Modest Epistemology

by

Kevin Dorst

B.A., Washington University in St. Louis (2014)

Submitted to the Department of Linguistics and Philosophy
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2019

Author .....
Department of Linguistics and Philosophy
August 3, 2019

Certified by .....
Roger White
Professor of Philosophy
Thesis Supervisor

Accepted by .....
Bradford Skow
Laurence S. Rockefeller Professor of Philosophy
Chair of the Committee on Graduate Students

Modest Epistemology

by

Kevin Dorst

Submitted to the Department of Linguistics and Philosophy
on August 3, 2019, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Thinking properly is hard. Sometimes I mess it up. I definitely messed it up yesterday. I'll likely mess it up tomorrow. Maybe I'm messing it up right now.

I'm guessing you're like me. If so, then we're both *modest*: we're unsure whether we're thinking rationally. And, it seems, we *should* be: given our knowledge of our own limitations, it's rational for us to be unsure whether we're thinking rationally. How, then, should we think? How does uncertainty about what it's rational to think *affect* what it's rational to think? And how do our judgments of people's (ir)rationality change once we realize that it can be rational to be modest? This dissertation makes a start on answering those questions.

Chapter 1 introduces a general framework for modeling situations in which you are rational to be unsure what the rational opinions are. I first show how this framework allows us to precisely formulate the questions from the "higher-order evidence" literature. I then use it to argue that rational modesty is needed to explain the epistemic force of peer disagreement, and therefore that any theory of such disagreement must be based on a general theory of rational modesty. Many have suggested that such a theory can be easily formulated based on the *enkratic intuition* that your first-order opinions must "line up" with your higher-order ones. But I argue that this is incorrect: whenever modesty is rational, so too is epistemic *a*krasia. We need to look elsewhere for a general theory of rational modesty.

Chapter 2 offers one. I build a theory that—in a precise sense—allows as much modesty as possible while still guaranteeing that rationality is a guide. The key principle—which I call Trust—formalizes the truism that *it's likely that what the evidence supports is true.* I show that Trust permits modesty, ensures that rational opinions are correlated with truth, and is necessary and (in a wide class of scenarios) sufficient to vindicate the truism that you should always prefer to respond rationally to your evidence. In sum, Trust establishes that there is a principled way for rational people to be modest.

Chapter 3 applies this theory of rational modesty to the psychology of human reasoning. In particular, a wide range of studies suggest that people have a tendency to *predictably polarize* in the face of conflicting evidence: to gather and interpret evidence in a way that leads them to predictably strengthen their prior beliefs. This "confirmation bias" is standardly taken to be a hallmark of human irrationality. It

need not be. I first prove that whenever modesty can be rational, so too can predictable polarization. I then argue, further, that this abstract possibility may play a role in the actual polarization we observe. In particular, given common structures of rational modesty generated by the process of *cognitive search*, rational agents who care only about the truth should sometimes exhibit confirmation bias.

So, I say, epistemology can simultaneously learn from mathematics and inform psychology. That is part of a broader narrative. This dissertation makes the argument that epistemology can be rigorous while also being relevant; that it can be formal while also being applicable; and that it can be abstract and principled, while also teaching us about ourselves. That is the hope, at least. Read it, and perhaps you will agree.

Thesis Supervisor: Roger White
Title: Professor of Philosophy

Acknowledgments

"Would you rather be fifty feet tall, or fifty feet wide?" "How fast do clouds move?" "Why did they say 'if and only if'—isn't the 'only if' redundant?" With childhood questions like these, my parents might've known I'd wind up a philosopher.

Then again, probably not: it takes a village—and a whole lot of luck—after all. Well, 15 years on, I've had more than my fair share of both.

Without a doubt, my biggest intellectual debt belongs to my brother: Chris Dorst. Many people wish for a second try in their career choices; I had the good fortune of having my first try in some ways be a second. Remarkably similar to him, I found myself again and again choosing a path that followed in his footsteps: asking my first philosophy (of religion) questions in high school; going to Washington University; becoming incredibly (overly?) academically focused; majoring in philosophy and then applying to PhD programs; completing a dissertation; going on the academic market and getting a job—every step, Chris took first. It is impossible for me to say how much this influenced both my choices and the outcomes that followed. But suffice it to say that very few of those steps would have been taken—or have been taken nearly as well—if not for his constant guidance and encouragement. Thanks, Chris, for never asking for the spotlight, and yet somehow always showing me the way.

On the road to and through graduate school, I've had more mentors and supporters than I can count. Alec Wright helped awaken my love of intellectual conversations. Frank Lovett showed me that it was *thinking*, not arguing, that I loved. Andrew Rehfeld challenged and encouraged me at one of my most fragile academic moments. Charlie Kurth offered me seemingly unlimited time and energy. Roy Sorensen taught me how to see a project through to completion. Gillian Russell helped me to start another, big project—the one that eventually became my dissertation. Bernhard Salow and Jack Spencer taught me the value and joy of thinking *together*. They were the paradigms of "MIT philosophers," showing me that we are all better off—professionally and intellectually—when we focus less on our own projects and more on each others'. Nilanjan Das, Ginger Schultheis, Matthias Jenny, Thomas Byrne, David Builes, and Haley Schilling continued this trend—thanks for hanging around the lounge and always being willing to talk philosophy with me; you are what makes the department at MIT such a special place. Harvey Lederman offered guidance at just the right times: he awakened my interest in the subfield that became my home, and years later gave me the reference to the key paper that led to the key theorem of the key chapter of my dissertation—it would not be the same (or as good) without him. Cosmo Grant showed me the value of a wide-ranging curiosity and the patience to follow it through, and went with me outside our comfort zones to learn things we otherwise never would have. Kevin Richardson offered a combination of good advice and hard questions that kept me from getting complacent. Kayla Dorst, my sister-in-law, always supported me in difficult times. I owe much to all of them.

The faculty at MIT made me feel at home from day one; I will miss their family-like dynamics that I have come to know and love. Roger White taught me how to be comfortable in silence—to take my time to think and figure things out properly. His constant attention to detail and drive for perfection—in both philosophy, and (surprise) wood-working—were exactly what I needed in an advisor. Bob Stalnaker kept his door open and always made time to stop and talk about my "big questions." After five years, I have come to understand why 'Bob's in his office' is every MIT grad's favorite example sentence. Kieran Setiya's gentle guidance always kept me grounded, both philosophically and personally. I am grateful to have had a mini mid-life crisis—and have talked myself down by reading his *Midlife*—well before it was due. Branden Fitelson—though officially based at Northeastern—managed to be as involved in my graduate school career as anyone at MIT. He always championed me without qualification, and he gave me far more opportunities than I had any right to expect. Jack Spencer and Miriam Schoenfield weren't officially on my committee, but my dissertation benefited so much from their time and energy that it feels as if they were. Thanks also to Agustín Rayo, for offering me his untempered enthusiasm; to Caspar Hare, for showing me the value of attention to style; and to Brad Skow, for asking me "Why thresholds?"—a question that took up more of my graduate career than you'd believe. Finally, thanks to Alex Byrne, Sally Haslanger, Justin Khoo, Vann McGee, Tamar Schapiro, and Steve Yablo—for meetings, wisdom, encouragement, and making MIT such a wonderful place to "grow up", philosophically. I could not ask for a better intellectual family.

Finally, there are a few people who deserve far more than I can say.

Thanks to Quinn White: for seeing the best in me, throughout the thick and thin of graduate school; for supplying cheer in the good times, and offering insight in the bad; and for giving me the unconditional love and concern of a true friend.

Thanks to Brooke Husic: for waking up the the little mathematician inside my head, showing me that I *could* write a dissertation like this; for always inspiring me to become a better version of myself, even when I was comfortable where I was; and, well, for everything—including all the unexpected, fortunate things that have happened along the way.

Thanks, finally, to my parents: Stan Dorst and Maggie Yoest. To my dad, for always asking what I thought, and for patiently engaging with my childhood questions. (I maintain my answer: "Fifty feet wide. If you were fifty feet tall, you'd trip and kill yourself.") To my mom, for always asking what I *felt*, and for helping me learn the value and meaning of a different sort of conversation. And to both of them—for giving so much to, and asking so little of, their two young philosophers.

Contents

Chapter 1

Higher-Order Uncertainty

Abstract

You have *higher-order uncertainty* iff you are uncertain of what opinions you should have. I defend three claims about it. First, the higher-order evidence debate can be helpfully reframed in terms of higher-order uncertainty. The central question becomes how your first- and higher-order opinions should relate—a precise question that can be embedded within a general, tractable framework. Second, this question is nontrivial. Rational higher-order uncertainty is pervasive, and lies at the foundations of the epistemology of disagreement. Third, the answer is not obvious. The Enkratic Intuition—that your first-order opinions must "line up" with your higher-order opinions—is incorrect; epistemic *a*krasia can be rational. If all this is right, then it leaves us without answers—but with a clear picture of the question, and a fruitful strategy for pursuing it.

Introduction

Here is one of my main claims:

> *Thesis:* Epistemic akrasia can be rational.

(Don't worry, just yet, about what it means.) I am confident of *Thesis*, for I have a variety of arguments that I take to be good evidence for it.

But—now that I think about it—*whenever* I sit down to write a paper, I'm confident of that paper's thesis. In fact, that confidence usually has a similar basis: I have a variety of arguments that I take to be good evidence for it. And yet I've later found—all too often—that my arguments weren't so good after all; that I've been *over*confident in my past theses.[1] Having meditated on these facts, I'm still confident

[1] 'Overconfident' here—as in natural language—means being more confident than you *should* be; not to having some confidence in something that's false. If you're 50-50 that this fair coin that I'm about to toss will land heads, then you're not overconfident—even if, in fact, it will land tails.

of *Thesis*, for I still think that I have good arguments for it. However, here's another proposition that I now consider possible:

> *Doubt:* I should not be confident of *Thesis*.

I'm not confident of *Doubt*—but nor do I rule it out: I leave open that maybe I *shouldn't* be confident of *Thesis*.

Question: how should my attitudes toward *Thesis* and *Doubt* relate? *Thesis* is a claim about some subject-matter. *Doubt* is a claim about what opinion I ought to have about that subject-matter. Let's call my opinion about *Doubt* a **higher-order opinion**—an opinion about what opinion I should have. Since I am uncertain about what opinions I should have, I have *higher-order uncertainty*. Let's call my opinion toward *Thesis* a **first-order opinion**—an opinion about something other than what opinions I should have. Generalizing our question: how should my first-order and higher-order opinions relate? For example: if I become more confident that I shouldn't be confident of *Thesis*, should that lead me to be less confident of *Thesis*? Or: if I have a lot of higher-order uncertainty about how confident I should be in *Thesis*, can I nevertheless be fairly confident of it? I will not give a full answer to such questions—but I will take three steps toward one.

First step. Many have asked similar questions. But they have often framed it as a question of how one *body* of evidence—your *first-order evidence*—interacts with another body of evidence—your *higher-order evidence*.[2] My first claim:

> Reframing: We should reframe the question: Given your *total* evidence, how should your first- and higher-order *opinions* relate?

I defend Reframing by showing how to build a general framework for studying the relationship between first- and higher-order uncertainty (§1.1), and then putting it to work (§§1.2–3).

Second step. So reframed, our question is nontrivial:

> Modest Truism: Your total evidence often warrants being uncertain what opinions your total evidence warrants, and (hence) being *modest*: uncertain whether you're rational.[3]

I defend the Modest Truism by arguing that rational modesty—i.e. rational higher-order uncertainty—is needed to account for the epistemic force of disagreement (§1.2).

Third step. Many have pointed out that it seems irrational to believe that *my Thesis is true, but I shouldn't believe it*. The inferred explanation has been that your

[2]E.g. Feldman (2005); Christensen (2010a, 2016); Horowitz (2014); Schoenfield (2015a, 2016); Sliwa and Horowitz (2015).

[3]I will assume a *single* normative notion that privileges certain opinions. I'll call them the opinions that "you *should* have," that "your (total) *evidence warrants*," or that "are *rational*." If you think these normative notions come apart, please replace my expressions with your preferred, univocal one.

first-order opinions must "line up" with your higher-order opinions. Call this the **Enkratic Intuition**. Many theories defend (or presuppose) it as the answer to our question.[4] My final claim—my *Thesis*—is that the Enkratic Intuition is wrong:

AKRATIC: If modesty is rational, so too is epistemic akrasia.

I defend AKRATIC by using the above framework to precisify the Enkratic Intuition and show that it is inconsistent with higher-order uncertainty (§1.3; cf. Titelbaum 2015).

That is the plan. Here is the picture. Higher-order uncertainty is pervasive and important. There is a general, tractable framework for studying it. Many open questions remain.

1.1 Two Problems

Recall that I'm confident of:

Thesis: Epistemic akrasia can be rational.

But I also suspect that:

Doubt: I should not be confident of *Thesis*.

In thinking about cases like this, the standard operating procedure is to make (something like) the following distinction:

(1) My *first-order evidence* about *Thesis* is the evidence that bears directly on it. (Example: my current arguments.)

(2) My *higher-order evidence* about *Thesis* is the evidence that bears *in*directly on it by bearing directly on claims like *Doubt*. (Example: the flaws in my past arguments.)

Authors making this distinction often presuppose that we can meaningfully speak of two distinct *bodies* of evidence—my first-order evidence, and my higher-order evidence.[5] Distinction made, the standard question goes something like this. Given my first-order evidence, I should have some opinion about *Thesis*. Now add my higher-order evidence. Should my opinion in *Thesis* change? If so, how? Since we are to imagine two interacting bodies of evidence, call this the **Two-Body Problem**.

My first claim is:

[4] E.g. Feldman (2005); Gibbons (2006); Christensen (2010b); Huemer (2011); Smithies (2012, 2015); Greco (2014b); Horowitz (2014); Titelbaum (2015); Sliwa and Horowitz (2015); Littlejohn (2015); Worsnip (2015); Salow (2017).

[5] E.g. Feldman (2005); Christensen (2010a, 2016); Horowitz (2014); Schoenfield (2015a, 2016); Sliwa and Horowitz (2015).

REFRAMING: We should reframe the question: Given your *total* evidence, how should your first- and higher-order *opinions* relate?

In other words: instead of two interacting *bodies* of evidence, we have two interacting *levels of opinions* warranted by a single, total body of evidence. We have a **Two-Level Problem**, not a Two-Body one.

I have no short, knock-down argument for REFRAMING. Instead, what I have to offer is (my own) confusion generated by the Two-Body Problem, and clarity generated by the Two-Level one. Perhaps you will share them.

Confusion first. One question: what exactly does it mean for a bit of evidence to bear "indirectly" on *Thesis*? In some sense, the claim that Jones has constructed an argument for *Thesis* does so. But this is not the sense has been meant in the higher-order evidence discussion, which focuses on the possibility of rational errors (Christensen 2010a)—sleep deprivation, hypoxia, irrationality pills, and the like. So perhaps evidence bears indirectly on *Thesis* when it bears on whether I've made a rational error in forming my opinion about *Thesis*? But suppose that I haven't yet formed any opinion about *Thesis*, and then the oracle informs me that *Doubt* is true. Surely this is still higher-order evidence, even though it says nothing about a rational error on my part.

Another question: how to these two bodies of evidence agglomerate? Suppose F is first-order evidence for q and H is higher-order evidence for q; what is the conjunction $F \wedge H$? It clearly bears directly on q, so it seems that it should be first-order evidence. But this means that when I go to base my beliefs about q on my first-order evidence, I will thereby base them on $F \wedge H$, bringing in the higher-order information H. So maybe those beliefs should be based on my *purely* first-order evidence. But what does it mean for a bit of evidence to be *purely*, *directly* about q? Consider q itself—surely this proposition is purely, directly about q if anything is. But consider the proposition:

Not-Known: My first-order evidence does not put me in a position to know $\neg q$.

Not-Known is a paradigm case of the sort of proposition that higher-order evidence about q works "through"—if a proposition p bears on *Not-Known*, it bears indirectly on q. But q *implies Not-Known*. Thus even q itself does not bear *purely* directly on q!

This is not meant to be clarifying. Nor is it meant to be a precise argument against the Two-Body Problem. What it is meant to be is an illustration of how easy it is to find oneself confused with this problem. My goal in the rest of the paper is to argue that the Two-*Level* Problem leads to a clearer framing of the questions and their potential answers.

1.1.1 Framing the debate

Assume that you have a single total body of evidence that determines what opinion you should have in any given proposition. Some of these propositions will be like *Doubt*—claims about what opinions you should have, i.e. about what opinions your single, total body of evidence warrants having. Your opinions about such propositions are higher-order opinions. Other propositions will be like *Thesis*—claims that aren't about what opinions you should have. Your opinions about such propositions are first-order opinions. Here is an interesting question: how do the *first-order* opinions warranted by your total evidence relate to the *higher-order* opinions warranted by your total evidence?[6] For example, how confident of *Doubt* can my evidence warrant being before it necessarily warrants being less than confident of *Thesis*? Or: if we minimally change my evidence so that it warrants more confidence in *Doubt*, will it thereby warrant being less confident of *Thesis*? If so, how much?

We can state things more precisely. Let 'C' be a definite description for my actual degrees of belief—whatever they are. $[C(q) = t]$ is the proposition that I'm t-confident of q—it's true at some worlds, false at others. Let 'P' be a definite description for the credences I *should* have, given my (total) evidence. For simplicity, assume *unique precision*: my evidence always warrants a unique, precise probability function P.[7] $[P(q) = t]$ is the proposition that my (current, total) evidence warrants being t-confident of q. So at any given world w, there's a particular probability function that I ought to have—let 'P_w' be a rigid designator for the function initialized by w. (Unlike the definite descriptions 'P' and 'C', 'P_w' refers to a particular probability function whose values are fixed and known.) Since I can (rationally) be unsure which world I'm in, I can (rationally) be unsure which probability function my credences should match: if the open possibilities are $w_1, w_2, ...$ then I can leave open whether $[P = P_{w_1}]$ (the rational credence function is P_{w_1}) or $[P = P_{w_2}]$ (the rational credence function is P_{w_2}), or ....

With this notation in hand, here's how we can regiment my attitudes toward *Thesis* and *Doubt*. For simplicity, suppose I'm confident of q iff my credence in q is at least 0.7, and I leave open q iff my credence in q is nonzero. We can treat *Thesis* as a primitive proposition. On the other hand, *Doubt* is the proposition that I should not be confident of *Thesis*, i.e. that the rational credence in *Thesis* is less than 0.7. So $Doubt = [P(\textit{Thesis}) < 0.7]$. Thus my attitudes:

I'm confident of *Thesis*: $[C(\textit{Thesis}) \geq 0.7]$, and I leave open that this confidence

[6] I'm certainly not the first to approach the issue in this way—see Williamson (2000, 2014, 2018); Christensen (2010b); Elga (2013); Lasonen-Aarnio (2015), and Salow (2017).

[7] It would be fairly straightforward to generalize the framework to drop this assumption. It's also worth noting that the models I use only presuppose int*ra*personal uniqueness: there is a uniquely rational credence function *for each agent*, given their information and standards of reasoning. For the (de)merits of these assumptions, see White (2005, 2009a); Joyce (2010); Schoenfield (2014), and Schultheis (2017).

is rational: $[C(P(\mathit{Thesis}) \geq 0.7) > 0]$.

I leave open that I shouldn't be confident of *Thesis*: $[C(P(\mathit{Thesis}) < 0.7) > 0]$.

What's distinctive about my epistemic situation is that I'm unsure which opinions my (total) evidence warrants: I think maybe it warrants having a credence of at least 0.7, and maybe it warrants having a credence below 0.7. If we further assume that I'm sure of my *actual* credences—so $[C(C(\mathit{Thesis}) \geq 0.7) = 1]$—it follows that I am unsure whether I'm rational; or, as I will say, I am *modest*.[8] For since I'm certain that I'm confident of *Thesis* and I leave open that I shouldn't be, I thereby leave open that I'm not rational: $\big[C\big(C(\mathit{Thesis}) \neq P(\mathit{Thesis})\big) > 0\big]$.

We might expect that if such higher-order doubts are warranted, then they should constrain my confidence in *Thesis*. The Two-Level Problem is whether, why, and to what extent this is so: how are rational opinions constrained by rational opinions about what opinions you should have? Notice that this is a question about how my higher-order doubts *should* affect my first-order opinions: it is a question about P (the credences I should have), *not* about C (my actual credences).

Stating the question more precisely: how is the value of $P(\mathit{Thesis})$ modulated by the varying values of $P(P(\mathit{Thesis}) = t)$ for various t? Using only the resources already specified, here are a host of natural answers that we could give to this Two-Level Problem:

> Access Internalism: $[P(q) = t] \rightarrow [P(P(q) = t) = 1]$
> If you should be t-confident of q, you should be certain that you should be t-confident of q.
>
> Graded Access: $[P(q) \geq t] \rightarrow [P(P(q) \geq t) \geq t]$
> If you should be at least t-confident of q, you should be at least t-confident that you should be at least t-confident of q.
>
> Degreed JJ: $[P(P(q) \geq t) \geq s] \rightarrow [P(q) \geq ts]$
> If you should be at least s-confident that you should be at least t-confident of q, you should be at least $t \cdot s$-confident that q.
>
> Reflection: $P(q|P(q) = t) = t$
> Conditional on the rational credence in q being exactly t, you should adopt credence exactly t in q.
>
> Simple Trust: $P(q|P(q) \geq t) \geq t$
> Conditional on the rational credence in q being at least t, you should adopt a credence of at least t in q.

[8] I follow Elga (2013) in the "modesty" terminology; note that it is orthogonal to the sense of "immodesty" used in the epistemic utility theory literature (Lewis 1971)

The goal of the Two-Level Problem is to assess principles like these for plausibility and tenability. Do they allow rational modesty? If so, do they nevertheless enforce plausible connections between first- and higher-order attitudes—or do they let such attitudes split radically apart? The answers are often surprising. Access Internalism obviously rules out higher-order uncertainty, and implies each of the other principles (by trivializing them). Surprisingly, Reflection implies that you must always be certain of Access Internalism (as we will see in §1.3). Graded Access rules out most cases of higher-order uncertainty (see Williamson 2018). On the other hand, Simple Trust implies Degreed JJ, and both of these principles allow massive amounts of higher-order uncertainty—meaning that Simple Trust is much weaker than Reflection.

It is not my aim here to explain or justify these particular assessments of these particular principles (see Dorst 2019a). I mention them to give a sense of the terrain—for my aim is to explain why the Two-Level Problem is a fruitful and tractable strategy for exploring the notion of higher-order evidence.

To do that, I need to do three things. First, I need to address the foundational questions of how to model and interpret rational higher-order uncertainty (§1.1.2). Second, I need to argue that the solution to the Two-Level Problem is not the trivial one given by Access Internalism—that higher-order uncertainty is often rational (§1.2). Finally, I need to argue that the solution to the Two-Level Problem is not the obvious one given by the Enkratic Intuition or Reflection (§1.3).

1.1.2 Modeling it

We want to model a particular agent (say, me) at a particular time (say, now) who's uncertain about a particular subject matter (say, *Thesis*).

I should be uncertain about *Thesis*. How do we model that? By saying that I should match my opinions to a probability function that's uncertain of which world it's in—it assigns positive probability to *Thesis*-worlds and positive probability to $\neg$*Thesis*-worlds.

I should also be uncertain about whether *I should be confident of Thesis*. How do we model *that*? The same way. By saying that I should match my opinions to a probability function that's uncertain which world it's in—it assigns positive probability to *I-should-be-confident-in-Thesis*-worlds, and positive probability to *I-should-not-be-confident-in-Thesis*-worlds. That is, just as *Thesis* expresses a proposition, so too *I should be confident of Thesis* expresses a proposition. What we need is a systematic way to represent such propositions.

Here's how.[9] Let W be a (finite) set of epistemic possibilities that capture the

[9] I'm drawing on the probabilistic epistemic logic literature, though it usually assumes you know your own probabilities (cf. van Ditmarsch et al. 2015); some exceptions: Samet (1997); Williamson (2000, 2008, 2014); Lasonen-Aarnio (2015); Salow (2017).

distinctions relevant to my scenario. Propositions are modeled as subsets of W. Truth is modeled as membership, so q is true at w iff $w \in q$. Logical relations are modeled as set-theoretic ones, so: $\neg q = W - q$; $q \wedge r = q \cap r$; etc.

'C' is a definite description for my *actual* degrees of confidence—whatever they are. It can be modeled as a function from worlds w to credence functions C_w—for simplicity, suppose C_w is always a probability function over W. (Note that while 'C' is a definite description that picks out different functions at different worlds, 'C_w' is a rigid designator for the credence function I have at world w.) Using it we can define propositions (subsets of W) about what I actually think. For any proposition $q \subseteq W$ and $t \in [0, 1]$, let $[C(q) = t]$ be the proposition that I'm actually t-confident of q: $\boldsymbol{[C(q) = t] =_{df} \{w | C_w(q) = t\}}$.[10]

'P' is a definite description for the degrees of confidence I *should* have—whatever they are. It too can be modeled as a function from worlds w to probability functions P_w over W, thought of as the credences I ought to have at w. What's crucial for modeling higher-order uncertainty is that we can use P to define propositions about what I should think. For any proposition $q \subseteq W$ and $t \in [0, 1]$, $[P(q) = t]$ is the proposition that I should be t-confident of q: $\boldsymbol{[P(q) = t] =_{df} \{w | P_w(q) = t\}}$. Since we have identified facts about rational credences as propositions (sets of worlds), your (rational) credences are thereby defined for any higher-order claim about what credences you should have—that is, (rational) higher-order opinions fall right out of the model.

In sum, we can model my epistemic situation with a **credal-probability frame** $\langle W, C, P \rangle$ capturing the relevant possibilities (W), what I *actually* think in those various possibilities (C), and what I *should* think in those various possibilities (P).

I know that I should have credences that match P. I also—perhaps—know what my actual credences are. Higher-order uncertainty slips in because I may not know whether what I *actually* think (C) lines up with what I *should* think (P). If we assume that rational agents know their actual credences, such higher-order uncertainty can be rational iff there can be agents who are *in fact* rational—$[C = P]$—but who are *modest*: they are not certain that they are rational—$[C(C = P) < 1]$.

To get a grip on how this machinery works, let's construct a toy model of my case: I'm confident of *Thesis*, but I am uncertain whether I'm rational to be confident of *Thesis*. Suppose I know that I should either be 0.7 or 0.6 confident of *Thesis*. Letting T abbreviate *Thesis*, here is what we would like to say about my case:

(1) I should be sure that I should either be 0.7 or 0.6 confident of *Thesis*: $[P\big([P(T) = 0.7] \vee [P(T) = 0.6]\big) = 1]$.

(2) In fact I should be 0.7 confident of *Thesis*: $[P(T) = 0.7]$

[10] Similar definitions apply to other claims about my confidence, e.g. that I'm more confident in q than r: $[C(q) > C(r)] =_{df} \{w | C_w(q) > C_w(r)\}$.

(3) I should leave open that I should be 0.7, but also leave open that I should be 0.6: $[P(P(T) = 0.7) > 0]$ and $[P(P(T) = 0.6) > 0]$.

(4) I'm in fact 0.7 confident of *Thesis*, and I should be certain that I am: $[C(T) = 0.7]$ and $[P(C(T) = 0.7) = 1]$.

(5) My credences are in fact warranted by my evidence: $[C = P]$.

Figure 1-1 is a credal-probability frame that makes (1)–(5) true at worlds a and c.

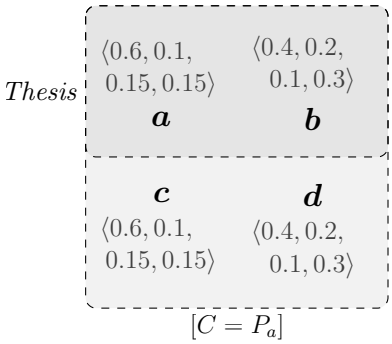


Figure 1-1: *Thesis* Uncertainty

There are four relevant epistemic possibilities: $W = \{a, b, c, d\}$. *Thesis* is true at worlds a and b, so $Thesis = \{a, b\}$ (hence $\neg Thesis = \{c, d\}$). Since I know my actual credences, C is a constant function: at each world w, C_w matches the credences that are rational at world a; $C_w = P_a$ for all w. (Indicated by the label for the shaded region covering all worlds.) The sequences next to each world w indicate the credences I should have in the various possibilities, in alphabetical order. So the '$\langle 0.6, 0.1, 0.15, 0.15\rangle$' next to a indicates that $P_a(a) = 0.6$, $P_a(b) = 0.1$, $P_a(c) = 0.15$, and $P_a(d) = 0.15$. This in turn specifies the rational credences to have in any proposition by summing across worlds: $P_a(T) = P_a(\{a, b\}) = P_a(a) + P_a(b) = 0.7$. By our definitions, $[P(T) = 0.7] = \{w | P_w(T) = 0.7\} = \{a, c\}$—at a and c I *should* be 0.7 confident in my thesis—while $[P(T) = 0.6] = \{b, d\}$—at b and d I should be 0.6.

Here's the crucial point. At worlds a and c I should be 0.7 confident of *Thesis*. Yet at those worlds I should also assign positive credence to b and d—where I should instead be 0.6 confident of *Thesis*. This means I should have higher-order uncertainty: I should be 0.7 confident of *Thesis*, but I should leave open that I should instead be 0.6 confident of it. Precisely, (1)–(5) are true at worlds a and c for the following reasons:

(1) $[P(T) = 0.7] = \{a, c\}$ and $[P(T) = 0.6] = \{b, d\}$, so $\big([P(T) = 0.7] \vee [P(T) = 0.6]\big) = W$. So $[P\big([P(T) = 0.7] \vee [P(T) = 0.6]\big) = 1] = [P(W) = 1] = W$.

(2) $[P(T) = 0.7] = \{a, c\}$.

(3) Every world w is such that $P_w(\{a, c\}) > 0$ and $P_w(\{b, d\}) > 0$, so $[P(P(T) = 0.7) > 0]$ and $[P(P(T) = 0.6) > 0]$ are true everywhere.

(4) $[C(T) = 0.7]$ is true everywhere since $[C = P_a] = W$ and $P_a(T) = 0.7$. Since $[C(T) = 0.7] = W$, $[P(C(T) = 0.7) = 1] = [P(W) = 1] = W$.[11]

(5) $[C = P]$ is true at $\{a, c\}$ since $[C = P_a]$ is true everywhere and $[P = P_a] = \{a, c\}$.

This is the framework within which I'm proposing we study higher-order evidence: the framework of higher-order uncertainty, modeled using (credal-)probability frames. It provides the formal backbone to REFRAMING. Most of my argument will consist of putting it to work.

Before doing so, two final notes. First, the framework does not presuppose anything about how the warranted credences at various worlds are related to each other (for instance, they do not have to be recoverable by conditioning from a common prior). Thus it is consistent with the view that higher-order evidence should lead you to "bracket" some of your information (Christensen 2010a)—or in some other way provides counterexamples to conditionalization. Second, W is a set of *epistemic* possibilities. Thus there is no formal problem with using such a framework to model (higher-order) uncertainty about logic. If you are unsure whether an argument is valid, we can simply add epistemic possibilities where it is (isn't)—so long as we treat the claim that the argument is valid as an atomic proposition, no formal problems will arise. Difficult interpretive questions will arise, of course—but those exist for all approaches to modeling logically non-omniscient agents.

For an initial application, let me illustrate how we can use this sort of "total-evidence" framework to define notions that correspond fairly well to the intuitive ideas of what's warranted by your first- and higher-order evidence. Intuitively, the opinions warranted by my first-order evidence (my arguments) are simply the opinions that someone who was fully informed about evidential matters—who has no doubts about what was warranted by my evidence—would think. The reason I'm unsure what my first-order evidence warrants (and, therefore, what my total evidence warrants) is that I *do* have

[11] Here is a very subtle point. At world b, the credence in *Thesis* warranted by my evidence is 0.6. Nevertheless, at b the credence warranted in the claim that my actual credence in *Thesis* is 0.7 is 1. That is, at b: $[P(T) = 0.6]$ and $[P(C(T) = 0.7) = 1]$. Though puzzling, this is correct. For in b I can tell what my actual credences are—I have overwhelming evidence that my credence in *Thesis* is in fact 0.7, thus my evidence warrants being certain of this claim. In b those credences are not warranted by my evidence—instead I should have 0.6 credence in *Thesis*. What this shows is that we can't understand P_b as giving the credences that *a rational agent* at b *would* have. Arguably, no rational agent could have credence 0.6 in *Thesis* while having my same (overwhelming) evidence that she has credence 0.7 in it—in adopting credence 0.6 in *Thesis* she would make it so that she had *different* evidence than I have (compare Salow 2017). In short, P_b captures the opinions that are warranted (rational) *given my evidence*—not necessarily the opinions that *would* be warranted if I were to *conform* to my evidence, since in so conforming I may *change* my evidence (e.g. my evidence about my beliefs).

such higher-order doubts about evidential matters. Thus to determine what my first-order evidence warrants, we can ask: what opinions would I be warranted in having *if all my higher-order doubts were removed?* That is, if I were to learn what the rational credence function was (i.e. what it was before I learned what it was), what would be the rational reaction to this information?[12]

Let's apply this thought to Figure 1. Notice that the credence function warranted at world a assigns 0.7 probability to *Thesis*—but it does so, in part, because it has higher-order uncertainty: it assigns 0.25 credence to being at b or at d, where a different credence in *Thesis* is rational. In other words, the rational 0.7 credence is modulated by higher-order doubts. At a, what would the rational opinions be if my higher-order doubts were removed? Let $\widehat{P}$ capture these opinions: $\boldsymbol{\widehat{P}_w(\cdot) =_{df} P_w(\cdot|P = P_w)}$, and $\boldsymbol{[\widehat{P}(q) = t] =_{df} \{w|\widehat{P}_w(q) = t\}}$ (cf. Stalnaker 2017). Since $\widehat{P}$ captures what the rational credences would be if higher-order doubts were removed, it can plausibly be understood as what my first-order evidence warrants. In Figure 1-1, $[P = P_a] = \{a, c\}$, so $\widehat{P}_a(T) = P_a(T|P = P_a) = \frac{P_a(T \wedge [P=P_a])}{P_a(P=P_a)} = \frac{P_a(a)}{P_a(\{a,c\})} = \frac{0.6}{0.75} = 0.8$. Hence $[\widehat{P}(\textit{Thesis}) = 0.8]$ is true at a and c, while a similar calculation shows that $[\widehat{P}(\textit{Thesis}) = 0.4]$ is true at b and d. So at a and c the first-order evidence strongly supports *Thesis* (my arguments are good), while at b and d it actually tells *against Thesis* (my arguments are bad). Yet both of these opinions are modulated by higher-order doubts to the more moderate opinions of 0.7 and 0.6.

We have the opinions warranted by your total evidence (P) and those warranted by your first-order evidence ($\widehat{P}$); what about the opinions warranted by your higher-order evidence? Intuitively, the opinions warranted by your total evidence should be "factorable" into the various possibilities you leave open for what your first-order evidence warrants, and your higher-order opinions about how likely those possibilities are to be actual. To see this, consider how we might alternatively represent Figure 1. There are two possibilities for what the first-order evidence warrants—$\{a, c\}$ and $\{b, d\}$. In this frame, each world agrees on the probability distribution *within* such cells: conditioned on $\{a, c\}$ or $\{b, d\}$, every P_w has the same distribution. The differences between the P_w are due to their distributions *across* such cells: the worlds in $\{a, c\}$ are split 75-25 between $\{a, c\}$ and $\{b, d\}$, while those in $\{b, d\}$ are split 50-50. Thus this frame can be equivalently represented using numbers *within* cells to indicate the first-order support there, and labeled arrows *between* cells to indicate the probability that the (total) evidence gives to being in each cell. That yields Figure 1-2.

In this picture, the credence in *Thesis* warranted by the total evidence at world a can be calculated by averaging the support of the two first-order-evidence cells, with

[12]Careful here. If I learn the values of P, and P had higher-order uncertainty, then I learn something that P didn't know. Thus, as we're about to see, the rational reaction to learning the values of P may be different from P itself (Elga 2013; Hall 1994, 2004).

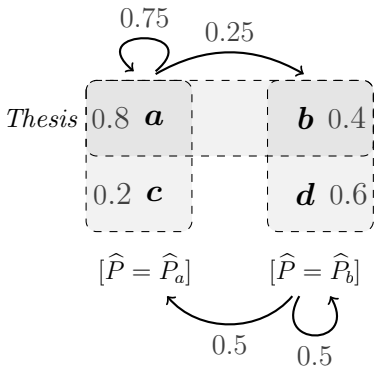


Figure 1-2: First- and Higher-Order Support

weights determined by how confident P_a is of each cell: $P_a(T) = 0.75 \cdot 0.8 + 0.25 \cdot .4 = 0.7$. Similarly, $P_b(T) = 0.5 \cdot 0.8 + 5 \cdot 0.4 = 0.6$. In fact, the reason we can redraw Figure 1-1 as Figure 1-2 is precisely because in Figure 1-1 this equality holds generally: the rational credence in q equals the rational expectation of the credence in q warranted by the first-order evidence. What do I mean? The rational expectation $\mathbb{E}_P[X]$ of a quantity X is a weighted average of the various possible values of X, with weights determined by how confident you should be in each. If you should be $\frac{1}{3}$ confident I have 3 hats and $\frac{2}{3}$ confident that I have 6 hats, then your rational expectation of my number of hats (the number you most expect to be *near* correct) is $\frac{1}{3}(3) + \frac{2}{3}(6) = 5$. Formally, $\mathbb{E}_P[X] = \sum_t (P(X = t) \cdot t)$. In the case at hand, the quantity we're estimating is the first-order support for q, $\widehat{P}(q)$, so $\mathbb{E}_P[\widehat{P}(q)] = \sum_t (P(\widehat{P}(q) = t) \cdot t)$. Given this, the principle that allows us to redraw Figure 1-1 as Figure 1-2 is:

> HIFI: $P(q) = \mathbb{E}_P[\widehat{P}(q)]$
> The rational credence in q ($P(q)$) equals the rational expectation of: the credence in q that's warranted by the first-order evidence ($\mathbb{E}_P[\widehat{P}(q)]$).

I call this principle HIFI because it captures the idea that the opinions warranted by your total evidence are factorable into your HIgher-order expectations of your FIrst-order evidence. In particular, when this principle holds it is natural to identify the opinions warranted by your *higher-order evidence* as simply the distribution across first-order-evidence cells that is warranted by your total evidence, i.e. $P(\widehat{P} = \widehat{P}_w)$ for various w. (This distribution is represented in the labeled arrows between cells in Figure 1-2.)

The upshot of this discussion is that for probability frames that validate HIFI, we can define well-behaved precisifications of the idea that some opinions are warranted by your "first-order evidence", and others are warranted by you "higher-order evidence."[13] Which frames do so?

[13] Admittedly, I have not told you what it means for a proposition to *be* first- or higher-order evidence. I have no idea how (or whether) that can be done.

HiFi turns out to be equivalent to the New Reflection principle proposed by Adam Elga (2013). New Reflection starts with the observation that if your evidence warrants being uncertain of what your evidence warrants, then if you *learn*[14] what your evidence warrants, you have gained new information that was not already entailed by your evidence. (If $P(P = P_w) < 1$, then $P(\cdot|P = P_w)$ is more informed than $P(\cdot)$.) So what should you do when you learn what opinions your evidence warrants? Elga says: adopt the opinions that your evidence *would* warrant if it were to be updated with what you've just learned. Precisely:

> New Reflection: $P(\cdot|P = P_w) = P_w(\cdot|P = P_w)$
> Upon learning the opinions warranted by the evidence, react to this information in the way that you (now) know the evidence would warrant.

New Reflection sounds truistic. It is one way of making precise the idea that your opinions should be guided by your opinions about what your evidence warrants. And it is what allows us to "factor" your total evidence into first- and higher-order components:

Fact 1. *A probability frame* $\langle W, P\rangle$ *validates* HiFi *iff it validates* New Reflection.[15]

Nevertheless, there are objections to New Reflection (Lasonen-Aarnio 2015). It is not my goal here to defend the principle, but instead merely to show that it represents a choice-point in our ability to vindicate a version of the first/higher-order evidence distinction. (And to argue—below in §1.3.2—that New Reflection and HiFi should not be seen as the *solution* to the problem of higher-order evidence.)

This concludes my proposal for how to think about higher-order evidence—the details behind Reframing. The rest of the paper applies it. §1.2 defends the Modest Truism that higher-order uncertainty is often rational, while §1.3 argues that the obvious principles for connecting first- and higher-order opinions do not succeed.

1.2 Rational Modesty

In this section I argue for:

> Modest Truism: Your total evidence often warrants being uncertain what opinions your total evidence warrants, and (hence) being *modest*: uncertain whether you're rational.

[14] New Reflection is strictly about conditional beliefs, not learning. For ease of exposition I'll switch between talk of the two—but we could reformulate everything in terms of conditional beliefs.

[15] Proof in the Appendix; cf. Stalnaker (2017). A probability frame is a credal-probability frame without C. A frame *validates* a principle iff it makes the principle true at all worlds for all instantiations on which it is well-defined.

Your evidence warrants higher-order uncertainty iff for some q and all t: $P(P(q) = t) < 1$. So long as you know what your actual opinions are, you have higher-order uncertainty iff you are modest, so I will treat modesty and higher-order uncertainty together.

Isn't it obvious that we often do—and *should*—have such self-doubts? Intuitive cases abound. *Bias*: I'm inclined to think that Kim's job talk wasn't great; but, knowing the literature, I have good reason to suspect that I have implicit bias against her—I'm probably underappreciating her talk. *Impairment*: the answer to the test's "challenge problem" seems obvious; but I'm running on four hours of sleeping—I'm probably missing something. *Disagreement*: I thought the evidence supported the defendant's innocence; but *you* thought it supported his guilt—perhaps I've misassessed it. And so on.

Clean cases can also be found (Christensen 2010a; Elga 2013; Schoenfield 2016):

> HYPOXIA
> Flying your plane, you've done some reasoning and become confident that 10,000 feet is a safe altitude (*Safe*). Then over the radio you're told there's a good chance you're hypoxic, in which case your opinions may be slightly irrational. You know, given all this information, that you should be either *somewhat* or *fairly* confident of *Safe*. In fact you are fairly confident of *Safe*.

Isn't it obvious that in HYPOXIA you should be uncertain whether (1) your fair confidence is rational, or (2) you should instead be only somewhat confident in *Safe*? That is, isn't it obvious that you shouldn't be certain of what is warranted by your *total* evidence (including the radio announcement)?

From one perspective, it certainly seems so. Being rational is *hard*. Very often we don't live up to the challenge. We know this about ourselves. So very often we should think that maybe *right now* we're not living up to the challenge—we should be unsure what it's rational to think.

But from another perspective, to admit such rational higher-order uncertainty is to give up the game. For—the thought goes—*getting to the truth* is the hard part, and the job of epistemology is to provide us with in-principle-accessible rules that ensure we do the best we can. If we allow higher-order uncertainty, we will have to deny a form of this "in-principle-accessible" claim:

> ACCESS INTERNALISM: $[P(q) = t] \rightarrow [P(P(q) = t) = 1]$
> If you should be t-confident of q, you should be certain that you should be t-confident of q.

To deny this principle is to say that sometimes you are required to have an opinion even though you can't be sure that you are so required. This can seem unacceptable:

failing to live up to requirements is grounds for criticism; how could you be legitimately criticized if you couldn't tell what was required of you? Those attracted to this line of thought will want a different way to think about our cases.

Let's focus on Hypoxia. Suppose that in this context you're *fairly confident* iff your credence is 0.7, and you're *somewhat confident* iff your credence is 0.6. Then the natural reading of the case is that you should be uncertain whether the rational credence in *Safe* is 0.6 or 0.7: $[P(P(\mathit{Safe}) = 0.7) > 0]$ and $[P(P(\mathit{Safe}) = 0.6) > 0]$. We can use the same model of me wondering about *Thesis* (from §1.1.2) to model you wondering about *Safe* (Figure 1-3). All of the above discussion applies equally well

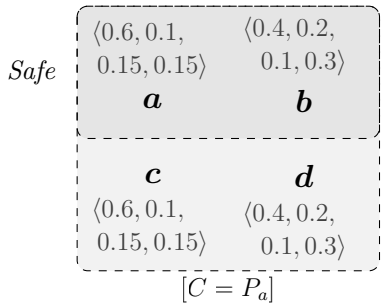


Figure 1-3: Hypoxic Uncertainty

to this Hypoxia case—you're 0.7 confident of *Safe* and should be sure that you are, you are (and should) be unsure whether you should instead be 0.6, etc.

Question: is there a recipe for generating an internalist-friendly reading of cases like this? The main strategy I know of goes as follows.[16] It's intuitive to say, of a case like Hypoxia, that "You should be uncertain of what you should think." If we interpret both those 'should's in the same way, then this says you should have higher-order uncertainty. But we needn't interpret them that way. Instead, we can interpret them as picking out different normative notions: there's (1) what you should think *given* your cognitive imperfections, and (2) what you 'should' think in the sense of what an *ideal* agent (with your evidence) would think. Thus the true reading of the sentence is: "You should (*given your imperfections*) be uncertain of what you should (*ideally*) think." Moreover, you should (ideally) know what you should (ideally) think; and you should (given your imperfections) know what you should (given your imperfections) think. Instead of higher-order uncertainty *within* a normative notion, these cases reveal first-order uncertainty *across* normative notions.

So far, so fair. But one more bit of explanation is needed: *Why* does each normative notion have no higher-order uncertainty? It's not too hard to get a sense for why you should ideally be certain of what you should ideally think—ideal agents are special, after all. But what explains why you should *given your imperfections* be certain

[16]The strategy is inspired by Stalnaker (2017)—though he may not agree with my formulation of it.

of what you should *given your imperfections* think? The line of reasoning from above is just as intuitive. Properly accounting for our imperfections is *hard*. Very often we don't live up to the challenge. We know this about ourselves. So very often we should (given our imperfections) think that maybe *right now* we're not properly accounting for our imperfections—we should (given our imperfections) be unsure what we should (given our imperfections) think. We should be modest.

An internalist may reply as follows. When we are uncertain of what we should (ideally) think, what we should (given our imperfections) do is to match our opinions to our *expectation* of what we should (ideally) think. Since we know this, we *do* know what we should (given our imperfections) think.

This strategy faces a dilemma. To illustrate, interpret "you should (ideally) have credence t" as "your *first-order evidence* warrants credence t", as defined in §1.1.2. On that definition, the first-order evidence warrants t-confidence in q ($\widehat{P}(q) = t$) iff the rational credence to have *once your higher-order doubts are removed* is t. What happens if we run the internalist reasoning using P and $\widehat{P}$? It goes as follows:

(1) You should (given your imperfections) be uncertain of what you should (ideally) think about *Safe*: $P(\widehat{P}(\textit{Safe}) = t) < 1$, for all t.

(2) You know that you should (given your imperfections) match your credence in *Safe* to your expectation of the credence $\widehat{P}(\textit{Safe})$ that you should ideally have.

(3) Therefore, you should (given your imperfections) be certain of what you should (given your imperfections) think: $[P(q) = t] \rightarrow [P(P(q) = t) = 1]$.

There is a mistake in this reasoning. Premise (2) can be interpreted in two ways. On one interpretation, it is false. On the other, it is true—but our frame validates it and (3) does not follow.

As we've seen, expectations of quantities are weighted averages of their possible values. Your *actual* expectations come from using your actual credences as the weights: $\mathbb{E}_C[X] = \sum_t (C(X = t) \cdot t)$. Premise (2) says that you should match your credence in *Safe* to your expectation of what the first-order evidence supports about *Safe*, i.e. to $\widehat{P}(\textit{Safe})$. But that premise can be interpreted in two ways, depending on whether the 'should' (represented with a '$\Box$') takes narrow- or wide-scope:

> NARROW: $[\mathbb{E}_C[\widehat{P}(q)] = t] \rightarrow \Box[C(q) = t]$
> If your expectation of the credence you should ideally have in q is *in fact* t, then you should (given your imperfections) have credence t in q.
>
> WIDE: $\Box\big([\mathbb{E}_C[\widehat{P}(q)] = t] \rightarrow [C(q) = t]\big)$
> You should (given your imperfections) be such that: if your expectation of the credence you should ideally have in q is t, then you have credence t in q.

NARROW says that whatever your actual expectation of the first-order evidence happens to be, that should determine your credence. If we assume that you know what

your actual expectations are, this rules out modesty—as the internalist hoped. But NARROW is false. Suppose we alter the description of HYPOXIA so that—for no reason at all—you expect the first-order evidence to warrant being certain of *Safe*. In this alternate case, would it follow that you should be certain of *Safe*? Of course not—what would follow is that you should change your expectation of the first-order evidence. So if we interpret Premise (2) as NARROW, (3) follows but (2) is false.

What's plausibly true is WIDE: you should be such that your actual credence lines up with your expectations of what the first-order evidence warrants. Given our assumption of unique precision, WIDE is equivalent to our familiar claim HIFI:

> HIFI: $P(q) = \mathbb{E}_P[\widehat{P}(q)]$
> The rational credence in q ($P(q)$) equals the rational expectation of: the credence in q that's warranted by the first-order evidence ($\mathbb{E}_P[\widehat{P}(q)]$).

As discussed above, Figure 1-3 validates HIFI. Thus this interpretation of premise (2) permits higher-order uncertainty: the expectations you should have ($\mathbb{E}_P[\cdot]$) depend precisely on what credences you should have (P)—therefore you cannot use knowledge of the former to gain knowledge of the latter.

Upshot: the "expectational" strategy for generating internalist-friendly readings of our cases does not succeed. Of course, this is not a refutation of ACCESS INTERNALISM. But it shifts the burden: the intuitive cases *do* put pressure on internalism.

Here, then, is how I see the dialectic. In many cases it's natural to think that you should be modest—and we have no internalist recipe for re-describing them. Nevertheless, internalists may make the principled claim that rational requirements must be accessible—that there must be *some* way to faithfully re-describe these cases. Nothing I've said so far is meant to dislodge this principled stand.

But the problem with principled stands is that they are brittle. Find a *single* exception, and the principle is shattered—the floodgates open. That is what I'll try to do. I'll argue that the epistemic force of peer disagreement cannot be accounted for without allowing rational higher-order uncertainty. With one case established, there's no reason to resist the natural hypothesis that rational higher-order uncertainty is *pervasive*.

1.2.1 Disagreement

I'll now argue that higher-order uncertainty is needed to make sense of the epistemic force of disagreement. First, the big picture.

Consider a situation in which you know that you and a peer have received independent, disjoint bodies of evidence about whether q. You are not confident of q, but then you discover that your peer *is* confident of q. All should agree that in this case you should increase your confidence in q. For you previously should've been unsure what your peer's (disjoint, independent) evidence supported. Your peer's opinion

provides you with evidence that her evidence supports q; thus you've received some evidence that *were you to pool your evidence*, the resulting more informed body of evidence would support q—you have received "evidence of evidence" for q (Feldman 2005). In general, you should defer to more informed bodies of evidence. So since you are now more confident that your pooled body of evidence supports q, you should increase your confidence in q.

Now consider a more standard peer disagreement case. You know that you and a peer have received the *same* body of evidence about whether q. You are not confident of q, but then you discover that your peer *is* confident of q. What should happen next? According to the higher-order uncertainty picture, the situation is precisely parallel. Since you previously should've been modest, you should've been unsure what your peer's (and your!) evidence supported. Your peer's opinion provides evidence that your (shared) evidence supported q; thus you've received some evidence that your evidence supported q—you have received "evidence of evidence" for q. In general, you should defer to what your evidence supports. So since you are now more confident that your evidence supported q, you should increase your confidence in q.

In short, we can use higher-order uncertainty to give a natural picture of the epistemic force of peer disagreement—one that is continuous with other widely acknowledged "evidence of evidence" effects. In fact, I'll argue that we *must*: if peer disagreement is to have the epistemic force that it's standardly taken to have, then higher-order uncertainty must be rational. I have two arguments.

The first begins with a highly circumscribed—yet paradigm—case of peer disagreement. The key features are these. First, you should be sure that you and your peer have the same relevant evidence with respect to some proposition q. Second, you should be sure that neither of you have any *non*-evidential connection to the truth.[17] My claim is that—in cases where these conditions hold—if learning that your peer disagrees with you should have *any* affect on your opinion about q, then you must have higher-order uncertainty.

To fix judgments, here's a concrete case:

> JUDGES
> You know all of the following. You and your peer Judy are judges who have served on the same court for years. You've both just heard the same case, in which you were presented with the same (relevant) total evidence. You have the same (relevant) standards of reasoning, have studied the same (relevant) cases, laws, precedents, and so on. Neither of you has any way to get to the truth, except by means of your evidence (no occult powers or special intuitions). You each have gone to form your own opinions before

[17] This is to screen off the possibility that a your credences could be further evidence for q even after the we know what opinions your evidence warrants; (cf. Levinstein 2017; Titelbaum and Kopec 2017).

convening.

Claim: in JUDGES, learning that Judy is less confident than you are that the defendant is liable (*Liable*) should lead you to change your credence that he's liable.[18] To get this verdict, you must have higher-order uncertainty.

First, the intuitive argument. Suppose you *don't* have higher-order uncertainty—you're certain of which opinion in *Liable* is warranted by your (total) evidence. Since you should be sure that you and Judy share relevant evidence, you should be sure that this opinion is likewise the opinion warranted by *Judy's* evidence. So since you're certain of what opinion Judy's evidence warrants—and you're certain she has no non-evidential connection to the truth—you will simply ignore her *actual* opinion. Contraposing: since you *shouldn't* ignore Judy's actual opinion, you *shouldn't* be certain of which opinion in *Liable* is warranted by your evidence.

Now the precise argument. What should your epistemic state look like, before you and Judy convene? Let 'P^y' be a definite description for the credences you (now) should have, and 'P^j' be one for the credences Judy (now) should have. Let 'C^j' be a definite description for Judy's actual credences. Let L be the proposition that the defendant is liable. Four premises.

First: since you know that you share evidence and standards of reasoning, you should be sure that the credence you should have in *Liable* is the same as the credence Judy should have:

Same: $[P^y\big(P^y(L) = P^j(L)\big) = 1]$

Second: since you and Judy are not perfect duplicates, you should leave open that her credence will be lower than yours:

Disagree: For all t: if $[P^y(L) = t]$, then $[P^y(C^j(L) < t) > 0]$

Third: since you know that Judy can only get to the truth via her evidence, if you learn[19] what credence she *should* have in *Liable*, then further learning what credence she *actually* has doesn't provide any evidence for or against *Liable*. So learning which credence Judy should have in *Liable* should screen off her actual credence:

Screening: For all t, s: $P^y\big(L\big|[P^j(L) = t] \wedge [C^j(L) = s]\big) = P^y\big(L\big|P^j(L) = t\big)$

Finally: if you should have a given credence t, then upon learning that Judy's credence is *lower* than that, you shouldn't simply ignore her opinion.

[18] This is a very weak claim about the force of peer disagreement. It does not presuppose any view about how your credence should shift—it simply presupposes that it *should* shift. It is compatible with the Equal Weight View (Elga 2007), the Total Evidence View (Kelly 2010), "synergistic" views (Easwaran et al. 2016), and pretty much any formal proposal for pooling credences (cf. Brössel and Eder 2014; Pettigrew 2017). I believe only Right Reasons views (Titelbaum 2015) would deny it.

[19] Again, read my talk of "learning" as shorthand for conditional beliefs.

Budge: For all t: if $P^y(L) = t$, then $P^y(L|C^j(L) < t) \neq P^y(L)$

These premises jointly imply that you should have higher-order uncertainty:

Fact 2. *If Same, Disagree, Screening, and Budge true at a world in a probability frame, so too are* $[P^y(P^y(L) = t) > 0]$ *and* $[P^y(P^y(L) \neq t) > 0]$ *for some* t.[20]

Upshot: to respect the epistemic force of peer disagreement in cases where you know that you and your peer share evidence and have no special access to the truth, higher-order uncertainty must be rational. If this is right, the principled internalist stand is shattered.

Internalists may respond by biting the bullet: grant that it is intuitive that in this case you shouldn't ignore Judy's disagreement, but insist that in fact you should. They may tell a debunking story: in *most* cases of peer disagreement, other dynamics are present which lead to rational conciliation—yet in the highly circumscribed cases described, these dynamics are gone.

I do not think such a response can work. The cases cannot be contained: to get the correct verdicts in *usual* cases of disagreement, higher-order uncertainty must be rational. This is my second argument.

Granted, *rarely* should you be certain that the credence you ought to have in q is *identical* to the credence your peer ought to have. However, *almost always* you should consider it possible that you and your peer will disagree in a way that implies that one of you was irrational. Maybe your opinions will be very far apart (*too* far apart); maybe the manner in which you disagree will reveal that you were thinking very differently (*too* differently) about a piece of evidence; maybe something else.

Suppose you are on a jury with an equally smart, equally informed peer Pete. All the evidence has been presented, but you have not yet convened to share your opinions. Let g be the proposition that the defendant is *g*uilty. Let 'P^y' and 'P^p' be definite descriptions for the opinions *y*ou and *p*ete *should* have (before convening). Let 'C^y' and 'C^p' be definite descriptions for the opinions that *y*ou and *p*ete *actually* have, before convening. Let *Disagree* be the proposition that one of your opinions (before convening) was irrational, i.e. $\textit{Disagree} =_{df} ([C^p(g) \neq P^p(g)] \vee [C^y(g) \neq P^y(g)])$. Three premises.

[20] *Proof:* Given a probability frame $\langle W, P^y \rangle$, suppose Same, Disagree, Screening, and Budge are true at w. (I won't add functions P^j and C^j; assume that the relevant propositions obey the expected logical relations.) For reductio, suppose the consequent of Fact 2 is false; so for some t', $[P^y(P^y(L) = t') = 1]$ is true. Recalling that probability frames are finite, by finite additivity, Disagree implies that there are values $s_i < t$ such that $P^y(C^j(L) = s_i) > 0$. By total probability, $P^y(L|C^j(L) < t)$ is a weighted average of the values of $P^y(L|C^j(L) = s_i)$ (with some weights possibly 0); so to establish that $P^y(L|C^j(L) < t) = P^y(L)$ it will suffice to show that $P^y(L|C^j(L) = s_i) = P^y(L)$ for each s_i. Since $P^y(C^j(L) = s_i) > 0$, by our hypothesis that $P^y(P^y(L) = t') = 1$, we have $P^y(L|C^j(L) = s_i) = P^y(L|[P^y(L) = t'] \wedge [C^j(L) = s_i])$. By Same, this equals $P^y(L|[P^j(L) = t'] \wedge [C^j(L) = s_i])$. By Screening, this in turn equals $P^y(L|P^j(L) = t')$. By Same again, this equals $P^y(L|P^y(L) = t')$, which by hypothesis equals $P^y(L)$. It follows that $P^y(L|C^j(L) < t) = P^y(L)$, contradicting Budge.

First : before you convene with Pete, you should leave open that you two will disagree. (After all, you can't be certain that *Pete* is rational.)

Open: $P^y(\mathit{Disagree}) > 0$

Second: since we are responding to Access Internalism, we may safely assume that you should be certain of your actual credence that the defendant is guilty.

Actual: $[C^y(g) = t] \rightarrow [P^y(C^y(g) = t) = 1]$

Third: what should you think if you learn that you and Pete *do* disagree? You two are equally smart—hence, initially, equally likely to be (ir)rational. If you disagree, it follows that one of you *was* irrational. It would be arbitrary (and immodest) to assume that it must have been him. So upon learning *Disagree*, you should not be certain that Pete was irrational, nor certain that you were. Let 'P^y_D' be a definite description for the rational credences you should have upon learning that you and Pete disagree, i.e. $P^y_D(\cdot) =_{df} P^y(\cdot|\mathit{Disagree})$. Then:

Uncertain: $P^y_D(C^p(g) \neq P^p(g)) < 1$ and $P^y_D(C^y(g) \neq P^y(g)) < 1$.

If Open, Actual, and Uncertain correctly describe your scenario, it follows that before you convene with Pete you should have higher-order uncertainty:

Fact 3. *If Open, Actual, and Uncertain are true at a world in a credal-probability frame, so too are* $[P^y(P^y(g) = t) > 0]$ *and* $[P^y(P^y(g) \neq t) > 0]$ *for some* t.[21]

If I am right that these premises correctly describe your scenario in *typical* cases of peer disagreement, it follows that such cases involve higher-order uncertainty.

I believe the only way to resist this argument is to deny Uncertain—claiming that in cases where you and a peer share similar evidence, if you learn that your peer disagrees with you (i.e. that one of you was irrational), then you must be certain that you were rational and they were not. This is a desperate move.

For one, we can make the case more extreme. Suppose that we re-run the scenario thousands of times and that in the cases where you *Disagree*, each one of you is

[21] *Proof:* Given a credal-probability frame $\langle W, C^y, P^y\rangle$, suppose Open, Actual, and Uncertain are all true at some world. (Again, I won't formalize C^p and P^p; assume they obey the expected logical relations.) There will be some t for which $[C^y(g) = t]$ is true. We first show that $P^y(P^y(g) = t) > 0$. Notice that $P^y(\mathit{Disagree} \wedge [P^y(g) = C^y(g)]) = P^y(\mathit{Disagree}) \cdot P^y_D(P^y(g) = C^y(g))$. By Open, the first multiplicand is > 0. The second multiplicand equals $1 - P^y_D(P^y(g) \neq C^y(g))$. Since by Uncertain the subtracted term is < 1, it follows that $P^y_D(P^y(g) = C^y(g)) > 0$ as well. Combined, we have that $P^y(\mathit{Disagree} \wedge [P^y(g) = C^y(g)]) > 0$, and so $P^y(P^y(g) = C^y(g)) > 0$. Since by Actual and our supposition that $[C^y(g) = t]$ we have $P^y(C^y(g) = t) = 1$, it follows that $[P^y(P^y(g) = t) > 0]$ is true. Next we show that $[P^y(P^y(g) \neq t) > 0]$ is also true. By parallel reasoning (through Open and Uncertain), we have that $P^y(\mathit{Disagree} \wedge [P^p(g) = C^p(g)]) > 0$. Note that every world in which $\mathit{Disagree} \wedge [P^p(g) = C^p(g)]$ is true is one in which $[P^y(g) \neq C^y(g)]$; thus $P^y(P^y(g) \neq C^y(g)) > 0$. Since $P^y(C^y(g) = t) = 1$, we've established that $[P^y(P^y(g) \neq t) > 0]$, as desired.

(ir)rational equally often. Now we run the experiment again, and you discover that you disagree. Should you be certain that *this* time it was your peer who was irrational? That seems absurd. For two, unlike my first argument, the scenario described by this second argument is pervasive—it happens every time you should think you might fundamentally disagree with a peer about morality, religion, or philosophy.

If that's right, the internalist stand is shattered. Rational higher-order uncertainty is possible—in fact, pervasive.

1.3 Enkrasia

So far I have defended Reframing (that we should think of higher-order evidence in terms of higher-order uncertainty—as a Two-Level Problem) and a Modest Truism (the Two-Level Problem is nontrivial—higher-order uncertainty can be rational). If this is right, the Two-Level Problem is well-formed and nontrivial.

But does it have a simple solution? Many have seemed to suggest so. They point out that the following states seem to be irrational: (1) believing that *my Thesis is true, but I shouldn't believe it*; (2) being confident that *my Thesis is true, but I shouldn't be confident of it*; and (3) believing *Thesis* while being agnostic on whether that belief is rational. The inferred explanation has standardly been that rationality requires your first-order opinions to "line up" with your higher-order ones—that your first-order opinions must be sanctioned (or, at least, *not disavowed*) by your higher-order opinions. Call this the **Enkratic Intuition**. Many theories of higher-order evidence have been built on top of it.[22] The Enkratic Intuition can be given a precise characterization within the higher-order-uncertainty framework. So if it is correct, our Two-Level Problem admits of a simple solution.

But it is not correct. Here I defend (cf. Titelbaum 2015):

> Akratic: If modesty is rational, so too is epistemic akrasia.

My strategy is as follows. I'll first argue that the Enkratic Intuition has a precise consequence for the relationship between your credence in q and your opinions about the rational credence in q. Then in §1.3.1 I'll show that this consequence is inconsistent leaving open the rationality of higher-order uncertainty. Since we often *should* leave open that higher-order uncertainty is rational, we should often be akratic. This does not show that (1)–(3) can be rational, of course. What it shows is that *if* they can't be, a different explanation is needed.

What does the Enkratic Intuition imply about the relationship between first- and higher-order credences? Suppose you're 0.5 confident that it'll rain tomorrow, yet

[22]Including Feldman (2005); Gibbons (2006); Christensen (2010b); Huemer (2011); Smithies (2012, 2015); Greco (2014b); Horowitz (2014); Sliwa and Horowitz (2015); Titelbaum (2015); Worsnip (2015); Littlejohn (2015); Rasmussen et al. (2016); Salow (2017), and perhaps Vavova (2014).

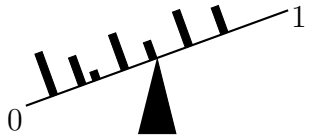
you're symmetrically uncertain whether this credence is overconfident, just right, or underconfident:

$C(\mathit{Rain}) = 0.5$, while:
$C(P(\mathit{Rain}) = 0.4) = 0.35$
$C(P(\mathit{Rain}) = 0.5) = 0.3$
$C(P(\mathit{Rain}) = 0.6) = 0.35$

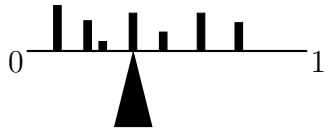
Then your 0.5 credence seems perfectly-well sanctioned by your higher-order opinion—the "pressure" from your higher-order beliefs to change your first-order credence is balanced.

Why is that? Here's a helpful metaphor (see Bertsekas and Tsitsiklis 2008, p. 83). The Enkratic Intuition suggests that to the degree you think the rational credence is higher than yours, that should "pull" your credence upward; and to the degree you think the rational credence is lower than yours, that should "pull" your credence downward. So imagine a bar labeled '0' at one end and '1' at the other is resting on a fulcrum:

Now imagine attaching a block on each spot $t \in [0, 1]$ along the bar with weight proportional to $C(P(q) = t)$—your credence that the rational credence in q is t. This will tip the scale:



Question: where would you have to put the fulcrum to *balance* the scale, so that it's not leaning left or right? (Where would you have to put your credence to balance the pull of your higher-order doubts?) Answer: you must place it at the *center of gravity* c:



This is the point at which the sum of the torques from the weights to the left of the fulcrum ($t < c$) is equal to the sum of the torques from the weights to the right ($c < t$). The torque of a block weighing $C(P(q) = t)$ at a distance $|t - c|$ is simply the weight times the distance: $C(P(q) = t) \cdot |t - c|$. A bit of algebra shows that this means that $c = \sum_t \big(C(P(q) = t) \cdot t\big)$. This formula should be familiar: it is the mathematical expectation of $P(q)$, calculated relative to C. So our metaphor leads to the plausible

conclusion that your credence in q is enkratic only if it equals your expectation of the rational credence in q. Recalling (from §1.2) that $\mathbb{E}_C[X] =_{df} \sum_t \big(C(X = t) \cdot t\big)$ is your actual expectation of X, we have:

ENKRATIC: Your credence in q is enkratic only if $C(q) = \mathbb{E}_C[P(q)]$

I claim that ENKRATIC captures what the Enkratic Intuition requires of your degrees of belief.[23] Of course, ENKRATIC is a precisification of a intuitive principle that has primarily been motivated by appeal to the irrationality of certain *outright* beliefs, so there is a risk of a terminological impasse here. My point is simply that the intuitive *motivation* for the claim that believing *p but I shouldn't believe it* is irrational is equally motivation for ENKRATIC. Thus if ENKRATIC must be rejected—as I'll argue it must—then we cannot look to the Enkratic Intuition to solve the Two-Level Problem.

Suppose ENKRATIC is correct. Then enkrasia is a rational norm only if the *rational* credence in q ($P(q)$) equals the *rational* expectation of the rational credence in q ($\mathbb{E}_P[P(q)]$):

RATIONAL ENKRASIA: $P(q) = \mathbb{E}_P[P(q)]$

Notice that this is *not* HIFI, for we are here calculating the expectation of the rational credence—not the expectation of the credence warranted by the first-order evidence. (More on this in §1.3.2.) RATIONAL ENKRASIA is a consequence of the simple REFLECTION principle discussed above: that upon learning that the rational credence in q is t, you should adopt credence t in q; $P(q|P(q) = t) = t$. So what do RATIONAL ENKRASIA and REFLECTION require?

1.3.1 Enkratic? Immodest.

They require that you be certain that you should be immodest.

If RATIONAL ENKRASIA is true, that's because it is a structural requirement that helps constrain the rational response to higher-order doubts. So if enkrasia is a rational requirement and modesty is rational, it should be possible to *know* that enkrasia is a rational requirement while being modest. (We shouldn't endorse a view on which the *only* times you can be modest are when you are uncertain whether you should be enkratic.)

Problem: it turns out that if an agent knows that they should be enkratic, they must be certain that they should be immodest. Recall that a principle is *valid* on a probability frame iff it is true at all worlds for all instantiations of its free variables. If you know RATIONAL ENKRASIA, it must be valid on the probability frame that represents your epistemic situation. What do such frames look like? Letting Sq be

[23] This conclusion is not original to me (Christensen 2010b; Sliwa and Horowitz 2015; Rasmussen et al. 2016; Salow 2017).

the proposition that you should be *S*ure of q, i.e. $Sq =_{df} [P(q) = 1]$, and 'π' be a rigid designator for a particular probability function:

Fact 4 (Samet 1997)**.** *If a probability frame validates* Rational Enkrasia*, then it validates* $S\big([P = \pi] \leftrightarrow S[P = \pi]\big)$.[24]

$[P = \pi] \leftrightarrow S[P = \pi]$ is true for all instantiations of π at a world w iff whatever the rational credence function is at w, you should be certain *that* it's the rational credence function—iff you should be immodest. So to say the frame validates $S\big([P = \pi] \leftrightarrow S[P = \pi]\big)$ is to say that you should be certain that you should be immodest. So if Rational Enkrasia is correct, anyone who knows that it is must be certain that they should be immodest. But they *shouldn't* be certain that they should be immodest. So Rational Enkrasia is incorrect.

Why does Rational Enkrasia rule out higher-order uncertainty? The basic reason was discussed in conjunction with Elga's (2013) New Reflection principle above. When higher-order doubts are rational, then learning what the rational credences are provides new evidence, and so *changes* the rational credences (Elga 2013). Metaphorically, Rational Enkrasia enjoins you to aim at a moving target.

To see why, consider the following case. You and Selena share evidence but disagree: Selena is self-confident—she's 0.9 confident that she's rational—while you are modest—you're 50-50 on whether she or you is rational. Schematically, if s is the set of possibilities where Selena is rational, y is the set where you are, and 'C^y' is a definite description for your actual credences, we have Figure 1-4.

$\boldsymbol{s}$	$\boldsymbol{y}$
$\langle 0.9, 0.1\rangle$	$\langle .5, .5\rangle$

$[C^y = P_y]$

Figure 1-4: Simple modesty

This frame illustrates why higher-order uncertainty is incompatible Rational Enkrasia. If you're rational, then the rational credence that Selena is rational ($P(s)$) is 0.5. This is equivalent to the rational expectation of the truth-value $T(s)$ of s (0 or 1): at y, $P_y(s) = \mathbb{E}_{P_y}[T(s)] = 1 \cdot P_y(s) + 0 \cdot P_y(y) = 0.5$. In contrast, the rational expectation of the *rational credence* in s is: $\mathbb{E}_{P_y}[P(s)] = 0.9 \cdot P_y(P(s) = 0.9) + 0.5 \cdot P_y(P(s) = 0.5) = 0.9 \cdot P_y(s) + 0.5 \cdot P_y(y) = 0.7$. This expectation is higher than your expectation of the truth-value of s due to the fact that the rational credence is affected by higher-order doubts. You think it's 0.5 likely that Selena's rational—in which case the rational credence of 0.9 is slightly below the truth-value of 1; but you also think it's 0.5 likely that you're rational—in which case the rational credence

[24]This result is due to Samet (1997) (although he has a different intended interpretation of P). In the Appendix I give a less mathematically-involved proof than his.

of 0.5 is well above the truth-value of 0. These two divergences are asymmetric, so they do not cancel out—which is why your credence of 0.5 that Selena is rational is below your expectation of 0.7 for the rational credence. The crucial point that given higher-order uncertainty, you should *not* be trying to get your credence as close to the rational credence as you can—for sometimes the rational credence is modulated by higher-order doubts.

This is subtle, but can become intuitive. What is not intuitive—but what Fact 4 shows—is that the example generalizes completely: there is *no* way to "balance" your higher-order doubts to respect the requirements of Rational Enkrasia, short of the trivial case of higher-order certainty. I do not know of a way of making this result seem intuitively obvious.[25] And I take that fact to be evidence that higher-order uncertainty is subtle, and that we do well to explicitly test our principles in a model theory like that of probability frames.

1.3.2 A Reply?

Perceptive readers sympathetic to the Enkratic Intuition may wonder: if the problem with "trying" to get close to the rational credence is that it's plagued with higher-order doubts, why don't we reformulate the enkratic requirement to aim at what your *first-order* evidence supports? Doing so would yield HiFi:

> HiFi: $P(q) = \mathbb{E}_P[\widehat{P}(q)]$
> The rational credence in q ($P(q)$) equals the rational expectation of: the credence in q that's warranted by the first-order evidence ($\mathbb{E}_P[\widehat{P}(q)]$).

As discussed above, *this* principle permits higher-order uncertainty. Why isn't *it* the proper precisification of the Enkratic Intuition?

Because it does not explain the cases that motivate that intuition. Granted, it *does* explain why attitudes like being confident that *it'll rain, but my first-order evidence doesn't support that* are irrational. But it *doesn't* explain why the following attitudes are irrational: (1) being confident that *it'll rain, but my <u>total</u> evidence doesn't support that* or (2) being very confident that *it'll rain, but my total evidence warrants being very confident it won't.* If any attitude is epistemically akratic, these ones are. Yet HiFi allows them. Example: two sycophants, Sybil and Phan, are each confident that the other person is the rational one. Sybil is 0.9 confident that Phan is rational, while Phan is 0.9 confident that Sybil is. If these opinions could be rational, we'd have a probability frame like Figure 1-5 (s is the possibility where Sybil is rational, and p is the possibility where Phan is). Since both P_s and P_p agree on everything

[25]Here's my best attempt. Expectations are estimates. $\mathbb{E}_P[P(q)]$ is the rational expectation of $P(q)$. If Rational Enkrasia is valid on a frame, then at all worlds, the rational estimate of $P(q)$ is always equal to $P(q)$ itself. But how could you *know* that your estimate of a quantity should exactly equal that quantity, unless you know that you will be able to know what that quantity is?

Figure 1-5: Sycophants

when they update on the claim that one of them is rational, this frame validates New Reflection and HiFi. Yet it gives rise to a paradigm case of akrasia:

> $P_s(p \wedge [P(\neg p) \geq 0.9]) \geq 0.9$
> At s they should be very confident that *Phan is rational but we should be very confident that he's not.*

Such an attitude looks akratic, if anything does. Any principle that allows it—like HiFi—cannot capture the Enkratic Intuition.

The upshot is clear: the Enkratic Intuition cannot form the foundation of a theory of higher-order uncertainty. We need some *other* principle to explain the irrationality of such radical splits between first- and higher-order attitudes.

1.4 Proposal

I've defended three claims. The first is Reframing: the problem of higher-order evidence is best formulated as one of higher-order *uncertainty*—as a Two-Level Problem about how first- and higher-order opinions should relate. The second is Modest Truism: such higher-order uncertainty is often rational—the solution to the Two-Level Problem is not trivial. The third is Akratic: the Enkratic Intuition is incorrect—the solution to the Two-Level Problem is not straightforward.

Where does this leave us? There are many reasons to want a general, strong theory of higher-order uncertainty. Such a theory would provide a foundation to the epistemologies of disagreement, debunking, and self-doubt. It would do so by formulating law-like principles connecting rational first- and higher-order opinions. It would use a framework like ours to both show that such principles are tenable and illustrate their consequences.

I have proposed such a theory elsewhere (Dorst 2019a). But the framework proposed here is compatible with many, many alternatives. I hope to have done enough to show that it provides fruitful terrain—terrain that is well worth exploring.[26]

Appendix

Fact 1. *A probability frame* $\langle W, P \rangle$ *validates* HiFi *iff it validates* New Reflection.

[26]Thanks to Bernhard Salow, Miriam Schoenfield, Mattias Skipper, Bob Stalnaker, and Roger White, and an anonymous referee for helpful discussion and feedback.

Proof. We prove a LEMMA: in any frame, $[P = P_x] = [\widehat{P} = \widehat{P}_x]$. If $y \in [P = P_x]$, then $P_y = P_x$, so $\widehat{P}_y = P_y(\cdot|P = P_y) = P_x(\cdot|P = P_x) = \widehat{P}_x$, so $y \in [\widehat{P} = \widehat{P}_x]$. If $y \notin [P = P_x]$, then since $\widehat{P}_y(P = P_y) = 1$, $\widehat{P}_y(P = P_x) = 0$ while $\widehat{P}_x(P = P_x) = 1$, so $\widehat{P}_y \neq \widehat{P}_x$, so $y \notin [P = \widehat{P}_x]$.

Now suppose that NEW REFLECTION is valid. Taking an arbitrary world w and proposition q, since $\{[P = P_x]\}$ forms a partition:

$$\begin{aligned}
P_w(q) &= \sum_{P_x} P_w(P = P_x) \cdot P_w(q|P = P_x) && \text{(total probability)} \\
&= \sum_{P_x} P_w(P = P_x) \cdot P_x(q|P = P_x) && \text{(NEW REFLECTION)} \\
&= \sum_{P_x} P_w(P = P_x) \cdot \widehat{P}_x(q) && \text{(definition)} \\
&= \sum_{P_x} P_w(\widehat{P} = \widehat{P}_x) \cdot \widehat{P}_x(q) \;=\; \mathbb{E}_{P_w}[\widehat{P}(q)] && \text{(LEMMA, DEFINITION)}
\end{aligned}$$

So HIFI is valid.

For the converse, suppose that NEW REFLECTION is *not* valid, so there is a w, q, x such that $P_w(q|P = P_x) \neq P_x(q|P = P_x)$. Consider the proposition $q \wedge [P = P_x]$. $P_w(q \wedge [P = P_x]) = P_w(P = P_x) \cdot P_w(q \wedge [P = P_x]|P = P_x) =$

$$P_w(P = P_x) \cdot P_w(q|P = P_x) \tag{1}$$

Meanwhile, $\mathbb{E}_{P_w}[\widehat{P}(q \wedge [P = P_x])] = P_w(P = P_x) \cdot \widehat{P}_x(q \wedge [P = P_x]) =$

$$P_w(P = P_x) \cdot P_x(q|P = P_x) \tag{2}$$

Since by hypothesis $P_w(q|P = P_x) \neq P_x(q|P = P_x)$, it follows that (1) $\neq$ (2), and hence that $P_w(q \wedge [P = P_x]) \neq \mathbb{E}_{P_w}[\widehat{P}(q \wedge [P = P_x])]$, so HIFI is not valid. □

Fact 4 (Samet 1997)**.** *If a probability frame validates* RATIONAL ENKRASIA*, then it validates* $S\big([P = \pi] \leftrightarrow S[P = \pi]\big)$.

I will write $\mathbb{E}_w[P(q)] =_{df} \sum_{x \in W} \big(P_w(x) \cdot P_x(q)\big)$ for the values of the expectation of $P(q)$ at w. Given a probability frame $\langle W, P\rangle$, there is an induced binary relation $\boldsymbol{R}$ such that wRx iff $P_w(x) > 0$ and $R_w = \{x|P_w(x) > 0\}$.

Lemma 4.1. *If* $\langle W, P\rangle$ *validates* RATIONAL ENKRASIA*, then* R *is transitive: if* wRx *and* xRy *then* wRy.

Proof. Suppose wRx and xRy but $w\not Ry$. Since xRy, $P_x(y) > 0$. Since wRx, $P_w(x) > 0$. Therefore $\mathbb{E}_w[P(y)] \geq P_w(x) \cdot P_x(y) > 0$. Since by hypothesis $w\not Ry$, $P_w(y) = 0$, contradicting RATIONAL ENKRASIA at w. □

Lemma 4.2. *If $\langle W, P\rangle$ validates* Rational Enkrasia*, then R is shift-reflexive: if wRy, then yRy.*

Proof. Suppose the frame validates Rational Enkrasia, so by Lemma 4.1 R is transitive. Suppose for reductio that wRy but $y\not Ry$. Since R is transitive, for all $z_i \in R_y$, $P_{z_i}(R_y) = 1$. And since $y \notin R_y$ but $P_y(R_y) = 1$ and $P_x(y) > 0$, we have

$$\begin{aligned}\mathbb{E}_w[P(R_y)] &\geq \sum_{z_i \in R_y} (P_w(z_i) \cdot P_{z_i}(R_y)) \ + \ P_w(y) \cdot P_y(R_y) \\ &= \sum_{z_i \in R_y} P_w(z_i) \cdot 1 \ + \ P_w(y) \cdot 1 \ > \ \sum_{z_i \in R_y} P_w(z_i) = P_w(R_y)\end{aligned}$$

Contradicting Rational Enkrasia at w. □

Lemma 4.3. *If $\langle W, P\rangle$ validates* Rational Enkrasia*, then R is shift-symmetric: if wRy and yRz, then zRy.*

Proof. Suppose the frame validates Rational Enkrasia, and so by Lemmas 4.1 and 4.2 R is transitive and shift-reflexive. Suppose for reductio that wRy and yRz but $z\not Ry$. By transitivity, all $z_i \in R_z$ are such that $P_{z_i}(R_z) = 1$. By shift-reflexivity, $P_y(y) > 0$ and $z \in R_z$, so $P_y(R_z) > 0$. Finally, $y \notin R_z$. Combining these facts:

$$\begin{aligned}\mathbb{E}_y[P(R_z)] &\geq \sum_{z_i \in R_z} (P_y(z_i) \cdot P_{z_i}(R_z)) \ + \ P_y(y) \cdot P_y(R_z) \\ &= \sum_{z_i \in R_z} P_y(z_i) \cdot 1 \ + \ P_y(y) \cdot P_y(R_z) > \sum_{z_i \in R_z} P_y(z_i) = P_y(R_z)\end{aligned}$$

Contradicting Rational Enkrasia at y. □

Lemma 4.4. *If $\langle W, P\rangle$ validates* Rational Enkrasia*, then for all $w \in W$: if wRy then $P_y(P = P_y) = 1$.*

Proof. Suppose the frame validates Rational Enkrasia, so by Lemmas 4.1 and 4.3, R is transitive and shift-symmetric. Suppose for reductio that wRy but $P_y(P = P_y) < 1$. By transitivity, if yRz and zRx, then yRx; equivalently: if yRz then $R_z \subseteq R_y$. By shift-symmetry, if yRz, then zRy; so by transitivity $R_y \subseteq R_z$ as well. Combined: if yRz, then $R_y = R_z$.

Since $P_y(P = P_y) < 1$, there must be a proposition q such that $P_y(P(q) = t) < 1$ for all t. Since W is finite, there is a set $T = \{t_1, ..., t_n\}$ such that for all t_i, $P_y(P(q) = t_i) > 0$, with at least two distinct $t_i \neq t_j$ in T. Relabel so that $t_1 < t_2 < ... < t_n$. There must be some $z \in R_y$ such that $P_z(q) = t_n$. By the above reasoning, $R_z = R_y$,

meaning T is also the set of values s such that $P_z(P(q) = s) > 0$. Then:

$$\begin{aligned}\mathbb{E}_z[P(q)] &= \sum_{t_i \in T} (P_z(P(q) = t_i) \cdot t_i \\ &= \sum_{t_i < t_n} P_z(P(q) = t_i) \cdot t_i \ + \ P_z(P(q) = t_n) \cdot t_n\end{aligned}$$

This is a weighted average with the highest possible value t_n. Since there are at least two values in T, the left summand has positive weight; thus the average is $\mathbb{E}_z[P(q)] < t_n$. Since $P_z(q) = t_n$, this contradicts Rational Enkrasia at z. □

Fact 4 is an immediate consequence of Lemma 4.4.

Chapter 2

Evidence: A Guide for the Uncertain

Abstract

Assume that it is your *evidence* that determines what opinions you should have. I argue that since you should take peer disagreement seriously, evidence must have two features. (1) It must sometimes warrant being *modest*: uncertain what your evidence warrants, and (thus) uncertain whether you're rational. (2) But it must always warrant being *guided*: disposed to treat your evidence as a guide. It is surprisingly difficult to vindicate these dual constraints. But diagnosing why this is so leads to a proposal—Trust—that is weak enough to allow modesty but strong enough to yield many guiding features. In fact, I argue that Trust is the *Goldilocks principle*—for it is necessary and sufficient to vindicate the claim that you should always prefer to use free evidence. Upshot: Trust lays the foundations for a theory of disagreement and, more generally, an epistemology that permits self-doubt—a modest epistemology.

2.1 A Modest Guide

Here is a spoon:

And here is another:

Which spoon is longer? Let *Top* be the claim that the top one is. There are opinions you should have about *Top*—perhaps you should be confident of it. There are opinions you should have about many other propositions as well. I will assume that it is your *evidence* that makes it so. But I am neutral beyond that—all I assume is that your evidence determines what opinions you should have. Thus: You should (would be rational to) have an opinion iff your evidence warrants having that opinion.[1] Epistemology studies evidence. And this paper studies its structure.

Two structural features, in fact.

One. Your evidence determines what *you* ought to think. Yet you are a fallible critter—and you know it. Even in simple cases, you might misjudge the force of your evidence: although you're pretty confident of *Top* (let's suppose), maybe it should be obvious that the top spoon is longer (you're underconfident)—or maybe you're being tricked by its shape (you're overconfident). You're aware of your fallibility, so you have evidence that requires being unsure whether you've properly conformed to your evidence. That's our first structural feature:

> Modest Truism
> Your evidence sometimes requires being *modest*: uncertain what your evidence requires.

Let 'P' be a definite description that picks out the credences—whatever they are—that are warranted by your evidence. P captures the opinions you should have about any relevant proposition; thus it captures the first-order opinions you should have about *Top*, as well as the *higher-order opinions* you should have about what first-order opinions you should have about *Top*. Our Modest Truism says that such first- and higher-order opinions should sometimes come apart—sometimes you should have credence t in p but be less than certain that you should: $P(p) = t$, but $P(P(p) = t) < 1$. Maybe you should be 0.7 confident of *Top*: $P(\textit{Top}) = 0.7$. But you definitely shouldn't be *certain* that you should be—you should leave open that you should be 0.6 or 0.8 instead: $P(P(\textit{Top}) = 0.6) > 0$ and $P(P(\textit{Top}) = 0.8) > 0$.[2]

[1] *Assumption Alert*: (1) I will use precise probabilities to model rational opinions (cf. White 2009a; Elga 2010; Joyce 2010); and (2) I will assume that your evidence always warrants a unique opinion (cf. White 2005; Feldman 2007; Schoenfield 2014). These are modeling choices: (1) *however* we represent rational opinions, it will be rational to be unsure what the rational opinion is; and (2) *whatever* features (evidence, priors, etc.) collectively determine a uniquely rational opinion, they will need the structural features I discuss. I understand these assumptions to be *descriptive* idealizations—like point-particles on frictionless planes—not normative ones. They are intended to provide simplified models of how *you* and *I* ought to think, rather than exact models of how our ideal counterparts would.

[2] Our Modest Truism allows rational agents to know what their credences are. Let 'C' be a definite description for your actual credences. Then you may in fact be rational ($[C = P]$ is true) and be certain of what your actual credences are ($[C(p) = t] \rightarrow [C(C(p) = t) = 1]$ holds), but be unsure whether your actual credences are rational: $C(C = P) < 1$. I will suppress talk of C, since it is the rational credences P (and so, primarily, wide-scope norms) we are interested in. Note: 'P' is descriptive even when embedded in larger constructions; '$P(P(p) = 0.5) = 0.3$' is to be read, 'The

Two. Your evidence determines what you *ought* to think. Arguably, this means that you must have reason to treat your evidence as a guide: your higher-order opinions about what you *ought* to think about *Top* should constrain your first-order opinions about *Top*. That's our second (not yet fully precise) structural feature:

> Guiding Truism
> Your evidence always warrants being *guided*: disposed to treat your evidence as a guide.

There are many things you should be disposed to treat as guides—chances (Lewis 1980), gurus (Elga 2007), future selves (van Fraassen 1984). But evidence must be a *modest* guide—a guide that is not sure that it is a guide. That's...

The Problem: We must vindicate both our Modest and Guiding Truisms. This is a problem of modulation: our Modest Truism allows you to leave open various possibilities for what the rational opinions might be; our Guiding Truism requires that these various possible rational opinions are correlated with truth. It's a hard problem. Many have argued that it can't be solved—that there's no principled center of gravity that allows modesty but requires guidance.[3] The "higher-order evidence" literature contains three responses to this challenge. *Splitters* accept our Modest Truism and so deny our Guiding one—they allow your first- and higher-order opinions to split radically apart.[4] *Mergers* accept our Guiding Truism and so deny our Modest one—they require you to be certain of what you should think.[5] *Bridgers* try to find a middle way—they search for principles weak enough to vindicate our Modest Truism but strong enough to vindicate our Guiding one.[6] Short story: Bridging is in trouble—the current proposals are all either too strong or too weak (Lasonen-Aarnio 2014, 2015; Dorst 2019b).

The Project: Show that Bridging succeeds, after all. There *is* a principled center of gravity that allows modesty while requiring guidance. We *can* construct an epistemology that makes room for modest critters—like you and me—that have a rational dose of self-doubt.

The Plan: After making the case that we should want a Bridging principle (§2.2), I'll explain why the standard one is too strong (§2.3) and propose a fix (§2.4). The rest

---

rational credence function (whatever it is) assigns 0.3 credence to the claim that the rational credence function (whatever it is) assigns 0.5 credence to p.'

[3]Smithies (2012, 2015); Lasonen-Aarnio (2014, 2015); Titelbaum (2015); Salow (2017); Horowitz (2018).

[4]Williamson (2000, 2014); Lasonen-Aarnio (2010, 2014, 2015); Coates (2012); Hazlett (2012); Wedgwood (2012).

[5]Smithies (2012, 2015); Greco (2014a); Titelbaum (2015); Salow (2017); cf. Tal (2018).

[6]Feldman (2005); Gibbons (2006); Elga (2007, 2013); White (2009b); Christensen (2010a,b, 2016); Huemer (2011); Horowitz (2014); Pettigrew and Titelbaum (2014); Vavova (2014, 2016); Littlejohn (2015); Schoenfield (2015a,b, 2016); Sliwa and Horowitz (2015); Worsnip (2015).

of the paper argues that this is the correct fix: it has the modest and guiding features we're after (§2.5), avoids the paradoxes of weaker proposals (§2.6), coincides with an independent characterization of our Guiding Truism (§2.7), and promises fruitful applications (§2.8).

2.2 A Disagreement

I've offered an intuitive narrative in favor of Bridging. I'll now offer a precise argument: if you should take peer disagreement seriously, our Modest and Guiding Truisms must both be true. Thus Bridging lies at the foundations of the epistemology of disagreement.

Take a paradigm case. Looking at the spoons, you and your colleague Disa know that you share *Top*-related evidence, and so should have the same opinion about *Top*. (If this seems implausible, stipulate that the Epistemology Oracle has announced it.) Suppose that you start out confident of *Top*—and that, in fact, such confidence is warranted by the evidence. Now you discover a disagreement: Disa is *not* confident of *Top*. Claim: this should make you less confident of *Top*. More generally: if you're rational and know that your peer has the same relevant evidence, then your confidence should drop when you learn that she's less confident than you.[7]

First fact: this requires that you be modest. To see why, suppose you were *im*modest: you were confident of *Top*, and were certain that you should be. Since you know that you share *Top*-related evidence, you'd infer that *Disa's* evidence warrants confidence in *Top*. So you'd think, "Disa's not confident of *Top*. But given her evidence, she *should* be. Since her only route to the truth is through her evidence (it includes all facts that determine what she should think, after all), her lack of confidence doesn't tell me anything about *Top*. So I should simply ignore her unwarranted opinion and maintain my confidence in *Top*." Thus if you were immodest, you'd ignore her opinion. Since you *shouldn't* ignore her opinion, you *shouldn't* be immodest. That's our Modest Truism.

More rigorously, let 'P_D' and 'C_D' be definite descriptions for the credences Disa should have and in fact has, respectively. Your evidence warrants being sure that you and Disa ought to have the same opinion about *Top*, i.e. $P\big(P(\mathit{Top}) = P_D(\mathit{Top})\big) = 1$. Since the only way Disa can connect her credence with the truth of *Top* is through her evidence[8], once you know which opinion her evidence warrants, further learn-

[7]This claim presupposes only that—in such highly circumscribed cases—you should not *completely ignore* your peer's lack of confidence. Although Right Reasons theorists may object (Titelbaum 2015), most others will agree (e.g. Christensen 2007, 2010a; Elga 2007; Feldman 2007; Roush 2009; Kelly 2010; Lasonen-Aarnio 2013; Vavova 2014). Moreover, my claim is compatible with "synergistic" views (Easwaran et al. 2016) since what you are learning is that Disa is *less* confident than you—not that she has a particular credence which is slightly less.

[8]Titelbaum and Kopec (2017) and Levinstein (2017) give cases where your peer has non-evidential routes to the truth, but we can stipulate that our case is one where you know this isn't so.

ing[9] her *actual* opinion won't affect the probability of *Top*: $P(\mathit{Top}|P_D(\mathit{Top}) = s) = P(\mathit{Top}|[P_D(\mathit{Top}) = s] \wedge [C_D(\mathit{Top}) = t])$. Finally, you should leave open that Disa may have a lower credence than you: $P(C_D(\mathit{Top}) < 0.7) > 0$. It follows that if you should be immodest $(P(P(\mathit{Top}) = 0.7) = 1)$, then you should ignore Disa's lack of confidence: $P(\mathit{Top}|C_D(\mathit{Top}) < 0.7) = P(\mathit{Top})$.[10] Since you shouldn't ignore her lack of confidence, you shouldn't be immodest.

So suppose you are modest: you are uncertain what your evidence warrants. Since Disa is a peer, learning that she's not confident of *Top* gives you reason to think that your shared evidence *doesn't* warrant confidence—despite what you initially thought. Second fact: in order for this change to require you to lower your credence in *Top*, you must be guided by the evidence. For if your opinions about your evidence didn't constrain your opinions about *Top*, then the change in higher-order opinion induced by Disa's disagreement needn't lead to a change in first-order opinion.

To illustrate, consider one way our Guiding Truism could fail. Suppose your evidence could warrant the following "akratic" state: being confident that *Top is true, but I shouldn't be confident of it*; $P(\mathit{Top} \wedge [P(\mathit{Top}) < 0.7]) \geq 0.7$.[11] Then even if Disa's disagreement made you *certain* that you shouldn't be confident of *Top*, you should still maintain your confidence in *Top*—after all, you're confident that possibilities where you shouldn't be confident in *Top* are ones where it's true![12] Generalizing our puzzle:

> **Misguided Evidence:** $\exists p, t: \ P(p \wedge [P(p) < t]) \geq t$
> You should have confidence that: *p but I shouldn't have confidence that p.*

If Misguided Evidence were possible, then in a parallel case (in which you start off t-confident and then discover that Disa's credence is lower), you needn't lower your credence in p. So if you must *always* take such disagreement seriously, Misguided Evidence must be impossible. More generally, your opinions about whether your evidence supports p must always constrain your opinions about p. That's our Guiding Truism.

Upshot: To account for the force of peer disagreement, we must build a Bridging theory that vindicates both our Modest and Guiding Truisms. Construction ahead.

---

[9] Strictly, all the principles I will discuss are about conditional beliefs, not about learning—but for ease of exposition I'll elide this distinction.

[10] $P(\mathit{Top}|C_D(\mathit{Top}) < 0.7)$ is an average of $P(\mathit{Top}|C_D(\mathit{Top}) = t)$ for the various $t < 0.7$. Since $P(P(\mathit{Top}) = 0.7) = 1$, we know that each such $P(\mathit{Top}|C_D(\mathit{Top}) = t) = P(\mathit{Top}|[P(\mathit{Top}) = 0.7] \wedge [C_D(\mathit{Top}) = t]) = P(\mathit{Top}|[P_D(\mathit{Top}) = 0.7] \wedge [C_D(\mathit{Top}) = t]) = P(\mathit{Top}|P_D(\mathit{Top}) = 0.7) = P(\mathit{Top}|P(\mathit{Top}) = 0.7) = P(\mathit{Top})$.

[11] This is a probabilistic version of what many have called "epistemic akrasia"—the paradigm instance being a belief in *p but I shouldn't believe it* (Smithies 2012; Horowitz 2014; Titelbaum 2015). But there are reasons to think that akrasia is not the distinctive feature of such attitudes (Dorst 2019b).

[12] If $P(\mathit{Top} \wedge [P(\mathit{Top}) < 0.7]) \geq 0.7$, then $P(\mathit{Top}|P(\mathit{Top}) < .7) = \frac{P(\mathit{Top} \wedge [P(\mathit{Top}) < 0.7])}{P(P(\mathit{Top}) < .7)} \geq 0.7$.

2.3 A Reflection

Looking at the spoons, you should be modest: there are various epistemic states that you should leave open might be the rational one. Here is a metaphorical (and literal) way to conceptualize this scenario (cf. Elga 2013; Hall 1994). Imagine you are a member of an epistemic panel: a group of candidates who share your evidence and have different opinions in response. (Literally: a set of credence functions.) Everyone knows what everyone else's opinions are. (Literally: each credence function is certain of the values of the other functions.) One of the candidates—the most diligent one—is *the expert*. (Literally: one of the credence functions is the one warranted by the evidence.) You *should* have an opinion iff the expert *does* have that opinion. (Literally: you should have an opinion iff your evidence warrants that opinion.) But by our Modest Truism, the expert may be unsure who the expert is—the most diligent person on the panel may not *know* they're the the most diligent person on the panel. (Literally: the credence function warranted by the evidence may be uncertain which credence function is warranted by the evidence.) Different candidates have different opinions about who the expert might be: you're sure it's either Disa, Carl, or yourself; Carl is sure it's either Disa or Betty, etc. (Literally: different credence functions assign different probabilities to claims about who the expert is.) Yet—by our Guiding Truism—all the candidates will defer when they learn about the expert's opinions. (Literally: each credence function will have conditional probabilities that defer to facts about the evidence.) The question is how.

Our story begins with the (seemingly) obvious answer. I'll say that the expert has a given **opinion**[13] about p iff their credence falls in a contextually specified range: $P(p) \in [l, h]$. Obvious answer: *defer to expert opinions* (cf. Skyrms 1980; van Fraassen 1984; Gaifman 1988; Christensen 2010b). That is, your opinions should be a reflection of the expert's: conditional on the expert being 0.6 confident of *Top*, be 0.6 confident of it; conditional on the expert being between 0.5 and 0.8 confident of *Top*, be between 0.5 and 0.8 confident of it; and so on. Letting $P(p|q)$ be the rational credence in p conditional on q:

> **Reflection:** $P(p|P(p) \in [l, h]) \in [l, h]$[14]
> Upon learning that your evidence warrants a given opinion, have that opinion.
> Slogan: *defer to expert opinions*.

Reflection says to treat the expert as a guide by simply adopting whatever opinions

[13] *Convention:* technical terms to be used in the statements of principles and theorems are bolded when defined; their definitions are collected in Appendix 2.C.

[14] Look strange? Some might expect to see different probability functions on the inside and outside, e.g. with your actual credences C deferring to the rational credences: $C(p|P(p) \in [l, h]) \in [l, h]$. But that principle is false—your *actual* credences can be whatever you like. What's true is that you *ought* to have credences that obey this principle; letting '$\Box q$' mean 'it ought to be that q': $\Box(C(p|P(p) \in [l, h]) \in [l, h])$. But your credences ought to be the rational credences: $\Box(C(p|q) = t) \Leftrightarrow [P(p|q) = t]$; hence our principle.

you find out they have. It clearly vindicates our Guiding Truism. What could go wrong?

Modesty could. Reflection is inconsistent with our Modest Truism: it requires you to be certain that the expert is *im*modest. Consider an example. Looking at the spoons, you have two people on your panel—Imani and yourself. Imani is immodest: she's certain she's the expert. You are modest: you're 50-50 on whether you or Imani is the expert. This simple case is inconsistent with Reflection. For conditional on the expert being 0.5 confident that Imani's the expert, how confident should you be that Imani's the expert? Reflection says to adopt the expert's credences: be 0.5. But that's wrong. Imani is certain that she's the expert. You are 0.5 confident that she is. So if the expert (whoever it is) is 0.5 confident that Imani's the expert, then the expert (whoever it is) is *not* Imani—it's you. Conditional on the expert being 0.5 confident that Imani's the expert, you should have credence *0* that she is: $P(\mathit{Imani}|P(\mathit{Imani}) = 0.5) = 0$. Reflection fails.

Surprisingly, the example generalizes completely: Reflection requires you to be certain that you ought to be immodest. Letting *Immodest* be the proposition that the rational credence function is certain of what the rational credence function is:

Fact 2.3.1. *If a probability frame validates Reflection, it validates* $[P(\mathit{Immodest}) = 1]$.[15]

Upshot: although Reflection vindicates our Guiding Truism, it is incompatible with our Modest one. A weaker principle is needed.

2.4 On Trust

Fact 2.3.1 shows *that* Reflection is too strong. But—in order to refine Reflection—we need to know *why* it's too strong. There are two parts to the explanation.

First: if you should be uncertain what your evidence warrants, then learning facts about your evidence can give you *new* evidence—and so can *change* what it's rational to think. This is what happens in the Imani case: when you learn that the expert is 0.5 confident that Imani's the expert, you learn that *you* are the expert. When the expert (i.e. you) formed their 0.5 credence, they didn't know that. So after learning something about the expert's opinions, you now have information that the expert didn't have when they formed those opinions. Thus you should react to this information *not* by adopting the opinion they *had*, but rather by adopting the opinion they *would* have *were they to learn what you've learned*. When you know more than

[15] A **probability frame** is a structure for modeling the opinions—including higher-order opinions—that a given agent should have in a given scenario. Details are in Appendix 2.A. A probability frame **validates** a principle iff the principle is true at all worlds for all well-defined instantiations of its free variables. Samet (1997), Williamson (2000, 2014), Elga (2013), and Dorst (2019b) prove similar results.

the expert, you should react to your information *as you know the expert would* (Elga 2013).

But when the expert learns that their opinion about p was rational, why would that lead them to *change* their opinion? This is the second part of the explanation: the expert will think that *the rational credence in p is correlated with the truth-value of p*, and therefore learning that their opinions were rational can sometimes be evidence *against p*. This implies that Reflection must fail.

Return to the Imani case. Letting p be the proposition that Imani's the expert, the expert is in fact 0.5 confident of p: $P(p) = 0.5$. Moreover, the expert is certain that the expert's credence in p (whatever it is) is either 0.5 or 1: $P\big([P(p) = 0.5] \vee [P(p) = 1]\big) = 1$.[16] The expert will think that the rational credence is correlated with truth—so since there are only two possible values of the rational credence (0.5 and 1), they will think that p is more likely to be true if the rational credence is higher than if it's lower:

$$P(p|P(p) = 0.5) \;<\; P(p|P(p) = 1)$$

But this implies a Reflection failure. Since the expert's original opinion $P(p)$ is an average of the conditional opinions $P(p|P(p) = 0.5)$ and $P(p|P(p) = 1)$, it follows that upon learning that the rational credence in p was the lower value, the expert will *drop* their credence below its original value of 0.5:

$$P(p|P(p) = 0.5) \;<\; P(p) \;=\; 0.5$$

In short: learning that the expert has a given opinion about p can sometimes be evidence that the rational credence is *lower* than you expected it to be, and therefore—since rational credence is correlated with truth—can be evidence *against p*. (Similarly for any upper-bounded opinion $P(p) \in [l, h]$—in the Imani case learning that the rational credence in p is between 0.5 and 0.8 would *also* be evidence against p: $P(p|P(p) \in [0.5, 0.8]) \;<\; P(p) \;=\; 0.5$.)

So Reflection fails because learning that the expert has a given opinion about p can sometimes be evidence against p. To refine it, we must find a type of information about the rational credence in p that is *never* evidence against p. What could it be?

Consider the claim that the expert is *at least t-confident* of p: $P(p) \geq t$. When this is so, I'll say that the expert **judges** that p (to the contextually specified degree t). We could say that judgments are special cases of opinions ($P(p) \in [l, h]$ where

[16]For aficionados: *this* is the point at which the argument I'm about to give would fail as an argument against (say) the Principal Principle (Lewis 1980). Since the Principal Principle connects two different probability functions—the rational credence and chance—it can forbid the possibility that the rational credence equals the lowest possible chance. However, principles that connect rational credence *with rational credence* must allow that whenever you know what the (finite) range of possible rational credences is, there will be a possibility where the actual rational credence is the lowest possible rational credence—as in our case where $P(p) = 0.5$ while $P\big([P(p) = 0.5] \vee [P(p) = 1]\big) = 1$.

$h = 1$). Or we could say that opinions are conjunctions of judgments.[17] However we carve it up, judgments are important. Why?

Because the claim that the rational credence in p is at least t can only provide evidence that the rational credence is *higher* than you originally thought, and so—since rational credence is correlated with truth—can *never* be evidence against p.

More precisely, suppose—for reductio—that learning that the expert judges p provides evidence *against* p:

$$P(p|P(p) \geq t) \;<\; P(p)$$

From this it follows (by total probability) that learning that the rational credence is *at least* t should lead you to *lower* your credence in p, while learning that the rational credence is *less than* t should lead you to *raise* your credence in p:

$$P(p|P(p) \geq t) \;<\; P(p) \;<\; P(p|P(p) < t)$$

In other words, you should think that possibilities where the rational credence in p is *higher* are *less* likely to be ones where p is true—you should think that the rational credence is *not* correlated with truth! But that's wrong. Contraposing: since you should think that rational credence *is* correlated with truth, learning that the rational credence in p is above a given threshold can never provide evidence against p. Precisely:

$$P(p|P(p) \geq t) \;\geq\; P(p)$$

This is the crux of our story, for it shows us how to refine Reflection.[18] Suppose you learn that the expert judges that p: $P(p) \geq t$. You should react to this information as you know the expert would—so how would the expert react? You've learned that they were originally at least t-confident of p. By the above reasoning, if *they* were to learn what you learned (namely, that the expert judges that p) this wouldn't provide them with any evidence against p—*their credence wouldn't drop*. Since you know that they were originally at least t-confident of p, you can infer that upon learning what you've learned the expert would react by *still* being at least t-confident of p. You should react to your information as you know the expert would. So *you* should be at least t-confident of p. Precisely:

> Simple Trust: $P(p|P(p) \geq t) \geq t$
>
> Upon learning that your evidence warrants judging that p, judge that p.
>
> Slogan: *take expert judgments on trust.*

Simple Trust is what we get when we restrict Reflection to apply to expert opinions of the form $[t, 1]$—opinions that are never evidence against p.

[17] Since $P(p) = 1 - P(\neg p)$, an $[l, h]$-opinion that p is equivalent to the conjunction of an l-judgment that p and a $(1-h)$-judgment that $\neg p$: $P(p) \in [l, h] \;\Leftrightarrow\; ([P(p) \geq l] \wedge [P(\neg p) \geq 1-h])$.

[18] The following line of reasoning is formalized in §2.5.

Summing up: For evidence to be a modest guide, rational credence must be correlated with truth. This means two things. First, it means that learning that the expert's credence in p falls *within a range* can be evidence *against* p. That is why Reflection fails. Second, it means that learning that the expert's credence in p falls *above a threshold* can *never* be evidence against p. That's why Simple Trust holds. The rest of the paper defends this solution.

Or rather: a solution *like* Simple Trust. We need one final piece of bookkeeping—one that applies equally to Reflection. No matter what bit of information q you learn, you should *still* treat your evidence (updated on q) as a guide. Thus our deference principles should apply not only to the unconditional attitudes warranted by the evidence, but also the conditional ones. Let $\boldsymbol{P_q(p)}$ be the credence (whatever it is) that your evidence warrants having in p *conditional on* q. Our deference principles should apply with 'P_q' substituted for 'P'. Of course, doing so yields our original principles as a special case (let $q = p \vee \neg p$); so since Reflection is already too strong, so too is its generalization.

But for principles that are *not* too strong—like Simple Trust—this generalization is exactly what we need. Thus we arrive at the promised principle:

> **Trust:** $P_q(p|P_q(p) \geq t) \geq t$
> Upon learning that your evidence warrants reacting to q by judging that p, react to q by judging that p.
> Slogan: *take expert judgments on trust.*

I claim that Trust is the key to making evidence a modest guide. But I won't ask you to take it on trust—the rest of this paper makes the case.

2.4.1 Trust me

Trust isn't just a solution—it's a natural, intuitive one. That may look doubtful. But I wouldn't waste your time (or mine) with a gerrymandered formal principle. Right off the bat, there are four things you need to know.

One: Trust is symmetric. Say that a principle **holds** at a world w iff all of its instances (well-defined instantiations of free variables) are true at w. Then:

Fact 2.4.1. *In any probability frame: Trust holds at a world iff* $P_q(p|P_q(p) \leq t) \leq t$ *does.*

Upon learning that the expert judges that p ($P(p) \geq t$), you should judge that p; and upon learning that the expert *doesn't* judge that p ($P(p) \leq t$), you *shouldn't* judge that p.

Uh oh. Does this imply Reflection? I say that (Trust:) upon learning that the expert is at least 0.7 confident, be at least 0.7; and upon learning that they're at most 0.8 confident, be at most 0.8. Does it follow that (Reflection:) upon learning

that the expert is at least 0.7 *and* at most 0.8 confident, you should be at least 0.7 and at most 0.8? No.

Two: Trust does not imply Reflection.

Fact 2.4.2. *There are probability frames that validate Trust in which Reflection fails at all worlds.*

This is possible because probabilistic support is non-monotonic: learning one thing can push your credence above 0.7, even if further learning would pull it lower. And it is actual because of the crux of our story: an expert judgment is never evidence against p, while an expert opinion sometimes is. Example: recall Immodest Imani, who is certain that Imani's the expert (I). If you learn that the expert judges that Imani's the expert to degree 0.7 ($P(I) \geq 0.7$), you should raise your credence to at least 0.7—for you know the expert would react to this information by doing so. But if you further learn that the expert both judges that Imani's the expert to degree 0.7 *and* does *not* judge that she is to degree 0.8 ($[P(I) \geq 0.7] \wedge [P(I) < 0.8]$), you should drop your credence to *0*—for you know that Imani *does* judge that she's the expert to degree 0.8. In short, Trust works because—and only because—it applies to judgments. It does not imply Reflection.

Three: Trust implies special cases of Reflection. Suppose you should be certain that the expert's opinion is in a given range: $P(P(p) \in [l,h]) = 1$. Then—since updating on these bounds doesn't provide any new information—Trust constrains your *un*conditional credences to be in that range.[19] More generally, when you should be sure that *if* the evidence warrants a given opinion, *then* it warrants certainty that it does so, you should obey Reflection. Letting $\boldsymbol{Sp}$ ($=_{df} P(p) = 1$) mean that your evidence warrants being *S*ure of p:

Fact 2.4.3. *In any probability frame, if Trust holds, and for a given $l, h \in [0,1]$ it's true that $P\big(S(P(p) \in [l,h]) \big| P(p) \in [l,h]\big) = 1$, then $P\big(p \big| P(p) \in [l,h]\big) \in [l,h]$.*

Four: paraphrased into natural language, Trust is truistic. That may seem doubtful, for Trust governs only judgments—and even if they are theoretically important, judgments may seem a contrivance.

They are not. Our natural-language talk of confidence trades in judgments. We talk about being $\left|\begin{smallmatrix} very \\ fairly \\ sorta \end{smallmatrix}\right|$ *confident* of *Top*, about it being $\left|\begin{smallmatrix} really \\ pretty \\ somewhat \end{smallmatrix}\right|$ *likely* to be true, about $\left|\begin{smallmatrix} leaving\ open \\ suspecting \\ thinking \end{smallmatrix}\right|$ that it's true, and so on. All of these terms share an important logical feature: they are preserved under increases in probability. If you are sorta confident of *Top* and then you become more confident, you are *still* sorta confident

[19] Precisely, $l \leq P(p|P(p) \geq l) = P(p) = P(p|P(p) \leq h) \leq h$. This is a synchronic version of a "conglomerability" constraint (Easwaran 2013). The generalization of Trust in §2.7 implies a more standard diachronic version: the current rational credence is bounded by the possible future rational credences.

of it (though you may now also be very confident); if it's somewhat likely to be true and then becomes more likely, it is *still* somewhat likely to be true (though it may now also be really likely); if you leave open that it's true and then you become more confident, you *still* leave open that it's true (though you may now also think that it is).[20] It follows that these terms cannot denote (proper) opinions, for such opinions are bounded above and below—if you have a middling opinion that *Top*, then increasing your confidence can lead you to lose it. Instead, these terms must denote *judgments*. Combining this observation with the fact that natural language expresses conditional probabilities as probabilities of (indicative) conditionals (cf. Stalnaker 1970; Adams 1975; Edgington 1995), we see that Trust is really just a set of truisms:

> You should think it $\left|\begin{smallmatrix}\text{really}\\ \text{pretty}\\ \text{somewhat}\end{smallmatrix}\right|$ likely that if the evidence makes it $\left|\begin{smallmatrix}\text{really}\\ \text{pretty}\\ \text{somewhat}\end{smallmatrix}\right|$ likely that p, then p.
>
> You should be $\left|\begin{smallmatrix}\text{very}\\ \text{fairly}\\ \text{sorta}\end{smallmatrix}\right|$ confident that if the evidence warrants being $\left|\begin{smallmatrix}\text{very}\\ \text{fairly}\\ \text{sorta}\end{smallmatrix}\right|$ confident that p, then p.
>
> You should $\left|\begin{smallmatrix}\text{think}\\ \text{suspect}\\ \text{leave open}\end{smallmatrix}\right|$ that if the evidence warrants $\left|\begin{smallmatrix}\text{thinking}\\ \text{suspecting}\\ \text{leaving open}\end{smallmatrix}\right|$ that p, then p.

And if Lockeans are right that belief reduces to sufficiently high credence (Foley 1992, 2009; Sturgeon 2008; Leitgeb 2013; Dorst 2019c):

> You should believe that if the evidence warrants believing that p, then p.

Upshot: Trust is an elegant, well-motivated weakening of Reflection.

2.4.2 Trust Trust

The progression of this paper suggests two hypotheses. The first:

Trust vindicates our Modest and Guiding Truisms.

I claim that I can establish this first hypothesis—that Bridging succeeds, after all. In fact, this success is so resounding that it suggests that Splitters and Mergers should jump ship. Precisely, say that you are *higher-order coherent* iff there is a body of evidence that would make your first- and higher-order opinions rational. The second hypothesis:

You are higher-order coherent iff you obey Trust.

The rest of this paper defends these two hypotheses. §2.5 shows that Trust both vindicates our Modest Truism and has many guiding features. However, it also has

[20] Saying "It's rather likely to rain" suggests that it's *not* very likely to. But this is a pragmatic phenomenon—compare: "Jane is rather tall" implicates (but does not entail) that she's not very tall.

commitments—most notably, the *positive access* principle that if you should be sure of p, you should be sure that you should be. In light of this, some may seek a weaker principle. §2.6 replies that we *need* a strong principle like Trust to avoid paradoxical results. §2.7 goes further: an independent characterization of the Guiding Truism leads *exactly* to Trust. §2.8 closes with applications.

It is the results in §2.5 and §2.7 that establish the first hypothesis. And it is the entire story—from the necessity of Bridging, to the failures of alternatives, to the successes of Trust—that composes the argument for the second.

2.5 Of Trust

In this section we'll see the virtues of Trust: it allows you to be extremely modest, while still requiring that you treat your evidence as a guide. Here's what you need to know. On the modesty side: Trust allows you to be certain that the expert is modest, as well as to be virtually indifferent over which of an arbitrary set of opinions is rational. On the guiding side: Trust requires you to react to new information as you know the expert would; it requires that you think the expert's judgment is correlated with truth; it rules out our puzzling case of Misguided Evidence; and it requires that the more higher-order doubts you have, the more moderate your first-order opinions must be. This is all exactly what we want. But Trust also has controversial commitments: it implies the *surely-factivity* principle that you should be sure that your evidence only warrants being certain of p if p is true; and it implies the *positive access* principle that if you should be sure of p, you should be sure *that* you should be sure of p. We might worry about these commitments—but §§2.6–7 argue that we *need* them. For those mainly interested in the big picture, the rest of §2.5 can be skipped. For those interested in the details, follow me into the weeds.

Modesty first. Trust isn't just formally weaker than Reflection—it's *substantively* weaker. First, Trust allows (but does not require!) you to be certain that your evidence warrants modesty. Letting *modest* be the proposition that you should be uncertain what the rational credence function is:

Fact 2.5.1. *There are probability frames that validate both Trust and* $[P(\textit{modest}) = 1]$.

Of course, Fact 2.5.1 doesn't tell us *how* modest Trust allows you to be. But it turns out that Trust allows you to be extremely modest—you can be virtually indifferent as to which of an arbitrary set of opinions is rational:

Fact 2.5.2. *For any* $T = \{t_1, ..., t_n\} \subset [0, 1]$*, there are probability frames that validate Trust with a candidate*[21] π *and proposition* p *such that* $\pi(P(p) \in T) \approx 1$ *and for all* t_i*:* $\pi(P(p) = t_i) \approx \frac{1}{n}$.

[21] Unlike 'P', '$\boldsymbol{\pi}$' is a rigid designator for a probability function whose values are known. π is a **candidate** in a probability frame iff at some world you should leave open that π is the rational credence function.

Upshot: our diagnosis of Reflection was correct. The *reason* it ruled out modesty was because it failed to distinguish facts about evidential support that can and cannot be evidence against p. Trust is built upon this distinction—and as a result, it vindicates our Modest Truism.

What about our Guiding Truism? There are two separable components to being properly guided. Start with a metaphor. If you're going on a hike, you need two things—a guidebook and a compass. The guidebook has no information about your immediate surroundings. It's useful because it has a wealth of *conditional* information—"If you're on the path facing east, the road is to your left", and so on. The guidebook is valuable as a *map*—once you enrich it with your local knowledge, it gives fine-grained directions; it is an optimal handler of new information. In contrast, the compass can't give you fine-grained directions. Its purpose is simply to orient you toward something you care about—in this case, north. The compass is valuable as a *tracker*; it is a reliable indicator of north. Having a map but no tracker is liable to lead you astray—if you're turned around, the map will send you in exactly the wrong direction. Having a tracker but no map is liable to leave you in the dark—if you don't know the area, you'll walk right past your campsite. To be properly guided, you need *both* a map and a tracker.

End of metaphor. For your *evidence* to be a guide, it must have the features of both a map and a tracker. More precisely, it must be both (1) an *optimal handler of new information* and (2) a *reliable indicator of the truth.*

(1) First, an optimal handler of new information—a map. When the expert knows everything you know, you should defer to their opinions. This doesn't require Reflection—for when you learn about the expert's opinion you may know something that they don't. (Looking at your surroundings, you may know more than your map does.) Rather, what it requires is that whenever you get new information, you should react to it *as you know the expert would.* Imagine the following scenario. Sitting on the panel, everyone hears an announcement that q. (q could tell you anything—e.g. that Bill is not the expert.) After q is announced, the remaining panelists—the ones who still might be the expert—announce their new opinions in *Top.* How, then, should *you* react to q? Slogan: *react as you know the expert would.* If all remaining panelists react with opinions about *Top* in a given range, then *you* should have that opinion in that range. Precisely[22]:

> **Reaction:** If $P_q\big(P_q(p) \in [l,h]\big) = 1$, then $P_q(p) \in [l,h]$
> If you should be sure that your evidence warrants reacting to your information q with a given opinion, react with that opinion.
> Slogan: *react as you know the expert would.*

[22]Reaction is inspired by Elga's (2013) "New Reflection" principle (cf. Hall 1994), but Reaction is stronger. Moreover—unlike New Reflection (cf. Pettigrew and Titelbaum 2014)—Reaction is preserved under conditioning.

Reaction encodes an important component of treating your evidence as a guide, for it guarantees that your opinions are constrained by your opinions about the expert's opinions. But—as will become clear in §2.6—just like a map without a compass, Reaction by itself would allow your evidence to send you in exactly the wrong direction.

(2) So evidence must also be a *tracker*—the expert's best guesses should be a reliable indicator of the truth. (The pointer on the compass should be a reliable indicator of north.) Imagine the following scenario. Everyone on your panel is asked whether or not they judge that *Top* (i.e. whether or not their credence is above a salient threshold). You are then going to find out what the expert did: you'll see a green light if the expert guessed that *Top*, a red light otherwise. You have some prior opinion about *Top*. Then you see a green light—how should your opinion change? You now know that the expert judged that *Top* is true. If expert judgments track the truth—if their confidence in *Top* is correlated with its truth value—this is evidence in favor of *Top*. So you shouldn't decrease your confidence in *Top*. Precisely, and generalizing to conditional judgments:

> **Reliance:** $P_q(p|P_q(p) \geq t) \geq P_q(p)$
> Upon learning that your evidence warrants reacting to your information q by judging that p, you shouldn't decrease your confidence in p.
> Slogan: *take experts to be reliable.*

We have two components to being a guide: being an optimal responder to new information (a map) and being a reliable indicator of the truth (a tracker). Trust vindicates both. In fact, Trust vindicates *exactly* both:

Fact 2.5.3. *A probability frame validates Trust iff it validates both Reaction and Reliance.*

This is the formalization of the intuitive argument from §2.4: if we require that you react to new information as you know the expert would and that expert credences are correlated with truth, then Trust is precisely what we get. Upshot: Trust yields the general guiding features—of a map and a tracker—that we're after.

It also yields specific ones. First, Trust rules out our paradigm cases of a mismatch between first- and higher-order opinions. Recall Misguided Evidence: putative cases in which it's rational to have confidence that *p but I shouldn't have confidence that p* $(\exists p, t: \ P(p \wedge [P(p) < t]) \geq t)$. Trust prevents this:

Fact 2.5.4. *If Trust holds at a world in a probability frame, Misguided Evidence does not.*

Finally, Trust forces a robust bridge between your first- and higher-order attitudes:

Fact 2.5.5. *In any probability frame: if Trust holds then* $[P(P(p) \geq t) \geq s] \rightarrow [P(p) \geq t \cdot s]$ *does too. No stronger connection holds: for any* $t, s \in [0, 1]$*: there are probability*

frames that validate Trust and make both $[P(P(p) \geq t) \geq s]$ *and* $[P(p) = t \cdot s]$ *true at a world.*

Fact 2.5.5 implies that if you're confident that you should be confident of something, you should be at least somewhat confident of it. Example: if you should be at least 0.9 confident that you should be at least 0.9 confident of *Top*, then you should be at least 0.81 confident of *Top*. Conversely, if you should have significant higher-order doubts, your first-order opinion must be moderate. Example: if you should be at least 0.4 confident that you should be at least 0.6 confident ($P(P(p) \geq 0.6) \geq 0.4$) but *also* at least 0.4 confident that you should be *at most* 0.4 confident ($P(P(p) \leq 0.4) \geq 0.4$), then your opinion in p must be between 0.24 and 0.76. More generally, if you are to have an opinionated credence in p, you can't have many higher-order doubts. For instance, if your credence in p should be 0.1 ($P(p) = 0.1$), then: the maximal credence you can have that you should be at least 0.2 confident ($\max[P(P(p) \geq 0.2)]$) is $\frac{1}{2}$; the maximal credence you can have that you should be at least 0.3 confident is $\frac{1}{3}$; ... and the maximal credence you can have that you should be at least 0.9 confident is $\frac{1}{9}$. In contrast: if you have a moderate credence in p, your higher-order doubts can range much wider. For instance, if your credence in p should be 0.5 ($P(p) = 0.5$), then: the maximal credence you can have that your credence in p should be at least 0.6 is $\frac{5}{6}$; the maximal credence you can have that your credence in p should be at least 0.7 is $\frac{5}{7}$; ...and the maximal credence you can have that you credence in p should be at least 0.9 is $\frac{5}{9}$. Using Fact 2.5.5, these relationships are graphed for various values of $P(p)$ in Figure 2-1 below.[23]

So we know a lot about Trust. How? The key to many of the results stated in this section is that we can give an *exact* characterization of Trust within a natural class of models (Theorem 2.5.7, Appendix 2.A.1).[24] The upshot is that Trust imposes three substantive—but tenable—constraints on the structure of evidence.

The first constraint is *surely-factivity*. You should be sure that: you should be sure of p only if p is true. Recalling that Sp is the proposition that your evidence warrants being *S*ure of p: $S(Sp \to p)$. In other words, you should never say to yourself, "Maybe I should be certain of p even though it's false" (i.e. $P(Sp \wedge \neg p) > 0$). (For if you did, then upon learning that your evidence warrants certainty of p, you'd still be less than certain of it—you wouldn't trust your evidence.) Surely-factivity makes full-blown factivity (i.e. $Sp \to p$) natural, though not inevitable.

The second constraint is *positive access*. You should be sure of p only if you should

[23]The graphs do not plot what your higher-order opinions *must* be, given your first-order ones—they represent how Trust *constrains* the relation between the two. Note: the maxima are not always jointly satisfiable: if $P(p) = 0.1$ you (obviously) cannot have *both* $P(P(p) \geq 0.1) = 1$ and $P(P(p) \leq 0) = 0.9$.

[24]Those models are *prior frames*—a subclass of probability frames in which uncertainty about what you should think flows from uncertainty about what you should be sure of (i.e. condition on).

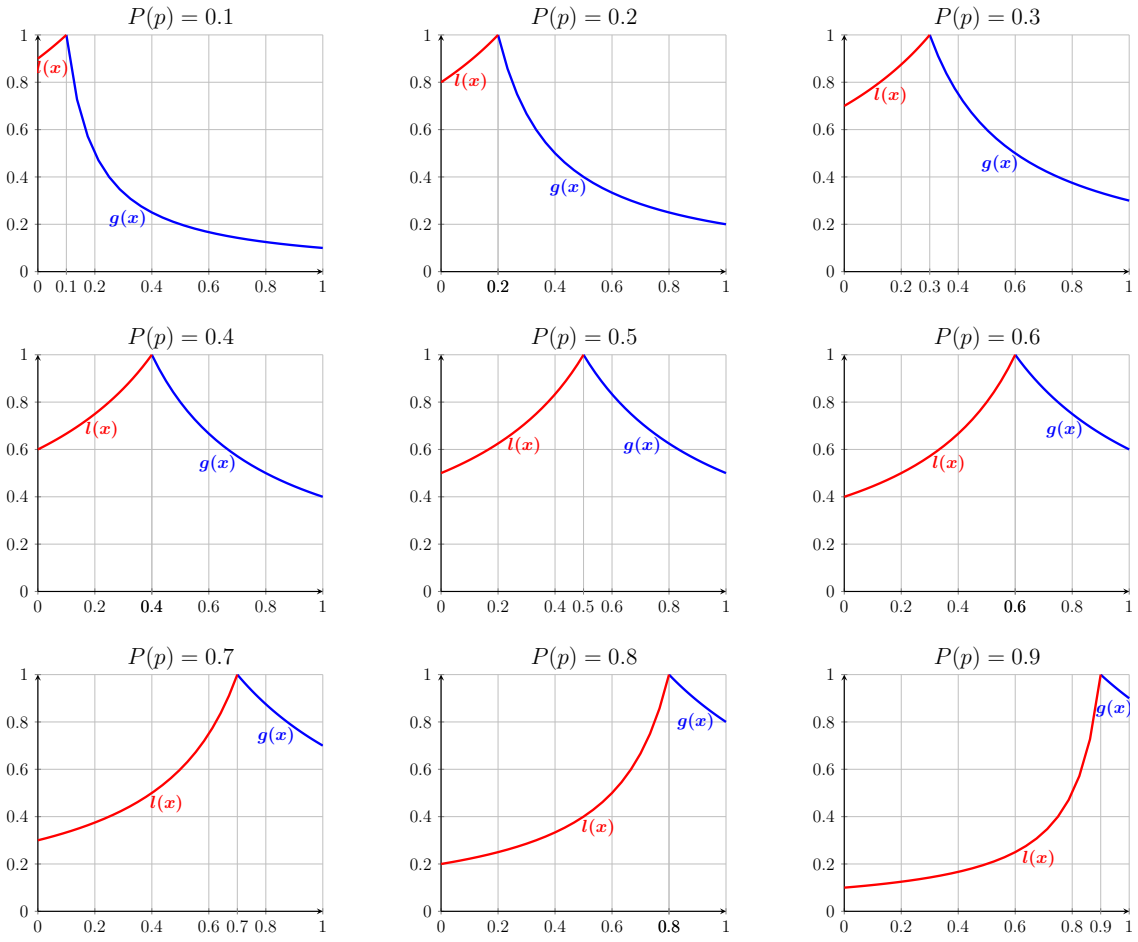


Figure 2-1: Possible higher-order doubts, given first-order credence.

be sure that you should be: $Sp \rightarrow SSp$.[25] In other words, you should never say to yourself, "p is true, but maybe I shouldn't be sure of it." (i.e. $Sp \wedge [P(\neg Sp) > 0]$). (For if you did, then upon learning that your evidence warrants being *un*sure of p, you'd still be sure of it—you wouldn't trust your evidence.)

Surely-factivity and positive access are the only non-probabilistic constraints that Trust imposes in full generality. But under the auxiliary assumption that your higher-order uncertainty comes from uncertainty about what you should be sure of, Trust is closely related to a third constraint. It is implied by (and *almost* implies) the constraint that we can define an indicative conditional operator '$\rightarrowtail$' that interacts sensibly with your conditional beliefs: in particular, (a) if you are certain of p condi-

[25]Trust does *not* require *negative access*, which says that if you shouldn't be sure of p, you should be sure that you shouldn't be: $\neg Sp \rightarrow S\neg Sp$. Unlike its positive counterpart, negative access rules out modesty (within prior frames) and is untenable for factive attitudes like knowledge (Stalnaker 2006).

tional on q, then you are certain of *if q then p*; and (b) for consistent q, if you are certain that *if q then p* then you are not certain that *if q then $\neg p$*. (See Appendix 2.A.1 for details.)

Here is where we are. Trust is the culmination of our search for a modest guide: it is a well-motivated refinement of Reflection, it allows plenty of modesty, it captures both the *map* and *tracker* aspects of being a guide, it rules out our puzzle cases, and it forces a sensible connection between your first- and higher-order attitudes. In short, we have reason to think that Trust is the Goldilocks principle.

But not conclusive reason. Trust imposes controversial constraints on the structure of evidence—most notably, positive access. Many will want to reject this principle—and will therefore be worried about Trust. Bridgers may try to endorse a weaker principle (cf. Elga 2013; Sliwa and Horowitz 2015; Christensen 2016), while Splitters may reject the search for such principles altogether (cf. Lasonen-Aarnio 2015; Williamson 2018). What of such alternative approaches—might they succeed? No.

2.6 In Judgment

This section argues that without a strong principle like Trust, paradox ensues. Though the argument is forceful, it is blunt—it cannot pinpoint Trust as the only solution. But a follow-up argument can: §2.7 proposes a characterization of our Guiding Truism in terms of the *value of evidence* (Good 1967), and then shows that it leads *exactly* to Trust.

Why won't a weaker principle do? Because there is another puzzle—one which, if allowed, would undermine the normative role of evidence. Consider:

> SYCOPHANTS
> Looking at the spoons, Sybil and Phan should be sure that one of them—the diligent one—has the credences warranted by their (shared) evidence. Conditional on it being either one of them, they agree on everything. But unconditionally they disagree: Sybil is 0.9 confident that Phan is the diligent one; Phan is 0.9 confident that Sybil is.

They are falling over themselves to give each other credit:

> Sybil: 'You're probably more diligent than me.'
> Phan: 'No, *you're* probably more diligent than *me*.'
> Sybil: 'No, YOU'RE probably more diligent than ME.'
> Phan: 'No!' [...]

Schematically, the scenario looks like this. There are two relevant possibilities: that Sybil is the diligent one (s) and that Phan is (p). What they should think depends on which of these possibilities they're in. If Sybil is the diligent one, they should be

0.9 confident that Phan is the diligent one and (so) 0.1 confident that Sybil is. If Phan is the diligent one, they should be 0.9 confident that Sybil is the diligent one and (so) 0.1 confident that Phan is. Letting an arrow labeled t from x to y mean that at possibility x they should assign credence t to being in possibility y, we have:

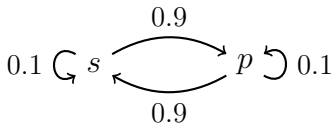


Figure 2-2: Sycophants.

We shouldn't allow Sycophants. For doing so would thereby allow you to be certain that your evidence is an anti-guide—that the expert has made an error in judgment. p is the proposition that Phan is the diligent one. If p is true, then their evidence warrants judging—as Phan does—that that p is false: $P(\neg p) \geq 0.9$. And if p is false, then their evidence warrants judging—as Sybil does—that p is true: $P(p) \geq 0.9$. Either way, their evidence warrants making the wrong judgment about p: at both possibilities the biconditionals $p \leftrightarrow [P(\neg p) \geq 0.9]$ and $\neg p \leftrightarrow [P(p) \geq 0.9]$ are true. Since Sybil and Phan should be sure that one of these possibilities is actual, they should be *sure* of these biconditionals. Thus they should be sure that *Phan is the diligent one iff we should judge that he's not*, and that *Phan is not the diligent one iff we should judge that he is*: $S(p \leftrightarrow [P(\neg p) \geq 0.9])$ and $S(\neg p \leftrightarrow [P(p) \geq 0.9])$. Generalizing, the puzzle is that you should be sure that the rational credence in p is anti-correlated with the truth:

> **Effacing Evidence:** $\exists p, t:\ S(p \leftrightarrow [P(p) < t])$ and $S(\neg p \leftrightarrow [P(p) > t])$.
> You should be sure that *p is true iff you should not have confidence in it*, and that *p is false iff you should have confidence in it.*

I claim that any theory that allows Effacing Evidence is too weak.

First, intuitively. Suppose we are trying to figure out what Phan thinks:

Phan: 'I've judged that Sybil is the diligent one.'

Us: 'What does *she* think?'

Phan: 'She's judged that I'm the diligent one. So I probably ought to judge that I am.'

Us: 'Well, why don't you?

Phan: 'I said I probably *ought* to judge that I'm the diligent one—not that doing so would probably be correct. It would most likely be incorrect.'

Us: 'But why do you think that it would most likely be incorrect? Is that what your evidence suggests?'

Phan: 'To the contrary. My evidence probably suggests that I'm the diligent one. That's why I think that Sybil is.'

Us: 'Wait—you're not listening to your evidence?'

Phan: 'I sure hope not! My evidence is wrong about who the diligent one is.'

Us: 'So why conform to your evidence?'

Phan: 'Exactly. Of course, I *ought* to—my evidence determines what I should think, after all. But it's a bad idea this time around. That's why I'm trying to avoid it.'

Us: 'No: why *ought* you conform to your evidence, if you know it'll lead you astray?'

Phan: '. . . I just ought to! My evidence tells me what I ought to think. Sometimes I know that what it tells me to think is false. But what am I to do? Life is hard.'

Something has gone wrong. A theory of rationality can't be made of just any arbitrary list of requirements—it must have the right structural features. In particular, the theory must be able to answer the question: 'Why conform to my evidence?' (Horowitz 2013; Schoenfield 2015b). No theory that allows Effacing Evidence can do so. For—as Phan's thoughts reveal—such theories sometimes allow you to know that your evidence is an anti-guide to what you actually value.

We can make this precise in two ways. Suppose you have Effacing Evidence with $t = \frac{1}{2}$: you should be sure that *p is true iff you should be less than $\frac{1}{2}$ confident of it*, and that *p is false iff you should be more than $\frac{1}{2}$ confident of it*. (The arguments generalize.)

First, you can be certain that choosing rationally will lead you to lose money. For I can offer you the following three options:

Bet 1: win \$1 if p, lose \$1 if $\neg p$

Bet 2: win \$1 if $\neg p$, lose \$1 if p

No Bet: \$0

It's rational (maximizes expected utility) to take Bet 1 iff you should be *more* than $\frac{1}{2}$ confident of p. It's rational to take Bet 2 iff you should be *less* than $\frac{1}{2}$ confident of p. But you should be certain that you should be more than $\frac{1}{2}$ confident of p (should take Bet 1) iff $\neg p$—i.e. iff you will lose that bet. And you should be certain that you should be less than $\frac{1}{2}$ confident of p (should take Bet 2) iff p—i.e. iff you will lose *that* bet. So you should be certain that acting rationally—taking the option you *ought* to take—will lead you to lose \$1 and be poorer than if you took No Bet! Upshot: theories that allow Effacing Evidence allow rational requirements to lead to a sure loss.[26]

Second, you can be certain that being rational will lead you to lose accuracy. For one of your doxastic options is to have credence $\frac{1}{2}$ in p, come what may. Yet you

[26]Many take it as a premise that rational requirements can do no such thing—most notably, Dutch Bookies (cf. Ramsey 2010; Skyrms 1966; Christensen 1991).

should be certain that it's rational to have credence *below* $\frac{1}{2}$ in p iff p is true, and that it's rational to have credence *above* $\frac{1}{2}$ in p iff p is false. That is, being rational will either lead you to have less confidence in a truth or more confidence in a falsehood than simply having credence $\frac{1}{2}$. So you should be certain that believing rationally—having the credence you *ought* to have—will lead you to be less accurate than having credence $\frac{1}{2}$. Upshot: theories that allow Effacing Evidence allow rational requirements to be accuracy-dominated by another available option.[27]

In short, theories that allow Effacing Evidence allow you to be certain that being rational—in and of itself—will prevent you from getting the things you actually value, like money and accuracy. They cannot vindicate our Guiding Truism.

But wait—is it really so surprising that sometimes you can know that being rational will make you worse off? Suppose I credibly threaten to drain your bank account iff your credences are rational. You then can be certain that being rational will prevent you from keeping the things you actually value (like money), right?

Right. But in this case there is a *cost* to being rational: doing so puts you in a bad situation where—no matter what else you do—you will be left with less money. Not so in my argument against Sycophants. In that argument there is was absolutely no cost to being rational: all the same options have all the same values, regardless of whether you are rational or not. (If you are rational, it is still true that *were you to take No Bet, you wouldn't lose any money.*) The only reason you can be certain that being rational will lead you to lose money is that you can be certain that being rational will lead you to choose the *worst* available option. The difference is crucial. All theories need to allow that sometimes you can know that being rational will lead you to be hurt. But no theory should allow that sometimes you can know that being rational will lead you to *hurt yourself.*

But wait—*couldn't* you have Effacing Evidence? Suppose God sets things up so that the rational opinions are inaccurate about p—and then she announces as much. Can you now be sure that being rational will lead you to have inaccurate opinions about p? No. Distinguish two versions of the case.

First version: God sets things up so that you are rational to be confident of p iff p is false: $[P(p) \geq 0.7] \leftrightarrow \neg p$. Then she announces this fact, so you update your beliefs on the claim that $[P(p) \geq 0.7] \leftrightarrow \neg p$. Do you now have Effacing Evidence? No. For what you've learned was that *before God told you anything*, the rational credence in p pointing in the wrong direction. This does not imply that *after* being told this, the updated rational credence is still pointing in the wrong direction. Formally, letting $q =_{df} ([P(p) \geq 0.7] \leftrightarrow \neg p)$, it is true that $P_q([P(p) \geq 0.7] \leftrightarrow \neg p) = 1$ but it does not follow (and is not plausible) that $P_q([P_q(p) \geq 0.7] \leftrightarrow \neg p) = 1$, as required for Effacing Evidence.

Second version: God offers the self-referential announcement, "The rational reac-

[27]Many take it as a premise that rational requirements can do no such thing—most notably, epistemic utility theorists (e.g. Joyce 1998, 2009; Pettigrew 2013, 2016).

tion *to this very announcement* is to be certain that: the rational credence in p is high iff p is false." In response to this announcement, do you have Effacing Evidence? No. For note that a parallel objection can be given against virtually any other rational norm—God could just as well announce: "The rational reaction to this very announcement is to be certain of $p \wedge \neg p$." Since this "counterexample" to the norm not to believe contradictions is not compelling, neither is the "counterexample" to a ban on Effacing Evidence.

I conclude, then, that we should reject theories that permit Effacing Evidence. What theories do so? Many. I cannot go into the details here, so a summary will have to do. First, the Sycophants model validates positive access ($Sp \rightarrow SSp$), negative access ($\neg Sp \rightarrow S\neg Sp$), and factivity ($Sp \rightarrow p$), so traditional versions of access internalism cannot rule it out (e.g. Smithies 2012). Second, since Sybil and Phan agree on everything upon learning which one of them is rational, the Sycophants model validates Reaction.[28] Finally, evidential versions of calibrationism (Sliwa and Horowitz 2015; Christensen 2016) can generate cases of Effacing Evidence, as can natural versions of Williamson's (2014) unmarked clock model. If I am right, no such theory can be correct. Instead, we need a principle strong enough to rule out Effacing Evidence. Trust fits the bill:

Fact 2.6.1. *If Trust holds at a world in a probability frame, Effacing Evidence does not.*

Upshot: despite its controversial consequences, we *need* a strong theory like Trust.

But how can we be sure that we've got that theory just right? What if Trust is too weak—permitting as-yet-unnamed puzzles? Or what if Trust is too strong—ruling out more than is required to solve them? I'm now going to offer a *proof*—of sorts—that it's not. No stronger theory is needed, and no weaker theory will do.

2.7 Of Value

What we need is *an* explication of our Guiding Truism that comes out true. I'm going to defend a proposal for the *minimal* such explication—the weakest principle that fully captures our Guiding Truism—and then show that it leads *exactly* to Trust. Thus Trust vindicates *an* explication—no stronger theory is needed. And Trust vindicates the *minimal* explication—no weaker theory will do.

The idea is this. Why should we care about gathering and conforming to evidence? What makes evidence something of value? I.J. Good (1967) gave a famous answer: evidence helps us make good choices. Suppose you face a decision problem—a set of options that may lead to different outcomes to which you assign different values. A body of evidence is *valuable* iff—supposing the evidence is free—the expected

[28]Hence it also validates Elga's (2013) New Reflection principle, mentioned in footnote 22.

value of taking the option warranted by the evidence (whatever it is) is higher than the expected value of any other particular option. (Loosely: iff you should prefer to *use* free evidence to make your decision, rather than ignore it.) Example: you have to decide whether to take a bet on *Top*. You could either (1) scrutinize the spoons more closely (obtain more evidence) or (2) simply decide now whether to take the bet (ignore that evidence). No doubt the expected value of (1) is higher than that of (2)—the evidence is valuable.

Good generalized that idea. He argued that—no matter what options and values you have—evidence is *always* valuable. In formulating this idea, many have have treated it diachronically: you should prefer to gather and use a *more* informed body of evidence to make your decision (cf. Skyrms 1990; Oddie 1997; Myrvold 2012; Huttegger 2014; Ahmed and Salow 2018). But once we allow modesty—uncertainty about your *current* evidence—there are in fact two instances of Good's idea. In making your decision, you should prefer: (1) to use your current evidence (whatever it is), rather than simply choose an option; and (2) to use a more informed body of evidence (whatever it is), rather than simply choose an option.

We can capture both instances under one schema. Say that a body of evidence is *at least as informed as yours* (for short: **informed**) iff it contains all your evidence, and maybe more. So (trivially) your evidence is at least as informed as your evidence, and (nontrivially) the evidence you'd have after scrutinizing the spoons is at least as informed as your evidence. Generalizing Good's idea (formalized in Appendix 2.A.2):

> **Value**
> You should always expect the option warranted by informed evidence to be at least as good as you should expect any other particular option to be.
> Slogan: *evidence is valuable.*

Example: I offer you a bet on *Top*—you must decide whether to take it or leave it. If you were 0.8 confident of *Top*, you'd take it; if you were 0.6 confident of *Top*, you'd leave it. Being 0.7, you're on the fence—the expected value of taking it and of leaving it are balanced. What about the expected value of doing what you *should* do (what the expert would do)—whatever that is? If you should be 0.8 (as perhaps you should), what you should do is take it; if you should be 0.6 (as perhaps you should), what you should do is leave it. Being unsure what you should think, you're unsure what you should do. But why care about doing what you *should* do? Why treat your evidence as a guide? Value says: because you should expect it to help you make good decisions.

Value applies to any informed body of evidence—not just you own. Following suit, we can generalize our principles; for example, our Guiding Truism:

> Informed Guiding Truism
> Your evidence always warrants being disposed to treat *informed* evidence as a guide.

Similarly for our deference principles. Let i and k be two (perhaps identical) bodies of evidence—subject to the constraint that k is at least as informed as i ($k \geq i$). Let P^i be the credences that are rational given evidence i (whatever they are), and P^k be those that are rational given evidence k. Then Trust (and its ilk) should be generalized:

> **Informed Trust:** $P^i_q(p|P^k_q(p) \geq t) \geq t$ $\qquad (k \geq i)$
> Upon learning that *informed* evidence warrants reacting to q by judging that p, react to q by judging that p.
> Slogan: *take informed expert judgments on trust.*

Having generalized our search for a modest guide, I'll now argue that Value is the *minimal explication* of the (Informed) Guiding Truism—it fully vindicates that truism; and no weaker principle would.

As with the argument against Effacing Evidence in §2.6, it is crucial to note that Value applies to being rational *in and of itself*. Three clarifications. (1) Value applies to the expected value of *freely* taking the option warranted by an informed body of evidence; thus it screens off any costs—monetary, psychological, computational, and so on—that using that evidence may have. (If you will be punished for using your evidence, doing so might not maximize expected value.) (2) Value applies to the expected value of *successfully* taking the option warranted by an informed body evidence; thus it screens off any risks of *mis*using the evidence that obtaining it may bring. (If trying to use the evidence will likely lead to mistakes, doing so might not maximize expected value.) (3) Value applies to the expected value of *simply* taking the option warranted by an informed body of evidence; thus it screens off any effects on you—such as changes in values—that obtaining that evidence may have. (If using the evidence will change your outlook on life (Paul 2014), doing so might not maximize expected value.) Although such costs, risks, and effects are ever-present in real life, Value explains why—when they become sufficiently small—you should always prefer to use the evidence.

Suitably clarified, Value is plausible. It is also applicable—it explains what's puzzling about our puzzles:

Fact 2.7.1. *In any dynamic probability frame: if Misguided Evidence or Effacing Evidence are true at a world, Value fails.*

This is because Value rules out the possibility that evidence could be an "anti-guide," and thus subsumes the argument against Effacing Evidence in §2.6. (A similar argument can be given against Misguided Evidence.)

Finally, here is a general argument that Value is the minimal explication of the (Informed) Guiding Truism—the claim that you should always be disposed to treat informed evidence as a guide. There is no doubt that it is *an* explication. For suppose that Value holds. Then no matter what decision problem you're facing—whether it's

what to do, or what to think—then if you *could* do what's warranted by an informed body of evidence, then you *should*. (Doing so would maximize expected value). Thus there's a perfectly good sense in which you should always treat informed evidence as a guide—our Guiding Truism comes out true. Conversely, no principle *weaker* than Value would vindicate our Guiding Truism. For suppose that Value fails: there is a decision problem in which you expect that doing what's warranted by an informed body of evidence will lead to a worse outcome than ignoring that evidence. So if you were given a choice between the two, you should prefer to ignore the evidence—for you should expect the expert (the person following the evidence) to make a worse choice than you! Thus there's *no* perfectly good sense in which you should *always* treat informed evidence as a guide—our Guiding Truism comes out false.

Suppose that this is right: we vindicate our Guiding Truism iff we vindicate Value. What does it take to vindicate Value? Famously, Good (1967) proved it under certain assumptions. But it turns out that those assumptions were inconsistent with our Modest Truism.[29] What happens when we loosen them?

Recall our story. In searching for a modest guide, we discovered that we must: permit modesty; forbid Misguided Evidence; permit Reflection failures; require Reaction; require Reliance; require surely-factivity; require positive access; forbid Effacing Evidence; require Trust.

Value runs the gamut. It permits modesty. It forbids Misguided Evidence. It permits Reflection failures. It requires Reaction. It requires Reliance. It requires surely-factivity. It requires positive access. It forbids Effacing Evidence. It requires Trust:

Theorem 2.7.2. *In any dynamic probability frame: if Value holds at a world, Informed Trust does as well.*

In fact—in (at least) a wide class of scenarios—it is *equivalent to* Trust:

Theorem 2.7.4 (Rough). *In (at least) a wide class of frames: Value $\Leftrightarrow$ Informed Trust.*

That wide class is the class of *prior* frames discussed in §2.5 and Appendix 2.A: models in which your higher-order uncertainty stems from uncertainty about what you should be sure of. Moreover, the restriction is on the proof—not the truth. There is reason to think that (Informed) Trust is equivalent to Value in full generality. I conjecture that it is (Conjecture 2.7.3, Appendix 2.A.2). But even if that conjecture fails, we now know that some *strengthening* of Trust will succeed. Upshot: we *know* that Trust is necessary—and in at least a wide class of cases, sufficient—to vindicate our Guiding Truism. And we have reason to think that Trust *characterizes* our Guiding Truism: that no stronger theory is needed, and no weaker theory will do.

[29] Precisely, he assumed that we could model you using a prior frame in which E is an equivalence relation (see Appendix 2.A.1). Any such frame validates $[P(p) = t] \leftrightarrow S[P(p) = t]$.

A final point: Theorem 2.7.4 is a "coincidence result." The progression of this project was not so prescient as the progression of this paper: I began with various puzzles of higher-order uncertainty, was led (through trial and error) to Trust, and characterized it over a class of prior frames. Only later did I discover that Geanakoplos (1989) had proven that (a principle like) Value was validated by *the exact same class of prior frames*. Imagine my surprise—shock even—upon seeing his theorem. And my satisfaction upon discovering it could be strengthened, yielding our coincidence. Personally, I think the fact that this convergence was serendipitous increases the plausibility of our destination. Perhaps you will agree.

This is the end of our story. Wanting to take disagreement seriously, we began searching for a modest guide. The obvious proposal—Reflection (*defer to expert opinions*)—was too strong, for it failed to acknowledge that learning an expert *opinion* can be evidence against p. The natural response—Trust (*take expert judgments on trust*)—had all the marks of the Goldilocks principle. And the measure of success—Value (*evidence is valuable*)—gave Trust a resounding confirmation. Our search for a modest guide has succeeded: those who are uncertain can take it on trust.

2.8 In Consequence

I'll close by sketching two applications of the theory developed here. Both deserve more discussion; my goal is just to sketch how Trust is rich in consequences for other debates.

2.8.1 To Disagreement

We began with a disagreement: you were confident of *Top*, while Disagreeing Disa was not. Intuitively, learning that she was not should lead you to lower your confidence in *Top*. And we saw that if this is to be so, then our Modest and Guiding Truisms must both be true. Having found a theory that reconciles them, we can vindicate the intuitive verdict.

You and Disa are looking at the spoons. For simplicity, suppose you should be sure that the rational credence in top is either *h*igh or *l*ow:

(1) $S([P(\mathit{Top}) = h] \vee [P(\mathit{Top}) = l])$ $\qquad (h > l)$

Suppose that the rational credence is *h*igh:

(2) $P(\mathit{Top}) = h$

Since Disa is smart, learning that she's t-confident of *Top* ($C_D(\mathit{Top}) = t$) provides you with some reason to think that your evidence warrants having credence t:

(3) $P(P(\mathit{Top}) = t | C_D(\mathit{Top}) = t) > P(P(\mathit{Top}) = t)$

Finally, since Disa's only route to the truth is through her evidence, once you learn what your shared evidence warrants, further learning Disa's actual opinion does not affect the probability of *Top*:

(4) $P(\mathit{Top}|P(\mathit{Top}) = t) = P(\mathit{Top}|[P(\mathit{Top}) = t] \wedge [C_D(\mathit{Top}) = s])$

This is a paradigm peer-disagreement scenario. Alone, (1)–(4) do not require you to lower your confidence when you learn that Disa disagrees; but given Trust, they do:

Fact 2.8.1. (1)–(4) *are consistent with* $P(\mathit{Top}|C_D(\mathit{Top}) < h) \geq P(\mathit{Top})$. *But given Trust,* (1)–(4) *imply* $P(\mathit{Top}|C_D(\mathit{Top}) < h) < P(\mathit{Top})$.

Clearly this is only the beginnings of a theory of how to respond to disagreement. But it is a possibility proof: it shows that there *is* a principled way to take disagreement seriously. The next step is to find general rules delineating in what situations and to what extent Trust requires doing so. Trust gives answers—we just have to find them.

Upshot: Trust lays the foundations for a theory of disagreement.

2.8.2 To KK

In the wake of Williamson (2000) it has become popular to combine two views:

> Knowledge First
> Knowledge plays the fundamental role in epistemology.
>
> No KK
> You can be in a position to know p without being in a position to know that you are ($Kp \wedge \neg KKp$).

Our theory puts pressure on this conjunction. For it is difficult to maintain Knowledge First without endorsing a close connection between knowledge and rational degrees of belief—after all, it is the latter that encode what you should think and determine what you should do. The natural way to forge this connection is through rational certainty: you should be sure of p iff you're in a position to know p ($Sp \leftrightarrow Kp$). But rationality must be something that we ought to treat as a guide. This, in turn, requires Trust and Value—both of which imply positive access:

Fact 2.8.2. *In any probability frame: if Value or Trust hold at a world, then* $Sp \rightarrow SSp$ *does as well.*

If knowledge is to be first, positive access in turn leads to KK: you are in a position to know p only if you're in a position to know that you are ($Kp \rightarrow KKp$).

Upshot: Trust forces a choice between Knowledge First and No KK.[30]

[30]Knowledge First relies on a systematic and wide-ranging epistemological program. No KK rests, primarily, on subtle margin-for-error-like arguments (Williamson 2000; Bacon 2013) that can be resisted on a variety of grounds (Sharon and Spectre 2008; Greco 2014b; Stalnaker 2015; Das and Salow 2016). Given that we face a choice, I say that No KK should be the one to go.

2.9 A Modest Goal

I have a modest goal. Literally. My goal is an epistemology informed by modesty. Because you and I are modest: we are constantly wondering whether we are thinking and doing as we should. And we are not irrational for that. Given the sort of critters we are—and the sort of mistakes we make—we *should* have a healthy dose of self-doubt. My goal is an epistemology that makes room for critters like us—a modest epistemology.

In this paper, I've tried to secure its foundations. I argued that in order to take disagreement seriously, evidence must be a modest guide (§2.2); that the intuitive guiding principle—Reflection—rules out modesty entirely (§2.3); that its natural refinement—Trust—allows modesty while guaranteeing a correlation between evidence and truth (§§2.4–5); that denying Trust risks allowing you to know that your evidence is *anti*-correlated with truth (§2.6); that Trust characterizes the platitude that you should prefer to use free evidence (§2.7); and that the resulting theory *does* require you to take disagreement seriously (§2.8). In short, Trust demonstrates that there is a coherent, rational way to have a healthy dose of self-doubt. There are two directions to take it from here.

The first is toward extensions. The theoretical arguments of this paper pinpoint a particular structure of evidence. Structure in hand, we face both formal and philosophical questions. What is the *general* relation between Value, Trust, and that structure? And what sort of thing must evidence *be*, if it is to have that structure?

The second is toward applications. We now know how to model higher-order probabilities in a principled, coherent way. What does this tell us about...

...how to analyze disagreement and debunking (Christensen 2010a)?

...how to model "ambiguous" evidence (Joyce 2010)?

...how to rely on people who don't know their own opinions (Roush 2016)?

...how to defer to chances that don't know their own values (Briggs 2009b)?

Foundations secured. Next step: developing modest answers to our new questions.[31]

2.A Formal Details

In Appendix 2.A I explain the formal details underlying the results stated in this paper. This section explains *probability frames* and how they encode higher-order information. §2.A.1 explains the special case of *prior frames*, how they can be used to characterize Trust, and how to interpret the tree-like structures that result. §2.A.2

[31]I can't thank everyone who has helped me with this long project. I received valuable feedback from Laura Callahan, Kenny Easwaran, Branden Fitelson, Brooke Husic, Harvey Lederman, Hanti Lin, Eric Pacuit, Miriam Schoenfield, Kieran Setiya, Jack Spencer, Steve Yablo, and many others. Special thanks to Bernhard Salow, Ginger Schultheis, Bob Stalnaker, and Roger White for sticking with me. And to Gillian Russell—for getting me started.

explains how to generalize such frames to capture multiple bodies of evidence, and uses this generalization to characterize Value and Informed Trust.

A **probability frame** $\langle W, \mathcal{P}\rangle$ consists of a finite set of worlds W and a function $\mathcal{P}$ from worlds w to probability functions $\mathcal{P}_w$ defined over the subsets of W (cf. Gaifman 1988; Samet 1997; Dorst 2019b). $\mathcal{P}_w$ represents the credences that you ought to have at w. Since rational uncertainty is modeled as uncertainty about which world you're in, and the rational credence function varies across worlds, you can have rational uncertainty about what the rational credence function is.

A proposition p is (any) set of worlds: $p \subseteq W$. p is true at w iff $w \in p$. Logical operations on propositions are captured with set-theoretic ones: $\neg p = W - p$; $p \wedge q = p \cap q$; etc. Probabilistic facts are captured in the obvious ways by using the probability functions associated with each world. Thus the proposition that *the rational credence in p is t* is the set of worlds w such that $\mathcal{P}_w(p) = t$: $[\boldsymbol{P(p) = t}] =_{df} \{w|\mathcal{P}_w(p) = t\}$. Similarly for other probabilistic propositions. Since these higher-order claims are simply sets of worlds, they get assigned probabilities like any other proposition.

Example. Let $W = \{a, b\}$ with $\mathcal{P}_a(a) = 0.6$ (so $\mathcal{P}_a(b) = 0.4$) and $\mathcal{P}_b(a) = 0.3$ (so $\mathcal{P}_b(b) = 0.7$). We can diagram this with an arrow labeled t from x to y indicating that $\mathcal{P}_x(y) = t$, as in Figure 2-3.

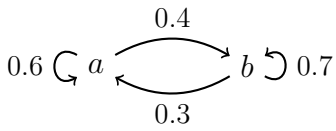


Figure 2-3: A probability frame.

In this frame, $[P(a) = 0.6] = \{a\}$ (since $\mathcal{P}_a(a) = 0.6$ and $\mathcal{P}_b(a) = 0.3$), so in turn $[P(P(a) = .6) = 0.6]$ is the set of worlds w such that $\mathcal{P}_w(P(a) = .6) = 0.6$, i.e. the set of w such that $\mathcal{P}_w(\{a\}) = 0.6$, i.e. $\{a\}$ itself: $[P(P(a) = .6) = 0.6] = \{a\}$. Meanwhile, $[P(a) = 0.3] = \{b\}$ (at b, you should be 0.3 confident you're at a); $[P(P(a) = .6) = 0.3] = \{b\}$ (at b, you should be 0.3 confident that you should be 0.6 confident that you're at a); $[P(P(P(a) = .6) = .6) = 0.3] = \{b\}$ (at b, you should be 0.3 confident that you should be 0.6 confident that you should be 0.6 confident that you're at a); and so on. As can be seen, in such frames there can be rational uncertainty all the way up.

2.A.1 Trust the Details

To characterize Trust we will make use of a natural, tractable subclass of probability frames. Given a probability frame $\langle W, \mathcal{P}\rangle$, say that the evidence at w *leaves open* w'—and write $\boldsymbol{wEw'}$—iff $\mathcal{P}_w(w') > 0$. The set of worlds that the evidence at w leaves open is $\boldsymbol{E_w} =_{df} \{w'|wEw'\} = \{w'|\mathcal{P}_w(w') > 0\}$. A *prior frame* is a probability frame in which there is a unique prior π such that the probability function $\mathcal{P}_w$ at each world

w can be recovered by conditioning π on E_w. Precisely: a **prior frame** $\langle W, E, \pi\rangle$ is a probability frame $\langle W, \mathcal{P}\rangle$ in which there is a regular probability distribution π over W (i.e. $\forall w \in W : \ \pi(w) > 0$)—the prior—such that for all w: $\mathcal{P}_w = \pi(\cdot|E_w)$ (cf. Williamson 2000, 2014, 2018; Cresto 2012; Lasonen-Aarnio 2015; Salow 2017).

Formally, since the variation in probabilities across worlds is all traceable to the binary relation E, prior frames are based on a standard Kripke **frame** $\langle W, E\rangle$. This allows us to use the ordinary tools of modal logic to render them tractable.[32] Philosophically, prior frames are natural given several different background pictures. These pictures all agree on two things: (1) rational opinion is separable into two components—(i) which standards of reasoning you should be using (π), and (ii) what you should take for granted (E_w); and (2) rational agents should know which standards of reasoning they should be using. Different approaches will fill out the components in different ways, but all will agree that what you should do is take that prior π and condition it on your evidence E_w. And all can allow modesty—for even if you know what prior you should use, you can be uncertain what you should condition on. (Moreover, even if we reject such pictures, prior frames still capture an important case of modesty—so are a useful tool.)

We can now characterize Trust. First, some definitions. When xEy, I'll say "x sees y." A frame $\langle W, E\rangle$ is **transitive** iff: if x sees y and y sees z, then x sees z: $(xEy \wedge yEz) \Rightarrow xEz$. A frame is **shift-reflexive** iff any world seen by anything sees itself: $xEy \Rightarrow yEy$. A frame is **shift-nested** iff whenever two worlds are seen, either one sees everything the other does or they see nothing in common: $wEx, y \Rightarrow$ ($E_x \subseteq E_y$ or $E_x \supseteq E_y$ or $E_x \cap E_y = \emptyset$). These conditions characterize Trust:

Theorem 2.5.7 (Trust Characterization)**.** *A prior frame $\langle W, E, \pi\rangle$ validates Trust iff $\langle W, E\rangle$ is transitive, shift-reflexive, and shift-nested.*

Such frames are composed of structures like Figure 2-4. (Circles drawn around worlds that see exactly the same worlds; transitive arrows omitted.)

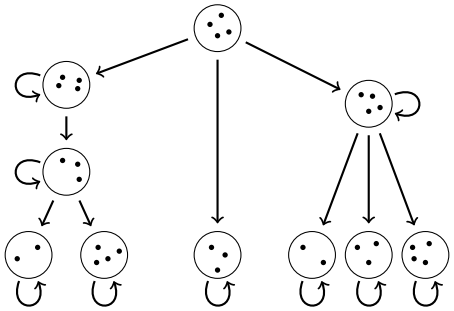


Figure 2-4: A transitive, shift-reflexive, and shift-nested frame.

[32]Note: in any prior frame $\langle W, E, \pi\rangle$, E must be serial: $\forall x \exists y : \ xEy$.

What we have is a *tree*.[33] The arrows can proceed down chains and splits, but they can never split and reconnect, as in Figure 2-5 below.

How do these constraints map onto the structure of evidence discussed in §2.5? Shift-reflexivity corresponds to surely-factivity: you should be sure that if you should be sure of p, then p is true; $S(Sp \to p)$. Transitivity corresponds to positive access: you should be sure of p only if you should be sure that you should be; $Sp \to SSp$.

Although shift-nesting does not correspond to an axiom that can be stated with the modal operator S, we can use other tools to get a grip on it. To see this, consider one interpretation of (a slight refinement of) our tree-structures. A frame is **reflexive** if every world sees itself: xEx. It is **nested** if for every pair of worlds, either one sees everything the other does or they see nothing in common: ($E_x \subseteq E_y$ or $E_x \supseteq E_y$ or $E_x \cap E_y = \emptyset$). We'll consider the set of transitive, reflexive, nested frames.

Think of the evidence as warranting not merely a set of conclusions, but also a *line of reasoning* to that set of conclusions. Lines of reasoning—like proofs—are path-dependent: "p; q; therefore, $p \wedge q$" is a different line of reasoning than "q; p; therefore, $p \wedge q$." On this picture, you should leave open that a state might be the rational one only if it is an *extension* of the line of reasoning you should take. Since no two lines of reasoning can diverge and still wind up in the same place, this is why we get the "branching" structure. Precisely: say that $\langle W, E\rangle$ is **reasoning-generable** iff there is a function f from worlds w to sequences of propositions $f(w)$ such that: wEx iff $f(w)$ is an initial segment of $f(x)$. (Think of $f(w)$ as the sequence of conclusions that represent the line of reasoning you should take, given your evidence.) Then:

Theorem 2.5.8. *$\langle W, E\rangle$ is reasoning-generable iff it is reflexive, transitive, and nested.*

This gives some sense to our tree structures; but *why* would your evidence warrant concluding p earlier in your line of reasoning than concluding q? Notice that if you have rationally concluded p but not q ($P(p) = 1$ but $P(q) < 1$), then conditional on only one of p and q being true, you should be certain it's p: $P(p|\neg(p \wedge q)) = 1$; you should think to yourself, "Supposing that at most one p or q is true, I'm sure it's p." In other words, your epistemic access to p is *more robust* than your epistemic access to q. *That* may be why your evidence warrants concluding p earlier in your line of reasoning than concluding q.

We can reinforce this point with some independent resources. On most theories of conditionals, being conditionally certain of p given $\neg(p \wedge q)$ (i.e. $P(p|\neg(p \wedge q)) = 1$) implies that you should be certain of the indicative conditional, *If at most one of p or q is true, then it's p.* Notice that our certainties in such conditionals encode facts about how *epistemically robust* their components are. I'm certain that Oswald shot Kennedy. But I'm also certain that (1) *if Oswald didn't shoot Kennedy, then someone*

[33] Formally: if we take equivalence classes under E and force it irreflexive, the resulting structure of every E_w becomes a *forest* in graph-theoretic parlance.

else did. My certainty in (1) is no trivial consequence of the fact that I'm certain that the antecedent is false—for notice that I do *not* believe that *if Oswald didn't shoot Kennedy, then no one else did*. Rather, my certainty in (1) encodes the fact that my certainty that someone shot Kennedy is more robust than my certainty that Oswald did—if you were to remove my certainty of the latter, I'd still be sure of the former. Thinking of indicative conditionals as capturing epistemic dependence in this way makes two (widely-endorsed) theses extremely natural. Let '$q \rightarrowtail p$' represent the indicative conditional *If q, then p*. Then:

a) If $P(p|q) = 1$, then $P(q \rightarrowtail p) = 1$.
   If conditional on q you should be sure of p, then you should be sure of *if q, p*.
b) If $P(q \rightarrowtail p) = 1$, then $P(q \rightarrowtail \neg p) < 1$ (for $q \neq \emptyset$).
   If you should be sure of *if q, p*, then you should not be sure of *if q,* $\neg p$.

(a) is an instance of the widely-attested Ramseyan thesis that the probability of an indicative conditional equals the corresponding conditional probability.[34] (b) is a consequence of the principle of "conditional non-contradiction" (not: *if q, then p* and *if q, then* $\neg p$), as well as the belief-revision principle that if you're certain of *if q, p*, then upon learning that q you should be certain of p (cf. Stalnaker 1984).

Say that a frame $\langle W, E\rangle$ is **conditionable** iff we can define a two-place propositional connective $\rightarrowtail$ satisfying (a) and (b). Then in order to have a sensible connection between conditional-beliefs and beliefs in conditionals, we need a tree-structure:

Fact 2.5.9. *A reflexive, transitive* $\langle W, E, \pi\rangle$ *is conditionable iff it is nested.*[35]

To see why this is so, it will help to illustrate how Fact 2.5.9 furnishes a response to an objection to nesting (Das 2017; Williamson 2018). One type of higher-order uncertainty arises with a good/bad-case asymmetry: looking at a real painting, you can tell it's not a fake; but looking at a fake, you should be unsure whether it's the real thing (Lasonen-Aarnio 2015). Thus in the *g*ood case you can tell you're in the good case, but in the *b*ad case you should be unsure which case you're in; omitting reflexive arrows, we have a structure like this: $b \rightarrow g$. Now suppose you are looking at *two* independent paintings. Shouldn't we expect a non-nested structure such as in Figure 2-5? (*gb* is where the first painting is real and the second is fake, etc; reflexive, transitive arrows omitted.)

[34] See Ramsey 1931; Stalnaker 1970; Adams 1975; van Fraassen 1976; Edgington 1995; Bennett 2003; Khoo 2013, 2016; Rothschild 2013; Bacon 2015. Note that there is no risk of triviality from (a).

[35] It follows from Fact 2.5.9 that no triviality results (à la Lewis 1976; Bradley 2000; Russell and Hawthorne 2016) can be proven from conditionability. Why? Because triviality results happen when you impose a connection between credences in conditionals and conditional credences for *all* probability functions (e.g. Bradley 2000). There is no threat if we only impose the connection for a *particular* probability function that is coordinated with the conditional (cf. Mandelkern and Khoo 2018)—as conditionability does.

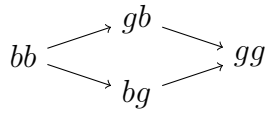
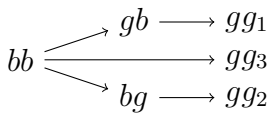


Figure 2-5: A non-nested frame.

I object to this model, for it implies inconsistent comparisons of epistemic robustness. Notice that at gb, conditional on one of the paintings being fake, you should be certain that the first one is real: $\mathcal{P}_{gb}(gb|\neg gg) = 1$. Thus, applying (a), at gb you should be certain of the conditional, *if only one of them is real, it's the first one*: $\mathcal{P}_{gb}(\neg gg \rightarrowtail gb) = 1$. Since you are certain of this conditional and you leave open that you are at world gg, the conditional must be *true* at gg. And since at gg you should be certain you're at gg, it follows that at gg you should be certain of the conditional: $\mathcal{P}_{gg}(\neg gg \rightarrowtail gb) = 1$. Of course, at gg you should be certain that the antecedent of this conditional is false. But—as we saw with the Oswald example above—that does not trivialize your attitude; rather, your certainty in the conditional means that at gg your certainty that the first painting is real should be *more robust* than your certainty that the second one is. Here lies the problem: the cases are symmetric! Parallel reasoning starting from bg would lead to the opposite conclusion that in gg you should be certain that *if only one of them it real, it's the second one*; thus your certainty that the first one is real is *less* robust than your certainty that the second one is. Contradiction.

Fact 2.5.9 implies that to make this case consistent with principles (a) and (b), we must divide the gg-possibilities into ones that make true different facts about epistemic robustness. For instance, we may have one possibility (gg_1) where your evidential access to the first painting is more robust than your evidential access to the second, one (gg_2) where vice versa, and one (gg_3) where neither is more robust than the other:



I conclude that there is plenty of motivation—from Trust, conditionals, and (as we will see) the value of evidence—to defend nesting against the model given in Figure 2-5.

2.A.2 Value the Details

To formalize and then characterize Value, we must generalize our frames to allow for multiple bodies of evidence. A **dynamic probability frame** $\langle W, \mathcal{P}^1, \mathcal{P}^2 \rangle$ is a generalization of a probability frame. It associates a world w with *two* probability functions $\mathcal{P}^1_w$ and $\mathcal{P}^2_w$. The intended interpretation is that $\mathcal{P}^1_w$ is the rational credence function

given body 1 of evidence, and $\mathcal{P}^2_w$ is the rational credence function given body 2 of evidence—with 2 at least as informed as 1. (Thus we expect 1 to defer to itself and 2, but 2 to defer only to itself.) Propositions about probabilities are now indexed to their respective body of evidence; for $i = 1, 2$: $[\boldsymbol{P^i(p) = t}] =_{df} \{w|\mathcal{P}^i_w(p) = t\}$. *Convention*: when using multiple variables to range over probability statements, I will use ***i* and *k*** with the constraint that $k \geq i$. Thus the statement $[P^i(P^k(p) = 0.7) \geq 0.4]$ expresses three claims: $[P^1(P^1(p) = 0.7) \geq 0.4]$, $[P^1(P^2(p) = 0.7) \geq 0.4]$, and $[P^2(P^2(p) = 0.7) \geq 0.4]$.

Given a dynamic probability frame $\langle W, \mathcal{P}^1, \mathcal{P}^2\rangle$, write $\boldsymbol{wE^iw'}$ iff $\mathcal{P}^i_w(w') > 0$. The set of worlds that evidence i leaves open at w is $\boldsymbol{E^i_w} =_{df} \{w'|\mathcal{P}^i_w(w') > 0\}$. A **dynamic prior frame** $\langle W, E^1, E^2, \pi\rangle$ is a dynamic probability frame $\langle W, \mathcal{P}^1, \mathcal{P}^2\rangle$ in which there is a regular probability distribution π over W such that for all w: $\mathcal{P}^i_w = \pi(\cdot|E^i_w)$.

Dynamic probability frames can interpret Value and Informed Trust. The latter is easy: it says $P^i_q(p|P^k_q(p) \geq t) \geq t$ (with $k \geq i$), and is valid iff all its instances are. How to formalize Value? Given a dynamic (probability or prior) frame F, we can enrich it with a **decision problem** $\langle O, U\rangle$: a set of options O and a real-valued utility function U.[36] $U_w(o)$ is the value of taking option o at world w. The *expected* value of an option o, according to evidence i, at world w, is an average of the various possible values of $U(o)$, with weights determined by how likely (according to $\mathcal{P}^i_w$) they are. Formally: $\boldsymbol{\mathbb{E}^i_w[U(o)]} =_{df} \sum_{w'} \mathcal{P}^i_w(w')U_{w'}(o)$. What about the expected value, according to evidence i, of taking the option that is warranted by evidence k (whatever it is)? Since what k warrants believing ($\mathcal{P}^k_w$) varies across worlds, what k warrants *doing* does as well: if you're at w, then k warrants taking an option o that maximizes expected value by the lights of $\mathcal{P}^k_w$. So just as we have a function $\mathcal{P}^k$ from worlds to probability functions that captures what you should *think* given evidence k, so too we'll have a function d that takes a body of evidence k and outputs a function d^k from worlds to options that captures what you should *do* given evidence k. Precisely: given a frame plus decision problem $\langle F, O, U\rangle$, $\boldsymbol{d}$ is (a variable over) any function from evidence k to a function d^k from worlds w to options $d^k_w \in O$ such that $d^k_w \in \arg\max_{o\in O}\left(\mathbb{E}^k_w[U(o)]\right)$.[37] Thus the expected value relative to evidence i of doing what evidence k warrants is the expected value of taking option d^k (whatever it is): $\boldsymbol{\mathbb{E}^i_w[U(d^k)]} =_{df} \sum_{w'} \mathcal{P}^i_w(w')U_{w'}(d^k_{w'})$.

With these definitions we can formalize propositions about expected values in the obvious ways, e.g. $[\mathbb{E}^i[U(d^k)] \geq \mathbb{E}^i[U(o)]] =_{df} \{w|\mathbb{E}^i_w[U(d^k)] \geq \mathbb{E}^i_w[U(o)]\}$. Thus:

Value: $\mathbb{E}^i[U(d^k)] \geq \mathbb{E}^i[U(o)]$ $\qquad (o \in O, k \geq i)$

[36] Technical constraint: if W is the set of worlds in frame F, then $\langle O, U\rangle$ must be such that for any π over W, there is at least one option with maximal π-expected utility: $\max_{o\in O}(\mathbb{E}_\pi[U(o)])$ is non-empty. A sufficient (but not necessary) condition for this is for O to be finite.

[37] Since d captures a strategy of responding to evidence, we impose the constraint that it cannot vary across worlds where the relevant evidence warrants the exact same beliefs: if $\mathcal{P}^i_x = \mathcal{P}^k_y$, then $d^i_x = d^k_y$.

You should always expect taking the option warranted by informed evidence to be at least as good as you expect any other particular option to be.
Slogan: *evidence is valuable.*

As with Trust, Value has three types of instances depending on whether $i = 1 = k$, $i = 1$ and $k = 2$, or $i = 2 = k$. Value *holds* at a world w in a dynamic frame F iff *for every decision problem* $\langle O, U\rangle$: all instances of Value are true at w in $\langle F, O, U\rangle$. Value is *valid* iff it holds at all worlds.

What does it take to validate Value? In full generality, we have partial answers:

Fact 2.7.1. *In any dynamic probability frame: if Misguided Evidence or Effacing Evidence are true at a world, Value fails.*

Theorem 2.7.2. *In any dynamic probability frame: if Value holds at a world, Informed Trust does as well.*

Thus we know that, in full generality, Trust is necessary for Value. The converse remains an open question:

Conjecture 2.7.3. *In any dynamic probability frame: if Informed Trust holds at a world, then Value does as well.*

But within our tractable subclass—dynamic *prior* frames—the question is closed. First, some generalizations of our definitions. A dynamic frame $\langle W, E^1, E^2\rangle$ is **transitive** iff $(xE^iy \wedge yE^iz) \Rightarrow xE^iz$. It is **shift-reflexive** iff $xE^iy \Rightarrow yE^ky$. It is **shift-nested** iff $wE^ix, y \Rightarrow (E^k_x \subseteq E^k_y$ or $E^k_x \supseteq E^k_y$ or $E^k_x \cap E^k_y = \emptyset)$. Finally, a dynamic frame is **shift-updating** iff if a world is seen by anything, it does not lose information between E^1 and E^2: $xE^iy \Rightarrow E^2_y \subseteq E^1_y$. Shift-updating corresponds to the surely-monotonicity axiom that you should be sure that if your evidence warrants being sure of p, a body of evidence at least as informed as your own will likewise warrant being sure of p: $S^i(S^ip \rightarrow S^kp)$.

These conditions characterize both Informed Trust and Value:

Theorem 2.7.4 (Value $\Leftrightarrow$ Trust)**.** *The following are equivalent:*

(1) *The dynamic prior frame* $\langle W, E^1, E^2, \pi\rangle$ *validates Informed Trust.*

(2) $\langle W, E^1, E^2\rangle$ *is transitive, shift-reflexive, shift-nested, and shift-updating.*

(3) *The dynamic prior frame* $\langle W, E^1, E^2, \pi\rangle$ *validates Value.*

Upshot: within dynamic prior frames, Trust characterizes the value of evidence.[38]

[38]The implication from (2) to (3) is a generalization of Geanakoplos (1989), Theorem 1—a result that was key to my discovery of the connection between Trust and Value.

2.B Proofs

Fact 2.3.1. *If a probability frame validates Reflection, it validates* $[P(\textit{Immodest}) = 1]$.

Proof. *Immodest* is true at x iff $\mathcal{P}_x(P = \mathcal{P}_x) = 1$. So it will suffice to take a probability frame that validates Reflection, suppose that $\mathcal{P}_w(P = \pi) > 0$, and then show that $\pi(P = \pi) = 1$. First suppose $\pi(P = \pi) = 0$; then $\mathcal{P}_w([P = \pi] \wedge [P(P = \pi) = 0]) > 0$, hence $\mathcal{P}_w(P = \pi | P(P = \pi) = 0) > 0$, contradicting Reflection. So $\pi(P = \pi) > 0$. We show that for an arbitrary world y such that $\pi(y) = t > 0$, it follows that $\pi(P(y) = t) = 1$. By Reflection, $\pi(y | P(y) = t) = t$, implying that $y \in [P(y) = t]$. But if $\pi(P(y) = t) < 1$, then $\pi(y | P(y) = t) = \frac{\pi(y \wedge [P(y)=t])}{\pi(P(y)=t)} = \frac{\pi(y)}{\pi(P(y)=t)} > \frac{\pi(y)}{1} = t$, i.e. $\pi(y | P(y) = t) > t$, contradicting Reflection. Since y was arbitrary: $\forall z \in W$: if $\pi(z) = t$, then $\pi(P(z) = t) = 1$. Since any probability function over W is fully determined by the probabilities it assigns to worlds, it follows that $\pi(P = \pi) = 1$. □

Remark 2.4.0. For any (finite) probability frame and any p, q, t, there is an $\epsilon > 0$ such that $[P_q(p) > t] = [P_q(p) \geq t + \epsilon]$. (Take the $w \in [P_q(p) > t]$ with minimal $\mathcal{P}_w(p|q)$ and set $\epsilon =_{df} (\mathcal{P}_w(p|q) - t)$.) So if $\langle W, \mathcal{P} \rangle$ validates Trust, it also validates $P_q(p | P_q(p) > t) > t$.

Fact 2.4.1. *In any probability frame: Trust holds at a world iff* $P_q(p | P_q(p) \leq t) \leq t$ *does.*

Proof. Note that $[P_q(p) \leq t] = [P_q(\neg p) \geq 1 - t]$. If Trust holds, then $P_q(\neg p | P(\neg p) \geq 1 - t) \geq 1 - t$, iff $P_q(\neg p | P(p) \leq t) \geq 1 - t$, iff $P_q(p | P(p) \leq t) \leq t$. Similarly for the converse. □

Fact 2.4.2. *There are probability frames that validate Trust in which Reflection fails at all worlds.*

Proof. The frame from Fact 2.5.1 validates Trust, yet it is easy to check that Reflection fails at all worlds. □

Fact 2.4.3. *In any probability frame, if Trust holds, and for a given* $l, h \in [0, 1]$ *it's true that* $P(S(P(p) \in [l, h]) \big| P(p) \in [l, h]) = 1$*, then* $P(p | P(p) \in [l, h]) \in [l, h]$.

Proof. Suppose Trust holds, and $P(S(P(p) \in [l, h]) \big| P(p) \in [l, h]) = 1$. Let $q =_{df} [P(p) \in [l, h]]$. Then $P_q(P = P_q) = 1$ (since conditional on q, you should be sure that the rational credence was already certain of q). Since Trust implies shift-reflexivity, the fact that $P_q(S(P(p) \in [l, h])) = 1$ implies that $P_q(P(p) \in [l, h]) = 1$; thus $P_q(P_q(p) \in [l, h]) = 1$. By Trust, $P_q(p) = P_q(p | P_q(p) \geq l) \geq l$, and $P_q(p) = P_q(p | P_q(p) \leq h) \leq h$, and so $P_q(p) = P(p | P(p) \in [l, h]) \in [l, h]$. □

Fact 2.5.1. *There are probability frames that validate both Trust and* $[P(\textit{modest}) = 1]$.

Proof. Brute force method. Let $W = \{a, b\}$, $\mathcal{P}_a(a) = 0.6$ (so $\mathcal{P}_a(b) = 0.4$), and symmetrically $\mathcal{P}_b(b) = 0.6$ (so $\mathcal{P}_b(a) = 0.4$). *modest* $=_{df} \{w | 0 < \mathcal{P}_w(P = \mathcal{P}_w) < 1\}$. Since both $\mathcal{P}_a$ and $\mathcal{P}_b$ assign positive probability to $[P = \mathcal{P}_a]$ and $[P = \mathcal{P}_b]$, *modest* $= W$, hence $[P(\textit{modest}) = 1] = W$. Moreover, this frame validates Trust. Since the frame is symmetric, it suffices to show that Trust holds at a. There are two nontrivial propositions: $\{a\}$ and $\{b\}$. If $q = \emptyset$, P_q is undefined at all worlds; if $q = \{a\}$ or $q = \{b\}$, $P_q(a) = 1$ or $P_q(b) = 1$ (respectively), so Trust holds trivially. So we need only show that for $q = W$, Trust holds. For $t_1 \leq 0.4$, $[P(a) \geq t_1] = W$; hence $\mathcal{P}_a(a|P(a) \geq t_1) = \mathcal{P}_a(a) = .6 \geq t_1$. For $t_2 \in (0.4, 0.6]$, $[P(a) \geq t_2] = \{a\}$; hence $\mathcal{P}_a(a|P(a) \geq t_2) = \mathcal{P}_a(a|a) = 1 \geq t_2$. For $t_3 \in (0.6, 1]$, $[P(a) \geq t_3] = \emptyset$, so $\mathcal{P}_a(\cdot|P(a) \geq t_3)$ is undefined. Similarly: For $t_1 \leq 0.4$, $[P(b) \geq t_1] = W$; hence $\mathcal{P}_a(b|P(b) \geq t_1) = \mathcal{P}_a(a) = .4 \geq t_1$. For $t_2 \in (0.4, 0.6]$, $[P(b) \geq t_2] = \{b\}$; hence $\mathcal{P}_a(b|P(b) \geq t_2) = \mathcal{P}_a(b|b) = 1 \geq t_2$. For $t_3 \in (0.6, 1]$, $[P(b) \geq t_3] = \emptyset$, so $\mathcal{P}_a(\cdot|P(b) \geq t_3)$ is undefined. Thus Trust holds at a and, by symmetry, is valid. □

Fact 2.5.2. *For any $T = \{t_1, ..., t_n\} \subset [0, 1]$, there are probability frames that validate Trust with a candidate π and proposition p such that $\pi(P(p) \in T) \approx 1$ and for all t_i: $\pi(P(p) = t_i) \approx \frac{1}{n}$. ($\boldsymbol{s \approx t}$ means that $s \in [t - \epsilon, t + \epsilon]$, for arbitrarily small $\epsilon > 0$.)*

Proof. Take any $T = \{t_1, ..., t_n\} \subset [0, 1]$ and $\epsilon > 0$. By Theorem 2.5.7, it will suffice to construct a shift-reflexive, transitive, shift-nested prior frame $\langle W, E, \pi \rangle$ with the desired property. Let $W = \{w, a_1, b_1, ..., a_n, b_n\}$. Define E so that $E_w = W$ while $E_{a_i} = E_{b_i} = \{a_i, b_i\}$. Clearly this frame is reflexive (every world sees itself), transitive (if xEy, $E_y \subseteq E_x$), and nested (if $E_x \not\subseteq E_y$ and $E_y \not\subseteq E_x$, then $E_x \cap E_y = \emptyset$), so any prior over it will validate Trust. In particular, let $\pi(w) = \epsilon$, and for each t_i, let $\pi(a_i) = \frac{t_i}{n}(1-\epsilon)$ and $\pi(b_i) = \frac{1-t_i}{n}(1-\epsilon)$. Then $\pi(W) = \pi(w) + \sum_{i=1}^{n}(\frac{t_i}{n} + \frac{1-t_i}{n})(1-\epsilon) = \epsilon + \sum_{i=1}^{n}(\frac{1}{n})(1-\epsilon) = \epsilon + (1-\epsilon) = 1$, so π is probability function.

Since $E_w = W$, $\mathcal{P}_w = \pi$; since $\mathcal{P}_w(w) > 0$, $\mathcal{P}_w = \pi$ will be our candidate. Let $p = \{a_1, ..., a_n\}$. For each t_i, $\mathcal{P}_{a_i}(p) = \mathcal{P}_{b_i}(p) = \pi(a_i|\{a_i, b_i\}) = \frac{\frac{t_i}{n}(1-\epsilon)}{\frac{1}{n}(1-\epsilon)} = t_i$; hence $[P(p) = t_i]$ is true at $\{a_i, b_i\}$ (and so $[P(p) = t_j]$ for $t_j \neq t_i$ is false at $\{a_i, b_i\}$). In general for all $t_i \in T$: $[P(p) = t_i] = \{a_i, b_i\}$ unless $\mathcal{P}_w(p) = t_j$ for some $t_j \in T$, in which case $[P(p) = t_j] = \{a_j, b_j, w\}$. Since $\pi(w) = \epsilon$, it follows that $\pi(P(p) \in T) \geq \pi(\neg\{w\}) = 1 - \epsilon$. And for any $t_i \in T$: $\pi(P(p) = t_i) \geq \pi(\{a_i, b_i\}) = \frac{1}{n}(1-\epsilon) = \frac{1}{n} - \frac{\epsilon}{n} \geq \frac{1}{n} - \epsilon$. Similarly, $\pi(P(p) = t_i) \leq \pi(\{a_i, b_i, w\}) = \frac{1}{n}(1-\epsilon) + \epsilon \leq \frac{1}{n} - \frac{\epsilon}{n} + \epsilon \leq \frac{1}{n} + \epsilon$, as desired. □

Fact 2.5.3. *A probability frame validates Trust iff it validates both Reaction and Reliance.*

Proof. **(⇒) :** Suppose Trust is valid. Consider Reaction: if $P_q(P_q(p) \in [l, h]) = 1$ is true at a given world, then $P_q(p) = P_q(p|P_q(p) \geq l)$, which by Trust is $\geq l$. Similarly, $P_q(p) = P_q(p|P_q(p) \leq h)$, which by Trust and Fact 2.4.1 is $\leq h$. So $P_q(p) \in [l, h]$.

Reaction is valid. Consider Reliance. Trust, Fact 2.4.1, and Remark 2.4.0 imply that $P_q(p|P_q(p) \geq t) \geq t > P_q(p|P_q(p) < t)$ whenever both are well-defined. By total probability, it follows that $P_q(p|P_q(p) \geq t) \geq P_q(p)$ whenever well-defined.

($\Leftarrow$) : Suppose Reaction and Reliance are valid. First note that Reaction requires that if $\mathcal{P}_w(x) > 0$, then $\mathcal{P}_x(x) > 0$ (shift-reflexivity). Two steps. First, if $\mathcal{P}_w(x) > 0$ and $\mathcal{P}_x(y) > 0$, then $\mathcal{P}_w(y) > 0$ (transitivity). For if not, then $\mathcal{P}_w(y|\{x,y\}) = 0$ even though $\mathcal{P}_w(P(y) > 0|\{x,y\}) = 1$. Now suppose $\mathcal{P}_w(x) > 0$ but $\mathcal{P}_x(x) = 0$. $\mathcal{P}_x(y) > 0$ for some y. If $\mathcal{P}_y(x) > 0$, by transitivity $\mathcal{P}_x(x) > 0$. If not, then $\mathcal{P}_w(P(x|\{x,y\}) = 0 \big| \{x,y\}) = 1$, yet since $\mathcal{P}_w(x) > 0$ then $\mathcal{P}_w(x|\{x,y\}) > 0$, contradicting Reaction. Now suppose $P_q(p|P_q(p) \geq t)$ is well-defined at an arbitrary x. Take arbitrary $w \in q \wedge [P_q(p) \geq t]$ such that $\mathcal{P}_x(w|q \wedge [P_q(p) \geq t]) > 0$. By shift-reflexivity, $\mathcal{P}_w(p|q \wedge [P_q(p) \geq t])$ is well-defined and by Reliance it is at least $P_w(p|q)$. Since $w \in [P_q(p) \geq t]$, it follows that $\mathcal{P}_w(p|q \wedge [P_q(p) \geq t]) \geq t$. Since w was arbitrary, $P_x(P_q(p|q \wedge [P_q(p) \geq t]) \geq t | q \wedge [P_q(p) \geq t]) = 1$. Letting $r =_{df} q \wedge [P_q(p) \geq t]$, that is: $\mathcal{P}_x(P_r(p) \geq t|r) = 1$, so by Reaction, $\mathcal{P}_x(p|r) = 1$, i.e. $P_q(p|P_q(p) \geq t) \geq t$ is true at x. Since x was arbitrary, Trust is valid. □

Fact 2.5.4. *If Trust holds at a world in a probability frame, Misguided Evidence does not.*

Proof. Suppose $P(p \wedge [P(p) < t]) \geq t$. Since $P(P(p) < t) \leq 1$, it follows that $\frac{P(p \wedge [P(p) < t])}{P(P(p) < t)} \geq t$, i.e. $P(p|P(p) < t) \geq t$, violating Trust (Remark 2.4.0 and Fact 2.4.1). □

Fact 2.5.5. *In any probability frame: if Trust holds then $[P(P(p) \geq t) \geq s] \rightarrow [P(p) \geq t \cdot s]$ does too. No stronger connection holds: for any $t, s \in [0,1]$: there are probability frames that validate Trust and make both $[P(P(p) \geq t) \geq s]$ and $[P(p) = t \cdot s]$ true at a world.*

Proof. Suppose Trust holds and $P(P(p) \geq t) \geq s$ is true. By Trust, $P(p|P(p) \geq t) \geq t$; by total probability, $P(p) = P(P(p) \geq t) \cdot P(p|P(p) \geq t) + P(P(p) < t) \cdot P(p|P(p) < t) \geq P(P(p) \geq t) \cdot P(p|P(p) \geq t) \geq s \cdot t$, so $P(p) \geq t \cdot s$.

Take any $t, s \in [0,1]$. Supposing $t, s \in (0,1)$, consider a prior frame $\langle W, E, \pi \rangle$ with $W = \{a,b,c\}$, $E_a = \{a,b,c\}$, $E_b = E_c = \{b,c\}$, and $\pi(a) = 1 - s$, $\pi(b) = t \cdot s$, and $\pi(c) = (1-t)s$. This frame is reflexive, transitive, and nested, so by Theorem 2.5.7 it validates Trust. Letting $p = \{b\}$, $\mathcal{P}_w(p) = t \cdot s$, so $[P(p) = t \cdot s]$ is true at w. And $\mathcal{P}_b(p) = \mathcal{P}_c(p) = \frac{t \cdot s}{s} = t$, hence $\{b,c\} \subseteq [P(p) \geq t]$. Since $\mathcal{P}_w(\{b,c\}) = s$, it follows that $[P(P(p) \geq t) \geq s]$ is also true at w. Finally, if we allow t, s to be extremal, it is routine to modify the frame (dropping worlds as needed) to ensure that π is regular and hence that we have a prior frame. □

Theorem 2.5.7 (Trust Characterization)**.** *A prior frame $\langle W, E, \pi \rangle$ validates Trust iff $\langle W, E \rangle$ is transitive, shift-reflexive, and shift-nested.*

Proof. A prior frame $\langle W, E, \pi\rangle$ is a special case of a dynamic prior frame $\langle W, E^1, E^2, \pi\rangle$, setting $E^1 = E^2$. In this case, (1) Informed Trust is equivalent to Trust and (2) shift-updating is trivially satisfied; hence the result follows from Theorem 2.7.4. □

Theorem 2.5.8. *$\langle W, E\rangle$ is reasoning-generable iff it is reflexive, transitive, and nested.*

Proof. **(⇒) :** Suppose W, E is reasoning-generable. *Reflexivity:* Since $f(w)$ is an initial segment of $f(w)$, wEw. *Transitivity:* If $f(w)$ is an initial segment of $f(x)$ and $f(x)$ is an initial segment of $f(y)$, then $f(w)$ is an initial segment of $f(y)$. *Nested:* Given transitivity, a frame is nested iff if xEz and yEz, then xEy or yEx. And if $f(x)$ and $f(y)$ are both initial segments of $f(z)$, then one (or both) must be an initial segment of the other.

(⇐:) For the converse, we need some definitions and a lemma.

Definition 2.5.8.a. E_y is an **expansion** of E_x iff $E_y \supset E_x$. And E_y is a **minimal expansion** of E_x iff E_y is an expansion of E_x, and for any expansion E_z of E_x, $E_z \supseteq E_y$.

Lemma 2.5.8.b. *If $\langle W, E\rangle$ is transitive and nested, then if E_x has an expansion, it has a minimal expansion.*

Proof. Suppose, for reductio, E_x has an expansion but has no minimal expansion. Thus:

$$(\forall E_y \supset E_x)(\exists E_z \supset E_x) \text{ such that } E_z \not\supseteq E_y \qquad (*)$$

W (hence E_x) is finite, so suppose $|E_x| = n$. Since E_x has a positive but finite number of expansions (since $\wp(W)$ is finite), there must be a $k > n$ such that $(\forall E_y \supset E_x)$ $|E_y| \geq k$ and $(\exists E_y \supset E_x)$ $|E_y| = k$. Take some such E_y with $|E_y| = k$. By (*), there is an $E_z \supset E_x$ such that $E_z \not\supseteq E_y$. (Hence $E_z \neq E_y$.) If $E_z \subseteq E_y$, then since $E_z \neq E_y$, $E_z \subset E_y$, and hence E_z is an expansion of E_x with $|E_z| < |E_y| = k$. Contradiction. Thus $E_z \not\subseteq E_y$. Since also $E_z \not\supseteq E_y$, since E is nested it follows that $E_z \cap E_y = \emptyset$. Yet E_z and E_y are both expansions of E_x, so $E_z \cap E_y \supseteq E_x \neq \emptyset$. Contradiction. □

Definition 2.5.8.c. Given Lemma 2.5.8.b, in transitive, nested frames we can define a function $\mathcal{M}$ such that given E_x, $\mathcal{M}(E_x)$ is E_x's minimal expansion (if it has an expansion). Define $\mathcal{M}^n$ by induction: $\mathcal{M}^1(E_x) = \mathcal{M}(E_x)$, and $\mathcal{M}^{n+1}(E_x) = \mathcal{M}(\mathcal{M}^n(E_x))$. Let $\mathcal{E}(x) =_{df} \langle E_x, \mathcal{M}^1(E_x), ..., \mathcal{M}^n(E_x)\rangle$ where $\mathcal{M}^n(E_x)$ has no expansion.

Lemma 2.5.8.d. *If $\langle W, E, \pi\rangle$ is transitive and nested, then every expansion E_z of E_x appears in $\mathcal{E}(x) = \langle E_0, E_1, ..., E_n\rangle$.*

Proof. For reductio, suppose not: $E_z \supset E_x$ but E_z does not appear in $\mathcal{E}(x)$. Since $E_z \not\supset E_n$ (since E_n has no expansion) but $E_z \supset E_0 = E_x$, there must be an i

such that $E_z \supset E_i$ but $E_z \not\supset E_{i+1}$. Since $E_z \neq E_{i+1}$, $\mathcal{M}(E_i) = E_y \neq E_z$. Since $E_y \cap E_z \supseteq E_x \neq \emptyset$, by nesting it follows that $E_y \supseteq E_z$ or $E_y \subseteq E_z$. Yet if $E_y \supseteq E_z$, then $E_y \supset E_z$—and since $E_z \supset E_i$, it follows that E_y is *not* the minimal expansion of E_i after all: $\mathcal{M}(E_i) \neq E_y$. Contradiction. And if $E_y \subseteq E_z$, then $E_y \subset E_z$. Yet $E_y = E_{i+1}$, meaning $E_{i+1} \subset E_z$, contradicting what was established above. □

We can now prove the right-to-left direction of Theorem 2.5.8. Suppose E is reflexive, transitive, and nested. Given any world x, define $f(x)$ to be the inverse of $\mathcal{E}(x)$, i.e. $f(x) = \langle \mathcal{M}^n(E_x), ..., \mathcal{M}^1(E_x), E_x \rangle$. First suppose xEy. By transitivity and reflexivity, $E_x \subseteq E_y$, so E_y is an expansion of E_x. By Lemma 2.5.8.d, E_y appears in $\mathcal{E}(x) = \langle E_x, ..., E_y, \mathcal{M}^1(E_y), ..., \mathcal{M}^n(E_y) \rangle$, and hence $f(y) = \langle \mathcal{M}^n(E_y), ..., E_y \rangle$ is an initial segment of $f(x)$. Conversely, suppose $f(x)$ is an initial segment of $f(y)$. Thus $f(y) = \langle \mathcal{M}^n(E_x), ..., E_x, ..., E_y \rangle$, and so $\mathcal{E}(y) = \langle E_y, ..., E_x, ..., \mathcal{M}^n(E_x) \rangle$, so E_x is an expansion of E_y: $E_y \subseteq E_x$, and so by reflexivity yEy and thus xEy. □

Fact 2.5.9. *A reflexive, transitive $\langle W, E, \pi \rangle$ is conditionable iff it is nested.*

Proof. **(⇒) :** Suppose we have a reflexive, transitive, non-nested frame; suppose, for reductio, that it is conditionable. Since nesting fails, we have x, y such that $E_x \not\subseteq E_y$ and $E_x \not\supseteq E_y$ and $E_x \cap E_y \neq \emptyset$. Let $C =_{df} E_x \cap E_y$. By reflexivity and transitivity we know $E_x \supset C \subset E_y$. Thus $\mathcal{P}_x(E_x|\neg C) = 1$, so by conditionability $\mathcal{P}_x(\neg C \rightarrowtail E_x) = 1$, hence (since $C \subseteq E_x$) $C \subseteq [\neg C \rightarrowtail E_x]$. By transitivity, for any $c \in C$, $E_c \subseteq C$, so $\mathcal{P}_c(C) = 1$, so **(α) :** $\mathcal{P}_c(\neg C \rightarrowtail E_x) = 1$. Meanwhile, since $\mathcal{P}_y(E_y) = 1$, $\mathcal{P}_y(E_y|\neg C) = 1$, and so $\mathcal{P}_y(E_y \cap \neg C|\neg C) = 1$, and thus $\mathcal{P}_y(\neg E_x|\neg C) = 1$, implying $\mathcal{P}_y(\neg C \rightarrowtail \neg E_x) = 1$. Since $C \subseteq E_y$, $C \subseteq [\neg C \rightarrowtail \neg E_x]$, so $\mathcal{P}_c(\neg C \rightarrowtail \neg E_x) = 1$; contradicting ($\alpha$) and the supposition that the frame is conditionable.

(⇐) : Supposing we have a reflexive, transitive, nested frame, we can define $\mathcal{E}(x)$ as in Definition 2.5.8.c. Given w, let F^w_q be the first element of $\mathcal{E}(w)$ such that $F^w_q \cap q \neq \emptyset$, or $F^w_q = W$ is there is none. Now for any q, p, let $[q \rightarrowtail p] =_{df} \{w | F^w_q \cap q \subseteq p\}$. We show this obeys the properties of conditionability. **(a):** Suppose $P_w(p|q) = 1$. Then $E_w \cap q \neq \emptyset$ and $E_w \cap q \subseteq p$. Now consider any x such $x \in E_w$. By transitivity, $E_x \cap q \subseteq p$. Since E_w appears in $\mathcal{E}(x)$ and $E_w \cap q \neq \emptyset$, by Lemma 2.5.8.d, F^x_q appears no later than E_w in $\mathcal{E}(x)$. Thus E_w is an expansion of F^x_q, so $F^x_q \cap q \subseteq p$, and hence $x \in [q \rightarrowtail p]$. Since x was arbitrary, $\mathcal{P}_w(q \rightarrowtail p) = 1$. **(b):** It'll suffice to show that for any x, if $x \in [q \rightarrowtail p]$ for $q \neq \emptyset$, then $x \notin [q \rightarrowtail \neg p]$. Supposing $x \in [q \rightarrowtail p]$, $F^x_q \cap q \subseteq p$. Since $F^x_q \cap q \neq \emptyset$, there is a $y \in p$: $y \in F^x_q \cap q$. So $F^x_q \cap q \not\subseteq \neg p$; hence $x \notin [q \rightarrowtail \neg p]$. □

Fact 2.6.1. *If Trust holds at a world in a probability frame, Effacing Evidence does not.*

Proof. Suppose $S(p \leftrightarrow [P(p) < t])$ and $S(\neg p \leftrightarrow [P(p) > t])$ are true. If $P(p) = 0$, then $P(\neg p) = 1$ and hence $P(P(p) > t) = 1$; but by Trust and Remark 2.4.0, $P(p|P(p) > t) >$

t, implying $P(p) > t \geq 0$—contradiction. So $P(p) > 0$, and hence $P(P(p) < t) > 0$, so $P(p|P(p) < t)$ is well-defined. But since $P(p \leftrightarrow [P(p) < t]) = 1$, $P(p|P(p) < t) = 1 \not< t$, violating Trust (Fact 2.4.1). □

Fact 2.7.1. *In any dynamic probability frame: if Misguided Evidence or Effacing Evidence are true at a world, Value fails.*

Proof. We prove the contrapositive. By Theorem 2.7.2, if Value holds at a world, then Informed Trust does as well. And the proofs of Facts 2.5.4 and 2.6.1 straightforwardly generalize to dynamic probability frames: if Informed Trust holds at a world, Misguided Evidence and Effacing Evidence do not. □

Theorem 2.7.2. *In any dynamic probability frame: if Value holds at a world, Informed Trust does as well.*

Basic idea: Given an Informed Trust failure, we construct a conditional bet that you expect the informed evidence to warrant making a poor decision on.

Proof. Contraposing, suppose Informed Trust fails: for some p, q, t, i, k with $k \geq i$, $P^i_q(p|P^k_q(p) \geq t) < t$ is true at some world z. Thus $P^i_q(p|P^k_q(p) \geq t) = t - a$ for $a > 0$. Define our decision problem such that $O = \{n, b\}$; the $n$ope option has 0 value at every world v, while the b is a conditional $b$et on p given q. For arbitrarily small $\epsilon > 0$:

$$U_v(n) = 0 \text{ for all } v \in W \qquad U_v(b) = \begin{cases} 0 & \text{if } v \notin q \\ 1 - t + \epsilon & \text{if } v \in p \cap q \\ -t & \text{if } v \in \neg p \cap q \end{cases}$$

We first establish $\boldsymbol{(\alpha)}$ **:** $[P^k_q(p) \geq t] \subseteq [d^k = b]$ $(=_{df} \{w | d^k_w = b\})$.

Proof. Take arbitrary w such that $\mathcal{P}^k_w(p|q) \geq t$. Since it is well-defined, $\mathcal{P}_w(q) > 0$. We must show that $\mathbb{E}^k_w[U(n)] < \mathbb{E}^k_w[U(b)]$. $\mathbb{E}^k_w[U(n)] = 0$, of course. On the other hand,

$$\begin{aligned} \mathbb{E}^k_w[U(b)] &= \mathcal{P}^k_w(q) \cdot \mathbb{E}^k_w[U(b)|q] \ + \ \mathcal{P}^k_w(\neg q) \cdot \mathbb{E}^k_w[U(b)|\neg q] && \text{(Total Expectation)} \\ &= \mathcal{P}^k_w(q) \cdot \mathbb{E}^k_w[U(b)|q] && \text{(Since } \neg q \subseteq [U(b) = 0]) \end{aligned}$$

So it'll suffice to show that $\mathbb{E}^k_w[U(b)|q] > 0$.

$$\begin{aligned} \mathbb{E}^k_w[U(b)|q] &= \mathcal{P}^k_w(p|q) \cdot (1 - t + \epsilon) \ + \ \mathcal{P}^k_w(\neg p|q) \cdot (-t) \\ &\geq \ t \cdot (1 - t + \epsilon) + (1 - t) \cdot (-t) \ = \ t\epsilon \ > \ 0 && \text{(Since } \mathcal{P}^k_w(p|q) \geq t) \end{aligned}$$

Hence $\mathbb{E}^k_w[U(b)|q] > 0$, so $\mathbb{E}^k_w[U(b)] > 0 = \mathbb{E}^k_w[U(n)]$, and thus $d^k_w = b$. □

Next we establish that $\boldsymbol{(\beta)}$ **:** $[P^k_q(p) < t] \subseteq [d^k = n]$.

Proof. Take any x such that $\mathcal{P}^k_x(p|q) < t$. Thus $\mathcal{P}^k_x(p|q) = t - d$ for some $d > 0$. (Since well-defined, $\mathcal{P}^k_x(q) > 0$.) Again, $\mathbb{E}^k_x[U(n)] = 0$; so we must show that $\mathbb{E}^k_x[U(b)] < 0$. Since $\mathbb{E}^k_x[U(b)|\neg q] = 0$ as before, we know $\mathbb{E}^k_x[U(b)] = \mathcal{P}^k_x(q) \cdot \mathbb{E}^k_x[U(b)|q]$. So it'll suffice to show that $\mathbb{E}^k_x[U(b)|q] < 0$.

$$\begin{aligned}\mathbb{E}^k_x[U(b)|q] &= \mathcal{P}^k_x(p|q) \cdot (1 - t + \epsilon) \ + \ \mathcal{P}^k_x(\neg p|q) \cdot (-t) \\ &= \ (t - d) \cdot (1 - t + \epsilon) + (1 - t + d) \cdot (-t) \ = \ \epsilon(t - d) - d\end{aligned}$$

As $\epsilon \to 0$ the left term vanishes and $\mathbb{E}^k_x[U(b)|q] < 0$. It follows that $\mathbb{E}^k_x[U(b)] < 0 = \mathbb{E}^k_x[U(n)]$, hence $d^k_x = n$. (Since $[P^k_q(p) < t]$ is finite, it follows that there is an ϵ small enough such that, for *all* $x \in [P^k_q(p) < t]$: $d^k_x = n$.) □

We now turn to showing that at our original world x, $\mathbb{E}^i_z[U(d^k)] < 0 = \mathbb{E}^i_z[U(n)]$, and hence that Value fails at z. Since $\mathbb{E}^i_z[U(d^k)|\neg q] = 0$, we know

$$\mathbb{E}^i_z[U(d^k)] = \mathcal{P}^i_z(q) \cdot \mathbb{E}^i_z[U(d^k)|q]$$

with $\mathcal{P}^i_z(q) > 0$. So it'll suffice to show that $\mathbb{E}^i_z[U(d^k)|q] < 0$. Since $[P^k_q(p) \geq t]$ and $[P^k_q(p) < t]$ partition the q-worlds assigned positive probability by z (since Value requires shift-reflexivity[39], and hence that the conditional probability is well-defined), we know

$$\begin{aligned}\mathbb{E}^i_z[U(d^k)|q] = \ & \mathcal{P}^i_z(P^k_q(p) \geq t|q) \cdot \mathbb{E}^i_z[U(d^k)|q \wedge [P^k_q(p) \geq t]] \\ & + \ \mathcal{P}^i_z(P^k_q(p) < t|q) \cdot \mathbb{E}^i_z[U(d^k)|q \wedge [P^k_q(p) < t]]\end{aligned}$$

(β) implies that $[P^k_q(p) < t] \subseteq [d^k = n]$, so $\mathbb{E}^i_z[U(d^k)|q \wedge [P^k_q(p) < t]] = 0$, hence the right summand drops out:

$$\mathbb{E}^i_z[U(d^k)|q] = \ \mathcal{P}^i_z(P^k_q(p) \geq t|q) \cdot \mathbb{E}^i_z[U(d^k)|q \wedge [P^k_q(p) \geq t]]$$

Thus it suffices to show that $\mathbb{E}^i_z[U(d^k)|q \wedge [P^k_q(p) \geq t]] < 0$. (α) implies that $[P^k_q(p) \geq t] \subseteq [d^k = b]$, so it'll in turn suffice to show that $\mathbb{E}^i_z[U(b)|q \wedge [P^k_q(p) \geq t]] < 0$. Since $[P^i_q(p|P^k_q(p) \geq t) = t - a]$ is true at z, $\mathcal{P}^i_z(p|q \wedge [P^k_q(p) \geq t]) = t - a$; hence

$$\begin{aligned}\mathbb{E}^i_z[U(b)|q \wedge [P^k_q(p) \geq t]] \ &= \ \mathcal{P}^i_z(p|q \wedge [P^k_q(p) \geq t]) \cdot (1 - t + \epsilon) + \mathcal{P}^i_z(\neg p|q \wedge [P^k_q(p) \geq t]) \cdot (-t) \\ &= \ (t - a) \cdot (1 - t + \epsilon) + (1 - t + a) \cdot (-t) \ = \ \epsilon(t - a) - a\end{aligned}$$

As $\epsilon \to 0$, the left term vanishes and $\mathbb{E}^i_z[U(b)|q \wedge [P^k_q(p) \geq t]] < 0$. It follows that $\mathbb{E}^i_z[U(d^k)|q \wedge [P^k_q(p) \geq t]] < 0$ and hence that $\mathbb{E}^i_z[U(d^k)|q] < 0$, and hence that

[39] If z sees w but w doesn't see itself, define a bet which pays off if $\neg\{w\}$ and has a huge cost $-N$ if w. Since $\mathcal{P}_w$ will warrant taking the bet no matter how large N is, eventually $E^i_z[U(d^k)] < 0$, and Value fails.

$\mathbb{E}^i_z[U(d^k)] < 0 = \mathbb{E}^i_z[U(n)]$. Value fails at z. □

Theorem 2.7.4 (Value ⇔ Trust)**.** *The following are equivalent:*

(1) *The dynamic prior frame* $\langle W, E^1, E^2, \pi\rangle$ *validates Informed Trust.*

(2) $\langle W, E^1, E^2\rangle$ *is transitive, shift-reflexive, shift-nested, and shift-updating.*

(3) *The dynamic prior frame* $\langle W, E^1, E^2, \pi\rangle$ *validates Value.*

It is immediate from Theorem 2.7.2 that (3) implies (1); so we must establish two lemmas: Lemma 2.7.4.2 that (1) implies (2), and Lemma 2.7.4.3 that (2) implies (3). First, we extend our definitions of transitive, reflexive (etc.) to apply to sets of worlds:

Definition 2.7.4.1. Given a set $q \subseteq W$ in $\langle W, E^1, E^2\rangle$ and arbitrary $x, y \in q$: q is **transitive** iff for any $z \in W$, $x \in E^k_z \Rightarrow E^k_x \subseteq E^k_z$; q is **reflexive** iff xE^kx; q is **updating** iff $E^2_x \subseteq E^1_x$; q is **nested** iff ($E^k_x \subseteq E^k_y$ or $E^k_x \supseteq E^k_y$ or $E^k_x \cap E^k_y = \emptyset$). Note that $\langle W, E^1, E^2\rangle$ is transitive, shift-reflexive, shift-nested, and shift-updating iff each E^i_w is transitive, reflexive, nested, and updating.

Now we prove that (1) implies (2) in Theorem 2.7.4.

Lemma 2.7.4.2. *The dynamic prior frame* $\langle W, E^1, E^2, \pi\rangle$ *validates Informed Trust only if* $\langle W, E^1, E^2\rangle$ *is transitive, shift-reflexive, shift-nested, and shift-updating.*

Proof. We show the contrapositive, using Definition 2.7.4.1: supposing there is a w with E^i_w not transitive (etc.), we show that Informed Trust fails.

Transitivity: Suppose $\exists x \in E^i_w$ such that $x \in E^k_z$ but $E^k_x \not\subseteq E^k_z$. By regularity, $[P^k(E^k_z) < 1]$ is true at x. Since $\mathcal{P}^k_z(E^k_z) = 1$ and zE^kx, $\mathcal{P}^k_z(E^k_z | P^k(E^k_z) < 1]) = 1$. Informed Trust fails at z (Fact 2.4.1).

Reflexivity: Suppose $\exists x \in E^i_w$ such that $x\not E^k x$. Let $p = W - \{x\}$, so $[P^k(p) = 1] \wedge \neg p$ is true at x. Since wE^ix, $\mathcal{P}^i_w(\neg p | P^k(p) = 1) > 0$, so $\mathcal{P}^i_w(p | P^k(p) \geq 1) < 1$. Informed Trust fails.

Updating: We know E^i_w is reflexive. Suppose $\exists x \in E^i_w$ such that $E^2_x \not\subseteq E^1_x$, so there is a y with xE^2y but $x\not E^1 y$. By the latter, $\mathcal{P}^1_x(y) = 0$. By the former, $[P^2(y) > 0]$ is true at x. By reflexivity, xE^1x, so $\mathcal{P}^1_x(P^2(y) > 0) > 0$. Combined, we have $\mathcal{P}^1_x(y | P^2(y) > 0) = 0$. Informed Trust fails.

Nesting: Suppose $\exists x, y, z \in E^i_w$ with $E^k_x \not\subseteq E^k_y$ and $E^k_x \not\supseteq E^k_y$, but $z \in E^k_x \cap E^k_y$. We know that E^i_w must be transitive, reflexive, and updating; we'll show that Informed Trust fails at w for $q = \{x, y, z\}$, $p = \{z\}$, and

$$t =_{df} \min_{v \in \{x,y,z\}} \Big[\mathcal{P}^k_v(z|q)\Big].$$

Since E^k is transitive, we know $x\not E^k y$, $y \not E^k x$, and $x, y \not E^k w$. By the definition of t, $[P^k_q(z) \geq t] \supseteq \{x, y, z\} = q \subseteq E^i_w$, so **($\boldsymbol{\alpha}$):** $q = q \cap [P^k_q(z) \geq t] \cap E^i_w$. Now, $x, y \notin E^k_z$

for otherwise zE^kx or zE^ky, and so (by transitivity) xE^ky or yE^kx—contradiction. Thus:

$$\mathcal{P}^k_z(z|q) = \mathcal{P}^k_z(z|\{x,y,z\}) = 1. \qquad (\beta)$$

Moreover, since $y \notin E^k_x$ and $x \notin E^k_y$:

$$\mathcal{P}^k_x(z|q) = \frac{\pi(z \cap E^k_x \cap q)}{\pi(E^k_x \cap q)} = \frac{\pi(z)}{\pi(\{x,z\})} \qquad (\gamma)$$

$$\mathcal{P}^k_y(z|q) = \frac{\pi(z \cap E^k_y \cap q)}{\pi(E^k_y \cap q)} = \frac{\pi(z)}{\pi(\{y,z\})} \qquad (\delta)$$

Combining (β), (γ), and (δ), and the definition of t, we know

$$\begin{aligned} t \;&\geq\; \frac{\pi(z)}{\pi(\{x,z\})}, \frac{\pi(z)}{\pi(\{y,z\})} \\ &>\; \frac{\pi(z)}{\pi(\{x,y,z\})} && \text{(by regularity)} \\ &=\; \pi(z|q) \;=\; \pi(z|q \cap [P^k_q(z) \geq t] \cap E^i_w) && \text{(by } (\alpha)\text{)} \\ &=\; \mathcal{P}^i_w(p\big|q \cap [P^k_q(z) \geq t]) \end{aligned}$$

That is, $P^i_q(z\big|P^k_q(z) \geq t) < t$ at w: Informed Trust fails. □

The final step for Theorem 2.7.4 is showing that that (2) implies (3):

Lemma 2.7.4.3 (cf. Geanakoplos 1989)**.** *If $\langle W, E^1, E^2\rangle$ is transitive, shift-reflexive, shift-nested, and shift-updating, then $\langle W, E^1, E^2, \pi\rangle$ validates Value.*

Basic idea: Show that if the frame is transitive, shift-reflexive, shift-nested, and shift-updating, then we can partition E^i_w into smaller "branches," with expectations from w an average of the expectations conditional on each branch. Then an induction on the size of E^i_w suffices to carry Value from the "leaves" up through the tree.

Definition 2.7.4.3.a (k-closed)**.** $q \subseteq W$ is k-closed iff for any $x \in q$, $E^k_x \subseteq q$. Note that if E^i_w is transitive and updating, E^i_w is k-closed. (Updating implies that for any $x \in E^i_w$, then $E^2_x \subseteq E^1_x$, so $E^k_x \subseteq E^i_x$, and by transitivity $E^i_x \subseteq E^i_w$.)

Definition 2.7.4.3.b (k-classes)**.** Given a k-closed $q \subseteq W$, let its set $\boldsymbol{\mathcal{N}^k}$ of **k-classes** partition q into worlds that see the same worlds under E^k: $\mathcal{N}^k =_{df} \{N \subseteq q|\forall x,y \in N:\ E^k_x = E^k_y\}$. The k-class of a world x is denoted $N^k_x = \{y \in q|E^k_y = E^k_x\}$. We let $\boldsymbol{A^k}$ denote the k-class whose members see *A*ll of q under E^k: $A^k =_{df} \{x \in q|E^k_x = q\}$. ($A^k$ may be empty.)

Fact 2.7.4.3.c (k-class accessibility)**.** If q is transitive, reflexive, and k-closed, and $N, M \in \mathcal{N}^k$, then $(\exists n \in N,\ m \in M:\ nE^km)$ iff $(\forall n \in N, m \in M : nE^km)$.

(Why? Suppose nE^km. $\forall m' \in M$: (reflexivity) $m'E^km'$, so (same k-class) mE^km', so (transitivity) nE^km'; so (same k-class) $\forall n' \in N:\ n'E^km'$.) Thus within E^i_w we can treat E^k as a relation between k-classes: for $N, M \in \mathcal{N}^k:\ NE^kM$ iff $\exists n \in N, m \in M:\ nE^km$, iff $\forall n \in N, m \in M:\ nE^km$. Similarly for the neighborhood of a class: $\boldsymbol{E^k_N} =_{df} \{x \in q | \exists y \in N:\ E^k_y = E^k_x\} = \{x \in q | \forall y \in N:\ E^k_y = E^k_x\}$. Note: by reflexivity and transitivity: NE^kM iff $E^k_M \subseteq E^k_N$; and $E^k_M \subset E^k_N$ iff NE^kM and $N \neq M$.

Definition 2.7.4.3.d (Maximal k-classes)**.** Given a transitive, reflexive, and k-closed q, the *maximal* k-classes of q are those that see strictly less than q under E^k but are not seen by any other k-classes that do so: $\{M \in \mathcal{N}^k | E^k_M \subset q \text{ and } \neg\exists K \in \mathcal{N}^k:\ E^k_M \subset E^k_K \subset q\}$.

Fact 2.7.4.3.e. *If q is transitive, reflexive, nested, and k-closed, and $M_1, ..., M_n$ are its maximal k-classes, then it is partitioned by $\{A^k, E^k_{M_1}, ..., E^k_{M_n}\}$.*

Proof. **Exhaustivity:** Take arbitrary $x \in q$. By reflexivity, $x \in N^k_x$. If $E^k_{N^k_x} \not\subset q$, then $E^k_{N^k_x} = q$, so $x \in A^k$, hence covered by $\{A^k, E^k_{M_1}, ..., E^k_{M_n}\}$. So suppose $E^k_{N^k_x} \subset q$; we show that $x \in E^k_{M_j}$ for some maximal M_j. By reflexivity $x \in E^k_x$, so $N^k_x E^k N^k_x$. Therefore there must be an node M_j that's maximal and $M_j E^k N^k_x$. For suppose not: there is no $K \in \mathcal{N}^k$ such that $E^k_K \subset q$, $KE^kN^k_x$, and (by definition of maximal) $\neg\exists K' \in \mathcal{N}^k:\ E^k_K \subset E^k_{K'} \subset q$, i.e.

$$\forall K \in \mathcal{N}^k:\ \text{if } E^k_K \subset q \text{ and } KE^kN^k_x \text{ then } \exists K' \in \mathcal{N}^k:\ E^k_K \subset E^k_{K'} \subset q. \qquad (\alpha)$$

But this blows up the size of q. Since q is finite, suppose $|q| = m$. Setting $K = N^k_x$, we have $E^k_{N^k_x} \subset q$ and $N^k_x E^k N^k_x$; therefore by (α) there is a K' with $E^k_{N^k_x} \subset E^k_{K'} \subset q$. Since $E^k_{N^k_x} \subset E^k_{K'}$, $K'E^kN^k_x$. But then setting $K = K'$ we have $E^k_{K'} \subset q$ and $K'E^kN^k_x$, so by (α) again we get a K'' such that $E^k_{K'} \subset E^k_{K''} \subset q$. By iterating this, we prove that $|q| > m$. Contradiction. Thus there must be a maximal node M_j that accesses N^k_x, and hence accesses x. Thus $x \in E^k_{M_j}$, as desired.

Exclusivity: If there is an $x \in A^k \cap E^k_{M_j}$, then $M_j E^k A^k$ so by transitivity $E^k_{M_j} \not\subset q$. Contradiction. So A^k is disjoint from all the $E^k_{M_l}$. Next, take any $M_l \neq M_j$, with $m_l \in M_l$ and $m_j \in M_j$. If $m_l E^k m_j$ or $m_j E^k m_l$, then either they access each other (so by transitivity $M_l = M_j$—contradiction) or only one accesses the other—WLOG, say $m_l E^k m_j$. Since $m_l E^k m_l$ but $m_j \not E^k m_l$, by transitivity $E^k_{m_j} \subset E^k_{m_l} \subset q$, contradicting the assumption that M_j is maximal. Thus m_l and m_j do not access each other, so by nestedness $E^k_{m_l} \cap E^k_{m_j} = \emptyset$, i.e. $E^k_{M_l}$ and $E^k_{M_j}$ are disjoint. □

Definition 2.7.4.3.f. Given a dynamic prior frame $\langle W, E^1, E^2, \pi\rangle$ and any random variable X, let $\boldsymbol{\mathbb{E}_q[X]}$ $(= \mathbb{E}[X|q])$ be the expectation of X relative to π conditional on q:
$\mathbb{E}_q[X] =_{df} \sum_t \pi(X = t|q) \cdot t$. And let $\boldsymbol{\pi_q} =_{df} \pi(\cdot|q)$.

Fact 2.7.4.3.g. *If q is k-closed, transitive, reflexive, and nested, then any k-closed $r \subseteq q$ is also transitive, reflexive, and nested.*

Proof. Let $x, y \in r$. *Transitive:* Suppose $\exists z \in W$ such that $x \in E^k_z$. Since $x \in q$ and q is transitive, if $xE^k y$ then $zE^k y$. *Reflexive:* Since $x \in q$ and q is reflexive, $xE^k x$. *Nested:* since $x, y \in q$, either $E^k_x \subseteq E^k_y$ or $E^k_x \supseteq E^k_y$ or $E^k_x \cap E^k_y = \emptyset$. □

Lemma 2.7.4.3.h. *If q is k-closed, transitive, reflexive, and nested, then for any decision problem $\langle O, U\rangle$ and any d and $o \in O$: $\mathbb{E}_q[U(d^k)] \geq \mathbb{E}_q[U(o)]$.*

Proof. We proceed by induction on the size of q. *Base case:* If $|q| = 1$, then $q = \{x\}$. Since q is k-closed and reflexive, $E^k_x = \{x\}$, so $\pi(\cdot|q) = \mathcal{P}^k_x$; so for any random variable X, $\mathbb{E}_q[X] = \mathbb{E}^k_x[X]$. Moreover $\pi(d^k = d^k_x|q) = 1$, so $\mathbb{E}_q[U(d^k)] = \mathbb{E}_q[U(d^k_x)] = \mathbb{E}^k_x[U(d^k_x)] = \max_{o\in O}\left(\mathbb{E}^k_x[U(o)]\right) = \max_{o\in O}\left(\mathbb{E}_q[U(o)]\right)$; hence $\mathbb{E}_q[U(d^k)] \geq \mathbb{E}_q[U(o)]$.

Induction case: Suppose $|q| = n$ and for all $r \subseteq W$ with $|r| < |q|$, the hypothesis holds. By Fact 2.7.4.3.e, if $M_1, ..., M_l$ are q's maximal k-classes, then q can be partitioned by $\{A^k, E^k_{M_1}, ..., E^k_{M_l}\}$. We thus can break down $\mathbb{E}_q[U(o)]$ as follows:

$$\mathbb{E}_q[U(o)] \;=\; \pi_q(A^k)\mathbb{E}_q[U(o)|A^k] \;+\; \sum_j \pi_q(E^k_{M_j})\mathbb{E}_q[U(o)|E^k_{M_j}] \qquad (\alpha)$$

Supposing $A^k = \emptyset$, the first summand drops out:

$$\mathbb{E}_q[U(o)] \;=\; \sum_j \pi_q(E^k_{M_j})\mathbb{E}_q[U(o)|E^k_{M_j}] \qquad (\beta)$$

Since each $E^k_{M_j} \subset q$ and q is transitive and k-closed, $E^k_{M_j}$ is k-closed. By Fact 2.7.4.3.g, $E^k_{M_j}$ is also transitive, reflexive, and nested. Since it is smaller than q, the inductive hypothesis holds and $\mathbb{E}[U(d^k)|E^k_{M_j}] \geq \mathbb{E}[U(o)|E^k_{M_j}]$. And since $q \cap E^k_{M_j} = E^k_{M_j}$, for any random variable X, $\mathbb{E}[X|E^k_{M_j}] = \mathbb{E}_q[X|E^k_{M_j}]$. Plugging these facts into (β) yields:

$$\leq \;\sum_j \pi_q(E^k_{M_j})\mathbb{E}_q[U(d^k)|E^k_{M_j}] \;=\; \mathbb{E}_q[U(d^k)]$$

That is, $\mathbb{E}_q[U(o)] \leq \mathbb{E}_q[U(d^k)]$, as desired.

Next suppose $A^k \neq \emptyset$, so we have some $w \in A^k$ such that $E^k_w = q$. Then $\pi_q = \mathcal{P}^k_w$, $\mathbb{E}_q[X] = \mathbb{E}^k_w[X]$; and so by the definition of d^k_w we have:

$$\begin{aligned}\mathbb{E}_q[U(o)] \;&=\; \mathbb{E}^k_w[U(o)] \leq \mathbb{E}^k_w[U(d^k_w)] \\ &=\; \mathcal{P}^k_w(A^k)\mathbb{E}^k_w[U(d^k_w)|A^k] \;+\; \sum_j \mathcal{P}^k_w(E^k_{M_j})\mathbb{E}^k_w[U(d^k_w)|E^k_{M_j}] \qquad (\gamma)\end{aligned}$$

Since for any $x \in A^k$, $\mathcal{P}^k_x = \mathcal{P}^k_w$, we know $d^k_x = d^k_w$. Hence $\mathbb{E}^k_w[U(d^k_w)|A^k] = \mathbb{E}^k_w[U(d^k)|A^k]$. And by parallel reasoning to above, for each $E^k_{M_j}$, $\mathbb{E}^k_w[U(d^k_w)|E^k_{M_j}] \leq$

$\mathbb{E}^k_w[U(d^k)|E^k_{M_j}]$. These facts imply a comparison with (γ):

$$\leq \mathcal{P}^k_w(A^k)\mathbb{E}^k_w[U(d^k)|A^k] + \sum_j \mathcal{P}^k_w(E^k_{M_j})\mathbb{E}^k_w[U(d^k)|E^k_{M_j}] = \mathbb{E}^k_w[U(d^k]$$

And since $\mathbb{E}^k_w[U(d^k)] = \mathbb{E}_q[U(d^k)]$, we have the desired result. □

We are finally in a position to complete the proof of Theorem 2.7.4 by establishing Lemma 2.7.4.3.

Lemma 2.7.4.3 (cf. Geanakoplos 1989)**.** *If $\langle W, E^1, E^2\rangle$ is transitive, shift-reflexive, shift-nested, and shift-updating, then $\langle W, E^1, E^2, \pi\rangle$ validates Value.*

Proof. Suppose $\langle W, E^1, E^2\rangle$ is transitive, shift-reflexive, shift-nested, and shift-updating, and consider an arbitrary prior frame $\langle W, E^1, E^2, \pi\rangle$ built on it. Consider an arbitrary world w. By Definitions 2.7.4.1 and 2.7.4.3.a, E^i_w is k-closed, transitive, reflexive, and nested. Thus Lemma 2.7.4.3.h applies: for any decision problem $\langle O, U\rangle$ and any d and $o \in O$: $\mathbb{E}[U(d^k)|E^i_w] \geq \mathbb{E}[U(o)|E^i_w]$; that is, $\mathbb{E}^i_w[U(d^k)] \geq \mathbb{E}^i_w[U(o)]$: Value holds. □

Recall our disagreement setup:

(1) $S([P(\mathit{Top}) = h] \vee [P(\mathit{Top}) = l])$ $\quad (h > l)$
(2) $P(\mathit{Top}) = h$
(3) $P(P(\mathit{Top}) = t|C_D(\mathit{Top}) = t) > P(P(\mathit{Top}) = t)$
(4) $P(\mathit{Top}|[P(\mathit{Top}) = t] \wedge [C_D(\mathit{Top}) = s]) = P(\mathit{Top}|P(\mathit{Top}) = t)$.

Fact 2.8.1. (1)–(4) *are consistent with $P(\mathit{Top}|C_D(\mathit{Top}) < h) \geq P(\mathit{Top})$. But given Trust,* (1)–(4) *imply $P(\mathit{Top}|C_D(\mathit{Top}) < h) < P(\mathit{Top})$.*

Proof. Our model of (1)-(4) will be based on the Sycophants frame. $W = \{s, p\}$. $\mathcal{P}(s) = \mathcal{P}_s$ such that $\mathcal{P}_s(s) = 0.1$ and $\mathcal{P}_s(p) = 0.9$. Similarly, $\mathcal{P}(p) = \mathcal{P}_p$ such that $\mathcal{P}_p(p) = 0.1$ and $\mathcal{P}_p(s) = 0.9$. Set $\mathit{Top} = p$ and $h = 0.9$. At world s, $S([P(p) = 0.9] \vee [P(p) = 0.1])$ and $[P(p) = 0.9]$ are true, so (1) and (2) are satisfied. Enrich this model with a function $\mathcal{D}$ from worlds to Disa's credences; and set $\mathcal{D}(w) = \mathcal{P}_w$ for each w. Then $S([C_D(q) = t] \leftrightarrow [P(q) = t])$ is valid. Thus $P(P(p) = t|C_D(p) = t) = 1 > .9 \geq P(P(p) = t)$, so (3) is satisfied. Finally, since $S([P(p) = t] \leftrightarrow [C_D(p) = t])$ is valid, $P(p|[P(p) = t] \wedge [C_D(p) = s]) = P(p|P(p) = t)$ if well-defined, so (4) is satisfied. Nevertheless, $P(p|C_D(p) < .9) = P(p|P(p) < .9) = 1 \geq .9 = P(p)$, establishing the first result.

Now suppose Trust holds and (1)–(4) are true. Let $q = [C_D(\mathit{Top}) < h]$. Then $P(\mathit{Top}|C_D(\mathit{Top}) < h) = P_q(\mathit{Top})$ can be broken down:

$= P_q(P(\mathit{Top}) = l)P_q(\mathit{Top}|P(\mathit{Top}) = l) + P_q(P(\mathit{Top}) = h)P_q(\mathit{Top}|P(\mathit{Top}) = h)$ [by (1)]
$= P_q(P(\mathit{Top}) = l)P(\mathit{Top}|P(\mathit{Top}) = l) + P_q(P(\mathit{Top}) = h)P(\mathit{Top}|P(\mathit{Top}) = h)$ [by (4)]

From (3) it follows that $P(P(\mathit{Top}) = l | C_D(\mathit{Top}) < h) = P_q(P(\mathit{Top}) = l) > P(P(\mathit{Top}) = l)$ and $P_q(P(\mathit{Top}) = h) > P(P(\mathit{Top}) = h)$. Since

$$P(\mathit{Top}) = P(P(\mathit{Top}) = l)P(\mathit{Top}|P(p) = l) + P(P(\mathit{Top}) = h)P(\mathit{Top}|P(p) = h)$$

$P_q(\mathit{Top})$ is weighted more towards $P(\mathit{Top}|P(p) = l)$ than $P(\mathit{Top})$ is. Since by (1), $P(\mathit{Top}|P(\mathit{Top}) = l) = P(\mathit{Top}|P(\mathit{Top}) \leq l) \leq l$ (by Trust), and $l < h \leq P(\mathit{Top}|P(\mathit{Top}) \geq h) = P(\mathit{Top}|P(\mathit{Top}) = h)$, it follows that $P_q(\mathit{Top})$ is more weighted towards the lower value than $P(\mathit{Top})$ is, so $P_q(\mathit{Top}) = P(\mathit{Top}|C_D(\mathit{Top}) < h) < P(\mathit{Top})$. □

Fact 2.8.2. *In any probability frame: if Value or Trust hold at a world, then* $Sp \rightarrow SSp$ *does as well.*

Proof. By Theorem 2.7.2, it suffices to show the result for Trust. For reductio, suppose Sp and $\neg SSp$ are true, so $[P(p) = 1]$ yet $P(P(p) = 1) < 1$. By the latter, $P(p|P(p) < 1)$ is well-defined; by the former, $P(p|P(p) < 1) = 1$, violating Trust. □

2.C Glossary

This glossary collects brief definitions of the technical terms used in the statements of principles and theorems, in alphabetical order.

- **Candidate:** π is a *candidate* in frame $\langle W, \mathcal{P}\rangle$ iff at some world you should think π might be the rational credence function: $\exists w \in W : \mathcal{P}_w(P = \pi) > 0$.
- **Conditionable:** A frame $\langle W, E\rangle$ is *conditionable* iff we can define a two-place propositional connective $\rightarrowtail$ such that (a) $P(p|q) = 1$ implies $P(q \rightarrowtail p) = 1$, and (b) for $q \neq \emptyset$, $P(q \rightarrowtail p) = 1$ implies $P(q \rightarrowtail \neg p) < 1$.
- $\boldsymbol{d}$**:** Given a frame plus decision problem $\langle F, O, U\rangle$, d is (a variable over) any function from evidence k to a functions d^k from worlds w to options $d^k_w \in O$ such that (1) $d^k_w \in \arg\max_{o\in O}\left(\mathbb{E}^k_w[U(o)]\right)$ and (2) if $\mathcal{P}^i_x = \mathcal{P}^k_y$, then $d^i_x = d^k_y$.
- **Decision Problem** $\boldsymbol{\langle O, U\rangle}$**:** A set of options O and a real-valued utility function U; $U_w(o)$ is the utility of $o \in O$ at w.[40]
- **Dynamic Prior Frame** $\boldsymbol{\langle W, E^1, E^2, \pi\rangle}$**:** A dynamic probability frame $\langle W, \mathcal{P}^1, \mathcal{P}^2\rangle$ in which there is a regular probability distribution π over W—the prior—such that for all w: $\mathcal{P}^i_w = \pi(\cdot|E^i_w)$.
- **Dynamic Probability Frame** $\boldsymbol{\langle W, \mathcal{P}^1, \mathcal{P}^2\rangle}$**:** A probability frame with an extra function $\mathcal{P}^2$; it associates each possibility w with a credence function $\mathcal{P}^1_w$ that's rational given evidence 1 and $\mathcal{P}^2_w$ that's rational given evidence 2.
- **Effacing Evidence:** $\exists p, t : S(p \leftrightarrow [P(p) < t])$ and $S(\neg p \leftrightarrow [P(p) > t])$.
- $\boldsymbol{\mathbb{E}^i_w[U(o)]} =_{df} \sum_{w'} \mathcal{P}^i_w(w')U_{w'}(o)$

[40]For $\langle O, U\rangle$ to enrich a frame with set of worlds W, $\langle O, U\rangle$ must be such that for any π over W there is at least one option with maximal π-expected utility: $\max_{o\in O}(\mathbb{E}_\pi[U(o)])$ is non-empty.

- $\mathbb{E}^i_w[\boldsymbol{U}(\boldsymbol{d}^k)] =_{df} \sum_{w'} \mathcal{P}^i_w(w')U_{w'}(d^k_{w'})$
- $\boldsymbol{wE^iw'}$**,** $\boldsymbol{E^i_w}$**:** Given a dynamic probability frame $\langle W, \mathcal{P}^1, \mathcal{P}^2\rangle$, wE^iw' iff $\mathcal{P}^i_w(w') > 0$ and $E^i_w =_{df} \{w' | \mathcal{P}^i_w(w') > 0\}$. In probability frames the superscripts are omitted.
- **Frame:** A structure $\langle W, E\rangle$ underlying a prior frame $\langle W, E, \pi\rangle$. W is finite; E is serial: $\forall x \exists y : \ xEy$.
- **Holds**: A principle *holds* at world w iff all of its instances (well-defined instantiations of free variables) are true at w.
- $\boldsymbol{i}$ **and** $\boldsymbol{k}$**:** *Convention*: when using multiple variables over bodies of evidence, i and k obey the constraint that $k \geq i$.
- **Informed:** A body of evidence is *informed* (with respect to yours) iff it contains all your evidence, and maybe more.
- **Informed Trust:** $P^i_q(p | P^k_q(p) \geq t) \geq t \qquad (k \geq i)$
- **Judgment:** You *judge* that p iff you have a credence in p above the given (contextually specified) threshold t.
- **Misguided Evidence:** $\exists p, t : \ P(p \wedge [P(p) < t]) \geq t$
- **Nested:** A frame $\langle W, E\rangle$ is *nested* iff for any x, y: either $E_x \subseteq E_y$ or $E_x \supseteq E_y$ or $E_x \cap E_y = \emptyset$.
- **Opinion:** You have a given *opinion* about p iff you have a credence in p in the given (contextually specified) range $[l, h]$.
- $\boldsymbol{\pi}$**:** A probability function whose values (unlike P) are known. $\boldsymbol{\pi_q} =_{df} \pi(\cdot|q)$.
- $\boldsymbol{\mathcal{P}_w}$**:** The probability function which is rational at world w; it's values are known.
- $[\boldsymbol{P^i(p) = t}]$ **:** Given a dynamic probability frame, $[P^i(p) = t] =_{df} \{w | \mathcal{P}^i_w(p) = t\}$. In probability frames the superscript is omitted.
- $\boldsymbol{P_q(p)}$ $(= P(p|q))$**:** The rational credence—whatever it is—in p conditional on q.
- **Prior Frame** $\boldsymbol{\langle W, E, \pi\rangle}$**:** A probability frame $\langle W, \mathcal{P}\rangle$ in which there is a regular probability distribution π over W (i.e. $\forall w \in W : \ \pi(w) > 0$) such that for all w: $\mathcal{P}_w = \pi(\cdot|E_w)$.
- **Probability Frame** $\boldsymbol{\langle W, \mathcal{P}\rangle}$**:** A structure consisting of a finite set of epistemic possibilities W and a function $\mathcal{P}$ that associates each possibility w with the rational credences $\mathcal{P}_w$ that you should have at w.
- **Reaction:** If $P_q\big(P_q(p) \in [l, h]\big) = 1$, then $P_q(p) \in [l, h]$
- **Reasoning-generable:** A frame $\langle W, E\rangle$ is *reasoning-generable* iff there is a function f from worlds w to sequences of propositions $f(w)$ such that: wEx iff $f(w)$ is an initial segment of $f(x)$.
- **Reflection:** $P(p | P(p) \in [l, h]) \in [l, h]$
- **Reliance:** $P_q(p | P_q(p) \geq t) \geq P_q(p)$
- **Reflexive:** A frame $\langle W, E\rangle$ is *reflexive* iff for all x: xEx.
- **Shift-Nested:** A dynamic frame $\langle W, E^1, E^2\rangle$ is *shift-nested* iff $wE^ix, y \Rightarrow (E^k_x \subseteq E^k_y$ or $E^k_x \supseteq E^k_y$ or $E^k_x \cap E^k_y = \emptyset)$. (For a frame $\langle W, E\rangle$, omit the superscripts.)
- **Shift-Reflexive:** A dynamic frame $\langle W, E^1, E^2\rangle$ is *shift-reflexive* iff $xE^iy \Rightarrow yE^ky$. (For a frame $\langle W, E\rangle$, omit the superscripts.)

- **Shift-Updating:** A dynamic frame $\langle W, E^1, E^2 \rangle$ is *shift-updating* iff $xE^iy \Rightarrow E^2_y \subseteq E^1_y$.
- ***Sp*:** Sp iff you should be Sure of p, iff $P(p) = 1$.
- **Transitive:** A dynamic frame $\langle W, E^1, E^2 \rangle$ is *transitive* iff $(xE^iy \wedge yE^iz) \Rightarrow xE^iz$. (For a frame $\langle W, E \rangle$, omit the superscripts.)
- **Trust:** $P_q(p|P_q(p) \geq t) \geq t$
- **Validates:** A frame *validates* a principle iff the principle is true at all worlds for all well-defined instantiations of its free variables.
- **Value:** $\mathbb{E}^i[U(d^k)] \geq \mathbb{E}^i[U(o)] \qquad (o \in O, k \geq i)$

Chapter 3

Rational Polarization

Abstract

Groups of people are disposed to divide into subgroups that *polarize* on a variety of topics: individuals in the same subgroup tend to converge in opinions, while individuals in different subgroups tend to diverge in opinions. This widely-confirmed empirical tendency is standardly taken to be a hallmark of human irrationality. It need not be. I'll first show that rational, predictable polarization is possible: whenever you face *ambiguous evidence*—evidence that you should be unsure how to react to—predictable polarization can be fully epistemically rational. This claim can be proven in a general Bayesian framework, as well as illustrated with a simple demonstration. I'll then argue, further, that this abstract possibility may play a role in the actual polarization we observe. One core contributor to predictable polarization is *confirmation bias*: roughly, the tendency for people to seek and interpret evidence in a way that is partial to their prior beliefs. And I'll argue that—given common structures of evidential ambiguity—rational agents who care only about the truth should sometimes exhibit confirmation bias.

Introduction

A sociological observation:

> Divide-and-Diverge: Groups of people have a predictable tendency to divide into subgroups that then *polarize* on a variety of topics:
>
> – Individuals in the same subgroup tend to converge in opinions; and
> – Individuals in different subgroups tend to diverge in opinions.

Examples are everywhere: schools divide into social circles, workplaces divide into factions, religions divide into congregations, academic fields divide into research programs, and so on. In each case, the resulting subgroups tend to diverge—leading to robust patterns of *global diversity* but *local conformity* in opinions and behaviors.

The result is a form of group polarization writ large.[1] Although this phenomenon is pervasive, let's take politics as our focal point.

The standard story for why DIVIDE-AND-DIVERGE happens is a thoroughly *irrationalist* story. It has two parts.

> Part One. The news has come in from social psychology that people display a number of irrational *reasoning biases* that lead them to predictably strengthen their prior beliefs when presented with ambiguous or conflicting evidence on a given question. The buzz-words: confirmation bias, motivated reasoning, biased assimilation of evidence, the backfire effect, etc.[2]

These psychological tendencies—to the extent the empirical results can be trusted[3]—have been with us since the beginning. So why has (e.g. political) polarization gotten so much *worse* in the past decades (Dimock et al. 2014)? That's where part two comes in:

> Part Two. These reasoning biases render people susceptible to the "informational traps" of the modern internet—echo chambers, filter bubbles, fake news, Daily Me's. The result? Groups of pig-headed people all paying attention primarily to information that confirms their prior beliefs, irrationally becoming increasingly confident that they are right and rational and that the other side is wrong and irrational.

You have probably heard some version of this standard story—it is hard to avoid, these days.[4]

I'm interested in pushing back against this story. In particular, I want to contest the claim that the driving force behind this process—the "biases" of Part One—are *biases* in the pejorative sense. What I want to explore is the extent to which these psychological tendencies could arise from epistemically rational processes. In particular, I think that predictable, epistemically rational polarization is possible, and—further—that rational mechanisms plausible help drive DIVIDE-AND-DIVERGE.

That, at least, is the big picture goal. Why do I want it to succeed—and why should you? Two reasons. First, because we want to be able to hold onto our political, religious, and social beliefs. And yet, if we come to agree with the standard irrationalist story, it is difficult to see how we can do so. After all, the "akratic" state of believing (say) *tax cuts are not good for the economy—but I formed that*

[1] For some analyses of the empirical phenomenon, see Bikhchandani et al. (1992); Mcpherson et al. (2001); Sunstein (2009, 2017); Easley and Kleinberg (2010); Dimock et al. (2014); Scott (2017).

[2] For summaries of much of the empirical work on these issues, see Myers and Lamm (1976); Lord et al. (1979); Kahneman et al. (1982); Isenberg (1986); Kunda (1990); Nickerson (1998); Fine (2005); Nyhan and Reifler (2010); Lazer et al. (2018); Mandelbaum (2018); Pennycook and Rand (2019) .

[3] See Gelman and Tuerlinckx (2000); Engber (2018); Guess et al. (2018) for reasons for skepticism.

[4] See Sunstein (2000, 2002, 2009, 2017); Jeffrey (2017); Nguyen (2018); Robson (2018); Lazer et al. (2018); Vosoughi et al. (2018); Piore (2018); Pennycook and Rand (2019) for examples.

belief through irrational mechanisms seems like a paradigm case of irrationality.[5] So we cannot reasonably both hold onto our beliefs and think that they were formed systematically irrationally. Thus if we do hold onto our beliefs, and we also buy into the standard irrationalist story, then we have to believe that this was caused by the "other side" being systematically irrational. My second motivation—and one that I hope you will share as well—is that this seems like a bad conclusion. It means that when we look across the political divide, we have to think that the other side is not only wrong, but also *dumb*. And *that* process of demonization—perhaps more than anything else—is the problem with political polarization.

What if it need not be so? What if we could think the other side is wrong, but not think they are dumb? What if a combination of rational and reasonable processes could take center stage in explaining DIVIDE-AND-DIVERGE? That is the possibility I want to try to make good on in this project, and to make a start on in writing this paper.

I do so with a bit of trepidation—for this rabbit hole goes deep, and I am far from having reached the bottom. In many ways, this paper is not ready. But I write it now—in an unfortunate thesis-submission-rush—in the hope of getting clearer on what it will take to make it so.

Here is the core reason that the paper is not ready. The polarization we see in politics and society at large has three features. (1) It is *predictable*, meaning that sensible people can see it coming, even as it starts to happen to them. (For example I could have predicted that going to college would make me more liberal, and that staying in small-town Missouri would make my childhood friends more conservative.) (2) It is *persistent*, meaning that the disagreements do not wash away once we learn that others disagree with us. And (3) it is *massive*, meaning that people can get seemingly arbitrarily confident in their (disagreeing) positions. This constellation of features is hard to make good on, rationally speaking.

Here is as far as I've gotten down the rabbit hole. Fully rational polarization can be predictable (§3.2). In fact, plausible general mechanisms of the way people do and should process information can lead to it (§3.3). Moreover, such polarization can be both predictable *and persistent*. We will see how in §3.4. But I cannot yet make out how the polarization could be predictable, persistent, and *massive* without some sort of irrationality or information-loss along the way. So what I will try to suggest is that the amount of information-loss we need in order to get this to work is surprisingly minimal and seemingly innocuous. If that's right, then DIVIDE-AND-DIVERGE could result from surprisingly rational mechanisms.

That is the gambit, anyways. But I fully admit that I do not yet fully understand the upshot of the rational modeling I am going to be doing here. Perhaps you will be able to help me do so.

[5] See the literature on "epistemic akrasia", e.g. Greco (2014b); Horowitz (2014); Dorst (2019a).

The plan: we begin with predictable polarization, forgetting about persistence and massiveness. §3.1 offers a simple example of how this could happen, and some empirical data supporting it. §3.2 offers a diagnosis, suggesting that what I'll call *ambiguous evidence* is the key to predictable polarization.[6] §3.3 argues that this picture shows that what is often called *confirmation bias* can in fact be fully rational. We will then turn to the question of whether the same models could help explain persistent (§3.4) and massive (§3.4.1) polarization—with limited success.

3.1 The Illustration

I am going to polarize you, my rational readers. Here's how. I'm about to flip a fair coin—it has a 50-50 chance of landing heads (H) or tails (T). I will divide you into two groups—the Headsers and the Tailsers—and the two groups will get different evidence. What's special about this evidence is the following: once I tell you what sort of evidence each group will be receiving, everyone—including you yourselves—can predict that the Headsers will end up on average more confident of H than the Tailsers will.

It's important to be clear on what that means. It does not mean that the Headsers will necessarily end up more confident, on average, in H than in T—that depends on how the coin lands. What it means is this. Everyone is currently 50-50 on whether H will be true. Yet we can now predict that wherever the average Tailser's confidence in H ends up upon receiving their evidence—be it 0.2 or 0.6—the average Headser's confidence will end up higher—be it 0.4 or 0.8. That means the evidence will result in predictable, rational divergence of opinions. (This will be made more precise in the next section.)

Now to divide you into groups. Consider your first name. It if contains an odd number of vowels, you are a Headser; if it contains an even number of vowels, you are a Tailser. ('Y' counts as a vowel.) Welcome to your team.

As I said, Headsers and Tailser will get different evidence. That evidence will come in the form of a *word completion task*. This is a task in which you are given a string of letters that contains some blanks, and in which you have (say) five seconds to try to answer the question of whether there is an English word that completes the string. I'll use Scrabble rules (no contractions or proper names); moreover, I won't use any fancy words—every word I use will be one that you are familiar with.

For example, you might see a string like this:

P _ A _ E T

[6] Psst—for those of you who have read the rest of this dissertation: ambiguous evidence is evidence that warrants being modest.

And the answer is... yes, there is an English word that completes that string. (Hint: What is Mercury?)

Or you might see a string like this:

_ E _ R T

And the answer is... yes, there is an English word that completes that string. (Hint: What is in your chest?)

Or you might see a string like this:

P _ G _ E R

And the answer is... no, there is no English word that completes that string.

That is the sort of evidence you will receive. Here is the rule for how to use it to determine whether the coin I'm about to flip landed heads or tails.

Rule:

> Headsers will see a completable string iff H.
> Tailsers will see a completable string iff T

Given this rule, what you should do is as follows. When I tell you, you will flip to the correct page number corresponding to your group (Headser or Tailser), and give yourself five seconds to look at your string. (I advise setting a timer on your phone.) Then write down your confidence (between 0 and 1) that there is a completion to the string you've examined. (If you're sure there's a completion, write '1'; if you're sure there's not, write '0'; if you're 50-50, write '0.5'; if you're somewhat confident there is, write '0.7', and so on.)

If you are a Headser, you're done: your confidence that there is a completion is equal to your confidence that H, since you know that one is true iff the other is.

If you are a Tailser, you know that there is a completion iff T, i.e. iff $\neg H$. Therefore to obtain your confidence in H you need to to take 1 minus your confidence that there is a completion. (If you wrote '1' for your confidence that there is a completion, write '0' $(= 1 - 1)$ for your confidence in H; if you wrote '0.7' for your confidence that there is a completion, write '0.3' $(= 1 - 0.7)$ for your confidence in H; and so on.)

Enough setup. You have been divided into groups, but everyone is still 50-50 on H. Now I'll flip the coin (...done), and you will diverge. If you are a Headser, please flip to page 95 (Figure 3-1) to see your string. If you are a Tailser, please flip to page 97 (Figure 3-2) to see your string.

...

Welcome back. What just happened? I can predict the rough outline of it. On average, those of you who are Headsers have ended up fairly confident in H; and those

of you who are Tailsers have ended up somewhat confident of H—but less so than the Headsers. Why? Well, the coin landed heads, and I have done this with a few dozen people; here are the averages of people's reported confidence in H after being presented with their word string:

Confidence in H (0–1 scale):

	Headsers	Tailsers
Overall:	0.59	0.39
H cases:	0.86	0.64
T cases:	0.34	0.16

Of course, everyone starts out 0.5 confident in H, and the coin lands heads about half the time. But the average Headser confidence in H after looking at their string is a bit higher (0.59), while the average Tailser confidence in H after looking at *their* string is a bit lower (0.39). We can see what's going on by separating the cases where the coin lands heads from those where it lands tails. In the H cases, Headsers are on average quite confident of H (around 0.86), while Tailsers are only somewhat confident of H (around 0.64). Meanwhile, in the T cases Headsers are on average only somewhat doubtful of H (around 0.34), while Tailsers are quite doubtful (0.16).

At a first pass, what's going on is straightforward. It is easier to recognize that there *is* a word that completes the string than to recognize that there's *no* word that completes the string. (To recognize that there is a completion, all you have to do is find one; to recognize that there's no completion, you have to rule out the possibility that any of the words you have not yet considered are completions. We'll come back to this in the next section.) Therefore Headsers are good at recognizing H cases (when they see a completable string) and bad at recognizing T cases (when they see an uncompletable string). Meanwhile, Tailsers are bad at recognizing H cases (when they see an uncompletable string) and good at recognizing T cases (when they see a completable string). Hence, on average, their credences diverge.

Predictably so, I say. And *rationally* so, I say. The next section explains why.

3.2 Predictable Polarization

What do I mean when I say that evidence is predictably polarizing? I don't mean that you can be sure that the evidence should lead you to raise your confidence in a given claim. That, I think, is impossible: if you can know beforehand that you'll be rational to raise your confidence, you should *now* raise your confidence.[7]

[7]This is a form of a "conglomerability" constraint—the current rational credence is spanned by the possibly future rational credences (Easwaran 2013). It follows from both the "Trust" and "Value"

_ E _ R N

Figure 3-1: Headser String

I also don't mean simply that it is *likely* that the evidence should lead you to raise your confidence. That is trivial. For instance, suppose I give you one of ten tickets in a fair lottery. You should have credence 0.9 that you will lose. I'm about to tell you whether you won or lost. So you should have 0.9 credence that the evidence you're about to receive should lead your credence to go up from 0.9 to 1 in the claim that you lost. So it's likely that the evidence should lead you to raise your confidence. Nevertheless, this is not a case of predictable polarization, since there's a small chance—0.1—that you won, in which case the evidence you're about to receive should make you drop your credence to 0 that you lost. Because of this, your prior rational *estimate* of the future rational credence is equal to your prior rational credence. More precisely, rational estimates of quantities are averages of the various possible values they could take, with weights determined by how confident you should be in each. Since you should beforehand be 0.9 confident that the rational credence will go up to 1, and 0.1 confident that it will go down to 0, you should now have an estimate for your future rational credence that equals $0.9 \cdot 1 + 0.1 \cdot 0 = 0.9$, i.e. that is equal to your current rational credence.

Evidence is predictably polarizing, I say, when *that* connection fails—when the current rational estimate of the future rational credence differs from the prior rational credence. Consider the rational credences—the ones warranted by your evidence—at two times, 1 and 2.[8] Let 'P^1' and 'P^2' be definite descriptions for those functions, whatever they are. (Note: these are the credence functions you *should* have; they say nothing about what credences you *do* have at the various times; more on that below.) Let 1 be the prior time and 2 be the later time after you've looked at your word string. Then $P^1(H)$ is the prior rational credence in H, $P^2(H)$ is the future rational credence in H, and $E^1[P^2(H)]$ is the prior estimate of the future rational credence in H, defined as the mathematical expectation relative to P^1. This is a weighted average of the various possible values $P^2(H)$ might take, with weights determined by how likely P^1 takes them to be. Thus for any numbers i, k and proposition q: $\mathbb{E}^{\boldsymbol{i}}[\boldsymbol{P^k(q)}] := \sum_t P^i(P^k(q) = t) \cdot t$. Thus in our lottery case, $P^1(P^2(\mathit{Lost}) = 1) = 0.9$,

principles defended in Dorst (2019a).

[8] I'm going to use "evidence" in a way that obeys the uniqueness thesis White (2005), but nothing hinges on this: if you think there are other factors besides your evidence that determine what degrees of confidence you should have (Schoenfield 2014), then please assume that those factors are also bundled into how I use the term "evidence". I'll also assume that rational degrees of confidence can be modeled with a unique, precise probability function. This is a modeling assumption—though there are reasons to worry about it in full generality (Joyce 2010), it is a harmless simplification for our purposes.

and $P^1(P^2(\mathit{Lost}) = 0) = 0.1$, so $\mathbb{E}^1[P^2(\mathit{Lost})] = 0.9 = P^1(\mathit{Lost})$.

Here's the crucial definition. Say that the future evidence is *predictably polarizing* on q iff either now or later you should have an estimate for the future rational credence in q that differs from your rational credence in q. Precisely:

Def. The evidence at time 2 is **predictably polarizing** on q iff $\mathbb{E}^i[P^2(q)] \neq P^i(q)$ for either $i = 1$ or $i = 2$.

Thus you evidence at time 2 is predictably polarizing iff either at time 1 you should have an estimate for the time-2 rational credence that differs from the prior credence you should have, or at time 2 you should have an estimate for the time-2 rational credence that differs from the time 2 credence you should have. So either before getting the evidence, or immediately upon getting it, if you are rational you will estimate that the time-2 evidence warrants having a credence above (or below) your actual credence.[9]

When I claim that the word completion task provides an example of predictable, rational polarization, this is what I mean. In particular everyone is rational to have a prior credence of 0.5 in H. But Headsers are rational to have an estimate that's above 0.5 for the future credence they should have in H, while Tailsers are rational to have an estimate that's below 0.5 for the future credence they should have. In fact, both should have these estimates for each other as well. Using subscripts h and t to indicate whether the rational credence or estimate under consideration is that of a Headser or a Tailer, we have: $\mathbb{E}^1_h[P^2_h(H)] = \mathbb{E}^1_t[P^2_h(H)] > 0.5$ and also $\mathbb{E}^1_h[P^2_t(H)] = \mathbb{E}^1_t[P^2_t(H)] < 0.5$

That's the claim. How could it be right? I take inspiration from Salow (2018), who shows that under certain modeling assumptions, externalist theories of evidence can allow for (in my terminology) predictably polarizing evidence.[10] We can extend those results further. Say that your evidence is *ambiguous* iff it makes it rational to be unsure of the rational response to your evidence. Ambiguous evidence is evidence

[9]Why define predictable polarization in terms of the prior rational credences P^1, instead of the prior *chances*? Both coincide in our word completion task (since you know the chances). In order for someone else to predict that you will polarize, all we need is for the chances to estimate that you will shift your (rational) credence. But in order for *you yourself* to be able to predict it, it needs to be rational for your estimate of the future rational credence to differ from your prior rational credence. I think it is important part of the phenomenon that we ourselves can predict it. I also suspect that in order to get *persistent* disagreement, in which people remain polarized upon learning of each other's opinions, they need to be able to predict it themselves. (But haven't yet made good on this suspicion.)

[10] He tries to use this as a *reductio* of such externalist theories, but I think things are more subtle than he makes them out to be. There turns out to be a crucial difference amongst theories that allow for predictably polarizing evidence. Some—like the ones he considers—allow for cases where you know that predictably polarizing evidence will *mislead* you. Others—like the one I will make use of here—do not. In particular, every model I write down will validate the *value of evidence* in that no matter what decision problem you face, you should always prefer to gather and use the evidence to help you make your decision. The background theory I will use here was developed in Dorst (2019a). More on this below.

T R _ P _ R

Figure 3-2: Tailser String

that does not wear its verdicts on its sleeve: it warrants having some opinion, but warrants being unsure whether it warrants having that opinion. Precisely:

Def. Evidence i is **ambiguous** iff there is a proposition q and number t such that $P^i(q) = t$ but $P^i(P^i(q) = t) < 1$.

We can then prove[11] that whenever it's possible to receive ambiguous evidence, it is possible to receive predictably polarizing evidence:

Fact 3.2.1. *If* $P^1([P^2(p) = t] \wedge [P^2(P^2(p) = t) < 1]) > 0$ *is true at some world in a dynamic probability frame*[12]*, then there is a proposition* q *such that* $\mathbb{E}^i[P^2(q)] \neq P^i(q)$ *is true at some world for* $i = 1$ *or* $i = 2$*.*

Conversely, assuming that rational agents update by conditionalization[13], we can prove that ambiguous evidence is also *necessary* in order to have predictably polarizing evidence:

Fact 3.2.2. *If a probability frame is such that for all* w *there is a proposition* q *such that* $P^2_w = P^1_w(\cdot|q)$*, then if there is a proposition* q *such that* $\mathbb{E}^i[P^2(q)] \neq P^i(q)$ *is true at some world, then* $[P^i(p) = t] \wedge [P^i(P^i(p) = t) < 1]$ *is true at some world for some* p, t, i*.*

Combined, Facts 3.2.1 and 3.2.2 show that—in abstract Bayesian models—ambiguous evidence is the key to predictable, rational polarization. To see its significance, return to our word completion task.

In that task you are presented with a string of letters and some blanks that may or may not be completable by an English word. The crucial feature about this task is that one of these kinds of evidence is easier to assess than the other. In particular, it is

[11] Unfortunately, I do not have the time to write out the proofs for this draft of the paper—most notably, because Fact 3.3.2 has been proved computationally, using Mathematica, and I do not yet know how to write an analytical proof. But trust me, the results stated in this paper are all true.

[12] A **dynamic probability frame** is a triple $\langle W, \mathcal{P}^1, \mathcal{P}^2 \rangle$ where W is a finite set of worlds and $\mathcal{P}^1$ and $\mathcal{P}^2$ are functions from worlds w to probability functions $\mathcal{P}^1_w$ and $\mathcal{P}^2_w$ over W. Propositions are sets of worlds; q is true at w iff $w \in q$; propositions about probabilities and estimates are defined in the obvious way using $\mathcal{P}^1$ and $\mathcal{P}^2$, e.g. $[P^i(q) = t] := \{w | \mathcal{P}^i_w(q) = t\}$ and $[\mathbb{E}^i[P^k(q)] = t] := \{w | \mathbb{E}^i_w[P^k(q)] = t\}$, where $\mathbb{E}^i_w[X]$ is the expectation of X relative to $\mathcal{P}^i_w$. See Dorst (2019a,b) for exposition and application of this formalism, as well as Gaifman (1988); Samet (1997); Williamson (2000, 2014, 2018); Lasonen-Aarnio (2013); Salow (2017, 2018) for precedents.

[13] In fact, other assumptions will do. So long as the rational update between times 1 and 2 satisfies the *value of evidence* (mentioned in footnote 10), evidential ambiguity is necessary for predictably polarizing evidence.

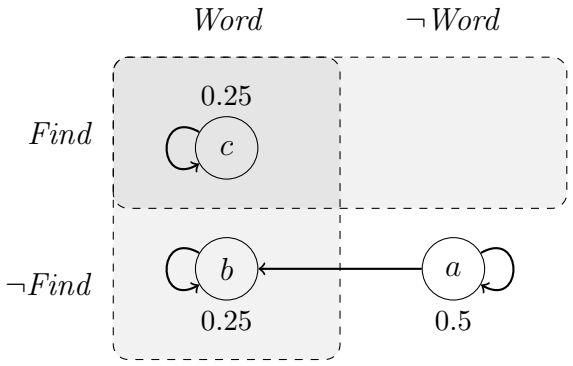
easier to recognize that there *is* a completion to a string than to recognize that there's *no* completion. To recognize that there is a completion all you have to do is find one; to recognize that there's no completion you somehow have to rule out the possibility that any of the strings you have not yet considered are completions. Recognizing that there is a completion is an existential task: find an instance. Recognizing that there's no completion is a universal task: show that there is no instance. In general, existential tasks are easier than universal tasks. (Compare: finding a proof vs. showing that there is no proof.)

I claim that this asymmetry in difficulty leads to an epistemic asymmetry. Assume, for simplicity, that you know that the words I will present will be easy enough such that, given 5 seconds to look at them, if there is a completion then you should be able to find it—if you don't, you've made a rational error, failing to use your evidence properly. (There clearly are such strings; consider 'C _ T'.) More precisely, say that a word is *accessible* to you iff given your knowledge, cognitive capabilities, and 5 seconds of attention, you should be able to recognize that it completes the string, i.e. failing to do so would be a failure to properly use your (lexical) evidence. The assumption I'm making is that you know beforehand that if there's a word, it'll be an accessible word.[14]

Here's why we get an epistemic asymmetry. Recognizing that there is a word is easier than recognizing that there's no word. So if we scale up the difficulty of our strings incrementally, there will come a point such that it is within your capabilities to find such a word, but it is not within your capabilities to recognize that there is no such word. That is the point at which we get our asymmetry. Precisely: if there *is* a word, then after looking at the string for 5 seconds you should be sure that there is. (That is true by definition of 'accessible word', and my assumption that you know that any words you'll be presented with are accessible.) However, if there's *no* word, then you won't find one—but you should be unsure whether that failure to find one is due to your evidence (namely, there is no word), or due to you (namely, you messed up; you missed it). Moreover, since you know that if there's a word you should be sure that there is, in cases in which you should be unsure whether there is a word it follows that you should be unsure whether *you should be sure* there's a word. That means that if there's no word, you have ambiguous evidence: you should be unsure that there's a word ($P^2(\textit{Word}) < 1$), but you should also be unsure whether you should be unsure that there's a word ($P^2(P^2(\textit{Word}) < 1) < 1$).

Here is a concrete model of the scenario. Either there is a word, or not; and either you find one after 5 seconds (when the timer goes off), or you don't. You can't find a word that's not there, so there are 3 possibilities:

[14]This is simply "for simplicity" because even if you don't know this, we can get the exact same epistemic asymmetry that follows using 'accessible word' in place of 'word'.



c is the possibility where there is a word and you find it; b where there is a word and you don't find it; and a is where there is no word and you don't find it. The arrows represent what's consistent with your evidence in the various possibilities: there is an arrow from x to y iff at world x your evidence (after looking at the string) leaves open that you are at world y.

Formally, that means the diagram should be read as follows. At c your evidence rules out both b and a, so it warrants being sure that you're at c. At b, your evidence rules out c and a, so warrants being sure that you're at b. At a, your evidence warrants leaving open that you're either at a or b. Since at a and b your evidence leaves open different possibilities (at a it leaves open $\{a, b\}$, at b it leaves open $\{b\}$), this means that at a you have *a*mbiguous evidence: your evidence leaves open a, but also leaves open that it doesn't leave open a.

Why is this the right diagram? Consider the possibilities. c is the possibility where there is a word and you find it before your timer goes off. If so, then by the time the timer goes off, you should be sure that you found a word before the timer went off—so you should be sure of *Word* (and so rule out a) and of *Find* (and so rule out b).

b is the possibility where there is a word but you don't find it before the timer goes off. What should you think, at the moment the timer goes off? Well, you should be sure that you didn't find a word in time, so you should rule out c. But there *is* a word—and, since we're assuming that you know that all words are accessible, it follows that you should *know* there's a word after 5 seconds. After all, it's there in your lexicon—'HEART', as the case may be. You know how it's spelled; you should know that it can complete the string—if you don't, you've failed to properly use your evidence. So what *should* you think at world b when the timer goes off? What would be the rational response to your evidence, at this stage? It would be to realize "Although I didn't find it before the timer went off, 'HEART' is a completion." Thus you should conclude that there's a word, and so rule out a.

What about a, where there's no word and you don't find one? You should know that you didn't find one in time, therefore you should rule out c. But you can't infer from the fact that you *didn't* find one to the conclusion that you *shouldn't have* found

one—sometimes you fail to draw conclusions that you should, after all; and in this case you are not in a position to tell that there's no word. So you should leave open that there's a word that you missed—you should leave open b. Of course, you should also leave open that there's no word and that your failure to find one was perfectly rational—so you should leave open a as well.

That, then, is why this is the proper (if simplified) model of your epistemic situation when you do a word completion task. In one possibility (a), it gives rise to ambiguous evidence. Now we turn to why that means your evidence is predictably polarizing.

The numbers next to possibilities represent the prior probabilities of each possibility, i.e. the values assigned to them by P^1 at every relevant world. Thus the prior rational credence that there's a word is $P^1(c) + P^1(b) = 0.25 + 0.25 = 0.5$. (Whether there's a word or not is known to be determined by a coin flip, so these numbers should add up to 0.5. For simplicity, I've divided credence equally between them.) The posterior rational credences are those that update this prior probability by conditioning it on the set of possibilities that remain consistent with the evidence. Thus at c, $P^2(c) = 1$; at b, $P^2(b) = 1$; and at a, $P^2(a) = P^1(a|\{a, b\}) = \frac{0.5}{0.5+0.25} = \frac{2}{3}$, while $P^2(b) = \frac{1}{3}$. In particular, at both possibilities where there is a word (b and c), you should be sure there is a word: if *Word* is true, then $P^2(\textit{Word}) = 1$ is true. But at the possibility where there is no word (a), you should still have some credence that there *is* a word: if $\neg$*Word*, then $P^2(\textit{Word}) = \frac{1}{3}$.

We know the prior rational credence that there is a word is 0.5. Now consider the prior rational estimate of the future rational credence that there's a word. This is a weighted average of the various possibilities: probability that there's a word, times the future rational credence if so, plus probability that there's *not* a word, times the future rational credence if so. That is: probability that there is a word, times 1, plus probability that there's not a word, times something positive. Thus the prior rational estimate of the future rational credence is equal to the prior rational credence that there's a word *plus a bit more*. Precisely, $\mathbb{E}^1[P^2(\textit{Word})] = P^1(\textit{Word}) \cdot 1 + P^1(\neg \textit{Word}) \cdot \frac{1}{3} = \frac{2}{3} > 0.5$.[15]

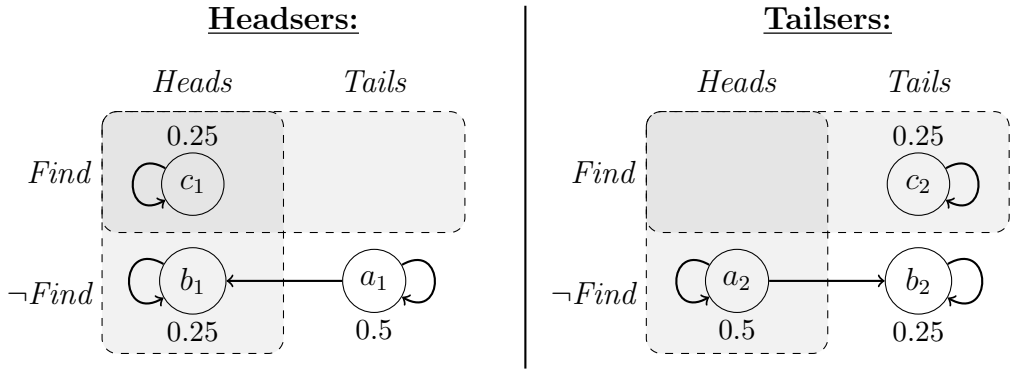
Here's another way to put it. If there is a word, the rational credence that there is will go way up (from 0.5 to 1); if there's not a word, the rational credence will go a *little* bit down (from 0.5 to $\frac{1}{3}$). Since the degree of movement in each direction is not symmetric, but the probability of each option is (0.5 each), these average out to an upward movement.

Upshot: when you face a word completion task, you should estimate the future rational credence that there is a word to be higher than (you know) the prior rational

[15]Note: the particular value of $\frac{2}{3}$ is a consequence of the fact that I set the prior probabilities between b and c evenly, at 0.25-0.25. But so long as b gets any positive prior probability (there is a possibility of failing to find a word that is there), there will be *some* divergence between $\mathbb{E}^1[P^2(\textit{Word})]$ and $P^1(\textit{Word})$—the greater the prior probability of b, the greater the divergence.

credence is—the evidence is predictably polarizing. That is how ambiguous evidence leads to predictable rational polarization.

Now return to Headsers and Tailsers. Headsers know that their string will have a word iff H; Tailsers know that *their* string will have a word iff T. Thus they each face symmetrical, inverted epistemic scenarios:



For the Headsers: if the coin lands heads, they should know it (for they should find a word), and if not, they should be unsure whether it lands heads or tails. For Tailsers: if the coin lands *tails* they should know it (for *they* should find a word), and if not, they should be unsure. What this means is that everyone should start out with credence 0.5 in H. But Headsers should be better at recognizing H cases; Tailsers should be better at recognizing T cases; and therefore everyone can predict that the two groups will rationally diverge.

In particular, in this simple model everyone should estimate the Headser's future rational credence in H to be $\frac{2}{3}$, and estimate the Tailser's future rational credence in H to be $\frac{1}{3}$. Moreover, we can in fact predict with certainty that the rational credence in H for the Headsers will be $\frac{1}{3}$ higher than the rational credence in H for the Tailsers. For if the coin lands heads, Headsers should have credence 1 and Tailsers should have credence $\frac{2}{3}$; and if it lands tails, Headsers should have credence $\frac{1}{3}$ and Tailsers should have credence 0. Since everyone starts out 0.5, everyone can predict that the distance between their rational credences will increase by $\frac{1}{3}$. That is predictable, rational polarization.

Upshot: I can divide you into groups and give you different evidence in a way such that everyone—yourselves included—can predict that you will rationally polarize. So predictable polarization can be rational.

3.2.1 Predictions and Objections

Before moving on, I want to pause to defend both the empirical applicability and the normative credentials of this model of the Headser/Tailser game.

First, empirical applicability. As you will have noticed, this model's predictions for the rational credences are in rough correspondence to the empirical data I've presented

about how people actually, on average, respond to this task. But so far I can't claim to have predicted much, since the model only represents *rational* credences, and the data we are collecting are people's *actual* credences—and the entire point of ambiguous evidence is that sometimes actual credences and rational credences can come apart. So how can we make predictions about actual credences?

Let 'C^1' and 'C^2' be definite descriptions that pick out your actual credence functions at times 1 and 2, whatever they are. Formally, we enrich a dynamic probability frame to get a *dynamic credal-probability frame* $\langle W, \mathcal{P}^1, \mathcal{P}^2, \mathcal{C}^1, \mathcal{C}^2 \rangle$ where $\langle W, \mathcal{P}^1, \mathcal{P}^2 \rangle$ is a dynamic probability frame and each $\mathcal{C}^i$ is a function from worlds w to probability functions $\mathcal{C}^i_w$ representing your actual credences at time i in world w. We then define propositions about C^i using $\mathcal{C}^i$ in the same way as we did for propositions about P^i using $\mathcal{P}^i$.

I think that on plausible ways of enriching our simple world-completion task model with actual credences, the task will also lead to predictable polarization of your *actual* credences, not merely of the rational ones. Importantly, the mechanism is going to have to be different. Predictable polarization of P^i happens when evidence is ambiguous, i.e. when P^i is not introspective: $P^i(q) = t$ but $P^i(P^i(q) = t) < 1$. Plausibly, at least to a first approximation, this is not what's going on with your actual credences: it seems like it is the norm to know (at least roughly) what you *actually* think, but be unsure whether that's what you *should* think—to know what C^i is but be unsure whether it matches P^i. How, then, can we generate actual polarization with introspective actual credences, i.e. such that if $C^i(q) = t$, then $C^i(C^i(q) = t) = 1$?[16]

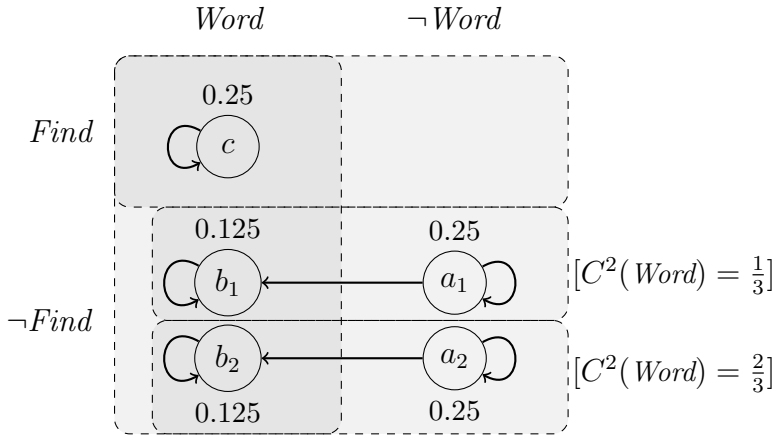
The basic idea is that this will happen precisely because you are unsure what the rational credences are, and therefore you are unsure how you should update your actual credences. In general, so long as your actual credences are *spanned* by the possible rational credences (you know your actual credences will be no more extreme than the rational credences), and you leave open that they might not always be at one of these extremes, then this sort of model will give rise to *actual* predicable polarization.

I am not yet certain of the best way to formulate this, but here is one way. Say that a proposition e is (at time 2) *self-evident* in a dynamic probability frame iff whenever e is true, $C^2(e) = 1$. Say that $\langle W, \mathcal{P}^1, \mathcal{P}^2, \mathcal{C}^1, \mathcal{C}^2 \rangle$ validates *spanning* iff: if you are beforehand certain that the rational credence given some self-evident e is in some range, then you are certain that your actual credence will be in that range. (Formally: for each (time 2) self-evident proposition e, if $C^1(P^2(q) \in [l, h] \,|e) = 1$, then

[16]There's an alternative option. It really is not so plausible that people know exactly what their actual credences are. Surely if you asked them several times what their confidence in *Heads* was, it would fluctuate. So what we're getting in the reports is more likely their best *estimate* of their actual credences. That leaves room for using the same mechanism (an actual-credence analogue of evidential ambiguity) to generate polarization of C^i. I'll set this aside for now.

$C^1(C^2(q) \in [l,h]|e) = 1.$)[17]) Say that $\langle W, \mathcal{P}^1, \mathcal{P}^2, \mathcal{C}^1, \mathcal{C}^2\rangle$ validates *openness* iff: if you are certain that given some self-evident e the rational credence in q is at least l, and leave open that might be higher, then you leave open that if e is true you might have a credence higher than l in q. (Formally: for each self-evident e, if $C^1(P^2(q) \geq l|e) = 1$ and $C^1(P^2(q) > l|e) > 0$, then $C^1(C^2(q) > l \ |e) > 0$.) Assuming that whether you in fact find a word or not (*Find*/$\neg$*Find*) are (time 2) self-evident propositions, and that the dynamic credal-probability frame that enriches our word-completion task validates spanning and openness, it follows that $\mathbb{E}^1[C^2(\textit{Word})] > 0.5$, i.e. that you can rationally estimate your future actual credences to go up.[18]

Here, for example, is a simple model of this that expands our above simple word completion model by breaking possibilities a and b each into two, depending on whether your actual credence in *Word* is $\frac{1}{3}$ or $\frac{2}{3}$, independent of whether *Word* is true within the $\neg$*Find* possibilities.



Though it is hard to specify all the components in the diagram, I'm imagining the following: $C^1 = P^1$ at all worlds, as specified by the numbers by worlds; $\mathcal{C}^2_c(c) = 1$; $\mathcal{C}^2_{a_1} = \mathcal{C}^2_{b_1}$, with both assigning $\frac{1}{3}$ to b_1 and $\frac{2}{3}$ to a_1; $\mathcal{C}^2_{a_2} = \mathcal{C}^2_{b_2}$, with both assigning $\frac{2}{3}$ to b_2 and $\frac{1}{3}$ to a_2. This means that $\mathcal{C}^2$ will be constant across worlds that leave each other open, so that your credences will be introspective (if $C^i(q) = t$, then $C^i(C^i(q) = t) = 1$). Moreover, since P^i_w always assigns positive probability to a subset of $\mathcal{C}^i_w$, it follows moreover that you *should* be sure of what your credences are: if $C^i(q) = t$, then $P^i(C^i(q) = t) = 1$.

In this model the prior rational credence in *Word* is 0.5. The (rational and actual) estimate of the future actual credence in *Word* is 0.625 ($\mathbb{E}^1[C^2(\textit{Word})] = 0.25(1)+$

[17]What is the status of constraints like this? I'm not yet sure. They are not simply rational constraints, since those constraints are imposed on P^i, the rational credences. I am thinking of them as something like "reasonableness" constraints: given what you are actually certain of about the rational credences, here's what a reasonable person will do. But I'm not sure on this point.

[18]*Proof:* Since *Find* $\subseteq [P^2(\textit{Word}) = 1]$, by spanning *Find* $\subseteq [C^2(\textit{Word} = 1]$. Since $\neg$*Find* $\subseteq [P^2(\textit{Word} \geq \frac{1}{3}]$, by spanning $\neg$*Find* $\subseteq [C^2(\textit{Word}) \geq \frac{1}{3}]$. Since $[P^2(\textit{Word}) > \frac{1}{3}]$ is true at $\neg$*Find* $\wedge$ *Word*, by openness there is some possibility in $\neg$*Find* in which $[C^2(\textit{Word}) > \frac{1}{3}]$. It follows that $\mathbb{E}^1[C^2(\textit{Word})] > P^1(\textit{Find}) \cdot 1 + P^1(\neg\textit{Find}) \cdot \frac{1}{3} = \frac{1}{4} + \frac{3}{4} \cdot \frac{1}{3} = 0.5$.

$0.375(\frac{2}{3})+0.375(\frac{1}{3}) = 0.625$.) The estimate of the future rational credence, conditional on there being a word, is 0.75. ($\mathbb{E}^1[C^2(\mathit{Word})|\,\mathit{Word}] = 0.5(1)+0.25(\frac{2}{3})+0.25(\frac{1}{3}) = 0.75$.) And the estimate of the future rational credence, given that there is no word, is 0.5. ($\mathbb{E}^1[C^2(\mathit{Word})|\neg\mathit{Word}] = 0.5(\frac{2}{3}) + 0.5(\frac{1}{3}) = 0.5$.) Those predictions leave something to be desired, in comparison with the empirical data. (In particular, we would want a model where the estimate is lower than 0.5 in the case where there's no word.) But it's a start, and I think it can be tweaked by building in correlations between $C^2(\mathit{Word})$ and *Word* within the $\neg Find$ possibilities. (Which, by the way, are empirically observed: among those who *don't* find a word, those who are looking at a completable string have an average credence of 0.51 that there's a word; those who are looking at an uncompletable string have an average credence of 0.33. Presumably this is because completable strings look more word-like than uncompletable ones—more on this below.) Unfortunately right now I don't have the time to tweak it; hopefully it is suggestive.

Let me move on to some more general considerations and objections about my rational model of the word-completion task. I believe that it is a proof of possibility that some forms of predictable polarization can be *fully* rational—I think there is no sound objection to the structure of evidence exhibited by the models used here. In particular, although it is in general easy to run yourself into puzzles by using intuitions to motivate particular models of ambiguous evidence (Williamson 2014; Lasonen-Aarnio 2015; Salow 2018; Dorst 2019a,b), provably no such puzzles will arise in these models. It is a consequence of Theorem 7.4 of Dorst (2019a) that each of these models validates the *value of evidence* in the following sense: no matter what decision problem you face (set of options and a utility function), the expected value of obtaining the evidence and using the rational credences to decide how to act (maximize expected utility relative to P^2) is always higher than the expected value of simply choosing an option. Given that, I claim that there is no basis on which to say that the update embodied in this frame is not a rational update. Although you can expect the word completion task to lead to be polarizing, that is because you expect it to polarize you *in the direction of truth*. One way to see this is simple: the posterior rational credence function P^2 *strongly accuracy-dominates* the prior rational credence function P^1, in the sense that the credences it assigns are strictly closer to the truth for every proposition in the model. If you *know* that P^2 is more accurate about every question, how could the update be objectionable from a rational point of view?

Nevertheless, I anticipate a certain form of objection, so it is worth addressing it head-on. The basic idea is that since this model allows for failures of an intuitive "Reflection" principle, it will lead to predictable exploitability, and therefore cannot represent a rational update. In particular, in my model in which the evidence is predictably polarizing, there are failures of the following principle:

Rational Reflection: $P^1(q|P^2(q) = t) = t$

Conditional on the future rational credence in q being t, adopt credence t in q.

This principle is violated in my word completion model—in particular, $P^1(\mathit{Word}|P^2(\mathit{Word}) = \frac{1}{3}) = 0$. Though surprising, this is as it should be. Whenever evidence is ambiguous, Reflection must fail (Elga 2013; Dorst 2019a). (This is in effect a consequence of Fact 3.2.1.) For when your evidence is ambiguous, the rational credence function is not certain of what the rational credence function is. Thus if you *learn* something about the rational credence function, you sometimes learn something that it didn't know. In particular, if you learn that $P^2(q) = t$, you may now have more evidence than P^2 had—and therefore it may be rational to adopt a credence other than t. In particular, when $P^2(\mathit{Word}) = \frac{1}{3}$, it is rational to have credence $\frac{1}{3}$ in *Word* only because it is rational to have credence $\frac{1}{3}$ that the rational credence in *Word* is *1*. (If *Word* is true, $P^2(\mathit{Word}) = 1$, after all.) Thus learning that the rational credence in *Word* is $\frac{1}{3}$ tells you that $P^2(\mathit{Word}) \neq 1$, and therefore that *Word* is false.

So Rational Reflection fails. It is well known that, formally speaking, a (diachronic) Dutch Strategy argument can be given against agents who instantiate probability functions that violate principles like Reflection (van Fraassen 1984). That means that if we assume that this model represents a rational update upon being presented with a word-completion task, then rational people can be subject to a sure loss. Some will take this to be a reductio of predictable, rational polarization. It is not, and trades on an equivocation on the term "sure loss." (I take inspiration here from Briggs (2009a); Roush (2016).)

In my model, we have $P^1(\mathit{Word}|P^2(\mathit{Word}) = \frac{1}{3}) = 0$, and also $P^1(P^2(\mathit{Word}) = \frac{1}{3}) = \frac{1}{2}$. Following the recipe from Briggs (2009a), the following sequence of bets is how one would use this information to construct a Dutch Strategy. First, at time 1 you are offered (and rationally accept) Bets 1 and 2 (you should be sure the first one will have no cost, and you should be 50-50 on the second one paying out):

Bet 1:
$-\$1$ if $\mathit{Word} \wedge [P^2(\mathit{Word}) = \frac{1}{3}]$
\$0 otherwise

Bet 2:
$\$\frac{1}{6}$ if $P^2(\mathit{Word}) = \frac{1}{3}$
$-\$\frac{1}{6}$ otherwise

Then at time 2, if (and only if) $P^2(\mathit{Word}) = \frac{1}{3}$, you are offered (and rationally accept) Bet 3:

Bet 3:
$\$\frac{2}{3}$ if *Word*
$-\$\frac{1}{3}$ if $\neg$*Word*

Bet 1 always pays out \$0, since *Word* is never true when $P^2(\mathit{Word}) = \frac{1}{3}$ is, so we can ignore it. If $P^2(\mathit{Word}) \neq \frac{1}{3}$, then Bet 3 is never made, and Bet 2 yields $-\$\frac{1}{6}$. If

$P^2(\textit{Word}) = \frac{1}{3}$, then Bet 2 pays out $\$\frac{1}{6}$ but since *Word* is false, Bet 3 pays out $-\$\frac{1}{3}$, yielding a total of $-\$\frac{1}{6}$. So there is a Dutch Strategy against an agent who updates like this: whatever happens, you will lose $\$\frac{1}{6}$. What to say?

The answer is straightforward. In order to perform this Dutch Strategy, the bookie has to know whether or not $P^2(\textit{Word}) = \frac{1}{3}$, in order to decide whether to offer the second bet. That is cheating. By stipulation, *you* know no such thing, even when it is in fact true. When the rational credence in *Word* is $\frac{1}{3}$, this is so precisely because it is rational to be unsure whether the rational credence in *Word* is $\frac{1}{3}$ or 1. So in order to carry out this Dutch Strategy in a way that leads to a *guaranteed* loss, the bookie must know more than you do. It is like me using the following strategy: if Bob has an even number of pens on his desk, offer you a 1:1 bet that he doesn't; if he doesn't, offer you a 1:1 bet that he does. If I have the chance to peek at Bob's desk beforehand, I can use this strategy to extract money from you. But that is no indictment of your rationality. The lesson here is that a Dutch Bookie cannot exploit facts about P^2, the rational credences, in deciding how to bet, when those are facts that the agent cannot know.

But wait. This reply works well for *Rational* Reflection. But I said above that we could use this model, enriched with facts about your actual credences, to lead to predictable polarization of your *actual* credences, not merely the rational ones. This implies that—even for someone who is in fact rational (rational at the actual world)—we will have a failure of the following Reflection principle:

> **Actual Reflection:** $C^1(q|C^2(q) = t) = t$.
> Conditional on your future actual credence in q being t, adopt credence t in q.

This principle must fail if we are to have expected actual polarization. In particular, in the above model I gave, $C^1(\textit{Word}|C^2(\textit{Word}) = \frac{2}{3}) = \frac{1}{3}$. We can then use Briggs's recipe to generate a sequence of bets that will lead to an agent with such credences receiving a sure loss. And in this case, since the agent is always certain of what her *actual* credences are, a parallel reply to the one above does not work. What to say?

Again, I think the answer is straightforward. (Here I follow Briggs 2009a.) Actual Reflection fails in this case because you do not know that your future credences will be rational. Again, this is essential: if you *did* know that your future credences would be rational, then (so long as you can know what they are), you could know what the rational credences are, and therefore would not have ambiguous evidence. And, in general, whenever you (rationally) have self-doubt—you are unsure whether you will be rational—then you can be exploited. When you have such self-doubt, you will rationally pay for insurance against being irrational. Then, if it turns out you *are* rational, the costs of the insurance are a loss. And if it turns out you are *not* rational, it's no surprise that you can be exploited!

More precisely, consider the bets we would use to generate a Dutch Strategy from this failure of Actual Reflection. At time 1 you will accept:

Bet 1:
$-\$\frac{2}{3}$ if $\mathit{Word} \wedge [C^2(\mathit{Word}) = \frac{2}{3}]$
$\$\frac{1}{3}$ if $\neg \mathit{Word} \wedge [C^2(\mathit{Word}) = \frac{2}{3}]$
\$0 if $C^2(\mathit{Word}) \neq \frac{2}{3}$
Bet 2:
$\$\frac{5}{24}$ if $C^2(\mathit{Word}) = \frac{2}{3}$
$-\$\frac{3}{24}$ if $C^2(\mathit{Word}) \neq \frac{2}{3}$

Then at time 2, if (and only if) $C^2(\mathit{Word}) = \frac{2}{3}$, you are offered (and accept) Bet 3:

Bet 3:
$-\$\frac{1}{3}$ if *Word*
$\$\frac{2}{3}$ if $\neg$ *Word*

This will lead to a sure loss of $-\$\frac{3}{24}$ no matter what. You might then object: "So this must mean that being disposed to update to a credence of $\frac{2}{3}$ in *Word* in this way must be irrational!" And I would reply: *Of course*. The model *entails* that being so disposed is irrational. If you're rational, you will have credence either $\frac{1}{3}$ or 1 in *Word*. So if you have credence $\frac{2}{3}$, you are (and know that you are) irrational. Unfortunately, you don't know whether you should move up to credence 1 or down to credence $\frac{1}{3}$—hence this knowledge provides you with no way out of your predicament.[19]

In sum, it is precisely because you think you might not be rational that you are exploitable. That's no surprise. Notice, in particular, that if (unbeknownst to you) we are only allowed to consider possibilities where you are *in fact* rational, then no Dutch Strategy can be given. After all, $C^1(\mathit{Word}|C^2(\mathit{Word}) = \frac{1}{3}) = \frac{1}{3}$, and likewise $C^1(\mathit{Word}|C^2(\mathit{Word}) = 1) = 1$.

I conclude, then, that this objection from "predictable exploitability" fails. It does not provide any reason to think that this model does not capture a genuinely rational way to update, given your limited powers to process your information.

3.3 Confirmation Bias

Let's now suppose that I've established that predictable polarization could, in principle, be rational. I now want to argue that this mechanism could in fact play a role in what causes real people to predictably polarize.

The first thing to note is that real people polarize without anyone manipulating what sort of evidence they will receive, as I did when I divided you into Headsers and Tailsers. In particular, a wide range of studies suggest that people engage in *confirmation bias*, which in the psychological literature is usually defined as: seeking or interpreting evidence in a way that is partial to your prior beliefs (Nickerson

[19][Note to self: If we require that C^2 obeys the value of evidence with respect to P^2, does that place further constraints on what it can be? I believe this model will satisfy that.]

1998). Though intuitive, that definition is problematic in a variety of ways. However, given the tools developed thus far, we can give a more precise statement of what confirmation bias amounts to.

Say that your strategy for gathering and interpreting evidence between times 1 and 2 is **confirmatory** for q iff it is rational to have an estimate beforehand for the future rational credence in q that is higher than the prior rational credence in q: iff $\mathbb{E}^1[P^2(q)] > P^1(q)$. Confirmation bias can then be defined as a a tendency to prefer strategies of gathering and interpreting evidence that are confirmatory for your prior beliefs: if you believe q, then other things being equal you should favor strategies that are confirmatory for q over those that are confirmatory for $\neg q$.[20]

Confirmation bias—we are told—is a big deal. Here are a couple representative quotes:

> Confirmation bias is perhaps the best known and most widely accepted notion of inferential error to come out of the literature on human reasoning. (Evans 1989, 41)

> If one were to attempt to identify a single problematic aspect of human reasoning that deserves attention above all others, the *confirmation bias* would have to be among the candidates for consideration... it appears to be sufficiently strong and pervasive that one is led to wonder whether the bias, by itself, might account for a significant fraction of the disputes, altercations, and misunderstandings that occur among individuals, groups, and nations. (Nickerson 1998, 175)

I am not going to contest any of the empirical results surrounding confirmation bias. What I'm going to contest is this normative interpretation of them. The claim I'll argue for in this section is that confirmation bias can be fully rational: rational people who care only about the truth will have a tendency to prefer strategies of gathering and interpreting evidence that are confirmatory for their prior beliefs. Therefore, people who rationally start out with opposite beliefs can predictably, rationally, diverge—even when presented with the same available evidence.

How could this be? We should start by getting clearer on the empirical phenomenon. Confirmation bias is often thought to comprise two separate tendencies:

> *Selective exposure:* People prefer exposure to new evidence that they expect to confirm their prior beliefs over that which they expect to disconfirm them.

[20]It is worth pausing here to emphasize something. Fact 3.2.2 entails that under standard Bayesian modeling assumptions (namely: probabilism and conditionalization), the *only* way one could expect that a strategy for gathering evidence is confirmatory is if that evidence is ambiguous. If this is right, then the basic idea behind confirmation bias (that some types of evidence, and ways of gathering and interpreting it, can be expected to confirm your beliefs) requires either a failure of the Bayesian model of evidence, or evidential ambiguity.

Biased assimilation: People tend to give greater weight to evidence that confirms their prior beliefs than to evidence that disconfirms them.

The classic examples of selective exposure are things like this (Frey 1986): after Watergate broke, studies showed that people who supported Nixon knew much less about the details of the scandal than those who opposed him. Explanation(?): one can expect beforehand that reading about a scandal will give you reason to think that Nixon is a bad president; so those who were inclined to think this ate up the reports, while whose who were inclined to deny it avoided them. For another example, consider how well you can predict peoples media habits (which news sites do you check?) based on their political views (Mitchell and Weisel 2014). Conservatives don't tend to watch MSNBC; liberals don't tend to watch Fox. Selective exposure would explain that.

Turn next to biased assimilation. These results are especially striking. The classic studies work like this. Take some people who believe that (say) that tax cuts Benefit the economy (B), and others who believe $\neg B$. Present them with symmetrical, conflicting pieces of information about B—for example, conflicting studies. Perhaps Study 1's abstract says, "We studied countries X,Y,Z over 10 years and observed that after they cut taxes, they had economic booms." And perhaps Study 2's abstract says, "We studied countries A,B,C over 10 years and saw that after they cut taxes they faced economic downturns." Looking only at the headlines, Study 1 supports B and Study 2 supports $\neg B$. Give these two studies to people, and give them time to think about them.

Here is the result you'll get (Lord et al. 1979; Plous 1991; Baron 1995; Kuhn and Lao 1996; Munro and Ditto 1997). On average, those who believed B will take the studies to (on the whole) tell in favor of B. They will recognize that Study 1 told in favor of B while Study 2 told against it, but they will claim that the former was a more convincing study than the latter, and so that on balance the new evidence tilts in favor of B. Meanwhile, those who believed $\neg B$ will do the exact opposite. They will take the studies to (on the whole) tell against B. They will recognize that Study 1 told in favor of B and Study 2 told against it, but *they* will claim that the latter was a more convincing study than the former, and so that on balance the new evidence tilts against B.

You might think—as many have—that there is no way that this could be rational (Lord et al. 1979; Plous 1991; Kuhn and Lao 1996; Munro and Ditto 1997; Fine 2005; Mandelbaum 2018). But I think you'd be wrong.

To make this argument, we need to take a closer look at what subjects actually do. (This has also been empirically documented. Here I follow an insightful paper by Kelly 2008.) Those who have strong prior beliefs are surprised by the study that disconfirms those beliefs, and unsurprised by the study that confirms them. This seems rational. (If you believe tax cuts don't benefit the economy, then a study claiming as much is exactly what you should expect; on the other hand, a study suggesting the reverse is

unexpected.) They have limited time and resources to think about and process these studies. So what they do is they spend more time scrutinizing the surprising study–i.e. the one that disconfirms their prior beliefs. When they are doing so, they are looking at the details of the study to try to find flaws in the arguments, or the data, or the inferences, and—more generally—looking for *alternative explanations* of the results: explanations that fit the study and explain away the data. (E.g. "Of course Study 1 saw economic growth—it was conducted during a global economic boom.") This—spending more time looking at the details of the surprising study—also seems rational. And, intuitively, it seems reasonable that such selective scrutiny *could* lead to predictable polarization. I am going to build a model that makes good on this intuition.

The first key observation (due to Kelly 2008) is this: people need to *process* the information they're presented with in order to know how to react to it. You cannot take in an entire study at once—it takes time an energy to sift through the details. This means that if people make different choices about *how* to sift through the details—about which study to scrutinize—then that will lead them to have different total bodies of evidence. For example, if I scrutinize Study 1 and you scrutinize Study 2, then I'll know all sorts of things about the details of Study 1 that you won't know, and you'll know all sorts of things about the details of Study 2 that I won't know. Hence, in the sense of "evidence" that's relevant to determining what we should think, you and I have different evidence.

The fact that we have to process our information in this way means that biased assimilation can in fact be explained in terms of selective exposure. While it at first seemed as though the subjects of this experiment were given the same evidence (they were presented with the same studies), the fact that they processed it in different ways implies that they end up with different total bodies of new evidence (one group looked closely at Study 1, the other looked closely at Study 2)—and, *that* is what drives the polarization in this case.

Of course, our explanation can't stop here. The fact that people are receiving different total bodies of evidence does not in itself mean that they can rationally expect that evidence to polarize them. In fact, this is exactly where Kelly (2008) draws the line: he concludes that so long as people are *aware* of their tendency to selectively scrutinize disconfirming evidence, they should "correct" for that fact, which (he thinks) will prevent them from predictably polarizing.

This is where Kelly and I part ways. I claim that people *cannot* "correct" for this tendency: fully rational people who are aware of their tendency to selectively scrutinize evidence will still be predictably polarized.

The key observation is this. Searching for an alternative explanation of a study is a form of *cognitive search*: it involves searching an accessible cognitive space to see if it contains an item of a particular profile—in this case, to see if there is an accessible explanation that fits the study and explains away the data.

That should sound familiar: it's a lot like searching your *lexicon* to see if it contains a word that fits a string. And because of this, it will generate evidential ambiguity in the same way that our word search task did.

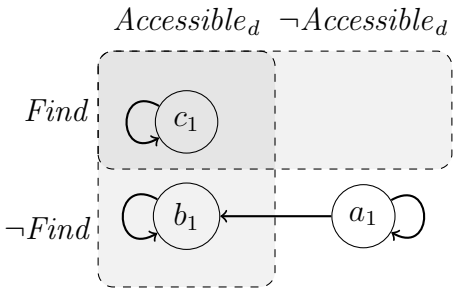
There is one issue, which this time around cannot be idealized away. Just because there is an alternative explanation of a study does not mean it is within your power to come up with it—perhaps it relies on information that you don't have, or is too complicated to formulate in the relevant time. But we can restrict attention. Say that an explanation X of a study is *accessible* to you, given your knowledge and cognitive limitations, iff upon scrutinizing the study you should come up with X—failing to do so would be a failure to properly make use of your evidence.

This yields a familiar epistemic asymmetry. It's easier to recognize that there *is* an accessible explanation than it is to recognize that there's *no* such explanation. To recognize that there is an accessible explanation, all you have to do is find one—it's an existential task. So if there *is* an accessible explanation, then you should be sure there is; you should find it; not doing so would be a mistake. (That is true by the definition of 'accessible'.) But to recognize that there's *no* accessible explanation you somehow have to rule out the possibility that any of the explanations you have not yet considered could fit the study. Thus if there is no accessible then you won't find one—but you should be unsure whether that failure to find one is due to your evidence (namely, there is no explanation), or due to you (namely, you messed up—you missed it). Thus we get our asymmetry: if there is an accessible explanation, you should know that there is; if not, you should be unsure whether there is.

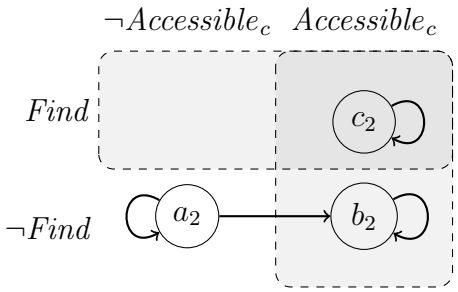
Suppose this is right. Then those who are presented with conflicting studies face a choice between two different cognitive search tasks: they can scrutinize the disconfirming study (the one that disconfirms their prior beliefs—whatever those are) or the confirming one. Let $Accessible_c$ be the claim that there is an accessible explanation of the *c*onfirming study, and let $Accessible_d$ be the claim that there's an accessible explanation of the *d*isconfirming study. Then those presented with conflicting studies face a choice between two cognitive search tasks—a choice between a familiar pair of epistemic situations[21]:

[21]Note: in this model it is important to realize that a_1, b_2, etc. are *sets* of possibilities, since we need to allow that B is true at some of them and false at others.

Scrutinize *Dis*confirming Study:



Scrutinize *Confirming* Study:



Assume that you have some prior rational credence function P^1 over the various possibilities; it is up to you whether you scrutinize the disconfirming study (end up in a possibility on the left) or the confirming one (the right); the arrows represent what will be consistent with your evidence after scrutinizing the study you choose to scrutinize; and you should update your prior rational credence by conditioning it on the set of possibilities that remain consistent with your evidence, whatever it is.

Suppose you scrutinize the disconfirming study. Then if there is an accessible explanation of it ($Accessible_d$), you should know it—and if not, you should be unsure whether there is one. (Just like: if you are a Headser and the coin lands heads, you should know it—and if not, you should be unsure whether it did.) Suppose you scrutinize the confirming study. Then if there is an accessible explanation of *it* ($Accessible_c$), you should know it—and if not, you should be unsure whether there is.

In order to prove that confirmation bias can be rational in this setting, we need to make some further assumptions about the structure of the model. Without loss of generality, suppose you start out believing B. (If we flip that, we should simply replace 'B' with '$\neg B$' in the below assumptions.)

First assumption: learning that there is an accessible explanation of the disconfirming study should raise your credence in B, since it explains away evidence against it. Moreover, this is true independently of whether you find such an explanation or not.

(A1) $P^1(B|Accessible_d) > P^1(B)$, and
$P^1(B|Accessible_d) = P^1(B|Accessible_d \wedge Find)$

(A bit more on the dynamics here. You start out with some prior rational credences. Then you get an initial volley of evidence—knowledge of the headlines and abstracts of each of these studies. The disconfirming headline should push your credence down, but the confirming one should push it up, therefore these forces should (roughly) cancel out. Then you are left with (roughly) your initial credence in B. This is the stage that the time 1 model is capturing, with P^1 the rational credences then. If you should then, at time 2, explain away the disconfirming study, that should remove the initial force pushing your credence down—and so, having only the confirming study remaining, your credence should rise.)

Second assumption: your choice of which study to scrutinize is not itself evidence for or against B.

(A2) $P^1(B|\textit{Scrutinize Disconfirming}) = P^1(B|\textit{Scrutinize Confirming})$

Given just these two assumptions, it follows that scrutinizing the disconfirming study is a confirmatory strategy on B:

Fact 3.3.1. *Given (A1) and (A2),* $\mathbb{E}^1[P^2(B)|\textit{Scrutinize Disconfirming}] > P^1(B)$.[22]

This is true for the exact same reason as in our word search. If there is an accessible explanation of the disconfirming study, your credence should go substantially up; if there is no accessible explanation, it should go down—but not as much, since you should still leave open that there *is* an accessible explanation that you missed. Thus, on average, the rational credence moves up upon scrutinizing the disconfirming study.

But is it *rational* to scrutinize the disconfirming study, if what you want is to get to the truth? Given some further assumptions, yes. Two of them are straightforward symmetry assumptions, one is more substantive.

Third assumption: if there is an accessible explanation, you're equally likely to find it either way; you're not better at finding accessible explanations of disconfirming studies than of confirming ones, or vice versa.[23]

(A3) $P^1(\textit{Find}|\textit{Accessible}_d) = P^1(\textit{Find}|\textit{Accessible}_c)$[24]

Fourth assumption: Learning of an explanation for either study should shift your credence by the same degree, and that degree is independent of whether you in fact find the study. (This assumption subsumes (A1).)

(A4) There is an $x \in (0, 1)$ such that:
$P^1(B|\textit{Accessible}_d) = P^1(B) + x$, and
$P^1(B|\textit{Accessible}_c) = P^1(B) - x$.
Moreover,
$P^1(B|\textit{Find} \wedge \textit{Accessible}_d) = P^1(B|\textit{Accesible}_d)$, and
$P^1(B|\textit{Find} \wedge \textit{Accessible}_c) = P^1(B|\textit{Accesible}_c)$

The final assumption is the crucial one, and is what drives the result. It says that beforehand you should think it more likely that the *dis*confirming study has an accessible explanation than that the confirming study does. The idea here is simple. You are presented with two conflicting studies: one telling in favor of your prior

[22] And of course from (A2) we get the same result with your conditional prior credence: $\mathbb{E}^1[P^2(B)|\textit{Scrutinize Disconfirming}] > P^1(B|\textit{Scrutinize Disconfirming})$.

[23] Insofar as this assumption breaks down, it seems plausible that you should expect yourself to be better at finding the accessible explanation of the disconfirming study, since you'll have reasons and hence knowledge about why your belief is true and how it can be defended. This will only push the dynamics more in favor of scrutinizing the disconfirming study.

[24] [Is this needed, given independence assumptions?]

(rational) belief, one telling against it. Supposing that one of the studies contains an error, which one do you expect it to be? Obviously the one that *dis*confirms your prior beliefs. After all, if your beliefs are true then it is the results of the disconfirming study that require a special explanation; so when other things are equal, to the extent that you think it likely that your belief is true, you should also think it likely that there is an alternative explanation of the disconfirming study—and, therefore, that there is an *accessible* such explanation. Formally, the assumption is[25]:

(A5) $P^1(\mathit{Accessible}_d|\mathit{Scrutinize\ Disconfirming}) > P^1(\mathit{Accessible}_c|\mathit{Scrutinize\ Confirming})$

Given assumptions (A2) – (A5), we can show that confirmation bias is rational in this case. In particular, it *maximizes expected accuracy* of the posterior rational credence in B, $P^2(B)$, to scrutinize the disconfirming study. To prove this formally I will measure accuracy using the most standard measure—the Brier score, A, which is the inverse of the squared distance between $P^2(B)$ and the truth-value of B.[26] Precisely:

Fact 3.3.2. *Given (A2) – (A5):*
$\mathbb{E}^1[A(P^2(B))|\mathit{Scrutinize\ Disconfirming}] > \mathbb{E}^1[A(P^2(B))|\mathit{Scrutinize\ Confirming}]$.[27]

The intuition behind this result is relatively straightforward. If there is an accessible explanation of whichever study you scrutinize, then you should be sure that there is; but if there is no such explanation, then you should be unsure whether there is. Since the rational credence function P^2 is more opinionated in the former case than the latter one, it is more expectedly accurate in the former. Thus, overall, it you are trying to get evidence that will warrant having the most accurate credence function, then you should try to avoid the ambiguous evidence that comes from trying-and-failing to find an alternative explanation. Therefore, you should look for an explanation where you should expect to find one—i.e. the disconfirming study.

Here's a way to get the intuition. Take our Headser/Tailser game, but modify it slightly. If the coin lands heads, I'll definitely show Headsers a completable string and I'll definitely show Tailsers an uncompletable one. But if the coin lands tails, I'll definitely show Headsers an uncompletable string and I'll *probably* show Tailsers a completable one—but I might show them an uncompletable one regardless. If what you care about is having an accurate opinion about how the coin landed, should you prefer to be a Headser or a Tailser?

[25][Perhaps it would be better to derive this assumption from something more basic, e.g. that $P^1(\mathit{Accessible}_d)$ is proportional to $P^1(\neg B)$, $P^1(\mathit{Accessible}_c)$ is proportional to $P^1(B)$, and $P^1(B) > 0.5$?]

[26]So $A(P^2(B)) = -(1 - P^2(B))^2$ if B is true, and $= -(0 - P^2(B))^2$ if B is false.

[27]Moreover, this expected accuracy is higher than that of the prior rational credence function: $\mathbb{E}^1[A(P^2(B))|\mathit{Scrutinize\ Disconfirming}] > \mathbb{E}^1[A(P^1(B))|\mathit{Scrutinize\ Disconfirming}))|\mathit{Scrutinize\ Disconfirming}]$, meaning it does not pay to ignore the evidence, despite it's predictably polarizing effect.

Well, if you see a completable string, you should be sure of how the coin landed—and so you should be perfectly accurate. If you see an uncompletable string, you should be unsure how the coin landed, and therefore should have some degree of inaccuracy. Therefore you should prefer to be in the group that is more likely to see a completable string—prefer to be a Headser.

This, then, is how the results from biased assimilation studies can be explained through fully rational mechanisms. What you should do when you see conflicting studies is scrutinize one of them. This is a form of cognitive search, and thus gives rise to ambiguous evidence. If you want to have accurate beliefs, you should prefer to avoid ambiguous evidence, other things being equal. Therefore it is rational to look for an alternative explanation of the study that it is rational to expect to contain one—namely, the one that disconfirms your prior (rational) beliefs. If you do so, this will lead, on average, to a strengthening of your prior beliefs.

Stepping back from biased assimilation in particular, what can we say about confirmation bias more generally? The idea behind this model generalizes, I think, quite a bit. Here's how.

People need to process their information in order to know what to do with it. Many of the most effective ways to do so take the form of a *cognitive search*: searching an accessible cognitive space to see if it contains an item of a particular profile.[28] Think, for example, of looking for a word that fits a string—a word completion task. Or looking for a word that fits a meaning—the "tip of the tongue" phenomenon. Or seeing a face you recognize and trying to recall the person's name. Or hearing someone report a pattern to you ("In situation X, things like Y tend to happen") and trying to recall such an example from your memory to confirm or disconfirm it. Or—maybe close to home—trying to formulate an argument for a fixed conclusion. If you are being presented by an argument for an implausible conclusion, what will you likely do? Search for a counterexample, an implausible consequence, a reductio. That process—in other words, much of the activity of philosophers and academics—is a cognitive search.

In all of these cases: if what you're looking for is there (it's accessible), you should find it; and if not, you should be unsure whether it's there. What Fact 3.3.1 says more generally is that whenever this is so, engaging in a cognitive search leads to predictably polarizing evidence. And what Fact 3.3.2 says is that since (other things being equal) it makes sense to look for things that you expect to find, it is rational to prefer to engage in *confirmatory* cognitive searches—ones which, if successful, lead to a strengthening of prior belief. (If you are surprised by my conclusion that predictable polarization can be rational, the sensible thing to do is to look not for buttressing examples, but for problematic consequences.)

[28] Cognitive scientists and psychologists have in recent years increasingly thought that cognitive search plays a crucial role in cognition; see Todd et al. (2012).

Upshot: when the best strategies for assessing a claim are cognitive search strategies, confirmation bias can be fully rational.

I closing, I want to say just a few words about the empirical plausibility of this explanation for confirmation bias. On the whole, I think it fits fairly well with the empirical picture emerging from the literature. Predictable polarization is not essentially driven by "hot cognition"—for it happens on topics as mundane as how comfortable dental chairs are (Baron et al. 1996). Predictable polarization is not essentially a social process—for it happens when people are thinking about their beliefs on their own (Tesser et al. 1995), when they repeat their opinion to themselves (Downing et al. 1992), and when they merely *expect* to debate someone (Fitzpatrick and Eagly 1981). In fact, there is a large literature suggesting that polarization is the result of simply thinking about a topic (see the citations in Kuhn and Lao 1996, 115). This all fits with a picture on which cognitive search drives the phenomenon in at least many cases. Moreover, studies of overconfidence show that people's confidence is (unsurprisingly) highly correlated with how many reasons they generate in favor of their belief (i.e. successful cognitive searches). The authors found that explicitly telling people to generate reasons for the *denial* of their view significantly reduced overconfidence (Koriat et al. 1980). That is fully predicted by this picture, for doing so will be a confirmatory strategy for the *negation* of your beliefs (that's Fact 3.3.1). And, moreover, will in general be rationally dispreferred (that's Fact 3.3.2)—which is why it takes special instructions to get people to do it. Finally, we are all familiar with individuals who are so smart that they can convince themselves of anything, for (as I would put it) they are so efficient at doing cognitive searches for evidence that confirms their beliefs. Studies confirm this anecdotal evidence (Kahan et al. 2012). In short, I think the mechanisms identified here plausibly play a role in the actual polarization we see, and—if the arguments of this section are correct—rationally so.

Final upshot: when the best strategies for assessing a claim are cognitive search strategies, confirmation can be—and plausibly sometimes *is*—fully rational.

3.4 Open Questions

That is as far down the rabbit hole as I've gotten. As promised, predictable polarization can be fully rational. But what about the other features of Divide-and-Diverge—namely, *persistence* and *massiveness*?

You might think that this will follow from what I've already shown—that the simple models examined here could be iterated without complication (e.g. repeating the Headser/Tailser game over and over) to get predictable, persistent, massive polarization. I did think so. But I was wrong.

Here is a proof. I promised that I would only write down structures of evidence that guarantee the *value of evidence* in the sense that you should always prefer to use

the evidence to guide your decision-making, rather than ignore it. Every model I've written down so far does so. But it is straightforward to show that there is no way to simply iterate these models (nest them inside each other) to obtain vastly more extreme versions of the same results. Suppose, for reductio, that we could do so. As I said, given a 50-50 coin, the expected value of the rational Headser posterior credence in H is $\frac{2}{3}$. If we could iterate this epistemic structure, then if we repeat this for an arbitrarily large number of coins, then (by the law of large numbers) with arbitrarily high probability the rational Headser will end up with an average credence of $\frac{2}{3}$ in the various claims of the form "Coin i landed heads." This is because for the coins that do land heads, the rational Headser will have credence 1 that they did so; and on the coins that land tails, the rational Headser will have a credence of $\frac{1}{3}$ that they land heads. With a roughly 50-50 occurrence rate of heads and tails, those numbers will average to roughly $\frac{2}{3}$.

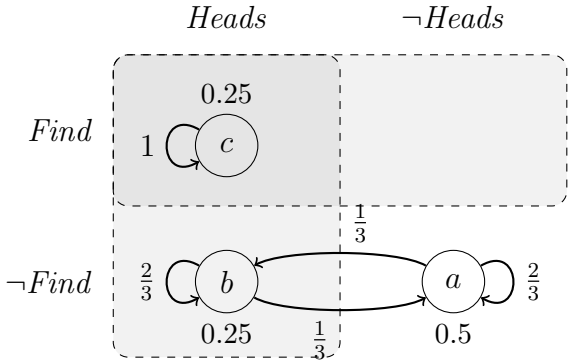
Now, having tossed all these coins, pick one at random—call it R. Ask the Headser what their credence in R landing heads is. Beforehand, like a sensible person, the Headser will be 50-50 confident that R lands heads. But (under our supposition that we can iterate these models) the Headser can herself predict beforehand, with arbitrarily high confidence, that *if* she follows this (putatively) rational pattern, she will end up with roughly $\frac{2}{3}$ credence in R landing heads. This means the value of evidence must fail. For example, suppose we offer the Headser a bet that pays out \$1 if R lands heads and costs $-$\$1.50 if it lands tails. Beforehand she thinks the bet is not worth it. But right now she can predict with near certainty that her future self *will* think the bet is worth it—and so will (ill-advisedly) take the bet. Thus she would rather make the decision now about whether to take this bet (and so decline it), than to wait for all the evidence to come in and then allow that evidence to guide her actions afterward. That means the value of evidence fails.

Thus that we cannot simply iterate the models used here without being careful about how we do so. So what to do? What of the other features of DIVIDE-AND-DIVERGE—persistent and massive polarization?

Persistence is, I think, rather straightforward to generate with a value-of-evidence frame. Go back to our Headser/Tailser case. Now suppose that you don't know that if there is a word, you should be able to find it—perhaps you leave open that it will be too difficult to do so. (And let's ignore talk of "accessible words", for simplicity.) Plausibly, however, completable strings tend to *look* more word-like than uncompletable strings. (As I mentioned above: consider those subjects who didn't find a word; those who were looking at a completable string averaged 0.51 credence that there was a word; those looking at an uncompletable string averaged 0.33.) So on this model of the situation, in the cases in which you don't find a word, you can never be sure that there is one—but if there *is* a word, you should be more confident that there is than you should be if there's not a word. In effect, we have a "graded" epistemic asymmetry: in the non-word case, you should be fairly confident that you

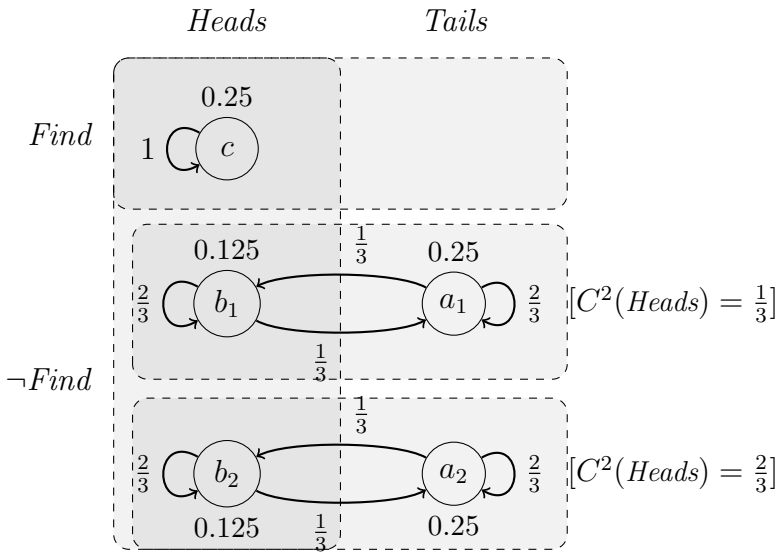
are in the non-word case; in the word case, you should be fairly confident you are in the word-case; but in neither case can you rule out the other.

In the lone Headser case, this leads to a model with this structure:

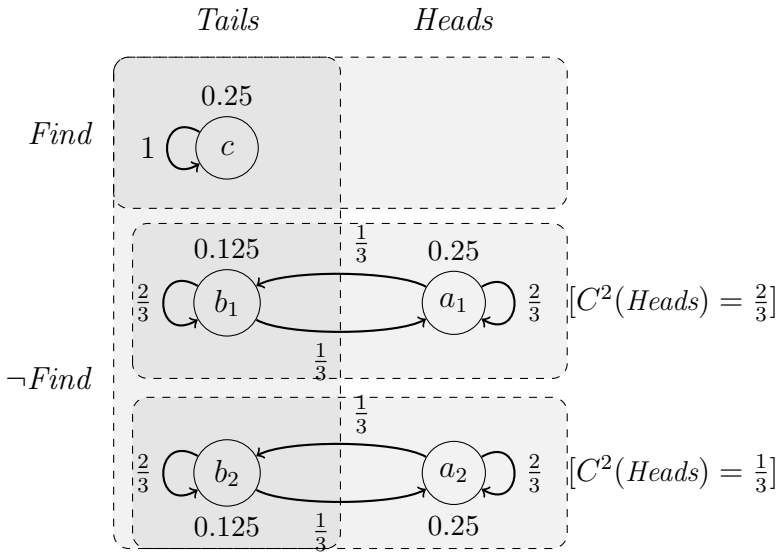


Here the arrows from possibilities represent the posterior rational credence to have in the various possibilities. So the arrow labeled '$\frac{1}{3}$' from a to b indicates that $P^2(b) = \frac{1}{3}$ is true at a. This model validates the value of evidence. (Though notice that it cannot be represented as the posterior rational credence updating from the prior one via conditionalization.)

Now let's add our Headser's *actual* credences to the model. We assume that C^i is introspective; that if they find a word, their actual credence in H is 1, and if they don't then their credence is either $\frac{1}{3}$ or $\frac{2}{3}$, independently of whether there is a word or not. (So when they don't find a word, half the time they are rational, half the time they are not.)



Now if we add a Tailser who's epistemic and doxastic state is inverted, we would have a model where everything is the same except that *Heads* and *Tails* are swapped, and therefore credences in *Heads* are inverted:



Now suppose that our Headser and Tailser share what credences they have with each other after they look at their strings. Since each of their credences is independent of H, this is equivalent to each of them learning whether or not the other one found a word. What happens?

In the case where the Headser finds a word, both should converge to a credence of 1 in H. In the case where the Tailser finds a word, both should converge to a credence of 0 in H. In this model each of these cases has a prior probability of $\frac{1}{4}$, for a total probability of $\frac{1}{2}$ that they should agree.

What about the other cases? Let's restrict attention to the credences they *should* adopt in each one. There's a $\frac{1}{4}$ probability that the coin lands heads but the Headser doesn't find a word (and of course the Tailser doesn't find a word); and there's a $\frac{1}{4}$ probability that the coin lands tails and the Tailser doesn't find a word (and of course the Headser doesn't).

In the former case, this model says that the Headser should initially be $\frac{2}{3}$ confident of H, and then upon learning that the Tailser did not find a word, she should raise her credence to $\frac{3}{4}$. The model says that the Tailser should also initially be $\frac{2}{3}$ confident of H, but when he learns that the Headser didn't find a word, his credence should drop to $\frac{1}{2}$. Why? The Headser should be $\frac{2}{3}$ confident initially because she didn't find a word, but the string looked word-like; learning that the Tailser didn't find a word is evidence that it landed heads, so this should cause her to raise her credence in H. Conversely, the Tailser should be $\frac{2}{3}$ in H because he didn't find a word, and his string didn't look especially word-like; but learning that the Headser didn't find a word is evidence that it landed tails, so this should cause the Tailser to lower her credence in H. Thus in this case, the Headser should end up $\frac{3}{4}$ and the Tailser should end up $\frac{1}{2}$ even after they learn of each other's opinions. We have rational, persistent disagreement.

In the second case, where it lands tails but the Tailser doesn't find a word, parallel reasoning shows that the Headser should wind up $\frac{1}{2}$ in heads while the Tailser should wind up $\frac{1}{4}$.

Thus in this model there is a total of a $\frac{1}{2}$ probability that the Headser and Tailser should have *persistent* disagreement: rational disagreement upon learning of the others' opinion. Since this model validates the value of evidence, we have fully rational, predictable, persistent polarization.

3.4.1 Massive Polarization?

The real problem is *massive* polarization. As we've seen above, there is no way to draw a model that guarantees the value of evidence on which a Headser can predict with great confidence that they will be rational to raise their credence substantially. Formally, this is because simply iterating the Headser/Tailser model in the simple way leads to a violation of the "tree-like" structure that Dorst (2019a) proves is necessary and sufficient for the value of evidence. Instead, if we iterate the model we will need to "split" various possibilities apart—and this will lead to a reduction in the expected degree of polarization at various later stages. Don't worry about the details; the crucial point is that in order to respect the value of evidence, once we have *some* evidential ambiguity on the scene, we need to be careful about how we add *new* layers of it.

What to do, then? Here is an idea. If there were a plausible explanation of why the evidential ambiguity would dissipate after looking at the first word completion task (without shifting your credence), *then* we could iterate the model from that new, unambiguous evidential state. And, plausibly, there is such an explanation. Why is it that 5 minutes after seeing an uncompletable word string, you are not certain that there's not a word? Obviously if you did nothing but stare at it for those 5 minutes and try to think of such completions, you *would* eventually become certain that there's no word. The reason you don't automatically do so in a real case is because you forget the details of what the string looked like, and therefore you are unable to keep doing cognitive searches for a completion. Once you've forgotten the string, you have no basis on which you *could* conclude that there is (or isn't) a completion. In other words, this sort of information loss (forgetting the details of the string you were examining) "flattens" the model of your epistemic situation so that you no longer have ambiguous evidence. In particular, all you know is that you ended up with (say) $\frac{1}{3}$ credence that there was a word, and then forgot what the string looked like. In that case, since you have no basis on which to change that credence, you should now be sure that $\frac{1}{3}$ is the rational credence for you now to have, given your current evidence (even though you leave open that maybe you *should have* found a word, and so become certain that there was a word).

Suppose that this is so: after a short time has passed, you lose the information

you have about the word string, and so at each world you become certain that the credence you *actually* had at time 2 is the one you *should* now have at (the later) time 3. Of course, the transition from time 2 to time 3 gives rise to a failure of the value of evidence: you lose information across that time. But, it seems, this is a rather innocuous such failure—and obviously is what actually happens for agents like us.

If this is what happens, then we *can* "iterate" the model. The picture would go like this. We begin with a cognitive search model (like the ones above) that validates the value of evidence, and represents both the rational and actual credences at times 1 and 2. Then at time 3, the model is "flattened", so that the rational credence at time 3 is known to be equal to the person's actual credence at time 2, and therefore evidential ambiguity disappears. There will be failures of the value of evidence here, in the sense that in some worlds the *rational* credence function P^2 will expect the update from time 2 to time 3 to lead to worse decisions that not updating. But, notably, since the model is being flattened to the person's *actual* credences, they themselves will never expect that the update from time 2 to time 3 will lead them to be worse off than simply using their actual credences to make their decisions. (Since C^2 assigns credence 1 to the claim that $P^3 = C^2$, of course C^2 will not think the update to P^3 will lead to worse choices than simply remaining at C^2.) Thus there is a sense in which even the information-loss steps, where we flatten the model, are not objectionable from the person's own perspective.

And *now*, given that at time 3 the person no longer has ambiguous evidence, we can have another simple cognitive search model representing the the actual and rational credences at times 3 and 4, which again validates the value of evidence. Then *that* model is "flattened", and we rinse and repeat.

This strategy will, I believe, allow for massive polarization. In particular, applied to the Headser/Tailser case, it will permit the naive reasoning used above to lead a Headser to predictably have a credence around 0.6 that a randomly chosen coin landed heads. Or, applying it to the confirmation bias models used in §3.4, if we can keep presenting someone with conflicting studies that would shift their credence in B to the same degree up or down were they to debunk ether one of them, then this iteration process will lead to them becoming arbitrarily confident in B. (At each stage, they should scrutinize the disconfirming evidence. This should sometimes lead them to drop their credence, and other times lead them to raise it; but on average it will lead them to raise it.) If we further represent people who disagree with you initially and undergo the opposite updates, I conjecture that this would lead to predictable, persistent, massive polarization.

If so, how interesting would that be? I'm not sure. On the one hand, we had to make use of a failure of the value of evidence to get it—and no one should be surprised that if you sometimes update irrationally, you will predictably, persistently, massively polarize. On the other hand, that failure was both innocuous-seeming and incredibly plausible—namely, that you forget the details of the word string or study that you

were doing a cognitive search on. If so, then it looks like I can claim the following. I can take a fully rational person. I can give them some evidence, cause them to lose a portion of that evidence, then give them some more evidence, cause them to lose a portion of *that*, and so on, such that: (1) in the end they will know strictly more than they did at the start; (2) at each stage they will expect that rationally responding to the change in evidence will make them no worse off, and maybe better off, when it comes to accuracy or decision-making; and yet (3) they can now predict that following such a sequence will lead them to become arbitrarily confident of a proposition that they are currently only ever-so-slightly confident in.

Perhaps those claims would be interesting enough? Or perhaps not. This is as far as I've gotten down the rabbit hole; I would love you help in figuring out where to dig from here.

3.5 Conclusion

What to make of all this? At least this. Predictable, fully rational polarization is possible. This is because sometimes evidence is *ambiguous*, and therefore rational people are unsure how they should react to it. This explanation plausibly plays at least some role in explaining the empirical results surrounding confirmation bias—at least sometimes, such a "bias" is fully rational. Moreover, it seems that if we take this explanation and add just a tiny bit of information-loss along the way—a bit that surely does happen—then we can get a story at which each stage people are either fully rational, or simply limited because of constraints on their memory; and yet, they end up predictably, persistently, and massively polarized.

If that's right, there are both theoretical and practical questions. Theoretical question: *how much* of DIVIDE-AND-DIVERGE can be explained by these rational mechanisms. Practical question: how would it *matter*, if much of it could?

To address the theoretical question: I don't know. Clearly more theoretical work is needed to set the bounds of possibility; then, empirical work will be needed to test how well the theory fits with what we observe as the actual drivers of polarization.

To address the practical question: it at least paints a politically palatable message. For it suggests that we can look at those on the "other side"—politically, religiously, or morally—and think they are wrong, but not think they are dumb. And that, I think, is an important step in any project of social reconciliation.

Bibliography

Adams, Ernest, 1975. *The Logic of Conditionals*, volume 86 of *Synthese Library*. Springer Netherlands.

Ahmed, Arif and Salow, Bernhard, 2018. 'Don't Look Now'. *British Journal for the Philosophy of Science*, To appear.

Bacon, Andrew, 2013. 'Stalnaker on the KK principle'. Manuscript.

———, 2015. 'Stalnaker's Thesis in Context'. *The Review of Symbolic Logic*, 8(1):131–163.

Baron, Jonathan, 1995. 'Myside Bias in Thinking About Abortion'. *Thinking & Reasoning*, 1(3):221–235.

Baron, Robert S., Hoppe, Sieg I., Kao, Chuan Feng, Brunsman, Bethany, Linneweh, Barbara, and Rogers, Diane, 1996. 'Social corroboration and opinion extremity'. *Journal of Experimental Social Psychology*, 32(6):537–560.

Bennett, Jonathan, 2003. *A Philosophical Guide to Conditionals*. Clarendon Press.

Bertsekas, Dmitri P and Tsitsiklis, John N, 2008. *Introduction to Probability*. Athena Scientific, second edition.

Bikhchandani, Saushil, Hirshleifer, David, and Welch, Ivo, 1992. 'A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades'. *Journal of Political Economy*, 100(51):992–1026.

Bradley, Richard, 2000. 'A preservation condition for conditionals'. *Analysis*, 60(267):219–222.

Briggs, R., 2009a. 'Distorted Reflection'. *Philosophical Review*, 118(1):59–85.

Briggs, Ray, 2009b. 'The Anatomy of the Big Bad Bug'. *Nous*, 43(3):428–449.

Brössel, Peter and Eder, Anna-Maria A, 2014. 'How to resolve doxastic disagreement'. *Synthese*, 191(11).

Christensen, D, 2007. 'Epistemology of Disagreement: The Good News'. *Philosophical Review*, 116(2):187–217.

Christensen, David, 1991. 'Clever Bookies and Coherent Beliefs'. *Philosophical Review*, 100(2):229–247.

———, 2010a. ‘Higher-Order Evidence’. *Philosophy and Phenomenological Research*, 81(1):185–215.

———, 2010b. ‘Rational Reflection’. *Philosophical Perspectives*, 24:121–140.

———, 2016. ‘Disagreement, Drugs, etc.: From Accuracy to Akrasia’. *Episteme*.

Coates, Allen, 2012. ‘Rational Epistemic Akrasia’. *American Philosophical Quarterly*, 49(2):113–124.

Cresto, Eleonora, 2012. ‘A Defense of Temperate Epistemic Transparency’. *Journal of Philosophical Logic*, 41(6):923–955.

Das, Nilanjan, 2017. ‘Externalism and the Value of Information’. Manuscript.

Das, Nilanjan and Salow, Bernhard, 2016. ‘Transparency and the KK Principle’. *Noûs*.

Dimock, Michael, Doherty, Carroll, Kiley, Jocelyn, and Oates, Russ, 2014. ‘Political polarization in the American public’. *Pew Reserach Center*.

Dorst, Kevin, 2019a. ‘Evidence: A Guide for the Uncertain’. *Philosophy and Phenomenological Research*, To appear.

———, 2019b. ‘Higher-Order Uncertainty’. In Mattias Skipper Rasmussen and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidenece: New Essays*, To appear. Oxford University Press.

———, 2019c. ‘Lockeans Maximize Expected Accuracy’. *Mind*, 128(509):175–211.

Downing, James W., Judd, Charles M., and Brauer, Markus, 1992. ‘Effects of repeated expressions on attitude extremity.’ *Journal of Personality and Social Psychology*, 63(1):17–29.

Easley, David and Kleinberg, Jon, 2010. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press.

Easwaran, Kenny, 2013. ‘Expected Accuracy Supports Conditionalization – and Conglomerability and Reflection’. *Philosophy of Science*, 80(1):119–142.

Easwaran, Kenny, Fenton-Glynn, Luke, Hitchcock, Christopher, and Velasco, Joel D, 2016. ‘Updating on the Credences of Others: Disagreement, Agreement, and Synergy’. *Philosophers’ Imprint*, 16(11).

Edgington, Dorothy, 1995. ‘On Conditionals’. *Mind*, 104:235–329.

Elga, Adam, 2007. ‘Reflection and Disagreement’. *Noûs*, 41(3):478–502.

———, 2010. ‘Subjective probabilities should be sharp’. *Philosophers’ Imprint*, 10(5):1–11.

———, 2013. ‘The puzzle of the unmarked clock and the new rational reflection principle’. *Philosophical Studies*, 164(1):127–139.

Engber, Daniel, 2018. ‘LOL something matters’. *Slate*, 8:1–18.

Evans, J. St. B. T, 1989. *Bias in Human Reasoning: Causes and Consequences*. Erlbaum.

Feldman, Richard, 2005. 'Respecting the Evidence'. *Philosophical Perspectives*, 19(1):95–119.

———, 2007. 'Reasonable religious disagreements'. In Louise Antony, ed., *Philosophers Without Gods: Meditations on Atheism and the Secular*, 194–214. Oxford University Press.

Fine, Cordelia, 2005. *A Mind of its Own: How Your Brain Distorts and Deceives*. W. W. Norton & Company.

Fitzpatrick, Anne R and Eagly, Alice H, 1981. 'Anticipatory belief polarization as a function of the expertise of a discussion partner'. *Personality and Social Psychology Bulletin*, 7(4):636–642.

Foley, Richard, 1992. 'The Epistemology of Belief and the Epistemology of Degrees of Belief'. *American Philosophical Quarterly*, 29(2):111–124.

———, 2009. 'Beliefs, Degrees of Belief, and the Lockean Thesis'. In Franz Huber and Christoph Schmidt-Petri, eds., *Degrees of Belief*, 37–47. Springer.

Frey, Dieter, 1986. 'Recent Research on Selective Exposure to Information'. *Advances in Experimental Social Psychology*, 19:41–80.

Gaifman, Haim, 1988. 'A Theory of Higher Order Probabilities'. In Brian Skyrms and William L Harper, eds., *Causation, Chance, and Credence*, volume 1, 191–219. Kluwer.

Geanakoplos, John, 1989. 'Game Theory Without Partitions, and Applications to Speculation and Consensus'. *Research in Economics*, Cowles Fou(914).

Gelman, Andrew and Tuerlinckx, Francis, 2000. 'Type S error rates for classical and Bayesian single and multiple comparison procedures'. *Computational Statistics*, 15(3):373–390.

Gibbons, John, 2006. 'Access Externalism'. *Mind*, 115(457):19–39.

Good, I J, 1967. 'On the Principle of Total Evidence'. *The British Journal for the Philosophy of Science*, 17(4):319–321.

Greco, Daniel, 2014a. 'A puzzle about epistemic akrasia'. *Philosophical Studies*, 161:201–219.

———, 2014b. 'Could KK be OK?' *Journal of Philosophy*, 111(4):169–197.

Guess, Andrew, Lyons, Benjamin, Nyhan, Brendan, and Reifler, Jason, 2018. 'Avoiding the Echo Chamber about Echo Chambers: Why selective exposure to like-minded political news is less prevalent than you think'. 25.

Hall, Ned, 1994. 'Correcting the Guide to Objective Chance'. *Mind*, 103(412):505–517.

———, 2004. 'Two Mistakes about Credence and Chance'. *Australasian Journal of Philosophy*, 82(1):93–111.

Hazlett, Allan, 2012. 'Higher-order epistemic attitudes and intellectual humility'. *Episteme*, 9(3):205–223.

Horowitz, Sophie, 2013. 'Immoderately rational'. *Philosophical Studies*, 167(1):41–56.

———, 2014. 'Epistemic Akrasia'. *Noûs*, 48(4):718–744.

———, 2018. 'Predictably Misleading Evidence'. In Mattias Skipper and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidence: New Essays*, volume To appear. Oxford University Press.

Huemer, Michael, 2011. 'The Puzzle of Metacoherence'. *Philosophy and Phenomenological Research*, 82(1):1–21.

Huttegger, Simon M, 2014. 'Learning experiences and the value of knowledge'. *Philosophical Studies*, 171(2):279–288.

Isenberg, Daniel J., 1986. 'Group Polarization. A Critical Review and Meta-Analysis'. *Journal of Personality and Social Psychology*, 50(6):1141–1151.

Jeffrey, Cassandra, 2017. 'Does Partisan Media Encourage a More Politically Polarized America?' *DNN Media*.

Joyce, James M, 1998. 'A Nonpragmatic Vindication of Probabilism'. *Philosophy of Science*, 65(4):575–603.

———, 2009. 'Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief'. In Franz Huber and Christoph Schmidt-Petri, eds., *Degrees of Belief*, 263–297. Springer.

Joyce, James M., 2010. 'A Defense of Imprecise Credences in Inference and Decision Making'. *Philosophical Perspectives*, 24(1):281–323.

Kahan, Dan M., Peters, Ellen, Wittlin, Maggie, Slovic, Paul, Ouellette, Lisa Larrimore, Braman, Donald, and Mandel, Gregory, 2012. 'The polarizing impact of science literacy and numeracy on perceived climate change risks'. *Nature Climate Change*, 2(10):732–735.

Kahneman, Daniel, Slovic, Paul, and Tversky, Amos, eds., 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Kelly, Thomas, 2008. 'Disagreement, Dogmatism, and Belief Polarization'. *The Journal of Philosophy*, 105(10):611–633.

———, 2010. 'Peer disagreement and higher order evidence'. In Alvin I Goldman and Dennis Whitcomb, eds., *Social Epistemology: Essential Readings*, 183–217. Oxford University Press.

Khoo, Justin, 2013. 'Conditionals, Indeterminacy, and Triviality'. *Philosophical Perspectives*, 27(1):260–287.

———, 2016. 'Probabilities of Conditionals in Context'. *Linguistics and Philosophy*, 39(1):1–43.

Koriat, Asher, Lichtenstein, Sarah, and Fischhoff, Baruch, 1980. 'Journal of Experimental Psychology : Human Learning and Memory Reasons for Confidence'. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2):107–118.

Kuhn, Deanna and Lao, Joseph, 1996. 'Effects of Evidence on Atittudes: is Polarization the Norm?' *Psycholohical Science*, 7(2):115–120.

Kunda, Ziva, 1990. 'The case for motivated reasoning'. *Psychological Bulletin*, 108(3):480–498.

Lasonen-Aarnio, Maria, 2010. 'Unreasonable Knowledge'. *Philosophical Perspectives*, 24(1):1–21.

———, 2013. 'Disagreement and evidential attenuation'. *Nous*, 47(4):767–794.

———, 2014. 'Higher-order evidence and the limits of defeat'. *Philosophy and Phenomenological Research*, 8(2):314–345.

———, 2015. 'New Rational Reflection and Internalism about Rationality'. In Tamar Szabó Gendler and John Hawthorne, eds., *Oxford Studies in Epistemology*, volume 5, 145–171. Oxford University Press.

Lazer, David, Baum, Matthew, Benkler, Jochai, Berinsky, Adam, Greenhill, Kelly, Metzger, Miriam, Nyhan, Brendan, Pennycook, G., Rothschild, David, Sunstein, Cass, Thorson, Emily, Watts, Duncan, and Zittrain, Jonathan, 2018. 'The science of fake news'. *Science*, 359(6380):1094–1096.

Leitgeb, Hannes, 2013. 'Reducing Belief Simpliciter to Degrees of Belief'. *Annals of Pure and Applied Logic*, 164:1338–1389.

Levinstein, Ben, 2017. 'Permissive Rationality and Sensitivity'. *Philosophy and Phenomenological Research*, XCIV(2):343–370.

Lewis, David, 1971. 'Completeness and Decidability of Three Logics of Counterfactual Conditionals'. *Theoria*, 17:74–85.

———, 1976. 'Probabilities of Conditionals and Conditional Probabilities'. *The Philosophical Review*, 85(3):297–315.

———, 1980. 'A subjectivist's guide to objective chance'. In Richard C Jeffrey, ed., *Studies in Inductive Logic and Probability*, volume 2. University of California Press.

Littlejohn, Clayton, 2015. 'Stop Making Sense? On a Puzzle about Rationality'. *Philosophy and Phenomenological Research*, To Appear.

Lord, Charles G., Ross, Lee, and Lepper, Mark R., 1979. 'Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence'. *Journal of Personality and Social Psychology*, 37(11):2098–2109.

Mandelbaum, Eric, 2018. 'Troubles with Bayesianism: An introduction to the psychological immune system'. *Mind & Language*, 1–17.

Mandelkern, Matthew and Khoo, Justin, 2018. 'Against Preservation'. *Analysis*, To appear.

Mcpherson, Miller, Smith-lovin, Lynn, and Cook, James M, 2001. 'Birds of a Feather: Homophily in Social Networks'. *Annual Review of Sociology*, 27:415–444.

Mitchell, A and Weisel, Rachel, 2014. 'Political Polarization and Media Habits'. *Pew Research Center*, (October).

Munro, Geoffrey D and Ditto, Peter H, 1997. 'Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information'. *Personality and Social Psychology Bulletin*, 23(6):636–653.

Myers, David G. and Lamm, Helmut, 1976. 'The group polarization phenomenon'. *Psychological Bulletin*, 83(4):602–627.

Myrvold, Wayne C, 2012. 'Epistemic values and the value of learning'. *Synthese*, 187(2):547–568.

Nguyen, C. Thi, 2018. 'Escape the echo chamber'. *Aeon*.

Nickerson, Raymond S., 1998. 'Confirmation bias: A ubiquitous phenomenon in many guises.' *Review of General Psychology*, 2(2):175–220.

Nyhan, Brendan and Reifler, Jason, 2010. 'When corrections fail: The persistence of political misperceptions'. *Political Behavior*, 32(2):303–330.

Oddie, Graham, 1997. 'Conditionalization, Cogency, and Cognitive Value'. *The British Journal for the Philosophy of Science*, 48(4):533–541.

Paul, L A, 2014. *Transformative Experience*. Oxford University Press.

Pennycook, By Gordon and Rand, David, 2019. 'Why Do People Fall for Fake News? Are they blinded by their political passions? Or are they just intellectually lazy?'

Pettigrew, Richard, 2013. 'Epistemic Utility and Norms for Credences'. *Philosophy Compass*, 8(10):897–908.

———, 2016. *Accuracy and the Laws of Credence*. Oxford University Press.

———, 2017. 'Choosing for Changing Selves'.

Pettigrew, Richard and Titelbaum, Michael G, 2014. 'Deference Done Right'. *Philosopher's Imprint*, 14(35):1–19.

Piore, Adam, 2018. 'Technologists are trying to fix the âĂIJ filter bubble âĂİ problem that tech helped create'. *MIT Technology Review*.

Plous, Scott, 1991. 'Biases in the assimilation of technological breakdowns: Do accidents make us safer?' *Journal of Applied Social Psychology*, 21(13):1058–1082.

Ramsey, F P, 2010. 'Truth and probability'. In Antony Eagle, ed., *Philosophy of Probability: Contemporary Readings*. Routledge.

Ramsey, Frank, 1931. *The Foundations of Mathematics and Other Logical Essays*. Kegan Paul, Trench, and Trubner & Co.

Rasmussen, Mattias Skipper, Steglich-Petersen, Asbjørn, and Bjerring, Jens Christian, 2016. 'A Higher-Order Approach to Disagreement'. *Episteme*, to appear.

Robson, David, 2018. ‘The myth of the online echo chamber’.

Rothschild, Daniel, 2013. ‘Do Indicative Conditionals Express Propositions?’ *Noûs*, 47(1):49–68.

Roush, Sherrilyn, 2009. ‘Second Guessing: A Self-Help Manual’. *Episteme*, 251–268.

———, 2016. ‘Knowledge of Our Own Beliefs’. *Philosophy and Phenomenological Research*, 93(3).

Russell, Jeffrey Sanford and Hawthorne, John, 2016. ‘General Dynamic Triviality Theorems’. *Philosophical Review*, 125(3):307–339.

Salow, Bernhard, 2017. ‘Elusive Externalism’. *Mind*, to appear.

———, 2018. ‘The Externalist’s Guide to Fishing for Compliments’. *Mind*, 127(507):691–728.

Samet, Dov, 1997. ‘On the Triviality of High-Order Probabilistic Beliefs’. https://ideas.repec.org/p/wpa/wuwpga/9705001.html.

Schoenfield, Miriam, 2014. ‘Permission to Believe: Why Permissivism is True and What it Tells Us About Irrelevant Influences On Belief’. *Nous*, 48(2):193–218.

———, 2015a. ‘A Dilemma for Calibrationism’. *Philosophy and Phenomenological Research*, 91(2):425–455.

———, 2015b. ‘Bridging Rationality and Accuracy’. *Journal of Philosophy*, 112(12):633–657.

———, 2016. ‘An Accuracy Based Approach to Higher Order Evidence’. *Philosophy and Phenomenological Research*, To Appear.

Schultheis, Ginger, 2017. ‘Living on the Edge: Against Epistemic Permissivism’. *Mind*, to appear.

Scott, John, 2017. *Social network analysis*. Sage, 4th edition.

Sharon, Assaf and Spectre, Levi, 2008. ‘Mr. Magoo’s mistake’. *Philosophical Studies*, 139(2):289–306.

Skyrms, Brian, 1966. *Choice and Chance: An Introduction to Inductive Logic*, volume 18. Dickenson Pub. Co.

———, 1980. ‘Higher Order Degrees of Belief’. In D H Mellor, ed., *Prospects for Pragmatism*, 109–137. Cambridge University Press.

———, 1990. ‘The Value of Knowledge’. *Minnesota Studies in the Philosophy of Science*, 14:245–266.

Sliwa, Paulina and Horowitz, Sophi, 2015. ‘Respecting *all* the evidence’. *Philosophical Studies*, 172(11):2835–2858.

Smithies, Declan, 2012. 'Moore's paradox and the accessibility of justification'. *Philosophy and Phenomenological Research*, 85(2):273–300.

———, 2015. 'Ideal Rationality and Logical Omniscience'. *Synthese*, 192(9):2769–2793.

Stalnaker, Robert, 1970. 'Probability and Conditionals'. *Philosophy of Science*, 37(1):64–80.

———, 1984. *Inquiry*. Cambridge University Press.

———, 2006. 'On the Logics of Knowledge and Belief'. *Philosophical Studies*, 128(1):169–199.

———, 2015. 'Luminosity and the KK Thesis'. In Sanford Goldberg, ed., *Externalism, Self-Knowledge, and Skepticism*, 1–19. Cambridge University Press.

———, 2017. 'Rational Reflection, and the Notorious Unmarked Clock'.

Sturgeon, Scott, 2008. 'Reason and the Grain of Belief'. *Noûs*, 42(1):139–165.

Sunstein, C, 2009. *Going to Extremes: How Like Minds Unite and Divide*. Oxford University Press.

Sunstein, Cass R, 2000. 'Deliberative trouble? Why groups go to extremes'. *The Yale Law Journal*, 110(1).

———, 2002. *Republic. com*. Princeton university press.

———, 2017. *#Republic: Divided democracy in the age of social media*. Princeton University Press.

Tal, Eyal, 2018. 'Self-Intimation, Infallibility, and Higher-Order Evidence'. *Erkenntnis*, To appear.

Tesser, Abraham, Martin, Leonard, and Mendolia, Marilyn, 1995. 'The impact of thought on attitude extremity and attitude-behavior consistency.'

Titelbaum, Michael, 2015. 'Rationality's Fixed Point (or: In Defense of Right Reason)'. In Tamar Szabó Gendler and John Hawthorne, eds., *Oxford Studies in Epistemology*, volume 5, 253–292. Oxford University Press.

Titelbaum, Michael G and Kopec, Matthew, 2017. 'Plausible Permissivism'. Oxford University Press.

Todd, Peter M, Hills, Thomas T, Robbins, Trevor W, and Lupp, Julia, 2012. *Cognitive search: Evolution, algorithms, and the brain*, volume 9. MIT press.

van Ditmarsch, Hans, Halpern, Joseph Y, van der Hoek, Wiebe, and Kooi, Barteld, 2015. *Handbook of Epistemic Logic*. College Publications.

van Fraassen, Bas, 1976. 'Probabilities of Conditionals'. In William Harper and Clifford Hooker, eds., *Foundations of probability theory, statistical inference, and statistical theories of science.*, 261–308. Reidel.

———, 1984. ‘Belief and the Will’. *The Journal of Philosophy*, 81(5):235–256.

Vavova, Katia, 2014. ‘Confidence, Evidence, and Disagreement’. *Erkenntnis*, 79:173–183.

———, 2016. ‘Irrelevant Influences’. *Philosophy and Phenomenological Research*, To appear.

Vosoughi, Soroush, Roy, Deb, and Aral, Sinan, 2018. ‘The spread of true and false news online’. *Science*, 359(6380):1146–1151.

Wedgwood, Ralph, 2012. ‘Justified Inference’. *Synthese*, 189:273–295.

White, Roger, 2005. ‘Epistemic Permissiveness’. *Philosophical Perspectives*, 19(1):445–459.

———, 2009a. ‘Evidential Symmetry and mushy credence’. *Oxford Studies in Epistemology*, 161–186.

———, 2009b. ‘On Treating Oneself and Others as Thermometers’. *Episteme*, 6(3):233–250.

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

———, 2008. ‘Why Epistemology Cannot be Operationalized’. In Quentin Smith, ed., *Epistemology: New Essays*, 277–300. Oxford University Press.

———, 2014. ‘Very Improbable Knowing’. *Erkenntnis*, 79(5):971–999.

———, 2018. ‘Evidence of Evidence in Epistemic Logic’. In Mattias Skipper and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidence: New Essays*, volume To appear. Oxford University Press.

Worsnip, Alex, 2015. ‘The Conflict of Evidence and Coherence’. *Philosophy and Phenomenological Research*, To Appear:1–42.