: when I-at-t_0 conditionalize on both of our future credences, what will I do?

If I don't treat one of us as experts, then there are times when I think at least one of us is in some ways predictably error-prone independently of the disagreement. Even if I didn't know whether a disagreement would occur—that's to say, even if we abstract from the input of the other agent—I would think I could correct for some epistemic mistake. We'll talk briefly about the general case after and in more formal depth in Appendix 3.B, but the case with two reliable agents disagreeing is the way to look at an updating policy without any confounding factors.

3.3.1 Fortune Favors the Bold

In this part of the paper, by taking a macro-approach, we can show what goes wrong with splitting the difference. Recall that we call an agent's credence *bold* (or synonymously *opinionated*) if it's close to 0 or 1. First, we show that weighted averaging (of which difference-splitting is a special case) will systematically result in underly bold credences. We then go on to show that when an agent hears both expert opinions, she should expect to end up with a bolder credence—on any relevant measure—than she would have had she only heard the views of one of them. Importantly, these results *won't depend on whether or to what extent the experts share evidence*. Below, we relate these results to the more general case of disagreement.[22]

3.3.1.1 Thrasymachus Was Right

Consider what you-at-t_0 with credence function $\mathfrak{b}_0$ think of the difference-split credence $\mathfrak{d} := \frac{A_1+B_1}{2}$. If you think A_1 and B_1 are peers, then you'll think $\mathfrak{d}(p)$ will be more accurate

[22]Because we're only concerned with a single proposition and time, I'll invoke the simplified notation laid out in § 3.1.2 on page 72.

than either $\mathfrak{a}_1(p)$ or $\mathfrak{b}_1(p)$. Nonetheless, you also expect—more surprisingly—that $\mathfrak{d}$ has some correctable deficiencies, which we'll go through informally here.[23]

First, $\mathfrak{d}$ *can't* be an expert for $\mathfrak{b}_0$. That is, there are some values $x \in [0,1]$ such that B_0 thinks it's possible $\mathfrak{d}(p) = x$ but $\mathfrak{b}_0(p|\mathfrak{d}(p) = x) \neq x$.

To get a handle on how this works, here's a frequential analog. Suppose that you (A) and I (B) are both long-run calibrated at predicting rain. A third weatherman D listens to both our forecasts and always adopts the difference-split credence. If we're on average equally accurate, then D is generally more accurate than either of us. However, D is also *un*calibrated.

Now, it turns out that not only does $\mathfrak{d}$ necessarily fail to be an expert, but we can also characterize how it tends to produce deficient credences. In particular, $\mathfrak{d}$ is always expected to be (1) underly bold, but (2) on the right track. Here's what I mean. As we show in Thm. (3.A.5), $\mathfrak{d}$ is expected to be too far from 0 or 1 and too close to $\mathfrak{b}_0(p)$. More formally: $\mathrm{E}(\mathfrak{b}_0^{\mathfrak{d}}(p) - \mathfrak{b}_0(p))^2 > \mathrm{E}(\mathfrak{d}(p) - \mathfrak{b}_0(p))^2$, where E is $\mathfrak{b}_0$'s expectation function.[24] However, $\mathfrak{d}$ is still expected to be on the correct side of $\mathfrak{b}_0(p)$. That is, $\mathfrak{b}_0$ expects $\mathfrak{b}_0^{\mathfrak{d}}(p)$ to be between $\mathfrak{b}_0(p)$ and p's truth-value. So, should $\mathfrak{d}$ be above $\mathfrak{b}_0(p)$, B_0 will tend to overshoot it, and if it's below, B_0 will tend to undershoot it.

It's important to emphasize that this relationship holds regardless of whether the experts share evidence. Even if they don't arrive at their views independently, and even if $\mathfrak{b}_0$ knows they don't, $\mathfrak{b}_0$ will still expect the average to be overly timid. Furthermore, the same goes not just for difference-splitting but for any method of linear pooling whatsoever regardless of the antecedent levels of expected accuracy. For instance, suppose instead of taking the straight average, we'd conditionalized on $\mathfrak{d}^* := {}^2/_3 A_1 + {}^1/_3 B_1$. We'd still expect $\mathfrak{d}^*$ to be overly timid but on the right track, and we couldn't treat $\mathfrak{d}^*$ as an

[23] See Thms. (3.A.5) and (3.A.6) for the formal details and for the more general case of n experts under an arbitrary scoring rule.

[24] We can generalize to a wide class of measures of distance from the prior. See Thm. (3.A.5).

expert.

Think about what it takes for this result to hold. No matter what linear weighting we use, the average is expected to be underly bold. That can only happen if the relative level of weight for an expert tends to *grow* the bolder she is. In other words, we have to follow the:[25]

THRASYMACHUS PRINCIPLE When adjudicating a disagreement between two experts, generally favor the more opinionated of the two.

Note how discordant the THRASYMACHUS PRINCIPLE seems initially with intuition. First, even if one of the experts was expected to be significantly less accurate than the other, we'll still end up on average giving her more weight should she turn out to be bolder.

Second, and more disconcertingly, we tend to shy away from policies that suggest fairly extremal credences and favor epistemic moderation. For instance, suppose I knew a fair coin were going to be flipped a million times. Each time, I announce my credence that it will land heads the next flip. Suppose Policy 1 advises me to have credence .5 every time. Policy 2 recommends I alternate between credence .4 and .6. On either policy, my credence in Heads averages out to .5, but Policy 2 is still clearly worse.

According to the THRASYMACHUS PRINCIPLE, however, even if I considered experts *A* and *B* epistemic equals under every rule, when they disagree, I'll give more weight on average to the bolder one. Shouldn't I generally play it safe, given my normal epistemic risk-aversion?

No. If I treat you as an expert, then I expect you to be responsible and risk-averse in forming your credences. When an expert has a fairly extremal credence, I take it she responsibly thought she could cull a good amount of information about whether *p*

[25] In Plato's *Republic*, Thrasymachus says, "Justice is nothing but the advantage of the stronger" (338c). The principle immediately below is an epistemic analog.

from her evidence. Thus, from my point of view, her credence alone carries quite a bit of information about p. When A and B are both experts, but A is more extremal, I'll generally think that A's credence carries on balance more information than B's credence carries. Therefore, the average of their credences will tend to give undue weight to B. When it comes to experts, we simply can't and shouldn't commit to treating all of them as epistemic equals regardless of what they say.

Indeed, despite the initial appeal of equal treatment for supposedly equal agents, in many cases, we can often see that the *ex post* response should depend on the credences reported. To begin, let's look at an extreme case. If you treat agents A and B as experts, and you first learn A is completely certain that p, then you become certain in p as well. If you then find out that B only has credence .5, you can't become any less confident in p, at least according to orthodox bayesianism. Presumably, slightly less extreme cases in which A is not exactly certain but 99.9999% sure that p should be treated nearly the same as this limiting case. It'd be quite odd if it was right for you to remain 100% sure that p in the former case but split the difference and up around 75% confident in latter.

In cases in which agents don't share evidence, it's often clear that relative boldness of credences should be important. Suppose you're unsure whether Dallas or San Diego has a bigger population, and you decide to ask two friends you antecedently think equally likely to know the answer. One friend reports that she's fairly uncertain, while the other says she's almost sure that San Diego is larger. Here, you should give more weight to the second friend supposing you have some trust in her epistemic responsibility.

It's still tempting, however, to think that perhaps different behavior is called for in cases in which evidence is largely shared. Indeed, when two agents with the same evidence disagree, one is often making some kind of cognitive error, and perhaps we should cast aside some of the prescriptions of orthodox bayesianism because of such a peculiarity. That sentiment, for all that's been said, may indeed be partially correct, but such special treatment can't work in the way most conciliationists think. To see that an intuitive case

in addition to a formal one can be made for this position, consider:

MURDER MYSTERY SH and HP are two equally competent and reliable detectives working on a case. To toy with police, the criminal left behind an exceedingly difficult riddle whose answer will reveal his identity. Both SH and HP are both unsure what the answer is despite spending days working on the puzzle. SH's assistant Dr. W enters and asks how the investigation is coming. As HP is explaining that they still don't know who the culprit is, SH shouts "Aha!" and announces that his arch-nemesis Professor M was responsible.

PROOF Since 1971, Mathematicians have worked without result on the question of whether $P = NP$. Fields Medalist TT announces on his blog that he has determined that it in fact does and is highly confident of his result.

In both of these cases, the various agents have overlapping—or perhaps even the same—evidence.[26] Imagine yourself first as Dr. W in MURDER MYSTERY. Even though HP and SH have talked through their evidence and much of their reasoning with each other and are equally competent, SH is bolder and thinks he's solved the case. Given your antecedent respect for both of them, you know that SH wouldn't become so confident prematurely. It's more likely that he's determined the key to solving the puzzle. Even if you yourself are no super sleuth and can't even follow his reasoning, it seems intuitive that you (and probably HP as well) should end up giving more weight to SH's judgment than to HP's agnosticism. On the other hand, had HP and SH both been unsure while leaning in slightly different directions, it's not nearly so clear that one should be preferred

[26] Philosophers will differ on the extent to which these cases count as ones with completely shared evidence, since it's unclear how much private inferences contribute to one's evidential base. The point is that the agents are given enough information in the beginning to determine the answer but because of cognitive limitations are at least temporarily uncertain.

over the other. Thus, the relative level of respect Dr. W should give to either detective generally grows with boldness.

In the PROOF case, all agents have evidence that *entails* whether or not $P = NP$, but mathematicians are nevertheless uncertain of its truth-value.[27] Generally, however, when a reliable and respected mathematician announces that he or she has determined the answer, then even before the rest of the community sees the proof, that person's claim is given serious credence. Again, reliability combined with boldness ought to result in additional respect.

Before closing this section, we mention one further fact: It also turns out that if we learn both A and B's opinion, under modest conditions, we have to expect $\mathfrak{b}_0^{AB}(p)$ to be strictly more opinionated than either agent (see Thm. (3.A.6)).[28]

So, in sum, we have the following:

- Straight averaging is expected to beat both experts when $\mathsf{Peers}(A, B; \mathfrak{b}_0)$.
- Straight averaging is expected to be headed in the right direction but overly timid—regardless of the relative expected accuracy of the experts—and can't be an expert itself.
- Learning what both the experts think is expected to lead to more opinionation.
- Regardless of any *ex ante* symmetry, we have to treat the experts differently *ex post* based on how bold they are.

Again, all of these facts hold regardless of whether the experts share evidence.

---

[27] According to a 2002 poll of 100 mathematicians and computer scientists, 61 believed $P \neq NP$, 9 thought $P = NP$, 22 were uncertain, and 9 thought the question was independent of the ZFC axioms (see Gasarch (2002)). We ignore this latter possibility.

[28] That's not surprising, since advisors' credences are evidence, after all, and we generally get more opinionated as we gain evidence. However, given the worries about spinelessness, it's worth emphasizing this point.

Such results once again speak to the fruitfulness of the macro-approach. There are some situations and particular values of $\mathfrak{d}$ for which you might end up deferring. Moreover, there are even cases in which you end up going in the opposite direction of $\mathfrak{d}$. It's only in expectation and in the long run that these relationships become apparent.

3.3.2 Disagreement in General

What about non-experts? As it turns out, we can in an important sense *reduce* disagreement between agents in general to disagreement between experts when we take a macro-approach. The process is a bit technically involved, so we relegate the details to Appendix 3.B. However, we have the ingredients to see what's wrong with difference splitting here.

Suppose now that A_1 and B_1 are no longer experts for $\mathfrak{b}_0$ but simply advisors. Since difference splitting is *underly* bold when experts are involved, $\mathfrak{b}_0$ could only possibly end up splitting the difference generally when it expected that at least one of the agents—let's say B_1—was already overly bold. That is, even if $\mathfrak{b}_0$ didn't get to hear from A_1 or learn of any disagreement, $\mathfrak{b}_0^{B_1}(p)$ was expected to be less bold than B was. In such a case, it's possible that $\mathfrak{b}_0$ might adopt a uniform difference-splitting policy if the details worked out just right, but the difference splitting had nothing to do with any symmetry stemming from disagreement. It simply worked out as a fluke to correct for B_1's predictable epistemic errors that $\mathfrak{b}_0$ already recognized (or thought it recognized) anyway. Thus, without disagreement-independent confounding factors, splitting the difference is generally spineless.

3.4 Epistemic Egalitarianism and Its Limits

At this point, it seems to me that we've laid the groundwork for a bolder and more flexible form of conciliationism. Once you're willing to make the initial step of conceding that you know people whom you expect to be about as accurate as you with respect to p,

there are serious constraints on what you can expect your updating behavior to look like. If there are lots of these sorts of propositions and many such advisors, then we should see you giving equal weight on average *and* becoming more opinionated.

These results seem to have serious implications for how many of us should handle actual disagreements. Speaking for myself, at least, I often expect many of the very smart people I know to be at least as accurate as I am over many different domains even when we have the same evidence. Nevertheless, we do often disagree, but as a matter of autobiographical fact, I'm not generally as conciliatory as the above arguments suggest I should be. What are my options?

3.4.1 You Can't Always Get What You Want

Here we can see how our values clash. When you're opinionated, you have to think you're doing a good job at getting your credence close to $v(p)$, so strong opinions require high self-regard. By extension, if you think a bunch of your friends will end up about as accurate as you are, they'd better tend to have pretty strong opinions too, since a moderate credence can only be but so accurate. But if you expect them to be opinionated *and* as accurate as you, you also can't expect that they'll generally be too far from your credence.[29] Think again of the weather forecaster paradigm case. It can't turn out that (1) we're both very accurate and usually give bold forecasts, but (2) we tend to make very different predictions. So, if I know I'll be giving a lot of bold predictions and I expect you to be as accurate on average, I have to expect that we'll generally be close to agreeing on the next day's forecast.

Therefore, if there's some domain of propositions over which you think you'll be quite opinionated, that will mean that it's pretty tough—by your lights—for somebody

[29]Compare: If we have two measuring devices that are both very accurate, they'll have to generally be in accord.

to be your peer. When there's a domain where strong disagreement is expected, you have some hard choices. In general, the more opinionated you think you'll be, the fewer people you'll be able to consider peers *ex ante*.

Worries that we'll be overly agnostic if we give peers equal weight can then be traced to a failure to stand by one's implicit evaluations. We don't get to be opinionated and humble. If we're inclined to reject humility, we may be right to do so, but that can't generally be good advice when disagreement is rampant. Indeed, the average of everybody's credence will be *more* accurate than average on any good measure.[30] If we're inclined to think lots of people are our peers over controversial propositions, we should recognize our erstwhile opinionated credences were epistemically reckless. If you're like me, this route seems generally more plausible, and we should probably be less opinionated than we are, since on reflection we're unwilling to accept the requisitely high self-regard necessary for boldness.[31] But the fault, dear peer, lies not in conciliationism, but in ourselves, for we are epistemic underlings. Maintaining strong opinionation will only tend to make us worse off.

3.A Results

First, we generalize from the Brier Score to a much more general class of inaccuracy measures. For simplicity, we stick to the single proposition case.

Definition 3. *A* scoring rule *is a function* $S : [0,1] \times \{0,1\} \to [0,\infty]$ *such that* $S(x,1)$ *and* $S(x,0)$ *are strictly monotonically decreasing and increasing respectively.*

We write $S(x)$ for the score a credence of x in p receives at the actual world. We then

[30] That follows from the convexity of strictly proper scoring rules.

[31] Indeed, that's what the empirical data suggest. By and large people are *massively* overly bold in making predictions. See Hoffrage (2004) for a nice and disturbing discussion.

say:

Definition 4. *A scoring rule S is* proper *if for all probability functions $\mathfrak{b}$ where $\mathfrak{b}(p)$ is defined and all $x \in [0,1]$, $\mathrm{E}_{\mathfrak{b}}(S(\mathfrak{b}(p))) \leq \mathrm{E}_{\mathfrak{b}}(S(x))$. S is* strictly proper *if S is proper and equality obtains only when $x = \mathfrak{b}(p)$.*

It's standard to restrict attention to strictly proper scoring rules.[32] We'll require in addition that $S(x,1)$ and $S(x,0)$ be continuous and differentiable in the first argument.

Definition 5. *A scoring rule S is* admissible *if S is strictly proper and $S(x,i)$ is continuous and differentiable in the first argument.*

As we've mentioned above, we're focused primarily on *expected* inaccuracy. So we'll use $\mathrm{ES}_{\mathfrak{b}}(A)$ to refer to the expected inaccuracy of advisor A (with respect to p). Keep in mind that the value of A's credence $\mathfrak{a}(p)$ may or may not be known to $\mathfrak{b}$. As before, we use $\mathfrak{b}^A$ to represent $\mathfrak{b}$'s credence conditional on A's credence.

As shown by DeGroot and Fienberg (1982, 1983), each admissible scoring rule S comes with an associated measure of *divergence* $D_S(x||y)$ that measures how "far" a credence of y is from a credence of x. In other words, from our starting point of measuring how far a credence is from a truth-value, we're able to induce a more general measure that compares the disparity between any two points in the unit interval. The gory mathematical details need not concern us, but one can think of $D_S(\mathfrak{b}(p)||y)$ as the *expected loss* of accuracy between $\mathfrak{b}(p)$ and y by the lights of $\mathfrak{b}$. When A's credence isn't yet known, $\mathrm{E}_{\mathfrak{b}}(D_S(\mathfrak{b}^A||A))$ denotes the *expected* divergence according to $\mathfrak{b}$ between A's credence and B's own as-yet-unknown credence conditional on what A thinks. We now have:

Theorem 3.A.1 (DeGroot and Fienberg)**.** *Suppose S is an admissible scoring rule, and A is an advisor. Then $ES_{\mathfrak{b}}(A) = \mathrm{E}_{\mathfrak{b}}(D_S(\mathfrak{b}^A||A)) + ES_{\mathfrak{b}}(\mathfrak{b}^A)$.*

[32] Cf. Joyce (2009); Predd et al. (2009).

In other words, you measure the expected inaccuracy of an advisor A simply by *summing* how far you expect your posterior credence to be from A's credence and how inaccurate you expect your credence *conditional* on A's credence to be. Note that $\mathrm{ES}_{\mathfrak{b}}(\mathfrak{b}^A) = \mathrm{E}_{\mathfrak{b}}(\mathrm{ES}_{\mathfrak{b}^A}(\mathfrak{b}^A))$, and since $\mathfrak{b}^A$ has strictly evidence than $\mathfrak{b}$, $\mathrm{ES}_{\mathfrak{b}}(\mathfrak{b}^A) \leq \mathrm{ES}_{\mathfrak{b}}(\mathfrak{b})$. Thus, on any admissible measure, the notion of expected inaccuracy *factors* into the expected level of respect you'll have for A and the expected evidential value of A's credence with respect to p.

From Thm. (3.A.1), we then get necessary and sufficient conditions for equal expected weight as an easy corollary.

Theorem 3.A.2. *For any scoring rule S and any advisors A and B,* $\mathrm{E}_{\mathfrak{c}}(D_S(\mathfrak{c}^{AB}||A)) = \mathrm{E}_{\mathfrak{c}}(D_S(\mathfrak{c}^{AB}||B))$ *iff* $\mathrm{ES}_{\mathfrak{c}}(A) = \mathrm{ES}_{\mathfrak{c}}(B)$.

Based on general convexity considerations, it's also easy to show:

Theorem 3.A.3. *Suppose A_1 and A_2 are advisors, $ES_{\mathfrak{b}}(A_1) = ES_{\mathfrak{b}}(A_2) < \infty$, and $\mathfrak{b}(\mathfrak{a}_1(p) = \mathfrak{a}_2(p)) < 1$. Then for hypothetical advisor $A_3 = \frac{A_1+A_2}{2}$, $ES_{\mathfrak{b}}(A_3) < ES_{\mathfrak{b}}(A_1) = ES_{\mathfrak{b}}(A_2)$.*

So, when you expect two advisors to do equally well (so long as they don't do infinitely badly and might disagree), you expect that splitting the difference beats deferring to either one completely.

For use final two results referenced in the main text, it'll be useful to invoke the main theorem from Greaves and Wallace (2006), which we state in slightly imprecise form:

Theorem 3.A.4 (Greaves and Wallace)**.** *Let S be an admissible scoring rule. Then for any probability function $\mathfrak{b}$, updating by conditionalization minimizes expected inaccuracy under S according to $\mathfrak{b}$.*

Now for the main result on linear pooling. We generalize here to the case with n experts and strengthen the result to generalized divergence.

Theorem 3.A.5. *For $i \in \{1, \ldots, n\}$, let A_i be experts (with respect to p) for $\mathfrak{b}$, and let $\mathfrak{b}(p) = b$. Fix constants $\lambda_i > 0$ such that $\sum \lambda_i = 1$, and let $A = \sum \lambda_i A_i$. Suppose further that there exist i and j such that $\mathfrak{b}(A_i \neq A_j) > 0$. We then have:*

(I) $\mathfrak{b}(\mathfrak{b}^A \neq A) > 0$. *I.e., A isn't an expert for $\mathfrak{b}$.*

(II) *Under any admissible rule S, $D_S(\mathfrak{b}^A || b) > D_S(A || b)$*

(III) *$EBS_{\mathfrak{b}}(A) < EBS_a(A)$, where $\mathrm{EBS}_{\mathfrak{c}}$ denotes the Expected Brier Score on credence function $\mathfrak{c}$.*

Proof. We start with (I). For the *reductio*, suppose A is an expert for $\mathfrak{b}$. Then $\mathrm{EBS}_{\mathfrak{b}}(A) = \mathrm{E}_{\mathfrak{b}}(v(p) - A)^2 = \mathrm{E}_{\mathfrak{b}}(A(1-A))$. (For convenience, we'll use E for $\mathrm{E}_{\mathfrak{b}}$ for the rest of the proof.) Since $A = \sum \lambda_i A_i$, then if A is an expert for $\mathfrak{b}$:

$$\mathrm{E}(A(1-A)) = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \mathrm{E}(A_i(1-A_j)) \tag{3.1}$$

We also have:

$$\begin{aligned} \mathrm{E}(A(1-A)) &= \mathrm{E}\left(\sum_{i=1}^{n} \lambda_i \left(v(p) - A_i \right) \right)^2 \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \mathrm{E}((v(p) - A_i)(v(p) - A_j)) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \mathrm{E}(v(p)^2 - A_i v(p) - A_j v(p) + A_i A_j) \end{aligned} \tag{3.2}$$

Since $v(p)$ is either 1 or 0, $\mathrm{E}(v(p)^2) = \mathrm{E}(v(p)) = b$. Furthermore, since each A_i is an expert, $\mathrm{E}(A_i) = b$, and it's easy to see that $\mathrm{E}(A_i v(p)) = \mathrm{E}(A_i^2)$. So, by (3.2):

$$\begin{aligned} \mathrm{E}(A(1-A)) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \mathrm{E}(A_i - A_i^2 - A_j^2 - A_i A_j) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \mathrm{E}(A_i(1-A_j)) - \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \mathrm{E}(A_i - A_j)^2 \end{aligned} \tag{3.3}$$

By assumption, there exist experts A_i and A_j such that $\mathfrak{b}(A_i \neq A_j) > 0$, so the last term in (3.3) is positive, i.e.,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j \mathrm{E}(A_i - A_j)^2 > 0 \tag{3.4}$$

Therefore (3.1) and (3.3) have different values, which completes the *reductio*.

(III) easily follows. $\mathrm{EBS}_{\mathfrak{a}}(A)$ is given by (3.1), whereas $\mathrm{EBS}_{\mathfrak{b}}(A)$ is given by (3.3). By (3.4), $\mathrm{EBS}_{\mathfrak{b}}(A) < \mathrm{EBS}_{\mathfrak{a}}(A)$.

For (II), by Thm. (3.A.4) $\mathrm{ES}_{\mathfrak{b}}(\mathfrak{b}^A) < \mathrm{ES}_{\mathfrak{b}}(A)$ since A isn't an expert. However, for every i, $\mathrm{ES}_{\mathfrak{b}}(A_i) < \mathrm{ES}_{\mathfrak{b}}(b)$, since each A_i is an expert. So, $\mathrm{ES}_{\mathfrak{b}}(\mathfrak{b}^A) < \mathrm{ES}_{\mathfrak{b}}(A) < \mathrm{ES}_{\mathfrak{b}}(b)$. Since D_S is a Brègman divergence, it's convex in its first argument. The conclusion follows by application of Thm (3.A.1). □

Note that (III) in the above theorem guarantees that $\mathfrak{b}$ expects to overshoot/undershoot A, since $\mathfrak{b}$ expects A to get a better Brier Score than A expects itself to get. Furthermore, note that whether any of the experts share evidence is irrelevant.

The last result is an easy consequence of Thm. (3.A.4).

Theorem 3.A.6. *For $i \in \{1,\ldots,n\}$, let A_i be experts (with respect to p) for $\mathfrak{b}$, and let $\mathfrak{b}(p) = b$. Let $\vec{A} := \langle \mathfrak{a}_1,\ldots,\mathfrak{a}_n\rangle$. Suppose that there's no k such that $\mathfrak{b}(\mathfrak{a}_k(p) \in \{0,1\}) = 1$, and there's no k such that $\mathfrak{b}(\mathfrak{b}^{\vec{A}} = \mathfrak{a}_k(p)) = 1$. Then under any proper scoring rule S, we have $D_s(\mathfrak{b}^{\vec{A}}||b) > D_S(\mathfrak{b}^{A_i}||b)$ for all i.*

3.B Expert Fix-Ups and Boldness

Here we show how to simulate a disagreement between two agents A and B as an isomorphic dispute between two hypothetical expert versions A^+ and B^+ of them and then show a surprising further consequence in terms of boldness in the case of peer disagreement.

To see how this works, it'll be enough to show how to generate one expert A^+ from non-expert A. First, recall Theorem (3.A.1). We saw that the expected inaccuracy of an advisor A under an admissible scoring rule S for an agent C was simply

$D_S(\mathfrak{c}^A||A) + \mathrm{ES}_{\mathfrak{c}}(\mathfrak{c}^A)$. From the agent's point of view, all that matters for her quest to determine whether p is the second term, since it's the metric by which she'll measure the expected epistemic utility of any kind of evidence whatever.

Let's go through the formal details of our calculation of the expected inaccuracy for an advisor A. First, we need a distribution $\mathfrak{c}_A$ over potential credences A might have. Next, for any x, we consider how we would update should A adopt credence x. That is, for each x, we determine $\mathfrak{c}(p|\mathfrak{a}(p) = x)$. From there, we can calculate A's expected inaccuracy by integration (or summation).[33] We'll now construct an "expert" version of A, whom we'll call A^+.

Since the posterior credence $\mathfrak{c}^A$ depends only which credence A reports, let $\mathfrak{c}(p|\mathfrak{a}(p) = x) = f(x)$. Take the distribution $\mathfrak{c}_A$, and let $c_A(x) = c_{A^+}(f(x))$. Since whenever A reports credence x, the agent updates to $f(x)$, she might as well pretend she heard some expert report $f(x)$. A^+ need not actually exist, but from a formal point of view, this maneuver allows us to use results about updating behavior when experts are involved to help us discover how updating works on advisors' opinions more generally. On any scoring rule $\mathrm{ES}_{\mathfrak{c}}(\mathfrak{c}^{A^+}) = \mathrm{ES}_{\mathfrak{c}}(\mathfrak{c}^A)$, so evidentially, A^+ and A are just as useful for the agent regardless of how you measure accuracy. A^+, in other words, has the exact same sensitivity-profile as A, where sensitivity measures how A's credence is thought (by C) to track p's truth-value. From the evaluative perspective, however, A and A^+ are markedly different. The agent expects the expert versions of non-experts to be *strictly more accurate* under any rule, as is easy to see from Theorem (3.A.1).

We now relate this idea back to peer and expert disagreement. Suppose an agent C-at-t_0 treats her future self C_1 as an expert. And suppose she thinks both A and B are peers

[33] You may be worried that (under the standard definition) conditional credences aren't defined for events of credence 0. I assume that for all credences A might have that are assigned non-zero density on $\mathfrak{c}_A$, we can arrive at a conditional credence either by taking conditional credence as primitive or by insisting that distributions be well-behaved enough to use standard tools of deriving $\mathfrak{c}(p|A = x)$ from $\lim_{\epsilon \to 0} \mathfrak{c}(p|\mathfrak{a}(p) \in (x - \epsilon, x + \epsilon))$.

(in the alethic sense) of C_1. That is, $\mathsf{Peers}(A, C_1; \mathfrak{c}_0)$ and $\mathsf{Peers}(B, C_1; \mathfrak{c}_0)$.

Suppose B is expert too. $\mathfrak{c}_0$ will follow the THRASYMACHUS PRINCIPLE and expects $\mathfrak{c}_0^{BC_1}$ to be bolder than both B and C_1.

Suppose that *A isn't* an expert even though she's taken to be a peer of C_1. In this case, we don't necessarily get to claim that splitting the difference will result in over-hedged credences because of disagreement-independent confounding factors. We do, however, have the resources to show that our agent should expect to be *even more opinionated* should a disagreement arise with A than should a disagreement arise with B.

Here's why: Consider A^+, the hypothetical expert version of A. As we've seen, $\mathfrak{c}_0$ expects that A^+ is more accurate than A is. Since A and C_1 are expected to be equally accurate, A^+ must be expected to be more accurate than both B and C_1 are. Thus, from the epistemically self-interested point of view of C_0, she'd rather learn the credence of some non-expert who's a peer of C_1 (*viz.*, A) than she would of some expert-and-peer of C_1 (e.g., B). In her quest to become more accurate, the former will—she expects—help her more than the latter.

C_0, then, has to expect that after updating on A's credence, she'd end up closer to the truth than she would after updating on B's credence. That will only happen if she expects to end up *more opinionated* by updating on A than she would by updating on B. Since both A^+ and B are experts, C_0 must in turn expect A^+ to be more extremal than she expects B to be. Again, we end up with a bolder form of conciliationism than is generally recognized.

Chapter 4

What Your Credence Tells Me About p

4.0 Introduction

A number of leading epistemologists have argued that factual disagreement often provides us with a peculiar sort of evidence that requires epistemic agents to revise their beliefs in special ways. When two agents share the same background evidence and nonetheless disagree, it looks like at least one of them has made some sort of rational error. Even if sometimes multiple doxastic attitudes are permissible, rationality is often restrictive enough to disallow a broad range of credences. So, when we disagree we sometimes get *higher-order evidence*, or evidence that we've made some sort of rational mistake. As David Christensen puts it at the start of his (2007):

> We all live out our lives in states of epistemic imperfection. Most obviously, this is true because the evidence on which we base our beliefs is limited. Only a little less obviously, we live in states of epistemic imperfection because we do not always respond to the evidence we have in the best way. Given that our epistemic condition consists in imperfect responses to incomplete evidence, part of being rational involves taking account of these sources of imperfection. (187)

Plausibly, our more orthodox theories of belief-revision, such as standard bayesianism, aren't equipped to handle higher-order evidence. It's often said that such theories apply

to ideal agents, and ideal agents don't countenance the possibility of their own rational error.

I think this sort of a response is unnecessary and has led the disagreement literature astray. One aim of this chapter is to defend treating evidence from disagreement as normal evidence that we can handle in standard ways. By and large, direct worries about our own rationality stemming from disagreement don't create any special problems for belief revision. Here's the basic idea. Instead of focusing on how disagreement provides evidence about what's rational to believe, we instead focus on what an advisor's credence tells us about p. That is, we should treat advisors as measurement devices of truth-values, or *alithometers*. Though such a perspective isn't new, I don't think it's been sufficiently appreciated, and I hope to show as much below.

The basic goal is to use this way of looking at doxastic attitudes to examine how evidence from disagreement should affect both an agent's credence in p and her evaluation of her advisor on a case-specific basis. By treating evidence from disagreement as normal evidence, I'll argue, we'll see that there can't be any simple, well-behaved updating procedure such as splitting the difference. Instead, how we should update hinges on specific details of the case, including the nature of our background evidence, fine-grained factors that go into our evaluation of our advisors, and the particular credence that our advisor reports. For contrast, I'll set this *variable weight view* against the *equal weight views* of Elga and Christensen, who think that our updated credence should depend, roughly, just on an agent's evaluation of her advisor going into the disagreement and their actual credences. By seeing where these views go wrong, we'll get a better and more complicated story about how evidence and evaluation work in disagreement cases. The VWV is neither steadfast nor conciliatory. Sometimes we should upgrade our respect for an advisor upon disagreement, and sometimes we should downgrade her. Though there won't be any recipe for how to figure out what to do, we can still get a rough picture of how ev-

idence works and give some advice to agents for handling evidence from disagreement.[1] This result may seem disappointing, but once we see disagreement as normal evidence, we shouldn't be surprised. We don't have any easy, actionable advice for handling evidence generally aside from very broad recommendations that tell agents to update by conditionalization on a good prior.

Here's the plan. After some initial stage-setting, I'll argue in §4.4 that treating evidence from disagreement as special is ultimately unworkable. First, I claim that higher-order evidence is everywhere and can't be separated off neatly from first-order evidence as Elga and Christensen require. Second, I'll argue that disagreement even under common knowledge of shared evidence can cause ideal agents who are certain that they're rationally perfect to revise their beliefs. In §4.5 we go into the trenches and examine more closely how evidence from disagreement works. We start by showing why the bootstrapping arguments against revising one's evaluation of an advisor based on her actual credence fail. This discussion will help us see how case specific our problem is and also how deeply epistemic evaluations are intertwined with expectations of what an advisor thinks. §4.5.3 identifies some key factors that determine how to update: *viz*., the relative resilience of one's evaluation on the one hand and of the background evidence on the other. In §4.5.4, we examine more rigorously how epistemic evaluation works on an alithometric perspective and check it against the informal discussion we've had till that point.[2] We wrap up in §4.5.5 with a discussion of the asymmetries between how an agent regards herself and her advisor as alithometers. Though there are some serious differences, I claim, we can use some of Elga's and Christensen's insights to gain a more neutral perspective that helps us learn from our own epistemic limitations.

---

[1]In the last chapter, we emphasized the reasons why the actual weight given post-disagreement has to depend on the credences reported in general and looked at constraints on updating over the course of many disagreements. In this chapter, we'll focus on conceptual reasons why the weight should vary in specific cases based on the nature of evidence we get from disagreement.

[2]Readers uninterested in the technical details may skip this section without much loss of continuity.

4.1 The Puzzle

To understand what has gotten the epistemological community so interested in disagreement, it's best to start with a few of the chestnut cases.[3]

RESTAURANT Five of us go out to dinner, and at the end of the night, the bill is $187.50. We agree to split the bill evenly and tip 20%. You and I do the math in our heads, and I become very confident that we each owe $43, while you're highly confident that we each owe $45. Since the two of us go way back, we're both aware that we're each equally talented at doing mental arithmetic, and neither of us is especially tired, or drunk, or distracted.

WEATHERMEN You and I are meteorologists. We were trained together and have a long track record that shows us to be pretty much equally accurate at predicting the next day's weather. We sit down and look at the same Doppler Radar images and privately arrive at different credences for rain tomorrow. I have credence .55, while you have credence .45.

In both of these cases, it looks like we have the same—or nearly the same—evidence. We at least share nearly all the evidence that seems relevant to the questions at hand. Furthermore, we consider the other equally competent as an epistemic agent over the relevant domain. So, it's natural to think that we should each give the other's credence quite a bit of weight. Indeed, it seems natural to say that we should give the other equal weight, since we have no apparent reason to take ourselves as more likely to have come to the right conclusion than the other.

But that's not all that's strange. As we see especially in the RESTAURANT case, we both have evidence that we've *mishandled* our evidence. After all, we both know exactly

[3]The following are adapted from Christensen (2007).

what the bill is and what the rules for dividing it up are. Given the rules of arithmetic, our evidence actually *entails* what the correct answer is.[4] The fact that you think we all owe $45 gives me reason to suspect my own response to the evidence in this case. That is, we have *higher-order evidence*: evidence that bears on the rational capabilities and performance of epistemic agents. Higher-order evidence looks peculiar and in need of special treatment, since it tends to undermine an agent's confidence in her own methods of coming to a judgment. Below, we'll examine whether this is so, at least in the case of disagreement.

4.2 Set-Up

The kinds of situations we'll look at are disagreements, which are fairly easy to understand, but it's worth going through them schematically. Two parties go into a disagreement over p with personal background evidence that bears both on p and on the epistemic capabilities of the other party. Then some news comes in: each agent learns the other's credence. They then update both their credences in p and their evaluations of the other person.

A few notes here: I'll be using 'disagreement' rather loosely. For our purposes, a disagreement occurs whenever the agents report different credences in p. Often, their background evidence will be largely or even entirely the same in cases of interest, but not always. Although this might depart from English usage, it seems like the natural epistemological category is this one.

Secondly, even though in real life, agents often report just categorical judgments instead of their actual credences (if such there be), it seems more natural to focus primarily

[4] RESTAURANT serves as a clean case of shared evidence but at the cost of requiring the agents not to be probabilistically coherent. This isn't, I think, worth worrying much about, since we can pull some technical trickery to still work within the spirit of the probabilist framework.

on the case where credences are shared. I'll be arguing that even at this level there isn't going to be any nice, well-behaved, and informative theory of how to update upon learning of a disagreement. If we focus on only categorical judgments, how we update will be even less well-behaved. Furthermore, by allowing more machinery into the mix, we'll allow ourselves a more nuanced perspective of how evidence in disagreement works. I will violate this policy periodically, but such cases will be the official focus of the discussion.

Third, I'll flag now that I'm understanding epistemic evaluative terms like *peer* based on purely alithometric criteria. That is, epistemic evaluation is based on how good of a truth-measurement device an agent takes her advisor to be. There will be a more thorough discussion of how this works below, but I hope the motivation is clear. We want to know how to revise our beliefs in cases of disagreement. It doesn't matter at the end of the day how smart or rational or clear-headed the advisor is unless these properties lead her to generally accurate credences. Epistemic agents gather evidence about p to get their credences closer to p's truth-value, so the evaluations that matter most for belief revision are truth-focused.

4.3 Conciliatory Views

Philosophers like Adam Elga and David Christensen advocate *conciliatory* or *equal weight* approaches, which I'll be discussing throughout this paper. According to this class of views, if you disagree with a peer, especially one who has the same evidence as you, you should give that person's opinion roughly equal weight to your own. The basic intuitive motivation—which I'll develop in more depth as we proceed—goes as follows. First, you took yourself and your advisor to be antecedently in equally good positions to determine whether p. Second, your disagreement after evidence sharing should cause you to doubt that you handled the evidence correctly, given your antecedent respect for your peer. So, it would be epistemically irresponsible for you not to worry that you're the one who

made the error, since she's just as good of a reasoning machine as you are. So, since your own reasoning based on the first-order evidence is now cast into doubt, you should reduce your reliance on it. Indeed, you shouldn't favor it at all, since you take your peer to be just as good as you at figuring out which sorts of reasoning should be relied upon. What you're left to go on, then, is just your assessment of the two of you pre-disagreement. Favoring yourself amounts to epistemic chauvinism, since you have no good epistemic reason for doing so. Favoring your peer suffers from the parallel vice of undue epistemic humility. Therefore, you should give yourself and your peer equal weight.[5]

One hard issue arises for conciliatory views when we see that there are clear cases in which I'm justified in downgrading your view despite my normally high opinion in your competence. Imagine, for instance:

Extreme Restaurant Just like before, except I think we owe $45 each, while you think we owe $450 each, which is more than the total bill.

School Play You and I are normally peers when it comes to judging acting talent. I see your kid in the school play and think she's terrible. You think she's wonderful. We compare notes.

In both of these cases, it would be odd to think I had to give your opinion roughly equal weight. Somehow or other, I'll need to be able to stop treating you as a peer in this instance once I learn your opinion.

The pressure from these cases is in tension with the countervailing pressure of epistemic humility. Especially in cases where we share evidence, my own rational abilities are called into question upon learning your opinion. Somehow or other, we have to

[5]One quick caveat: It's not always clear exactly what "equal weight" means. Generally, it means moving your credence significantly toward your peer's. Indeed, both Elga and Christensen think that your post-disagreement credence should often be the average of the two pre-disagreement credences. However, this view, which I'll call "Split the Difference", shouldn't be identified with conciliatory views generally or the equal weight views that Elga and Christensen ultimately endorse.

make due with our now epistemically indicted credence function that takes account of our own fallibility while finding a way not to give undue weight to opinions we should be able to identify *post hoc* as defective even if we thought the respective agents were our peers before we learned their actual opinions. Both Elga and Christensen provide some guidance. It's quite hard to cash out either proposal fully, so for now I'll just try to get the basic ideas on the table.

4.3.1 Christensen's Approach

Christensen (2007) tries to separate out EXTREME RESTAURANT and SCHOOL PLAY from normal cases of peer disagreement by appealing to what he refers to as an Independence criterion:

INDEPENDENCE In evaluating the epistemic credentials of another's expressed belief about p, in order to determine how (or whether) to modify my own belief about p, I should do so in a way that doesn't rely on the reasoning behind my initial belief about p. (From Christensen 2011)

The idea here is that since I went in to the disagreement thinking you were a peer, and since your disagreement with me calls my reasoning into question, I should only be able to downgrade you if I can find some reason to do so independent of why I arrived at my initial credence in p. In both SCHOOL PLAY and EXTREME RESTAURANT, I (arguably) have such independent reason. In one, I can reason that the actual share must not be higher than the original (which is not a premise I relied on essentially in my original calculation), and in the other, I can rely on the independent reasoning that you're likely biased in favor of your own child, since I know that people are generally biased in favor of their own children. INDEPENDENCE then is supposed to allow me to throw out or downgrade an opinion so long as I can do so without relying on reasoning that's now in serious question.

4.3.2 Elga's Approach

Elga provides similar criteria for downgrading advisors *ex post*, which he refers to as the "Equal Weight View." In order to remain more ecumenical and allow other versions of EWV, I'll change the name:

COARSENING Upon finding out that an advisor disagrees, your probability that you are right should equal your prior conditional probability that you would be right. Prior to what? Prior to your thinking through the disputed issue, and finding out what the advisor thinks of it. Conditional on what? On whatever you have learned about the circumstances of the disagreement. (From Elga 2007, 490)

Elga's suggestion is to take what you *would have thought* about how well each of us would do in forming an opinion about p if you hadn't yet thought through the issue and knew instead only about the "circumstances of the disagreement." As one would expect, it's painfully difficult to figure out what gets to count as circumstances of the disagreement, though Elga provides some guidance. Of especial import is that we don't get to count detailed facts about our actual reasoning as a circumstance of the disagreement. He, like Christensen, is concerned that if we get to rely on our antecedent reasoning too much, we'll end up over-privileging our own views by supporting them through what we should now worry is faulty reasoning and retain something too close to our initial opinion. As he puts it:

> In general, circumstances of disagreement should be individuated just coarsely enough so that the relevant conditional probability judgment is genuinely prior to your reasoning about the disputed issue. (This coarseness constraint is what makes the…view nontrivial. For otherwise—if the view simply required that one's new opinion should equal one's prior opinion, conditional on all of one's new information—the view would be tantamount to the requirement that one conditionalize on one's new information.) (490)

Facts about whether your answer in EXTREME RESTAURANT is greater than the total bill gets to count as a circumstance of the disagreement as do general facts about tenden-

cies to favor one's own children, since neither require anything like a detailed specification of the actual reasoning. On the other hand, facts about how I arrived at a given weather forecast are case-specific and require me to trust the actual methods I use too much.

While both Christensen's and Elga's are still underspecified, I hope the basic idea is clear enough to convey the basic spirit of the conciliatory policies advocated in the literature. The biggest thing to notice is that they both treat evidence from disagreement in a different way from standard evidence. Usually, you can appeal to your entire evidential base, but for both Elga and Christensen, some evidence has to be "bracketed" now that your own reasoning is suspect.

4.4 Higher-Order Evidence: The New Normal

In this section, I want to push the line that evidence from disagreement is normal evidence. First, I'll argue that higher-order evidence is ubiquitous and cannot be separated from first-order evidence. That is, higher-order evidence is everywhere, and we usually can't divide our evidence up into first- and higher-order component pieces. Therefore, it's misguided to insist on treating it in a categorically different way from normal evidence, since no rule that requires alternative treatment could be effectively applied. Second, I'll argue against the received view that a rationally perfect agent wouldn't update upon disagreement under shared background evidence. I argue that, at least in some cases, such an agent (even if she knows she's ideally rational) may even end up giving extra weight to her less rational counterpart, since the sub-ideal agent's credence could still be informatively sensitive to whether p. Disagreement, I claim, can be evidence for everyone.

The basic point here is that we need to separate sensitivity to whether p from rational credence in p. While the latter can be useful instrumentally, it's the former that really matters for the question of updating. Thus, while certain kinds of alithometers take

truth-value measurements through the proxy of finding a rational credence, the question of discovering what the rational credence is loses much of its theoretical import when we're determining how to update. Learning you have an irrational credence is similar to learning your credence doesn't match the chance of p. In both cases, you'll use such information to update, but there aren't any special problems that arise that require treatment through new theoretical means.

The upshot is that—by showing we can treat higher-order evidence as normal evidence—we can justify our use of standard tools for addressing the problem of incorporating information from disagreement and can bypass many of the difficulties afflicting the alternative views on disagreement that require special theoretical apparatus.[6]

4.4.1 The Ubiquity and Inextricability of Higher-Order Evidence

It's natural to think that the most interesting cases of disagreement arise when two agents consider themselves equally competent over the relevant domain and share the same evidence. I agree that this species of disagreement is of special interest, but I think our focus on it has led us astray. By turning attention more generally to the question of how learning an advisor's credence should affect your own posterior credence in p, we'll be able to gain new insights into the special case of interest and into the right way to treat higher-order evidence more generally.

When I learn your credence in a proposition, I don't also necessarily get to share all my evidence with you and learn all the evidence you have that I don't have. Instead, I just learn what you think about p. That is, as far as I know, you may know some things I don't, and I may also know some things you don't. When I learn what you think about p and I see your credence is different from my own, your credence gives me evidence

[6]Some such alternatives include the following: Christensen (2007, 2011); Elga (2007); Lehrer and Wagner (1981); Roush (2009).

both that I've made some rational errors and that you know some things I don't know. Consider:

SLEUTHS We're two detectives working on the same case but in different parts of the city and have interviewed overlapping but non-identical sets of people. We come back with very different credences that the butler did it, with me at .9 and you at .1. (As usual, I take you to be about as good as I am as a detective.)

In SLEUTHS, I have evidence that there's a bunch of exculpatory evidence I'm missing. However, your different credence gives me *some* evidence that I mishandled my original evidence. After all, you're generally a competent detective, and it's relatively rare that there's strong evidence for the guilt and innocence of the same person. It's even rarer that I would just happen to get the strong guilt evidence and you'd just happen to get the strong exculpatory evidence. What's more likely is that my evidence for guilt wasn't as strong as I'd taken it to be. So despite the fact that we have distinct sets of evidence, our different credences here still point both to new evidence and to our mishandling old evidence, since both partially explain the discrepancy in our judgments.[7]

Note that in SLEUTHS, there just isn't any good way of dividing the evidentiary significance of my knowledge of your credence up into the "higher-order" contribution and the "first-order" contribution. The fact that you think something different from me can't be carved up so easily along these lines. It's in this sense that first- and higher-order evidence are *inextricable*.[8]

Most real cases are like SLEUTHS. We typically have (near) common knowledge of

---

[7]Let H_1 be the claim that I mis-handled my evidence and H_2 be the claim that I'm missing a bunch of exculpatory evidence. Both H_1 and H_2 raise the probability of our having very different credences (E). So, when I learn E, I should raise my credence in both H_1 and H_2.

[8]One could try to do some fancy bayesian analysis of decomposing the evidentiary significance of your credence into how much is expected to come from differences between what first-order evidence we have and into a part that comes from expected second-order contributions. However, trying to work this out seems massively complex and, as I hope you'll agree, likely a wrong-headed approach.

some evidential base, but we're not entirely sure what the other person's total evidence is.[9] In these sorts of cases, there won't be any good or clean way of splitting the evidence from disagreement up along the lines of the hierarchy of first- and higher-order. Insofar as we can't separate evidence into these two different components, we should be wary of having separate policies for handling first- and higher-order evidence.

More generally, if we really push this line, we can see that in most circumstances when I have a credence in p that turns out to be far from p's actual truth-value, I gain some evidence that I have mishandled my original evidence. Handling evidence in the right way leads to general epistemic success. That is, agents who handle their evidence well tend to have credences that are close to actual truth values in the long run. So, suppose my credence in p when I know E is .2, and I'm not certain that I handled E correctly. I then learn that p at t'. At t' I should become more worried that I initially mishandled E, since I generally expect to be closer to p's actual truth-value when I handle evidence in the right way. But that means that even with a little self-doubt, I'm getting higher-order evidence all the time.

Since there's no nice and easy separation of cases into those where disagreement gives you second-order evidence from those in which you gain normal first-order evidence, no good theory of how to resolve the pure second-order cases (like RESTAURANT and WEATHERMEN) can be attractive unless it fits in nicely with a theory of how to update upon learning somebody's credence when you're unsure of her evidence.

This observation should serve as a bit of a warning to proponents of theses like INDEPENDENCE and COARSENING, which advise agents to update by, in essence, tem-

[9]Indeed, one might doubt that two people really could have exactly the same evidence, at least in practice, since it's generally quite difficult to demarcate which propositions one knows have any evidential relevance to a given proposition. Furthermore, two agents could have the same ur-prior and have conditionalized on exactly the same E but arrive at different posteriors if either has, say, Jeffrey-conditionalized. The problem is exacerbated further if one takes a "fine-grained" view of evidence that allows reasoning and inferences to be part of the evidential base. I won't pursue either of these worries below but highlight them here to cast additional doubt on the methodology of focusing on idealized cases with the exact same evidence.

porarily impoverishing the first-order evidential base and relying much more heavily on second-order considerations. Cases like SLEUTHS show that it's very difficult to separate first- and second-order considerations and that relying too heavily on that distinction is ill-advised. Instead, we have reason to try to assimilate our treatment of higher-order evidence to our treatment of first-order evidence, at least in these everyday sorts of cases.

We also should be wary of the idea that higher-order evidence should change our normal epistemic policies too drastically. We're nearly always acquiring some higher-order evidence in the course of normal learning, since we're usually not entirely sure we handled our initially evidence correctly. Once we learn more about whether p, we get new information about our own rational abilities. It would be quite odd if we had all been blind to the need for new methods of evidence-handling, especially given how successful many sub-ideal but still really smart people are epistemically. These sorts of considerations give some weight to views that are reluctant to let higher-order evidence boss the credence function around too much.

Of course, for all I've said, conciliation could still end up the right policy when we have shared first-order evidence and we still disagree. However, such views shouldn't be motivated through strong appeal to the peculiarities of higher-order evidence. Higher-order evidence is ubiquitous and mixes inextricably with first-order evidence in our normal epistemic lives.

4.4.2 Ideal Rationality and Disagreement

In this section, I'll argue that disagreement can provide useful evidence even for ideal agents. The idea is that, by treating others as alithometers, even ideal agents can exploit potential sensitivity to whether p that the advisor might have to gain information about whether p.

This claim—that disagreement can be evidence for everybody—goes against not just

Elga and Christensen but many of their opponents as well.[10] I think this received view stems from a conceptual confusion. The epistemologist's task is often to determine what the most rational thing to believe is, while an epistemic agent's task is to determine whether p. Even though epistemic agents might find facts about what's rational useful in determining p's truth-value, that's not their ultimate concern. By conflating questions about the agent's rationality with questions about her sensitivity to whether p, the importance of rational error ends up overhyped.

4.4.2.1 Uniqueness

This conflation manifests itself most prominently in the surrounding debate over the UNIQUENESS thesis, which for our purposes is:

UNIQUENESS For any agent with any body of total evidence E and for any proposition p, there's exactly one maximally rational credence for the agent to have in p.

Some, such as Kelly (2010); White (2005); Enoch (2010), and Christensen (2007), think that the fate of UNIQUENESS largely determines the viability of conciliatory views. I think that UNIQUENESS is an orthogonal issue that has little relevance to disagreement. By showing why UNIQUENESS doesn't matter, we'll be able to see more directly how disagreement can be evidence for any kind of agent.

First, suppose (1) we both know UNIQUENESS is false, (2) your credence in p is y and mine is $m \neq y$, (2) it's common knowledge that we have the exact same evidence E, and (4) we know we're both perfectly within the permissible range of credences. That is, even though we both have the same evidence and have different credences, we've both responded to E in a maximally rational way.

---

[10] See, e.g., Enoch (2010); Kelly (2005, 2010); White (2009).

Here's how I should reason about your credence. You are a fairly competent evaluator of this sort of evidence. If p is true, I should expect you to have a higher credence in p than if p is false because in general I think you're pretty good at determining whether p. That is, your credence is more or less responsive to whether p. For the sake of exposition, let's suppose $y > m$, and let $\mathfrak{b}$ be my initial credence function and let Y be a random variable representing your credence. The following is a perfectly reasonable thing for me to think: $\forall k \in (0,1)[\mathfrak{b}(Y > k) < \mathfrak{b}(Y > k|p)]$.[11] In English, that says that I think it's more likely that your credence is above a given fixed point k that's between 0 and 1 supposing p is true than I do unconditionally. Now, once I think this, then upon learning that $Y = y > m$, I should adjust my credence upward. Even without UNIQUENESS, and even when our credences are both in the permissible rational range, your credence still carries information about whether p.

Of course, that's not *always* the case. I could imagine that you were completely rational and could instantly identify the permissible range of credences in p for E and then would simply pick a credence from that range at random. In such a case, once I know what the rational range is, your credence won't matter to me, and I can stick to my guns. However, this sort of a case isn't common. In the real world, we're struggling to make our credences as close as we can to p's actual truth-value. Determining what sorts of responses to evidence are rational can help us along the way, but only instrumentally. That is, we expect that we'll do better, generally, by respecting the constraints that rationality imposes on our response to evidence than we will through some alternative means. Your credence doesn't just give me information about how rationally to respond to E; it also gives me information about whether p! It's because of what I take to be your general sensitivity to whether p—not just your sensitivity to what meets the norms of epistemic rationality—that I should in general be willing to update regardless of whether

[11] We can obviously weaken this condition and still have the desired result in many cases.

UNIQUENESS is true.

Now, we see that if two agents are both perfectly rational and UNIQUENESS fails, they still may update when they disagree even when they know they share exactly the same first-order evidence. *Pace* Elga and Christensen, even ideally rational agents who are certain that they're ideally rational will sometimes change their credence upon learning of a disagreement.

If UNIQUENESS holds, would an ideal agent certain of her abilities still update on disagreement? I think she would, at least sometimes. All that matters is that she thinks her interlocutor is sufficiently sensitive to whether p. She may know that the optimal method of credence formation will return her own credence with E as an input. When she learns of her advisor's disagreement, she'll know her advisor is using some inferior method. But an inferior method isn't a useless method that's insensitive to whether p. Therefore, she'll sometimes be able to exploit her advisor's sensitivity—inferior though it may be—to gain information as to whether p. Compare: We could have two measuring devises, one of which is top-of-the-line but still not 100% accurate, while the other is a bit outdated. When we get two different measurements from the devices on the same input, the inferior one still provides some useful data about the true-value of the quantity in question.

Indeed, I think that sometimes a rationally ideal agent may end up giving extra weight to an agent who she knows is following an epistemically inferior policy. Here's a case to show how this could work out supposing UNIQUENESS holds:

MAX AND EVE Max and Eve are two young but precocious children. Max was born and is aware that he was born with the uniquely most epistemically rational ur-prior. Since coming out the womb he has always and only updated by conditionalization, guaranteeing his continued maximally rational epistemic performance. Unfortunately for Max, he still doesn't have an especially accurate credence function, since he is quite young and hasn't acquired much information about mat-

> ters like human psychology and the physics of medium-sized dry goods. Eve, on the other hand, doesn't have the most rational urprior and instead is unjustifiably opinionated about these matters. Her creator, Evolution, knew lots of stuff about the actual world and deviously programmed her—unbeknownst to either child—to have an accurate but unjustified urprior in certain important physical and psychological facts. Max and Eve meet every day on the playground to bet their lunch money on things like how other children and physical objects will behave. To make matters fair, they share all their background evidence. Max initially scoffs at Eve's brazenly extremal credences, but over time he notices that she has been winning lots of his money.

In this case, Max rightly begins to give more and more weight to Eve despite her epistemic inferiority. He's learned through induction that, even though she doesn't have sufficient *epistemic* justification for her credences, she is still sensitive to the truth-values of the propositions under concern in a way that provides Max additional information.

I set up the details of this case under the assumption that evolutionarily programmed credences are, in the relevant sense, less rational than those arrived at by more standard recommendations of objective bayesians. However, these assumptions aren't mandatory. Once we concede that the rationally ideal credence function isn't extremal, we know that some other credence functions will as a matter of fact be more accurate. Those other credence functions may have some sort of epistemic imperfection about them, but that doesn't necessarily prevent the hapless ideal agent from learning (or at least gaining evidence) that they'll tend to beat her own credence function. At the least, it doesn't prevent her from thinking that these less rational agents will provide her with information about whether p even under a common evidential base E.

Of course, ideal agents might not always, or even usually, revise their credences in cases of disagreement, but when they do, it doesn't seem like their reasons have to be different in kind from our own. Both ideal and sub-ideal agents update based on the

sensitivity to p they expect their advisors to have.

4.4.2.2 Summary

We've seen (1) there's no good way of separating first-order evidence from higher-order evidence in many situations, (2) higher-order evidence is everywhere, and (3) disagreement can be evidentially useful even for rationally ideal agents. I think we have enough to be skeptical, at this point, of any treatment of disagreement as evidentially special.

4.5 Bootstrapping, Question-Begging, and Evidential Interaction

In this part, the discussion will focus on the evidential effect of an advisor's reported credence both on the assessor's posterior credence in p and on her posterior evaluation of the advisor. I'll argue that there won't be any well-behaved theory of how the antecedent expected epistemic performance of an advisor ought to determine an updated credence in p upon learning the advisor's actual credence. That is, the *ex ante* evaluation doesn't do all that much to constrain what $\mathrm{Cr}(p|Y = y)$ is for arbitrary y. This view, as stated before, will be neither steadfast nor conciliatory. Instead, I'll claim that many case-specific factors will determine how one should respond to a disagreement. Sometimes, for instance, one will end up giving an expected pre-disagreement peer little weight after learning her credence, while other times, one will end up nearly entirely deferring to her. Furthermore, in keeping with the internalist spirit advanced so far, these responses won't depend on what really is the right thing to do. We'll begin with a discussion of Elga's and Christensen's charges that changing one's evaluation of an advisor merely based on her credence are epistemically unjustified and then look at why those arguments fail. We then look more specifically at epistemic evaluation from formal perspective to get further insight on how it's tied to expected information about whether p. Next, we discuss

evidential resiliency and how it affects updating. Last, we discuss how and whether agents can regard themselves as alithometers.

4.5.1 Elga on Bootstrapping

I'll be arguing below that we can often change our evaluation of an advisor based solely and merely on what her credence turns out to be. At first glance, there's something funny about this view, and I think Adam Elga nicely expresses what he finds worrisome: it appears that using the disagreement to re-evaluate advisors allows agents to beg the question in their own favor and to bootstrap their way into extreme confidence in their own opinions.[12]

To see how such a worry gets spelled out, let's look at how Elga's bootstrapping objection against steadfast views goes:

> Suppose... you and your friend are to judge the truth of a claim, based on the same batch of evidence. Initially, you count your friend as an epistemic peer—you think that she is about as good as you at judging the claim.... Then the two of you perform your evaluations. As it happens, you become confident that the claim is true, and your friend becomes equally confident that it is false. When you learn of your friend's opposing judgment, you should think that the two of you are equally likely to be correct. ...
>
> If it were reasonable for you to give your own evaluation extra weight... then you would have gotten some evidence that you are a better evaluator than your friend. But that is absurd. (2007: 487)

The absurdity is brought out more perspicuously when we consider how a policy of giving oneself extra weight would play out:

> Suppose for *reductio* that whenever the two of you disagree, you should be, say, 70% confident that your friend is the mistaken one. It follows that over the course of many disagreements, you should end up extremely confident that you have a

[12]Other conciliationists, such as Christensen (2011) and Feldman (2007), rely on similar motivations. What I say below should answer their worries as well.

> better track record than your friend. As a result, you should end up extremely confident that you are a better evaluator. But that is absurd. Without some antecedent reason to think that you are a better evaluator, the disagreements between you and your friend are no evidence that she has made most of the mistakes. (*ibid*.)

The general spirit of this arguments is quite appealing, and it works well against views that allow for automatic self-favoring. The essential problem is that Elga only considers what I'll call *fixed weight views*. That is, he assumes that the weight one is supposed to give to an expected peer should stay roughly constant. We see this assumption come out most explicitly in Elga's *reductio*. He assumes that one is supposed to be x% confident in one's own initial view for some fixed x. He then argues that, for a peer, the only reasonable value for x is 50.

I instead favor a *variable weight view*. On variable weight views, the amount of weight given to a peer can vary with the background circumstances, the credence she reports, and the pre-disagreement evidence one possesses.[13] All of these will be explored in more detail below, but we can already see how to block the *reductio*. In repetitions, there's generally not a single level of confidence that an agent should retain in her pre-disagreement views, nor should her assessment of her advisor remain constant.

I'll resist the arguments for conciliatory views primarily by claiming that the evidential significance of disagreement can't be decomposed as neatly as its proponents believe. To see where we'll be parting paths here, we can look at the following passage from Elga (2007) where he explains his approach to the problem of disagreement:

> How should one take into account the opinions of an advisor who may have imperfect judgment? That question factors into two parts:
>
> 1. To what degree should one defer to a given advisor's judgment? For example, when should one count an advisor's judgment as completely worthless? Or

[13] Again, the formal reasons variable weight is necessary are covered in the previous chapter. I hope here to make a more conceptual case for variable weight that develops the view and also exposes in more detail where arguments for fixed-weight views go wrong.

as approximately as good as one's own? Or as better than one's own, but still less than perfect?

2. Given one's assessment of an advisor's level of competence, how should one take that advisor's opinion into account?

> On the first question, I have no substantive answer to offer here. My excuse is that the question concerns a huge, difficult, and domain-specific matter. How should one judge the epistemic abilities of weather forecasters, dentists, math professors, gossipy neighbors, and so on? This is a question with the same sort of massive scope as the question: "When does a batch of evidence support a given hypothesis?" Fearsome questions both, and worthy of investigation. But leave them for another day. Here I will focus on the second question. Assume that you defer to an advisor's judgment to a certain degree. Given that rating of the advisor's judgment, how should you take her opinions into account? (p. 483)

The problem, I claim, is that the question doesn't factor into two parts. One's assessment of an advisor's level of competence should sometimes be affected by the actual credence she reports. The remainder of this part of the essay will examine how this evidential web may work.

4.5.2 Expectations of Opinions

To see how accuracy-theoretic evaluations can be intermingled with expectations, it's best to start with an extreme case. Recall the following example from last chapter:

SUNRISE I ask Tom what his credence that the sun will rise tomorrow is.

I consider Tom—and indeed, nearly everybody I ever come in contact with—roughly a peer when it comes to the question of whether the sun will rise tomorrow. The reason I hold such an egalitarian view is almost entirely based on my expectations about what these people think. My distribution over their potential opinions has nearly all its weight right around 1, and for pretty much that reason alone, I think they'll be roughly as accurate as I am. So, if Tom reports a credence of .97, I will legitimately downgrade my respect for him because his credence is obviously too low.

From a formal perspective, it should also be clear that evaluations can depend on expectations of opinion. When I'm evaluating how accurate I think you are before I learn your credence, I have to take a weighted average of how I accurate I think you *would be* were you to have credence y discounted by how probable I think it is that you have credence y. If I think you are very likely to be within a small range of credences, then my overall evaluation of you—how accurate I expect you to be—can depend on this antecedent expectation about what you think.

But once we concede this point, we have the ingredients to undermine bootstrapping arguments like the one above. When my evaluation of you depends on expectations of what you think, learning that you think something else is all the information I need to change my evaluation. I was nearly sure Tom had credence at least .9999 that the sun will rise, and my evaluation of him as a peer was based primarily on this expectation, so I didn't engage in any sort of illegitimate bootstrapping.

4.5.2.1 The Spectrum of Cases

To see further how evaluations interact with expectations of opinions and constrain updating, we should look at some less extreme cases like the following:

Election Forecasting You and a friend Emily are both well-trained statisticians hired to forecast the probability that a given candidate BO will defeat his opponent MR. Though you don't know Emily's exact track-record, you have enough evidence to expect her to be about as good as you at the job. Each day, you both look at the same data and independently arrive credences that BO will win and then compare your verdicts. The recent polls vary to some degree, but BO has lately been hovering around 3 points ahead, and there are still two months before the election. Today, your credence is .5 that BO will win before learning of a disagreement.

Now, we can imagine two reasonable people arriving at somewhat distinct credences in this case. However, it would be clearly crazy to have credence close to 1 that MR will win. Presumably, even if you thought Emily was your peer going in, you'd feel no pressure to give her views equal weight were she certain MR would win. Of course, she shouldn't have to be that extreme for you to discount her opinion. If she had credence .9 that MR would win, then you presumably could discount her still. After all, it seems clear her credence goes well beyond the evidence. The problem is that we can sorites this example, and there won't be anything like a clear point between crazy and non-crazy. What's more, there clearly shouldn't be. In this case, there's some sort of smooth transition between discounting her opinion and treating her opinion as worthy of the normal respect you give to her. Your respect for her verdict in this particular case should *decay* as it gets closer to 1.

On our picture, the explanation is roughly that, conditional on Emily having a credence within a certain range, you're willing to give her a decent amount of respect. Since you know she's fairly well-trained, you expect her to have an opinion within some reasonable range of .5. You can then antecedently assign her a relatively high level of expected accuracy, but you'll gradually lower this expectation as her credence gets more extremal.

Let's compare this picture with the one we receive from conciliationists like Elga and Christensen. They want to disallow appeals to certain bits of evidence or reasoning for the purpose of changing levels of respect for an advisor. In cases like these, though, it's hard to see how we could get the gradual decay we desire with these prohibitions. For one, in a case like ELECTION FORECASTING, there is no clear line of reasoning from the data to your credence. You have a number of statistical analyses, but your ultimate judgment is based on your well-honed nose instead of a premise-conclusion argument. It's also not the case that you have any real argument against a fairly wide range of alternative credences; they merely strike you as a bit too high or too low. So, when you downgrade Emily when she's still in the generally sane range of credences, you have no independent

grounds for doing so.

We should remind ourselves that instead of respect decay, there can be respect growth with distance from one's own credence.[14] We can imagine an alternative case, where you and your friend have lots of—what seems to you—ambiguous data. You're not especially sure what to make of the data, so you set your credence that BO will win to close to .5. Emily, on the other hand, has credence .8 that BO will win. In this case, you might want to upgrade your respect for her. After all, she's a competent statistician, so it's unlikely that she would have such a high credence without good justification. Since you didn't know what to make of the data, you might give her extra weight, since you antecedently took her to form credences responsibly.

Indeed, cases with ambiguous but informative background evidence often do result in respect growth. A perspicuous example is the following:

Murder Mystery Redux SH and HP are two equally competent detectives working on a case. HP has credence .5 that the brilliant criminal known as Professor M is the mastermind behind the murder who left a series of tantalizing and mysterious clues to toy with the police. SH is going over the evidence with HP and stops mid-sentence to announce that he has just solved the case and determined that the Professor was indeed responsible.

HP, despite taking SH to be his peer antecedently, should give his colleague more than equal weight. SH is very competent at his job and would never become so confident without good reason.

In these sorts of cases, the actual credence reported ought to affect the level of respect given. However, it's not the case that respect will always rise or fall with the disagreement but will instead depend on further background beliefs about the evidence and the advisor.

---

[14]For much more, see the discussion of the Thrasymachus Principle in the last chapter.

We see, then, that merely thinking that somebody's a peer pre-disagreement does little to constrain our updating behavior. Of course, this line is generally compatible with the position that we should—on average—give equal weight to our peers. This latter claim is, however, more or less trivial on our analysis. We're defining *peer* as somebody you expect to be roughly as accurate as you yourself are. So, if you don't give equal weight to your expected peers in general, then you're either bad at forming opinions about what their judgments are likely to be (when your evaluation depends on these expectations), or you end up violating your pre-disagreement conditional credences.

Now that we've explored the interconnectedness of evaluations and expectations of opinion, we should look at other factors that have a large role to play in constraining updating. As I've said, I think that how one updates for any given disagreement should very much depend on the specifics of the case. Still, one important element—that of resilience—is worth remarking on.

4.5.3 Resilience, Decay, and Screening Off

In this section, we'll look at the interplay between the volatility of the level of respect one has for the advisor, on the one hand, and the degree to which one's evidence locks one in to a particular credence.

It will be useful here to talk more directly about the weight one gives an advisor. Suppose my prior in p is m, and I find out that $Y = y$, and I end up with some sort of weighted average of our two priors. We can represent my posterior credence as $\mathfrak{b}^*(p) = \lambda m + (1-\lambda)y$, where $\lambda \in [0,1]$.[15] The *lower* the value of λ, then, the *higher* the level of respect the assessor actually ends up giving to her advisor.

Elga and Christensen *inter alia* think that λ should be more or less fixed by one's *ex*

[15]Of course, sometimes I might end up with λ outside of this interval, but we can restrict attention to this more normal case.

ante expectations of an advisor's accuracy, but as I've argued, nothing that simple can be correct. Instead, on variable weight views, λ won't generally be a constant, but instead a function that depends on the actual value of Y. Two features that will do a lot to determine the behavior of λ are (1) the resilience of your background evidence E^- where E^- is your total evidence relevant to p aside from evidence about your advisor's epistemic performance, and (2) the resilience of your estimate of the epistemic performance of your advisor.

This needs some cashing out. Let's start with resilience. Some evidence can do more to "lock" you into a particular credence than others. To take an extreme case, suppose you know the actual chance of p is x. Then outside of pathological cases, you ought to set your credence in p to x regardless of what else you learn. It's very hard to get a rational agent to budge much after she learns what the chances are. In other cases, you can be easily shaken from your initial credence in p. Suppose I don't know whether a given coin is biased. I start out with credence .5 that on the next toss, it will land Heads. Because of my relative ignorance, a little information about the coin's past history can have serious effects on my credence. For instance, if I had a uniform distribution over all possible biases and discovered that the previous toss resulted in heads, I'd update to credence $2/3$ in Heads from my original $1/2$.

I don't here wish to enter into a technical discussion of measures of resilience, but I hope to make the intuitive concept clear: the resilience of a given piece of evidence is roughly the degree and extent to which it can screen off other kinds of evidence, weighted by the probability of that evidence. In other words, E is resilient with respect to p if for lots of potential evidence F, we have $\mathfrak{b}(p|E) \approx \mathfrak{b}(p|EF)$.[16]

As I've argued, we should suspicious of the "special status" of higher-order evidence, so as expected, the resilience of the evidence provided by an advisor's credence behaves

[16]See Skyrms (1980) for a thorough-going discussion.

similarly to normal evidence. Often, if you have lots of data about your advisor, then learning her credence can end up more or less screening off a bunch of other evidence you had before talking to her. If instead you don't know all that much about your advisor, your other evidence can allow you to discount her opinion more heavily than you otherwise would. Let's look at some cases to see how this might work.

Recall the WEATHERMEN from above. Here, we set up the case so that I have very extensive information about your track-record. We can suppose that I have a detailed record of all our past forecasts, and our average Brier score is the same. Let's add that there are also no odd facts about your performance record that might be of import to me. For instance, you don't ever take mind-altering drugs that make you give weird predictions every so often. You aren't particularly inaccurate when you forecast between .7 and .72 probability of rain. You don't ever get depressed and become overly pessimistic about the weather. Indeed, we can imagine you and I are both perfectly well calibrated, so that when we've forecast an x percent chance of rain, then $100x$% of the time it has rained. My background evidence E^- is moderately resilient: after years as a meteorologist, I get pretty good estimates of the chance of rain from Doppler forecasts, but there's still a fair amount of wiggle room as more evidence comes in. In this case, plausibly, the resilience of my respect level for you can nonetheless overpower this only moderate resilience of the background data. Because of my detailed knowledge of our relative track-records, I'll end up with a roughly constant λ value for you, since I know quite a bit about how you take measurements of the truth-values of propositions about rain in the near future.

Compare this case to the case of ELECTION FORECASTING above. In that case, you know Emily's a good statistician, but you don't have detailed knowledge of her track-record. Furthermore, you have polling data, which is somewhat robust, but you also know lots of other factors might influence the election outcome, like the fate of the European economy or the actions of belligerent nations. Both your evidence E^- and your estimate of Emily's abilities are moderately resilient. You can be pushed around

some, but it'd be tough for any new evidence to make you close to certain that a particular candidate will win or that Emily is complete idiot. You may reasonably be willing to nearly split the difference for credences that Emily reports between, say, .35 and .65, but since E^- is sufficiently robust, you won't come near to splitting the difference with Emily when her credence is .95.

We can then give a rough plot of these sorts of cases. We have along one axis the resilience of E^-: how much my evidence aside from facts I know about my advisor tends to lock me in to credences near my current one. Along the other axis, we have the resilience of my respect for my advisor: how close λ will be to a constant as my advisor's credence varies. Cases like WEATHERMEN allow for high respect resilience. Cases where I know the chance of p antecedently have high resilience for E^-. Cases like ELECTION FORECASTING are somewhere in the middle along both dimensions.

4.5.4 How Epistemic Evaluation Works

Above, we saw that sensitivity to whether p can come apart from what's rational to believe on the evidence. Though the latter might better track English usage of evaluative terms like *peer*, *inferior*, *guru*, etc., it doesn't fit well with our alithometric perspective. Instead, for our purposes, epistemic evaluation focuses solely on how accurate an advisor is or is expected to be.

Indeed, it's how accurate the advisor is *expected* to be that really matters when we're interested in how to update upon discovery of an advisor's credence. We don't actually know whether p, so actual accuracy isn't of much use. More specifically, we take someone to be a *peer/superior/inferior* based on whether we expect her to be roughly as/more/less accurate than we are when it comes to forming a credence in p. The formal details are worth looking at in a bit more depth to see how evaluation and updating can play off one another.

A formal analysis will be useful for a few reasons. First, we'll see more precisely how

evaluation and updating are connected. Second, we'll obtain a useful result that shows the relationship between overall expected accuracy, deference, and information. Third, we'll get a handle on differences between the way we can treat ourselves as alithometers and the way we treat others as alithometers.[17]

4.5.4.1 The Brier Score

To look at the issues of disagreement in any formal depth, it'll be useful to have a mathematically precise notion of accuracy. The most popular way to measure the *in*accuracy of a given credence x in p is with the Brier score:

$$\mathrm{BS}(x, p) := \begin{cases} (1-x)^2 & p \text{ is true} \\ x^2 & p \text{ is false} \end{cases} \tag{4.1}$$

The Brier score is the most intuitive half-way decent of inaccuracy, so for our purposes it can serve as the official measure.[18] It simply looks at the absolute distance between p's actual truth-value and x and squares it.[19] Nothing I say will depend on the choice of this particular measure, but I use it for the sake of concreteness.

Two caveats: (1) The lower the Brier score, the better. Agents with higher scores are less accurate. (2) Since no ambiguity will arise, I'll use $\mathrm{BS}(x)$ instead of $\mathrm{BS}(x, p)$.

4.5.4.2 Accuracy and Expected Accuracy

As explained above, we need to focus on *expected* inaccuracy. The people I'm pressured to take seriously are people whose opinions I expect to be accurate. To talk about this

[17] Much of this section recaps the more technical discussion in the previous chapter, but I include it again since the emphasis is somewhat different.

[18] I don't actually think the Brier score is the right measure of accuracy (see Levinstein (2012) or Chapter 2 of this dissertation). I use it here primarily for ease of use and exposition.

[19] The squaring is for technical reasons that would at present be a distraction to discuss.

notion, we'll have to get a little fancier with the formalism.

First, if B has credence function $\mathfrak{b}$ such that $\mathfrak{b}(p) = m$, she can calculate the expected accuracy of some other potential credence y as follows:

$$\begin{aligned} \mathrm{EBS}_{\mathfrak{b}}(y) :=& m(1-y)^2 + (1-m)y^2 \\ =& m(1-2y) + y^2 \end{aligned} \tag{4.2}$$

Equation (4.2) averages the possible scores a credence of y could get by B's credence that y will get those scores. If p is true, y will get the score $(1-y)^2$, which B believes to degree m will happen, and if p is false y will get the score y^2, which S believes to degree $1-m$ will happen.

First observation: B's own expected accuracy is a function just of her own credence. If we let $m = y$, we see that under the Brier score, an agent's own expected accuracy is always

$$\mathrm{EBS}_{\mathfrak{b}}(m, p) = m(1-m) \tag{4.3}$$

No factors other than her actual credence come into play. This fact will be important below, since one's own expected inaccuracy will behave differently from the expected inaccuracy of an advisor.

To figure out the antecedent expected accuracy of an advisor, though, we need to know how to measure the expected accuracy of somebody's opinion we're unaware of. Equation (4.2) only tells us how to measure the expected accuracy of a known opinion. However, it's important for our purposes to find out how accurate an agent expects her advisor to be before she knows the advisor's opinion.

Suppose you're my advisor. I don't yet know what you think, but I may have a high or low opinion of you beforehand. Let Y be a random variable representing your credence. Before learning what you think about whether p, I have some distribution over the possible values of Y. Suppose $\mathfrak{b}$ is my credence function. To figure out how accurate I expect you to be, I'll have to invoke my probability distribution function, which we

denote $b(Y)$, to average over the scores I'd expect you to get for different credences you might have.

Now, it's important to keep in mind that *my* credence in p will sometimes change conditional on your having a particular credence, as we've already seen. That is, to calculate the expected accuracy for Y, I need to build in the fact that I'll change my opinion in p once I know what you think. Later on, this mathematical fact will be useful since we'll be able to look more directly at the question of how I'll update can be related to how accurate I expect you to be.

Here's the formula for what I expect your accuracy to be antecedently:

$$\begin{aligned}\mathrm{EBS}_{\mathfrak{b}}(Y) &= \int_0^1 b(y)(b(p|y)[1-2y]+y^2)\,\mathrm{d}y \\ &= \int_0^1 b(y)(b(p|y)-y)^2\,\mathrm{d}y + \int_0^1 b(y)b(p|y)(1-b(p|y))\,\mathrm{d}y \end{aligned} \tag{4.4}$$

To get an intuitive handle on what EBS is doing, look at the second line. Suppose I learn your credence and nothing else. My updated credence is just $\mathfrak{b}(p|Y=y)=m^*$. We can measure my respect for you based on how far y is from from m^*. If I completely defer to you, m^* will be equal to y. If I don't rely on you completely, then there will be some distance between m^* and y. Clearly, how much I end up respecting you should correlate with how accurate I think you'll be and how good of an alithometer I think you are.

Look first at the first integral in (4.4):

$$\mathrm{ExRes}_{\mathfrak{b}}(Y) := \int_0^1 b(y)(b(p|y)-y)^2\,\mathrm{d}y \tag{4.5}$$

$\mathrm{ExRes}_{\mathfrak{b}}(Y)$ measures how much I expect that my posterior credence will deviate from your actual credence. It looks at what I expect to be the square of the distance between your credence and my credence upon learning yours. In other words, it serves as a measure of my *expected respect* for you.

Respect isn't all that goes into epistemic evaluation, however. I also care about how well-off I'll be epistemically after talking to you. If after talking to you, I end up really close to 1 or 0, then I think I'm doing really well. And I can sometimes exploit your sensitivity to whether p even if you don't actually form what I take to be good credences. For instance, in the extreme case, you could have credence 0 in all truths and credence 1 in all falsehoods. From my perspective, you're maximally useful to talk to because your credence will allow me to know whether p. So, even if your credences are not apt, I'll still give you some positive epistemic evaluation for your sensitivity, especially if it allows me to have a high expected *posterior utility*: i.e., if I expect that after talking to you *I'll* be well-off.

Now let's look at the second integral:

$$\mathrm{ExUt}_b(Y) := \int_0^1 b(y)b(p|y)(1 - b(p|y))\,\mathrm{d}y \tag{4.6}$$

We saw above that an agent's own expected accuracy is just a function of her credence, and for the Brier score, it's just $m(1-m)$. $\mathrm{ExUt}_b(Y)$ looks at what the agent *expects her own posterior expected accuracy to be*. That is, it measures how well she expects to expect to be doing after learning her advisor's credence.[20]

Finally, before moving on, we should note:

Second observation: From a formal perspective, even once we fix the value of $\mathrm{EBS}_b(Y)$, there aren't that many constraints on what the distributions $b(Y)$ and $b(p|Y)$ look like. In other words, we can have lots of different distributions that all end up integrating to the same value. So, an agent could assign the same pre-disagreement expected inaccuracy to two advisors represented by C and D but be disposed update differently for particular values of C and D. In particular, the assessor might assign very different values for ExRes and ExUt to the same agents. For instance, suppose C is an expert—i.e., somebody the

[20]This factoring of expected accuracy into expected utility and expected respect generalizes to all admissible scoring rules, as we saw in Appendix A of the previous chapter. For the original demonstration, see DeGroot and Fienberg (1982, 1983).

assessor should defer entirely to—but the assessor expects C's credence to be close to .5. D, on the other hand, is mis-calibrated according to the assessor but is very sensitive as to whether p. That is, D systematically is mildly biased, but his credence carries a lot of information about p. So once the assessor learns the value of D she'll be able to form a credence in p that's close to 1 or 0. So, while $\text{EBS}(C) = \text{EBS}(D)$, the assessor will update differently for $C = .6$ than she will for $D = .6$.

4.5.5 Internalism and the Impersonal Stance

So far, we've been pushing the view that evidence from disagreement behaves and can be treated like normal evidence. We have noted, however, that you're still not—on the models we've examined so far—treating yourself as an alithometer in the same way you treat your advisor as an alithometer. There's a strong asymmetry in the way I calculate expected accuracy for the two of us. I always have access, on this formal idealization, to my own credence in p, so I expect my accuracy to be $m(1-m)$. My own expected accuracy is simply a function of my credence. Once I have a credence in p, I don't get to take into consideration factors like my general alithometric abilities or my sensitivity to p or any other features about my own epistemic abilities. When calculate *your* expected accuracy, I do get to take all this into consideration. As we see from Equation (4.4), I don't calculate your expected accuracy solely as a function of your expected credence. Instead I take into account factors like your expected sensitivity to whether p and the shape of my subjective distribution over what you think. Additionally, it's built in to these sorts of self-calculations that I afford myself a minimum level of self-trust. The highest expected inaccuracy I can assign myself is .25, which I assign myself when and only when I have credence .5 in p. I can assign other people the worst possible expected score of 1, however.

And it's here, I think, where Elga, Christensen, and others are really onto something. Suppose I notice that I've historically been really bad at predicting Oscar winners, but

I also have credence .8 that Film A will win this year. If I just use my current credence function to calculate my expected inaccuracy, I'll get an expected score of $.8 \times .2 = .16$. But it seems like I should at least be given some pause when I realize that historically I've done much worse when my credence that a given film will win is .8 or greater. Let's suppose, for concreteness, that my average score has been around .5. If the data are robust enough, I may have good reason to lower my credence. Furthermore, the Principle of Reflection notwithstanding, I might be less than fully willing to defer to my future self if I think that my credences are systematically too extreme around the time of the Academy Awards. If I find out, for instance, that next year I'll have a high credence in a given film, then data about my past bad epistemic performance may rightly make me treat my future self as an expert. The problem when is that when considering how well she expects to be doing currently, the bayesian will *use* her current credence. I can calculate future and past expectations in a way that doesn't rely on the credences I used to or later will have, but my current expected inaccuracy is calculated with my own credence. Likewise, when I disagree with you, I'll evaluate myself with my credences, and I'll also evaluate you with my credences.

Using my current credence function to evaluate itself is different from using it to evaluate others in a fundamental and sometimes problematic way. It can think of other agents (or even me at other times) as being kinds of alithometers it cannot regard itself as being. That's simply *built in* to the way expectations are calculated and the way decent accuracy measures (or information measures more generally) are designed. It leaves us, unfortunately, with a sometimes unwarranted level of self-trust and expectation of our own performance.

On the other hand, it's impossible to escape entirely from this sort of predicament. Probabilities and expectations have to come from somewhere, after all. If I already thought some alternative credence function was better than my own, I'd be using it already. Since I'm the one who'll have to figure out how to respond to your credence, I'll

have to rely on myself in the end no matter what.

So, there's a clear tension here. We feel a pull to a more neutral or *impersonal* perspective. That is, we recognize that the standard machinery requires the assessor to treat herself in an uncomfortably asymmetric way from how she treats others. However, the assessor will have to rely on herself in the end no matter what, so the asymmetry can't entirely be done away with.

4.5.5.1 Schematic Reconstruction of Conciliationism

Let's look again at Christensen's and Elga's proposals. Both authors suggest that the way we should react to a case of disagreement is to abstract away from the particulars of our own reasoning and evidence. To determine our posterior credence, we should look at what we *would have thought* if we didn't have all of our actual evidence but instead knew only (i) some general facts about the circumstances of disagreement and (ii) epistemic abilities that we and our advisor possess. We don't use our actual prior in p at the time of the disagreement, but instead switch to some alternative prior we had or would have had before thinking through the issue. Since "thinking through the issue" can be treated as gaining extra evidence, the idea is that we have to de-conditionalize our current credence function by forgetting about some of our resources.[21]

Here's how to understand these proposals schematically. Suppose $\mathfrak{b}_R$ is the assessor's current credence function and E is her total evidence before learning her advisor's credence (where E includes potential inferences and reasoning she's gone through and so on). That is, $\mathfrak{b}_R$ is the assessor's real credence function, and E is her total evidence. When learning her advisor's credence, she first ought to banish some of E from her reasoning and instead rely on some strictly weaker evidential base E' that's entailed by E. Second, since her current credence function $\mathfrak{b}_R$ knows about E and is perhaps reasoning incor-

[21] See Titelbaum (2013) for a theory of how to update upon loss of evidence.

rectly, she should switch to some other less informed (but, we hope, less epistemically irrational) credence function $\mathfrak{b}_J$.

To see how this works, let's apply it to Elga's proposal. E' is some coarsened description of E that abstracts away from the particulars of the agent's reasoning. E' contains facts like how reliable the two agents are generally along with the circumstances of the disagreement. $\mathfrak{b}_J$ is the credence function the assessor would have had without having thought through whether p.

Since $\mathfrak{b}_J \neq \mathfrak{b}_R$, we can use $\mathfrak{b}_J$ to treat both the assessor and the advisor symmetrically. $\mathfrak{b}_J$ acts like a third party. In other words, it's *as if* we go to a judge J who gets to decide the issue. By invoking this quasi-external agent, J gets to do something the assessor herself can't do directly, *viz.*, treat the assessor as an alithometer in the same we she treats the advisor as an alithometer. Thus, the assessor's own actual credence in p is now evidence for or against p in the same way that the advisor's is.

On the standard bayesian models, an agent's own credences are always completely luminous to her, but by invoking $\mathfrak{b}_J$, we can now represent both the assessor's and the advisor's credences as random variables X and Y respectively whose actual values are unknown. On Elga- and Christensen-style conciliationism, reasonable agents are supposed to set $\mathfrak{b}_R(p|E, Y = y) = \mathfrak{b}_J(p|E', Y = y, X = x)$. That is, the assessor's post-disagreement credence should be set equal to what J would think about p conditional on the impoverished evidential base E', the fact that the assessor's pre-disagreement credence was x and advisor's was y.[22]

[22]We actually need a further abstraction to get the proposal to work along the lines Elga and Christensen desire. Since we want $\mathfrak{b}_J$ to form an opinion largely based on what it thinks of the alithometric properties of the agents, not on what it already thought about p, we can't make E' too resilient, nor can we let $\mathfrak{b}_J(p)$ be too set in its opinion. If we follow Elga at his word that $\mathfrak{b}_J$ is simply the prior we had before thinking through the issue, we might not find a judge J who's sufficiently un-opinionated. For instance, before thinking through the issue, two philosophers might still have firm but opposite opinions of theism. So, if philosopher DD realizes he disagrees with philosopher AP, DD won't end up with much of a different posterior opinion in theism by relying on earlier, pre-reflective credence functions. (Of course, less dramatic examples will work too. Even in idealized circumstances of having a super-baby who has yet to acquire any

4.5.5.2 Neutrality Partially Regained

In broad strokes, the problem with this proposal is that there won't in general be any good recipe for figuring out which parts of your evidence should be banished nor for figuring out what the alternative prior credence function $\mathfrak{b}_J$ should be.

However, it doesn't look like there's any good way of being able to treat yourself as an alithometer without abstracting away from your evidence and your actual credence function. So is there any way to regain this neutrality and avoid the problems of the conciliatory views on the table?

I think there is, but it will not, unfortunately, result in anything like an effective procedure for determining the right post-disagreement credence.

The basic idea is to treat the evidential significance of $\mathfrak{b}_J(p|E', x, y)$ as we would a normal statistic. By looking at what an alternative prior with a distinct evidential base would think about p we gain some useful information that *helps* us figure out a good posterior credence to have. While this information can be useful, there won't be a good way to specify anything like a precise relationship between what $\mathfrak{b}_J$ thinks and what we ought to think.

Compare: I learn that 1/2 of people in some reference class I belong to get disease D. What should my credence be that I'll get disease D? There's no good answer to this question. Sometimes, I should simply set my credence to 1/2. However, based on

---

empirical evidence, we don't represent agents as having entirely uninformative priors.)

So, to avoid this problem, we need to abstract away even further from the particular proposition we're looking at and instead look to something like the relevant domain of propositions. That is, J has to determine her posterior credence m^* not by looking at the particular proposition p but by looking only at some relevant broader domain of propositions. For example, in RESTAURANT, we don't ask J: How likely is it that we owe $43 given that Assessor has credence .9 and Advisor has credence .3 with a total bill of $187 and such-and-such policy for dividing up and tipping? Instead, J should be asked what her posterior credence that the answer to a particular arithmetic problem is given the background circumstances of judging and what she knows about our relative arithmetic abilities. Therefore, we need to add a further constraint on what $\mathfrak{b}_J$ should look like: $\mathfrak{b}_J(p|E', X, Y)$ should vary heavily based on the particular values of X and Y. That is, J's opinion conditional on E' and the opinions of the two agents needs be quite sensitive to what the disputants actually think before learning of the disagreement.

particulars of my situation, I might also be rational to ignore this statistic or to allow it to have only a small effect on my credence. So, while this information is potentially useful to me and may help me form a credence, there isn't much in general that we can say about how it ought to affect an my credence.

An advantage of this approach is that we don't now owe a theory of what exactly should be abstracted away from our total evidence, nor how to figure out what exactly $\mathfrak{b}_J$ is. Instead, we can take a more pluralistic view. There are also sorts of different abstractions we can look at, all sorts of different reference classes, and all sorts of different priors. Since these other priors are evaluating me as an alithometer, and I in turn use them to help me determine my own actual credence, I indirectly get to treat *myself* as an alithometer in a more neutral way.

So, when considering how to update in RESTAURANT, I may at my general ability at arithmetic, or my ability just at arithmetic involving small numbers in my head after 6 pm, or my problems with calculations involving decimals. I might also consider what different kinds of priors would think on these evidential inputs. All these statistics are useful, but there's no single one alone that I should look at. There simply isn't any general theory forthcoming of how to handle these facts, but that's as we should expect if disagreement provides normal evidence.

Bibliography

Ballantyne, N. and E. Coffman (2011). Uniqueness, evidence, and rationality. *Philosopher's Imprint 11*(18).

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78*, 1–3.

Briggs, R., F. Cariani, K. Easwaran, and B. Fitelson (ms). Individual coherence and group coherence.

Christensen, D. (2007). Epistemology of disagreement: The good news. *Philosophical Review 116*(2), 187–217.

Christensen, D. (2009). Disagreement as evidence: The epistemology of controversy. *Philosophy Compass 4/5*, 756–767.

Christensen, D. (2011). Disagreement, question-begging and epistemic self-criticism. *Philosophers' Imprint 11*(6).

de Finetti, B. (1964). Foresight: Its logical laws, its subjective sources. In H. Kyburg and H. Smokler (Eds.), *Studies in Subjective Probability*. Wiley.

de Finetti, B. (1974). *Theory of Probability*, Volume 1. John Wiley and Sons.

DeGroot, M. H. and E. Eriksson (1985). Probability forecasting, stochastic dominance, and the Lorenz Curve. In J. Bernardo, M. DeGroot, D. Lindley, and A. Smith (Eds.), *Bayesian Statistics 2*, Proceedings of the Second Valencia International Meeting, North-Holland, pp. 99–118. Elsevier Science Publishers B.V.

DeGroot, M. H. and S. E. Fienberg (1982). Assessing probability assessors: Calibration and refinement. In S. S. Gupta and J. O. Berger (Eds.), *Statistical Decision Theory and Related Topics* III, Volume 1. New York: Academic Press.

DeGroot, M. H. and S. E. Fienberg (1983). The comparison and evaluation of forecasters. In *Proceedings of the 1982 I.O.S. Annual Conference on Practical Bayesian Statistics*, Volume 32, pp. 12–22. Blackwell Publishing.

Elga, A. (2007). Reflection and disagreement. *Nous 41*(3), 478–502.

Enoch, D. (2010). Not just a truthometer: Taking oneself seriously (but not too seriously) in cases of peer disagreement. *Mind 119*(476), 953–997.

Feldman, R. (2007). Reasonable religious disagreements. In L. Antony (Ed.), *Philosophers without Gods*, pp. 194–214. Oxford University Press.

Gasarch, W. I. (2002). The P=?NP poll. *SIGACT News 33*(2), 34–47.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*(477), 359–378.

Greaves, H. and D. Wallace (2006). Justifying conditionalization: Conditionalization maximizes expected epistemic utility. *Mind 115*(632), 607–632.

Gutting, G. (1982). *Religious Belief and Religious Skepticism*. Notre Dame: University of Notre Dame Press.

Hájek, A. (2008). Arguments for—or against—probabilism? *British Journal for the Philosophy of Science 59*, 793–819.

Hájek, A. (ms). A puzzle about partial belief.

Hoffrage, U. (2004). Overconfidence. In R. Pohl (Ed.), *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement, and Memory*, Chapter 13, pp. 235–254. Psychology Press.

Hume, D. (2000). *A Treatise of Human Nature*. Oxford Philosophical Texts. Clarendon Press.

James, W. (1979). *The Will to Believe and Other Essays in Popular Philosophy*. Cambridge, MA and London: Harvard Univeristy Press.

Jeffrey, R. C. (1983). *The Logic of Decision* (2nd ed.). University of Chicago Press.

Jehle, D. and B. Fitelson (2009). What is the 'equal weight view'? *Episteme 6*, 280–293.

Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science 65*, 575–603.

Joyce, J. M. (2009). Accuracy and coherence: Prospects for an alethic epistemology of partial belief. In F. Huber and C. Schmidt-Petri (Eds.), *Degrees of Belief*, Volume 342, pp. 263–297. Springer.

Kelly, T. (2005). The epistemic significance of disagreement. In J. Hawthorne and T. Gendler (Eds.), *Oxford Studies in Epistemology*, Volume 1. Oxford: Oxford University Press.

Kelly, T. (2010). Peer disagreement and higher order evidence. In R. Feldman and T. Warfield (Eds.), *Disagreement*, pp. 111–174. Oxford: Oxford University Press.

Keynes, J. M. (1924). *A Tract on Monetary Reform*. Amherst: Prometheus Books.

Kierland, B. and B. Monton (2005). Minimizing inaccuracy for self-locating beliefs. *Philosophy and Phenomenological Research 70*(2), 384–395.

Kolodny, N. (2005). Why be rational? *Mind 114*, 509–563.

Lehrer, K. and C. Wagner (1981). *Rational Consensus in Science and Society*. Dordrecht-Boston: Reidel.

Leitgeb, H. and R. Pettigrew (2010a). An objective justification of bayesianism I: Measuring inaccuracy. *Philosophy of Science 77*, 201–235.

Leitgeb, H. and R. Pettigrew (2010b). An objective justification of bayesianism II: The consequences of minimizing inaccuracy. *Philosophy of Science 77*, 236–272.

Levinstein, B. A. (2012). Leitgeb and Pettigrew on accuracy and updating. *Philosophy of Science 79*(3), 413–424.

Levinstein, B. A. (ms). Rationality as expert?

Lindley, D. (1982). Scoring rules and the inevitability of probability. *International Statistical Review 50*, 1–26.

Lindley, D. (1991). *Making Decisions* (2nd ed.). Wiley.

Loewer, B. and R. Laddaga (1985). Destroying the consensus. *Synthese 62*(1), 79–95.

Maher, P. (2002). Joyce's argument for probabilism. *Philosophy of Science 96*, 73–81.

Nau, R. (1985). Should scoring rules be 'effective'? *Management Science 31*, 527–535.

Oddie, G. (1997). Conditionalization, cogency, and cognitive value. *British Journal for the Philosophy of Science 48*(4), 533–541.

Pettigrew, R. (2011). An improper introduction to epistemic utility theory. In R. de Henk, S. Hartmann, and S. Okasha (Eds.), *EPSA Philosophy of Science: Amsterdam 2009*, pp. 287–301. Springer.

Predd, J., R. Seiringer, E. H. Lieb, D. Osherson, V. Poor, and S. Kulkarni (2009). Probabilistic coherence and proper scoring rules. *IEEE Transactions on Information Theory 55*(10), 4786–4792.

Ramsey, F. P. (1931). Truth and probability. In R. Braithwaite (Ed.), *Foundations of Mathematics and Other Essays*, pp. 156–198. Routledge & P. Kegan.

Rosenkrantz, R. (1981). *Foundations and Applications of Inductive Probability*. Ridgeview Press.

Roush, S. (2009). Second-guessing: A self-help manual. *Episteme 6*(3), 251–268.

Savage, L. J. (1971). Elicitation of personal probabilities. *Journal of the American Statistical Association 66*, 783–801.

Schervish, M. (1989). A general method for comparing probability assessors. *The Annals of Statistics 17*, 1856–1879.

Scott, D. (1964). Measurement structures and linear inequalities. *Journal of Mathematical Psychology 1*(2), 233–247.

Seidenfeld, T. (1985). Calibration, coherence, and scoring rules. *Philosophy of Science 52*, 274–294.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics 1*, 43–62.

Silver, N. (2012). *The Signal and the Noise: Why So Many Predictions Fail But Some Don't*. New York: The Penguin Group.

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

Titelbaum, M. (2013). *Quitting Certainties: A Bayesian Modeling Framework*. Oxford: Oxford University Press.

van Fraassen, B. (1984). Belief and the will. *Journal of Philosophy 81*, 235–256.

White, R. (2005). Epistemic permissiveness. *Philosophical Perspectives 19*, 445–459.

White, R. (2009). On treating oneself and others as thermometers. *Episteme 6*, 233–250.

Winkler, R. L. (1996). Scoring rules and the evaluation of probabilities. *Test 5*(1), 1–60.

Curriculum Vitæ

Benjamin Anders Levinstein

Education

2007 – 2013 Ph.D. in Philosophy
Rutgers, The State University of New Jersey, New Brunswick, NJ

2007 B.A. in Philosophy with Mathematics minor
The University of Chicago

Positions

2007 – 2009 Graduate Fellow, Rutgers University