Rational Polarization

Kevin Dorst
Massachusetts Institute of Technology

1. A Standard Story

I owe a lot to a bench.

My friends and I had been using it to sneak out the window. To push the boundaries. To 'experiment'. But my luck held: I forgot it outside, my parents confronted me, I wasn't quick on my feet… and that was that.

For me. But my friends were quicker on their feet; their parents slower to see the problem; their luck sooner to run out. So we went our separate ways. While I got studious, they got disaffected. While I went

Far too many people helped me with this project to properly thank them all. I received feedback from audiences at MIT, the Prindle Institute for Ethics, Indiana University, the University of Pittsburgh, the National University of Singapore, the 2019 Pacific APA, the University of Oxford, the University of Missouri, the Pittsburgh Center for Philosophy of Science, the University of Lisbon, and USC. Bernhard Salow, Jack Spencer, Dmitri Gallow, Roger White, Sally Haslanger, Caspar Hare, Kieran Setiya, and Bob Stalnaker each played formative roles early in this project. I received feedback along the way from Riet van Bork, Martina Calderisi, Agnes Callard, Chris Dorst, Adam Elga, Jane Friedman, Brian Hedden, Branden Fitelson, Rachel Fraser, Jeffrey Friedman, Peter Gerdes, Dan Greco, Brian Hedden, Jay Hodges, Michael Hannon, Jean Janasz, Joshua Knobe, Harvey Lederman, Ben Levinstein, Annina Loets, Tim Maudlin, Travis McKenna, Aydin Mohseni, Pedro Passos, Steven Pinker, Drazen Prelec, Kevin Richardson, Mark Schroeder, Teddy Seidenfeld, Laura Soter, Tom Stafford, Kate Stanton, Daniel Stone, Mason Westfall, Kevin Zollman, two stellar referees, and many others—including several anonymous blog and social media commenters. Special thanks to Liam Kofi Bright, Thomas Byrne, Cosmo Grant, Matt Mandelkern, Miriam Schoenfield, Ginger Schultheis, and Quinn White for helping me pull the project together. And to some old friends—for sharing their stories, and setting me straight.

Philosophical Review, Vol. 132, No. 3, 2023
DOI 10.1215/00318108-10469499

to a liberal city, many of them stayed in conservative towns. While I was having my eyes opened, some of them were fighting for their lives.

Yet this isn't a story about how a bench changed my life. It's a story about how a bench changed my *beliefs*. So let me ask: What do you think happened to our politics? Who now is concerned about far-right militias, and who about Antifa? Who believes gun rights should be restricted, and who owns handguns for their own protection? Who voted for Biden, and who thinks Trump shook things up in a needed way?

I think you can guess.

That's no surprise. Most societies display both *local conformity* and *global disunity*: people's attitudes are predictable given their social group, despite varying widely across such groups (Mcpherson, Smith-Lovin, and Cook 2001). As a result, people who set out on different trajectories often polarize in ways that are profound, persistent, and *predictable* (Cohen 2000; Sunstein 2009). When I went a liberal university in a liberal city, I could predict—not with certainty, but with some confidence—that I would become more liberal (Lottes and Kuriloff 1994).

My question is why.

The standard story: predictable polarization is due to *epistemic irrationality*, the fact that people's beliefs are insufficiently constrained by evidence.[1] Instead, people glom onto the beliefs of their peers,[2] confirm and entrench those beliefs,[3] and become wildly overconfident in them.[4] Combined with the informational traps of the modern internet,[5] we have a simple explanation of the rise of polarization (see Iyengar et al. 2019; Boxell, Gentzkow, and Shapiro 2020).

[1] See Sutherland 1992; Lakoff 1997; Mills 2007; Lilienfeld, Ammirati, and Landfield 2009; Haidt 2012; Klein 2014; Brennan 2016; Achen and Bartels 2017; Bregman 2017; Carmichael 2017; Mercier and Sperber 2017; Lazer et al. 2018; Pennycook and Rand 2019; Finkel et al. 2020; Klein 2020.

[2] See Myers and Lamm 1976; Isenberg 1986; Baron et al. 1996; Sunstein 2000, 2009; Mcpherson, Smith-Lovin, and Cook 2001; Cohen 2003; Pronin 2008; Iyengar, Sood, and Lelkes 2012; Mäs and Flache 2013, Myers 2012: chap. 8, Baumgaertner, Tyson, and Krone 2016; Brownstein 2016; Mason 2018; Wilkinson 2018; Talisse 2019; Siegel 2021; Williams 2021.

[3] See Lord, Ross, and Lepper 1979; Frey 1986; Kunda 1990; Nickerson 1998; Jost et al. 2003; Fine 2005; Taber and Lodge 2006; Taber, Cann, and Kucsova 2009; Kahan et al. 2012; Kahan 2013; Kahan et al. 2017; Kahan 2018; Stanovich 2020.

[4] See Lichtenstein, Fischhoff, and Phillips 1982; Harvey 1997; Johnson 2009; Glaser and Weber 2010; Moore, Carter, and Yang 2015; Ortoleva and Snowberg 2015; van Prooijen and Krouwel 2019; Stone 2019.

[5] See Jamieson and Cappella 2008; Pariser 2012; Sunstein 2017; Nguyen 2018; Vosoughi, Roy, and Aral 2018.

Notice that this story combines components: empirical hypotheses about why people predictably polarize, and normative claims that they should not. The empirical hypotheses are (largely) true. I argue that the normative claims are false.

This requires rejecting Standard Bayesian assumptions. Though it is often overlooked, they imply that predictable polarization must be irrational, regardless of varying evidential standards (Schoenfield 2014), background beliefs (Jern, Chang, and Kemp 2014; Benoît and Dubra 2019), or distributions of trust (O'Connor and Weatherall 2018; Henderson and Gebharter 2021). For they require your current opinion to always match your estimate of your future rational opinion, meaning you cannot (rationally) do what we do all the time: predict the direction our actions will shift our opinions (see section 2).

But we *should* reject those assumptions, for they also imply that rational people can never be unsure whether they have been rational. Given *ambiguous evidence*—evidence that is hard to know how to interpret—such self-doubts can be rational. As a result, there are updates that satisfy the value of evidence (Blackwell 1953; Good 1967)—that are expected to improve your accuracy and cannot be Dutch booked—that nonetheless are predictably polarizing (see section 3). Indeed, common cognitive processes generate *asymmetric* ambiguities, making it easier to recognize evidence pointing in one direction than the other (see section 4). Each such update is expected to improve accuracy, despite the fact that a long series of them can predictably lead to profound polarization (see section 5). Moreover, this mechanism plausibly plays a role in the psychological processes that drive real-world polarization (see sections 6 and 7).

Although this story is built on a series of technical results, the main ideas can be understood without them. Thus I have partitioned the article: those interested in the story but not the technicalities can skip the formal subsections and footnotes without loss of continuity.

But what's the point? Why *want* a rational story? Consider the alternative. From the outside, it looks like my beliefs were just as predictable as my friends': long before I came to believe that (say) guns decrease safety, it was predictable that I would. That implies that if predictable polarization is due to irrationality, *my* beliefs are due to irrationality. Yet *I* cannot admit that, at least not while maintaining my beliefs: it is incoherent ('akratic') to believe "guns decrease safety, but it is irrational for me to believe that" (Horowitz 2014; Dorst 2020). So if I am not willing to give up my beliefs—as indeed I am not—I must resort to special pleading: “Their beliefs were predictable, but *mine* were not. *They* were the irrational ones, not me.” That’s desperate. It is also dubious. My friends were smarter (and quicker) than I was. My trajectory was more predictable than theirs was. Our divergence is due to our *circumstances*, not ourselves. A slight change in those, and I would believe everything they do—there but for a bench go I.

That’s the point. A rational story lets us to avoid both special pleading and incoherence. It lets us admit our own predictability, maintain the truth of our own deeply held commitments, and yet acknowledge the rationality of others’. Let me show you how.

1.1. The Idea

Here’s the idea. Sometimes evidence is clear—you should know exactly how to respond to it. Other times evidence is ambiguous—you should be unsure how to respond. Ambiguity asymmetries can make it easier to recognize evidence pointing in one direction than another. For example, is the following word search completable?

FR_ _L

If you find a completion, you know you should be 100% confident it is *c*ompletable (c). But if you do not find one, your evidence is ambiguous—you should be unsure how confident you should be (“Am I missing something?”), and so should stay near 50% (see section 4).

Notice two things. First, you expect this update to improve accuracy. When the evidence is clear, it leads you to the truth; when it is ambiguous, it leaves you where you were. So the (potentially ambiguous) evidence will not hurt, and might help.

Second, such *asymmetric accuracy increases* can drive polarization. Iterate this with many claims $c_1, \ldots, c_n$ that you are 50% confident in, and you can predict that your average rational confidence (‘credence’) will rise: the average of ‘rise a lot’ and ‘fall a little’ is ‘rise a little’. Thus you can predict that it will be rational to become confident in things you initially doubt: if your average confidence in the (independent) c_i becomes 60%, you must become confident that more than half are true. Predictable polarization amounts to an epistemic *diachronic tragedy* (Hedden 2015): taking steps that are each expected to make you more accurate predictably leads, in the long run, to opinions you (initially) think are wrong. Once we allow ambiguous evidence, all of this this can be proven in a Bayesian setting (see sections 3–5).

I further argue that it helps rationalize real-world polarization. Sound naive? Hasn't psychology shown that people are *ir*rational? Though many think so,[6] many do not: they critique the empirical replicability and normative interpretations of such work[7] and contrast it with the growing evidence that rational processes explain the mind's ability to perform intractably complex tasks that computers cannot.[8] I show how confirmation bias can be rational when your prior beliefs make it easier to recognize flaws in arguments against than in favor of them (section 6), and arguments can predictably persuade you by making the evidence favoring their side less ambiguous than the evidence opposing it (section 7).

The payoff? This story makes sense of our *own* polarization. When we scrutinize opposing viewpoints or check partisan news sources, we often think it is the best way to figure things out. According to my story: *we're right.* The problem is that locally optimal steps toward the truth can lead, in the long run, to a predictable drift away from it.

2. The Problem

What's the epistemic problem of 'predictable' polarization?

Many think: nothing. They point out that differences in background beliefs, networks of trust, and lived experiences (evidence) can easily lead rational Bayesians to persistently disagree, or polarize further upon seeing the same evidence.[9] Case closed?

[6] For example, Tversky and Kahneman 1974; Kahneman, Slovic, and Tversky 1982; Kahneman and Tversky 1996; Fine 2005; Ariely 2008; Hastie and Dawes 2009; Kahneman 2011; Thaler 2015; Mandelbaum 2018.

[7] For example, Cohen 1981; Gigerenzer 1991, 2018; Krueger and Massey 2009; Stafford 2015, 2020; Whittlestone 2017; Rizzo and Whitman 2019; Mercier 2020; Cushman 2020.

[8] For example, Anderson 1990; Oaksford and Chater 1994, 1998; Gopnik 1996, 2012, 2020; Tenenbaum and Griffiths 2006; Tenenbaum et al. 2011; Griffiths et al. 2012; Griffiths, Lieder, and Goodman 2015; Lieder and Griffiths 2019; Gershman 2021.

[9] For example, Feeney, Evans, and Clibbens 2000; Dixit and Weibull 2007; Austerweil and Griffiths 2011; Le Mens and Denrell 2011; Olsson 2013; Acemoglu and Wolitzky 2014; Jern, Chang, and Kemp 2014; Cook and Lewandowsky 2016; Angere and Olsson 2017; Pallavicini, Hallsson, and Kappel 2018; Benoît and Dubra 2019; Nimark and Sundaresan 2019; Nielsen and Stewart 2021; Henderson and Gebharter 2021; Bowen, Dmitriev, and Galperti 2023.

No. Distinguish different types of 'predictable' polarization. Extant models show that there can be two Bayesians P and P' and some *other* agent—who knows more than they do—who can predict how they will polarize further. For example, Jern, Chang, and Kemp 2014 and Henderson and Gebharter 2021 show that for two Bayesians who disagree about the likely causal paths or the reliabilities of sources, there can be a proposition E such that learning E will exacerbate their disagreement about q (so $P(q|E) > P(q) > P'(q) > P'(q|E)$). Yet they cannot predict how they will polarize: they cannot know whether they will learn E or $\neg E$, and learning the latter would push their opinions in the *other* direction ($P(q) > P(q|\neg E) > P'(q|\neg E) > P'(q)$).

This is no accident. Standard Bayesian models (including those in fn. 9) forbid a rational person from expecting a rational update to move their opinions in a particular direction. Let 'P' be the prior rational probability function, and let '$\widetilde{P}$' be the rational one after the update. (More on my rationality assumptions in section 3.) Since you can be unsure what evidence you will receive, $\widetilde{P}$ picks out different functions in different possibilities. Nevertheless, you can form an *estimate* of what your updated rational credence should be. On Standard Bayesian models, your initial credence in q ($P(q)$) must match your initial estimate for your updated rational credence in q (your estimate of $\widetilde{P}(q)$); thus you cannot estimate that it will be rational to move your opinion in a particular direction. This is intuitive. Rationally estimating that your more informed future self will be confident of q seems to make it rational to *now* be confident of q. If so, there cannot be a rational divergence between what you expect your future rational self to believe and what you now believe.

More precisely, a Standard Bayesian model is one on which $\widetilde{P}$ is always obtained by conditioning P on the true answer to a question, that is, the true cell of a finite[10] partition (see section A.3). (E.g., if the question is whether E, then the partition is $\{E, \neg E\}$; in E-worlds, $\widetilde{P} = P(\cdot|E)$; and in $\neg E$-worlds, $\widetilde{P} = P(\cdot|\neg E)$.) Any such model yields:[11]

> **Reflection (Martingale property):** Your prior rational credence in q must equal your rational estimate of your updated rational credence in q.

[10] I restrict attention to finite models.

[11] See Kadane, Schervish, and Seidenfeld 1996; Weisberg 2007; Briggs 2009; Salow 2018 for explanations. The 'Bayesian persuasion' literature (Kamenica and Gentzkow 2011) takes this constraint as axiomatic. As we will see, it need not.

For all q, $P(q) = \mathbb{E}_P(\widetilde{P}(q))$.[12]

This is the epistemic problem of predictable polarization: empirically, our beliefs violate Reflection, and hence (normatively) they are rational only if Standard Bayesianism is wrong. In this section I defend this empirical point, leaving the question of rationality for later.

Reflection violations are mundane. We can often predict how our actions will shift our beliefs, even when those actions provide no evidence about the issue. Not long ago, I had both Piketty's *Capital in the 21st Century* and Pinker's *Enlightenment Now* on my shelf. It wasn't hard to predict that reading Pinker would make me more optimistic about our economic system, and reading Piketty would make me less. (I read both.) Next up: I predict that Gessen's *Surviving Autocracy* will increase my credence that America's political woes are due to Republican authoritarian tendencies, while Lind's *The New Class War* will increase my credence that they are due to Democrats' distance from the working class. No surprises here—recall Pascal's (1660) advice: if you want to become religious, read religious thinkers and spend time with religious people. Likewise with other topics.

Another example is biased sources. Make an estimate of the number of extreme weather events there will be in the United States in the next 50 years. This is hard, but pick a number (say, 300). Now, which direction do you think your estimate will shift if you decide to be extremely biased in your climate news consumption, say, reading only the most dire, doomsday climate-change reports? Obviously you expect this would increase your estimate! You are aware that reading biased sources will bias your opinions. This is a familiar Reflection failure[13]—the sort that motivates us to try to be *un*biased in our news consumption (Worsnip 2019).

Here's a simpler case. Think of a bodily symptom that has puzzled you—a new pain, a bump where you don't remember one, and so on. I predict that if you spend an hour Googling possible causes, you will increase your credence that it is worrying. (And I suspect you predict as much too, which is part of why you *have not* Googled it.)

It is not just you. It is well documented that people tend to shift their beliefs in the direction they are searching for evidence (e.g., Isen-

[12] $\mathbb{E}_P$ captures the expectations of P: for any function from worlds to numbers X, $\mathbb{E}_P(X) = \sum_t P(X = t) \cdot t$.

[13] Reflection requires your estimate to equal your estimate of your future estimate: $\mathbb{E}_P(X) = \mathbb{E}_P(\mathbb{E}_{\widetilde{P}}(X))$.

berg 1986; Kunda 1990; Nickerson 1998; Kahan et al. 2017) and, moreover, that those who are aware of this tendency—so in a position to predict it—are still subject to it (Pronin 2008; Lilienfeld, Ammirati, and Landfield 2009).

Still skeptical? Granted, it can be hard to pinpoint a moment when Reflection clearly fails. But it must at some point or other, for you obey Reflection for each update in a sequence (P^1 to P^2 to… P^n) only if your initial opinion matches your initial estimate of the opinions you will have at the end.[14] Yet as the epistemology of 'irrelevant influences' emphasizes, this defies common sense.[15] The following is a standard example: in 1961, G. A. Cohen was choosing between Harvard and Oxford for graduate school. He had no idea whether the analytic/synthetic distinction was legitimate, but since most students at Oxford thought it was, while most at Harvard thought it was not, he could predict how his opinion would move given his choice. The choice *itself* was no evidence—upon choosing Oxford, he still had no opinion, but could now predict that he would increase his credence in the distinction's legitimacy.

Our politics is rife with such stories. Take me and an old friend, Dan. Consider a moment soon after we had parted ways—when our opinions had not moved, but our trajectories were clear. I had started studying at an urban university; he had started bartending in a rural town. Let P be my (rational) opinions then, and let $\widetilde{P}$ be those it would be rational to have 5 years later. Likewise for D and $\widetilde{D}$ for Dan. Let s be a partisan-coded claim, for example, that *guns increase safety*. Neither of us had any strong opinions about s—we were close to 50:50 on it. Yet we knew Republicans tended to believe it, while Democrats did not.[16] We knew living with liberals tends to make you liberal, and likewise for conservatives (Lottes and Kuriloff 1994; Brown and Enos 2021). And we had no reason to think we would be exceptions to this rule. Thus—regardless of what we *in fact* expected—we were *in a position* to expect that in 5 years time, Dan would be more confident of s, while I would be more doubtful. If this was rational, the following must be possible:

[14] If $P^i = \mathbb{E}_{P^i}(P^j)$ and $P^j = \mathbb{E}_{P^j}(P^k)$, then $\mathbb{E}_{P^i}(P^k) = \mathbb{E}_{P^i}(\mathbb{E}_{P^j}(P^k)) = P^i$. Iterating, $\mathbb{E}_{P^1}(P^n) = P^1$.

[15] For example, Cook 1987; Cohen 2000; White 2010; Schoenfield 2017; Vavova 2018. For empirical work, see Mcpherson, Smith-Lovin, and Cook 2001; Kossinets and Watts 2009; Sunstein 2009; Easley and Kleinberg 2010; De Cruz 2017.

[16] A 2018 poll found that 89% of Republicans agree with s, while only 29% of Democrats do (Murray 2018).

> **Expectable polarization:** Dan and I could both estimate that my rational credence in s would end up lower and his would end up higher.
> $\mathbb{E}_P(\widetilde{P}(s)) < P(s)$, and $\mathbb{E}_P(\widetilde{D}(s)) > D(s)$; likewise for $\mathbb{E}_D$.

This violates Reflection. Even though we knew we would receive radically different evidence, Standard Bayesianism forbids our expectable polarization. (When Reflection fails in this way, I will speak of a *single* person expectably polarizing.)

Some clarifications. First, I do not claim *politics* is predictable—it is hard to say how the Democratic party's platform will shift. What I claim is that since people often shift faster than parties, we can often say how a given person's opinions—even our own—will likely shift.

Second, estimates ('expectations') are not necessarily predictions. If I toss a fair coin 10 times, your *estimate* for the number of heads is 5, but you do not *predict* this, since you are pretty (76%) confident that it won't be exactly 5. Expectable polarization thus permits uncertainty about whether the rational posterior ($\widetilde{P}(s)$) will move in the expected direction; all it says is that you rationally think that *on average,* across the various possibilities, it will. Still, expectable polarization violates Reflection and so is all we need to generate the epistemic problem of predictable polarization. In response, I show that updates that are expected to make you more accurate about every subject matter and cannot be Dutch-booked can nonetheless expectably polarize you (sections 3–4).

But third: more is needed. In both Cohen's case and mine, polarization is *more* than expectable: we could reasonably *predict with confidence* that our opinions would move substantially in the expected direction. In an increasingly polarized society, there does not seem to be a principled limit on how strong these predictions could be. Thus if we aim to rationalize real-world polarization, we should consider whether the following (strictly) stronger type of polarization could be rational (section 5):

> **Predictable polarization:** Dan and I could both *predict with confidence* that my credence in s should substantially drop and his should substantially rise.
> $P(\widetilde{P}(s) \ll P(s)) \approx 1$, and $P(\widetilde{D}(s) \gg D(s)) \approx 1$; likewise for D.

You should balk at this—if rationality is a guide to truth, how could rational updates predictably radicalize you? The main theoretical result of this article (Theorem 5.1) is that they can: there can be a sequence of updates—each of which is expected to make you more accurate about a given subject matter and cannot be Dutch-booked on the basis of that subject matter—that nonetheless will predictably polarize you about that subject matter.

3. The (Im)possibility Theorems

What would it take for polarization to be epistemically rational? Being good Bayesians, assume that in any world w (at a given time), the rational opinions for you can be modeled with a probability function P_w. This assumes rational opinions are precise (White 2009; Schoenfield 2012), but it allows varying standards of reasoning across people (Schoenfield 2014) and times (Callahan 2019). Since what is rational to think (what you 'should' think) varies across worlds—with your evidence, priors, and so on—let 'P' be a *description* for 'the rational opinions, whatever they are': in w, it picks out P_w; in x, it picks out P_x; and so on.[17]

How is it rational to *change* opinions? I will not assume any particular mechanism (e.g., that a proposition comes in as evidence). Rather, let an *update* be a pair of (descriptions of) the prior and posterior rational opinions, $\langle P, \widetilde{P}\rangle$: at each world w, you should start out with P_w and end up with $\widetilde{P}_w$. This makes no assumption about mechanism; all it assumes is that the facts about you (priors, evidence, etc.) pin down rational probability functions at the two times (standard Bayesians assume this too). Think of it as 'black-box learning' (Huttegger 2014); we only model the input, P, and output, $\widetilde{P}$.

Our question is which updates $\langle P, \widetilde{P}\rangle$ represent *potentially rational* updates: which could be rational given some rational prior and some learning experience? Bayesians usually say one of three things. (1) Rational updates cannot be *Dutch-booked*: rationally choosing bets before and after the update cannot result in a foreseeable loss (Teller 1973). (2) Rational updates *improve accuracy*: the prior expects the posterior to be (at least or) more accurate than itself, on all reasonable ways of measuring accuracy (Oddie 1997; Greaves and Wallace 2006). (3) Rational updates satisfy the *value of evidence*: given any decision problem, the prior expects the posterior to make a decision that is (at least as good or) better than itself (Ramsey 1990; Blackwell 1953; Good 1967). There

[17] **Notation:** I will use uppercase Roman characters ('P', '$\widetilde{P}$', 'H',...) for descriptions that pick out different functions in different worlds. Their subscripted versions ('P_w', 'P_x',...) and lowercase Greek characters ('π', 'δ',...) will be rigid designators for functions whose values are known. See Schervish, Seidenfeld, and Kadane 2004; Williamson 2008; Dorst 2019.

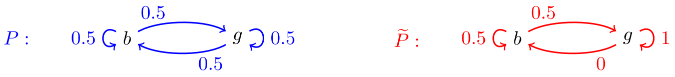
are various ways to formalize these constraints, but Dorst et al. (2021) show that, on arguably the most natural, they are equivalent. Say that P *values* $\widetilde{P}$ if and only if the update $\langle P, \widetilde{P}\rangle$ satisfies these constraints (appendix A.2)—if and only if, in other words, P prefers to give $\widetilde{P}$ power of attorney to make its decisions for it. I assume throughout—with a slight weakening in section 5—that:

> **Valuable rationality:** $\langle P, \widetilde{P}\rangle$ is a potentially rational update iff P values $\widetilde{P}$.[18]

I assume that a *sequence* of updates $\langle P^1, P^2\rangle$, $\langle P^2, P^3\rangle, \ldots$ is potentially rational if and only if each P^i values P^{i+1}. This offers a bright line between the updates that can and cannot be rational: rational ones are those that can be expected to improve accuracy and decision-making.

It is commonly thought that Value (or the avoidance of Dutch books) on its own entails Reflection and hence forbids expectable polarization. It does not:[19]

> **Example** There are two worlds, b and g. We can specify P and $\widetilde{P}$ by saying how, at each world, they distribute credence between b and g. At both, P is 50:50 between b and g. In the bad case (b), $\widetilde{P}$ remains 50:50, but in the good case (g), $\widetilde{P}$ becomes certain of g. We can diagram this by letting an arrow labeled t from x to y indicate (left) that $P_x(y) = t$ or (right) that $\widetilde{P}_x(y) = t$:



> Clearly P values $\widetilde{P}$: at all worlds, $\widetilde{P}$ is either equally accurate (at b) or strictly more accurate (at g) in all propositions. But Reflection fails: at both worlds, P is 0.5 in g, but its expectation of $\widetilde{P}(g)$ is 0.75.[20]

How do Standard Bayesians forbid this? In this model, at world g you learn that you are at g ($\widetilde{P}_g(\cdot) = P_g(\cdot|g)$), while at world b you learn nothing ($\widetilde{P}_b(\cdot) = P_b(\cdot|\{b, g\}) = P_b(\cdot)$). Standard Bayesians will insist that

[18] You might add: "... and there is no available update preferable to $\widetilde{P}$." If so, what I assume is that in my cases the only available updates are to stay with P, switch to a particular π, or switch to $\widetilde{P}$.

[19] This follows from Geanakoplos 1989 (Thm. 1) and is suggested by the assumptions imposed in Skyrms 1990; Huttegger 2014, but as far as I know was not explicit until Dorst 2020 (cf. Williamson 2000: chap. 10).

[20] $\widetilde{P}(g)$ is a random variable with possible values of 0.5 and 1, so, for example, at b its prior expectation is $\mathbb{E}_P(\widetilde{P}(g)) = \mathbb{E}_{P_b}(\widetilde{P}(g)) = \sum_t P_b(\widetilde{P}(g) = t) \cdot t = P_b(\widetilde{P}(g) = 0.5){\cdot}0.5 + P_b(\widetilde{P}(g) = 1){\cdot}1 = P_b(b){\cdot}0.5 + P_b(g){\cdot}1 = 0.5{\cdot}0.5 + 0.5{\cdot}1 = 0.75 \neq 0.5 = P_b(g)$.

the latter is an error: if sometimes you learn g, then when you do not learn g you learn something—namely, that *you did not learn* g. In other words, they assume that rational updates are *introspective*: you always can be rationally sure of what you (did or did not) learn. I will *not* assume that. It fails in the above model; $\widetilde{P}_b$ has *higher-order uncertainty*: it knows that at b it learned nothing, while at g it learned g, but since in fact it learned nothing (it is at b), it does not know what it learned! Thus it is 50:50 on whether $\widetilde{P}$ is 50% or 100% confident of g: $\widetilde{P}_b(\widetilde{P}(g) = 0.5) = \widetilde{P}_b(b) = 0.5$ and $\widetilde{P}_b(\widetilde{P}(g) = 1) = \widetilde{P}_b(g) = 0.5$.

Standard Bayesians may protest that this breaks Bayesianism. It does not. At each world, the rational credences are probabilistic at each time. And Value holds: P expects $\widetilde{P}$ to be more accurate and make better decisions than itself.[21] Mathematically, nothing is broken.

What about philosophically? How to interpret introspection failures? Recall that $\widetilde{P}$ is the posterior credence it is *rational* to have. When $\widetilde{P}$ is uncertain what $\widetilde{P}$ is, that means it is rational to be unsure what the rational opinions are—it is rational to have epistemic self-doubt.[22] Standard Bayesians assume that such self-doubts *could not* be rational:

> **No Ambiguity:** Rational opinions are always sure what the rational opinions are.
> Always, if $\widetilde{P} = \pi$, then $\widetilde{P}(\widetilde{P} = \pi) = 1$. That is, $\forall q, t$: if $\widetilde{P}(q) = t$, then $\widetilde{P}(\widetilde{P}(q) = t) = 1$.

'Ambiguity' is a fitting label. Evidence is *ambiguous* when it is hard to know what to make of it—when it is rational to be unsure what it is rational to think (Ellsberg 1961: 661). This higher-order model of ambiguity follows naturally from 'antiluminous' epistemology, which argues that we often cannot tell exactly what rationality requires of us (see chapters 4 and 10 of Williamson 2000 and Srinivasan 2015). If you endorse antiluminosity, you should permit ambiguity in this sense—and even if you

[21] Moreover, in this model posteriors result from conditioning—namely, on $\{b, g\}$ in b and on $\{g\}$ in g.

[22] Formally, $\widetilde{P}$ fails the axiom $[\widetilde{P}(q) = t] \to [\widetilde{P}(\widetilde{P}(q) = t) = 1]$. Despite doubts (Savage 1954; de Finetti 1977), higher-order uncertainty is mathematically nontrivial whenever this axiom fails (see section A.1 and Samet 2000) and philosophically nontrivial on many interpretations (Lewis 1980; Williamson 2008; Pettigrew and Titelbaum 2014; Salow 2018; Dorst 2019, 2020; Das 2022a, 2023; Levinstein 2022; Levinstein and Spencer 2022).

have doubts about antiluminosity in general, there is reason to permit ambiguity (Elga 2013; Dorst 2019; Carr 2020).[23]

Ambiguity is consistent with knowing your actual opinions: since $\widetilde{P}$ represents the *rational* posteriors, it is distinct from your *actual* posteriors $\widetilde{C}$. Even if you are rational ($\widetilde{C} = \widetilde{P}$ at the actual world) and know what your credences are ($\widetilde{C}$ knows what $\widetilde{C}$ is), you can doubt that your credences are rational ($\widetilde{C}$ leaves open worlds where $\widetilde{C} \neq \widetilde{P}$). See Dorst 2019.

No Ambiguity is the assumption that makes Value and Reflection equivalent (appendix A.3):

Theorem 3.1. *Given No Ambiguity, P values $\widetilde{P}$ iff P obeys Reflection toward $\widetilde{P}$.*

This is an impossibility result: any theory of rational (expectable) polarization must deny either Value or No Ambiguity.

I know of no proposals that deny No Ambiguity.[24] In fact, an update is Standard Bayesian—the result of conditioning a fixed prior on the true cell of a partition—if and only if it satisfies both No Ambiguity and Value (Theorem A.1). This is why none of the models in footnote 9 permit expectable polarization: they are Standard Bayesian, so they impose Reflection.

Meanwhile, extant models that *allow* expectable polarization do so using updates that violate Value, so they are subject to Dutch books and are expected to make you less accurate.[25] What to make of this? If

[23] Bayesians usually model ambiguity differently, either using an 'imprecise' *set* of probability functions (Levi 1974; Seidenfeld and Wasserman 1993; Joyce 2010; cf. Moss 2018) or positing an introspective $\widetilde{P}$ that is unsure about a *different*, more ideal (introspective) P^* (Camerer and Weber 1992; Klibanoff, Marinacci, and Mukerji 2005). Such models either violate Value (e.g., Kadane, Schervish, and Seidenfeld 2008; Baliga, Hanany, and Klibanoff 2013; Bradley and Steele 2016) or mimic standard Bayesianism (e.g., Das 2022b) in a way that yields Reflection.

[24] Salow 2018—who I take inspiration from—uses expectable polarization to argue *for* No Ambiguity.

[25] For example, Kadane, Schervish, and Seidenfeld 1996; Rabin and Schrag 1999; Hegselmann and Krause 2002; DeMarzo, Vayanos, and Zwiebel 2003; Halpern 2010; Flache and Macy 2011; Andreoni and Mylovanov 2012; Baliga, Hanany, and Klibanoff 2013; Wilson 2014; Baumgaertner, Tyson, and Krone 2016; Proietti 2017; O'Connor and Weatherall 2018; Fryer, Harms, and Jackson 2019; Loh and Phelan 2019; Singer et al. 2019; Stone 2020; van der Maas, Dalege, and Waldorp 2020; Weatherall and O'Connor 2020; Zollman 2021.

we allow *non*valuable updates to be 'rational', the standard storytellers might fairly complain that we have moved the goalposts. For example, some argue that allowing evidence to be *permissive*—open to multiple rational interpretations—nullifies worries about predictably polarizing influences.[26] Theorem 3.1 entails that such predictable shifts can be expected to make you less accurate. The natural complaint: what distinguishes this from irrational forms of (say) motivated reasoning?

The way around the impossibility result is to allow ambiguity (see appendix A.3):

Theorem 3.2 (Informal). *Whenever $\widetilde{P}$ is ambiguous but valued by some P, Reflection fails.*

This shows that the above example generalizes: *whenever* evidence is ambiguous, Reflection can fail for valuable updates. It is our possibility proof: expectable polarization *could* be valuable—hence (I say) rational.

The upshot: assuming that the rational updates are the valuable ones, there is a tight theoretical connection between rational expectable polarization and ambiguity—the former is possible if (Theorem 3.2) and only if (Theorem 3.1) the latter is.

Intriguingly, there is also a tight *empirical* connection between polarization and ambiguity. The intuitive cases of rational self-doubt—what I call 'ambiguity'—are ones in which people face complicated evidence, have peers who disagree with them, or have reason to doubt their own reasoning.[27] These are also the cases in which there is the strongest *psychological* evidence for expectable polarization. People are most inclined to engage in 'biased processing'—seeing evidence in ways that fit with their prior beliefs—when evidence is mixed, complex, or hard to interpret (e.g., Lord, Ross, and Lepper 1979; Kunda 1990; Kahan et al. 2017; see section 6). These effects are exacerbated by group discussions, where peer (dis)agreements have large effects on people's opinions (e.g., Isenberg 1986; see section 7). And when the evidence is made easier to interpret or discussion norms are altered, biased processing

[26] For example, Schoenfield 2014; Podgorski 2016; Simpson 2017; Callahan 2019; Ye 2019; Jackson 2021.

[27] See the 'higher-order evidence' literature, for example, Christensen 2010; Lasonen-Aarnio 2013, 2014, 2015; Horowitz 2014, 2019; Schoenfield 2015, 2018; Sliwa and Horowitz 2015; Fraser 2022; Dorst forthcoming gives a summary.

often disappears (Lundgren and Prislin 1998; Grönlund, Herne, and Setälä 2015; Anglin 2019).

In short, people tend to predictably polarize in exactly the situations where self-doubts seem rational. What if it's not a coincidence?

4. The Mechanism

In principle, ambiguous evidence could rationalize expectable polarization. But are there realistic mechanisms that generate it? And can they generate *predictable* polarization?

There are, and they can. Consider a *word-search task* (cf., Elga and Rayo 2022). Given a string of letters and some blanks, you have a few seconds to figure out whether there is an (English) completion. For example:

P_A_ET

And the answer is yes, there is a completion. Another:

P_G_ER

And the answer is no, there is no completion.

A word-search task involves *cognitive search* (Todd et al. 2012): searching an accessible cognitive space for a particular type of item. Other cases include searching your memory for an example, your reasoning for a flaw, or your knowledge for a proof. This involves calling on background knowledge. Intuitively, sometimes you know you have done this rationally, other times you do not. If you *find* a completion ('PLANET!'), you (often) know that it is rational to be certain there is a word (that $\widetilde{P}(\mathit{Word}) = 1$). But if you *do not* find a completion, you do not know how confident to be: "Maybe I should be doubtful (maybe $\widetilde{P}(\mathit{Word})$ is low), but maybe I'm missing something obvious (maybe $\widetilde{P}(\mathit{Word})$ is high)." I argue that this generates an ambiguity asymmetry between completable and uncompletable searches, rationalizing expectable polarization. In section 5, I turn to *predictable* polarization.

Meet Haley. She is wondering whether a fair coin landed heads. I will show her a word search determined by the outcome: if heads, it will be completable; if tails, it will be uncompletable. Thus her credence in heads equals her credence that it is completable. She will have 7 seconds, then she will write down her credence. She knows all of this.

Let H and $\widetilde{H}$ be the rational prior and posterior for Haley. She should initially be 50:50 on heads: $H(\textit{Heads}) = 0.5$. But I claim her estimate for her posterior rational credence should be *higher* than 50%: $\mathbb{E}_H(\widetilde{H}(\textit{Heads})) > 0.5$. Remember: estimates are not predictions, so she need not be confident her credence should go up. Rather, expectable polarization means that across many identical trials, she should be confident that the *average* posterior rational credence will be above 50%. Why? Intuitively, it is easier for her to assess her evidence when the string is completable (when the coin lands heads) than when not. So if heads, her credence should (on average) increase a lot; if tails, it should (on average) decrease a bit; and the average of 'increase a lot' and 'decrease a bit' is 'increase a bit'.

Standard Bayesians will balk. They will say that we must find the most fine-grained question (partition) Q that Haley can always answer with certainty, and that she is rational if and only if she conditions on the true answer to Q. It is as if she rummages around in her head for a completion; at the end *all she learns* is either that the search succeeded (*Find*) or failed ($\neg$*Find*), so $Q = \{\textit{Find}, \neg\textit{Find}\}$. (If she learns more, they will insist that there is a finer-grained Q to update on.) As we know from Theorem 3.1, such a model forbids expectable polarization. For example, suppose Haley thinks that if there is a word, she will find it half the time ($H(\textit{Find}|\textit{Word}) = 1/2$), and if there is not, she will never find one ($H(\textit{Find}|\neg\textit{Word}) = 0$). Then $1/4$ of the time she will learn *Find* ($1/2$ likely to be a word and $1/2$ likely to find if so), making it rational to be sure there is a word: $\widetilde{H}(\textit{Word}) = 1$. (And she will *know* this is the rational reaction: $\widetilde{H}(\widetilde{H}(\textit{Word}) = 1) = 1$.) The remaining $3/4$ of the time she will learn only $\neg$*Find*, making it rational to slightly lower her credence: $\widetilde{H}(\textit{Word}) = 1/3$.[28] (And since she will know all she learned is that she did not find one, she will *know* that this is the rational reaction: $\widetilde{H}(\widetilde{H}(\textit{Word}) = 1/3) = 1$.) Thus her expected future rational credence is $1/4 \cdot 1 + 3/4 \cdot 1/3 = 1/2$. There is no expectable polarization.

I object. It is implausible to insist that such a model is *always* correct. As I have argued, that does not follow from the justifications of Bayesianism (section 3). Moreover, it rules out the possibility of ambiguity, so it ignores the most salient feature of a word search: that it is easier to know what to make of your evidence when you have found a word than when you have not.

[28] $\widetilde{H}(\textit{Word}) = H(\textit{Word}|\neg\textit{Find}) = \frac{H(\textit{Word}\&\neg\textit{Find})}{H(\neg\textit{Find})} = \frac{1/4}{3/4} = 1/3$.

Reflect on your experience with another example:

_E_RT

When you have not found one, your mind is racing ('Beurt? No... teart? No...'), your credence is oscillating ('Probably... no wait, maybe not. Oh I got it! Wait, no...'), and you have the nagging sensation that maybe you are missing something obvious. If you have not found one when the 7-second timer goes off, your credence that there is a word may have gone down or gone up, but you will not (should not!) be willing to bet the farm that it has moved in the rational direction. After all, sometimes it does not: if your credence went down to 1/3, and then I whisper 'heart', you might think, 'Oh! I should've seen that...'. It was rational for you to have more than 1/3 credence in a completion; after all, you know that 'heart' is a word—you just failed to make proper use of that knowledge.

Given that sometimes you are irrational, what about when you have *in fact* been rational to lower your credence? You should still wonder whether you have been *ir*rational. For example, if you do not find a completion to ST_ _RE and so drop your credence to 1/3, you might still wonder whether there is a word and (so) wonder whether you should have a higher credence—even though in fact there is not, so in fact you should not. Rational people can doubt that they are rational, just as humble people can doubt that they are humble.

These are intuitions. If we could not make precise sense of them, perhaps they could be ignored. But we can—just introduce ambiguity. Here is one way to do so. There is more that Haley (is and) should be sensitive to than what she can settle with certainty. Beyond whether she found a completion, there is the question of whether the string is 'word-like'—whether it contains subtle hints that it is completable. If it does, she should increase her credence that it is completable; if it does not, she should decrease it. But—and here is where ambiguity comes in—she cannot always tell with certainty whether it is word-like, and hence she cannot always tell whether her credence should go up or down.

Here is a simple model (details in section 4.1). Suppose, as before, it is 1/2 likely that there is no word (and so she does not find one), 1/4 likely there is a word she finds, and 1/4 likely there is a word she does not find. Moreover, suppose she knows the string will be word-like if and only if there is a word. If she finds a word, she is rational to become certain there is one: $\widetilde{H}(\mathit{Word}) = 1$. If she does not find a word *and there is none* (so it is not word-like), she is rational to drop her confidence slightly:

$\widetilde{H}(\mathit{Word}) = 1/3$. So far this is just like the Standard Bayesian model. Yet suppose that if she does not find a word *but there is one* (so the string is word-like), she is rational to raise her credence slightly—she should suspect it is word-like: $\widetilde{H}(\mathit{Word}) = \frac{2}{3}$. This yields ambiguous evidence: if she does not find a word, she is rational to be unsure whether the rational posterior is $1/3$ or $2/3$: $\widetilde{H}(\widetilde{H}(\mathit{Word}) = 1/3) > 0$ and $\widetilde{H}(\widetilde{H}(\mathit{Word}) = 2/3) > 0$. (Which one it is depends on whether the string is word-like, but she is *also* rational to be unsure of that. There is no cognitive home; see section 4.1, and Williamson 2000; Srinivasan 2015.)

Two facts about this model. First, her prior H values her posterior $\widetilde{H}$. In fact, this ambiguous update is better than the Standard Bayesian one: if she finds a word, both update to credence 1; if there is no word, both update to credence $1/3$; but if there is a word she does not find, the Standard Bayesian updates to credence $\frac{1}{3}$, while the ambiguous model updates to $\frac{2}{3}$. The ambiguous update is never less accurate and is sometimes more accurate.[29] Thus neither is Dutch-bookable, and it is always rational to prefer the ambiguous one (see section 4.1).

Second, this update is expectably polarizing: Haley is initially 0.5 confident there is a word, but her estimate of the future rational credence is roughly 0.58.[30] Notice why. Uncompletable searches are more likely to generate ambiguity than completable ones. So although the rational opinions always move toward the truth, they (on average) move further if the string is completable than if it is not. It is *asymmetric increases in accuracy* that lead to polarization.[31]

This is just one simplified model of how word searches could generate ambiguity. Here, $\widetilde{H}$ may best be interpreted as the *average* rational credence to have across cases since in realistic models there would be a much wider range of possibly rational posteriors. Appendix A.4 proves

[29] Is the comparison unfair since the ambiguous posterior differs in more places than the Standard Bayesian one? Insisting it is unfair presupposes that if people can distinguish between two possibilities *at all*, they can distinguish them *with certainty* (Greaves and Wallace 2006; Huttegger 2013; Schoenfield 2017; Gallow 2021; Isaacs and Russell 2022; Zhang and Meehan 2022). That, in turn, forbids ambiguous evidence (since it implies that $\widetilde{P}$ is available only if its 'informed' version is; see appendix A.3, Theorem A.2). So although this is a way to object, if you are onboard with ambiguous evidence you should not worry about such 'unfairness'.

[30] $\mathbb{E}_H(\widetilde{H}(\mathit{Word})) = H(\widetilde{H}(\mathit{Word}) = 1/3) \cdot 1/3 + H(\widetilde{H}(\mathit{Word}) = 2/3) \cdot 2/3 + H(\widetilde{H}(\mathit{Word}) = 1) \cdot 1 = \frac{1}{2} \cdot 1/3 + \frac{1}{4} \cdot 2/3 + \frac{1}{4} \cdot 1 \approx 0.58$.

[31] $\mathbb{E}_H(\widetilde{H}(\mathit{Word}) | \mathit{Word}) = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 2/3 \approx 0.83$, while $\mathbb{E}_H(\widetilde{H}(\mathit{Word}) | \neg \mathit{Word}) = 1/3 \approx 0.33$, so if it is completable, the average rise is $0.83 - 0.5 = 0.33$, and if it is uncompletable, the average drop is $0.33 - 0.5 = -0.17$.

that a wide class of such models will lead to expectable polarization—so even if you object to the details, I hope you will agree that updates like this are possible.

I claim that these expectably polarizing updates can be rational. But I *also* claim (and will argue in sections 6 and 7) that they might drive polarization of *actual* opinions. How, in theory, could expectable polarization in the opinions that are *rational* for Haley ($\widetilde{H}$) lead to polarization in her actual opinions? There are a variety of answers, but the simplest is that if Haley is approximately rational, her actual opinions will be a noisy indicator of the rational ones—thus her actual opinions will expectably polarize too.[32]

In theory. How can we test the hypothesis that ambiguous evidence can polarize real people? Meet Thomas. Like Haley, he is about to see a word search determined by the (same) coin toss. But while she will see a completable string if and only if heads, he will see a completable string if and only if *tails*. By parallel reasoning, Thomas's opinion should expectably polarize in the opposite direction: it will be easier for him to assess his evidence if tails than if heads, so his average posterior rational credence in heads should be *lower* than 50%.

In fact, meet everyone. Half are *Headsers*: like Haley, they will see a completable string if and only if heads. The rest are *Tailsers*: like Thomas, they will see a completable string if and only if tails. Headsers get evidence that is easier to assess when the coin lands heads; Tailsers get evidence that is easier to assess when the coin lands tails. So if the coin lands heads, the average Headser should be confident it did, while the average Tailser should be unsure. If it doesn't land heads, the average Tailser should be doubtful it did, while the average Headser should be unsure. Since all start out 50%, they can predict that they will split apart.

Do they? I've tested this in two ways. The fun way: audiences. In 6 of 7 talks, Headsers had a higher average posterior in heads. The rigorous way: an experiment. Across trials, there was a significant (and large) difference in the average posterior credence in heads (Headsers: 57.7%, Tailsers: 36.3%; $p < 0.001$; $d = 1.57$; see section 4.2). This does not establish that the participants themselves could predict how they would

[32] If her actual opinion, $\widetilde{C}(\mathit{Word})$, is an unbiased estimator of the rational opinion (meaning $\forall t : \mathbb{E}_H(\widetilde{C}(\mathit{Word})|\widetilde{H}(\mathit{Word}) = t) = t$), then it will expectably polarize to the same degree. If it is a biased estimator, it may still polarize depending on the degree and direction of the bias.

polarize, but it does support a necessary precondition of that—namely, that *I* could predict it.

More work is needed. But we have now seen—in principle and perhaps in practice—how cognitive search could generate ambiguities that rationalize expectable polarization. What of *predictable* polarization? If you would like to jump to that argument, skip to section 5; for the technical (section 4.1) or experimental (section 4.2) details from this section, read on.

4.1. The Formalities

Figure 1 specifies the Standard Bayesian model of the word search in two forms: the left in a generalized-Kripke (or Markov) diagram, the right in stochastic-matrix notation. (See appendix A.1 for formal semantics.) w_1 and w_2 are where Haley does not find a word; w_3 is where she does. The rational prior is always $(1/2\ 1/4\ 1/4)$ over (w_1, w_2, w_3). In w_1 and w_2, the rational posterior shifts to $(2/3\ 1/3\ 0)$ (conditioning on $\neg$*Find*), and in w_3, it shifts to $(0\ 0\ 1)$ (conditioning on *Find*). No Ambiguity holds because the posterior is constant within worlds it leaves open: $\widetilde{H}_i(j) > 0$, then $\widetilde{H}_i = \widetilde{H}_j$.

The ambiguous model (figure 2) is identical except that in w_2 the rational posterior assigns higher credence to there being a word (the

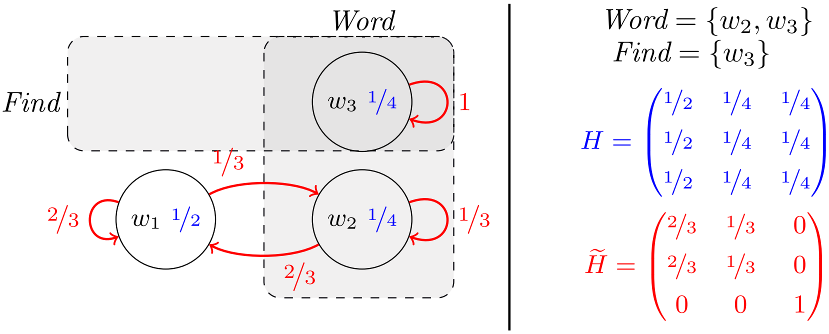
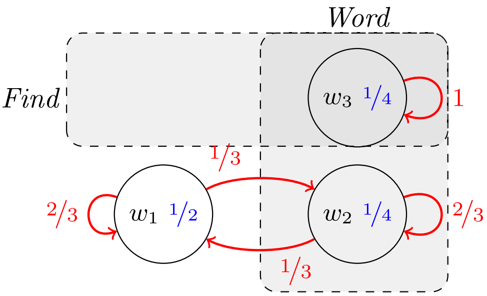

Word
Find
w_3 $1/4$
1
$1/3$
$2/3$
w_1 $1/2$
w_2 $1/4$
$1/3$
$2/3$
$Word = \{w_2, w_3\}$
$Find = \{w_3\}$
$H = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 1/4 & 1/4 \\ 1/2 & 1/4 & 1/4 \end{pmatrix}$
$\widetilde{H} = \begin{pmatrix} 2/3 & 1/3 & 0 \\ 2/3 & 1/3 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

Figure 1. (Color online.) Standard Bayesian model of a Headser's rational opinions. Left: Generalized-Kripke (Markov) diagram, in which blue numbers within circles represent the prior probabilities of possibilities, and labeled red arrows from circles represent the posterior probabilities *in* those possibilities. Right: The matrix H represents (constant) prior probabilities; the matrix $\widetilde{H}$ represents posteriors: row i column j is the probability, in world i, that it is rational to assign to being in world j. Thus the third row of $\widetilde{H}$ says what Haley's probabilities should be if she finds a word; the second row says what they should be if it is completable but she does not find one, etc.



$$Word = \{w_2, w_3\}$$
$$Find = \{w_3\}$$

$$H = \begin{pmatrix} ^1/_2 & ^1/_4 & ^1/_4 \\ ^1/_2 & ^1/_4 & ^1/_4 \\ ^1/_2 & ^1/_4 & ^1/_4 \end{pmatrix}$$

$$\widetilde{H} = \begin{pmatrix} ^2/_3 & ^1/_3 & 0 \\ ^1/_3 & ^2/_3 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Figure 2. (Color online.) Ambiguous model of a Headser's rational opinions. See figure 1 for interpretation.

string is word-like). Thus $\widetilde{H}_{w_1} \neq \widetilde{H}_{w_2}$, and since $\widetilde{H}_{w_1}(w_2) > 0$, the evidence is ambiguous: if Haley does not find a word, she should be unsure whether the rational credence is $^1/_3$ (as it is at w_1) or $^2/_3$ (as it is at w_2).

Four comments. First, the ambiguous update is preferable to the Standard Bayesian one, since it is identical at w_1 and w_3 and strictly more accurate at w_2—see footnote 29 for why the comparison is fair. (Appendix A.4 proves that this model validates Value—hence is not Dutch-bookable and is expected to improve accuracy.) Second, the ambiguous model violates Reflection: $\mathbb{E}_H(\widetilde{H}(Word)) = {}^1/_2 \cdot {}^1/_3 + {}^1/_4 \cdot {}^2/_3 + {}^1/_4 \cdot 1 = \frac{7}{12} \approx 0.58 > 0.5 = H(Word)$. Third, note that this update cannot be modeled by conditioning ($\widetilde{H}_{w_2}$ is the culprit). This is for simplicity; we can also generate valuable expectable polarization using conditioning updates, as we have seen in the example in section 3.[33]

Fourth, you might be puzzled: How is Haley in a position to be $^2/_3$-confident of *Word* in w_2, but only $^1/_3$ in w_1? Because she receives different signals in the two—'word-like' in w_2 and 'not world-like' in w_1. Why, then, can she not be *sure* that there is a word in w_2? Because she cannot be *sure* which signal she received—in w_2, she can only be $^2/_3$-confident that she received 'word-like'. Well, in w_2 can she be *sure* that she can be $^2/_3$-confident she received 'word-like'? No. Look at the model; she can only be $^2/_3$-confident that she can be $^2/_3$-confident she received 'word-like'. (And so on.) . . .Okay, but if she cannot be *sure* she received 'word-like',

[33] If we assume all updates happen by conditioning, ambiguity occurs if and only if the possible propositions that might be rational to condition on do not form a partition. See, for example, Geanakoplos 1989; Williamson 2000: chapter 10; Salow 2018; Dorst 2020; Das 2019; Isaacs and Russell 2022; Zendejas Medina Forthcoming.

how can she be sensitive to whether it is word-like? The same way you can be humble without knowing you are, or can understand my argument without being sure you have. It is only by implicitly assuming that facts about rationality are introspectable—that you can always know what the rational opinions are or what signals you received—that the puzzle arises.

4.2. The Experiment

Here I sketch an experiment that both suggests that cognitive search can cause people to polarize, and controls for a confound. (See appendix B for more details.)

The confound: ambiguous evidence is not simply *weak* evidence. Evidence is weak when it should not move your opinions very much; evidence is ambiguous when *you should not be sure how weak it is.* Highly ambiguous evidence must be weak (Dorst 2020: Fact 5.5), but evidence can be weak without being ambiguous. Figure 3 gives an example. Urn A contains one black and one red marble; urn B contains two red marbles. I flip a coin, grabbing A if heads and B if tails. Then I draw a marble and show you. A black marble is strong evidence: you should be sure I am holding A. A red marble is weak evidence: you should slightly boost your confidence that I am holding B. Either way, it is unambiguous: if it is black, you should know that you should be sure I am holding urn A; if it is red, you should know that you should be $2/3$ confident I am holding urn B. You should not have self-doubt.

The upshot is that the strong/weak asymmetry is not the unambiguous/ambiguous asymmetry. In the word search, both are present. If the string is completable, you can get unambiguous evidence that it is; if it is not, you get ambiguous evidence that it is not. But also, if the string is completable, you can get *strong* evidence that it is; if it is not, you get weak evidence that it is not. My theory predicts that the *ambiguity* asym-

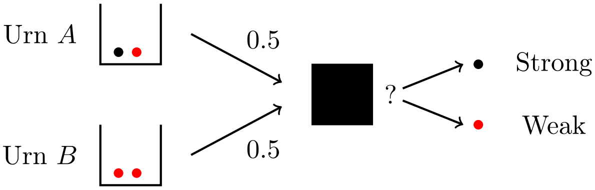


Figure 3. A case of unambiguous (but sometimes weak) evidence.

metry drives polarization, but what if the weak/strong asymmetry does? What if people polarize because they underreact to weak evidence?[34]

The experiment tested this[35] in a 2 × 2 design that independently manipulated both valence (Headser vs. Tailser) and ambiguity (ambiguous vs. unambiguous). Headsers sometimes got strong evidence when a coin landed heads and always got weak evidence when it landed tails (Tailsers vice versa). The ambiguous condition saw word-search tasks; the unambiguous condition saw marble draws from urns. I predicted more polarization in the ambiguous than unambiguous condition.

It worked. Each subject saw four bits of evidence, determined by four different coin flips. Figure 4 shows how the mean subject's average confidence in $Heads_1, \ldots, Heads_4$ evolved as they saw evidence about each flip: at 0, this is the average of their priors in each toss; at 1, this is the average of their posterior in the first toss (having seen the first bit of evidence) and their priors in the remaining three, and so on. The ambiguous condition polarized,[36] and did so significantly more than the unambiguous one.[37] Appendix B reviews evidence that ambiguity explains this effect—including the minor polarization in the 'unambiguous' condition.

5. The Predictable Theorem

We have seen (valuable, so I say) rational expectable polarization. But what about *predictable* polarization—the fact that when I went off to college, I could predict with confidence that I would come to doubt that guns increase safety? Since estimates are not predictions, this does not follow from what we have seen so far. Can we go further?

[34] There is indeed some evidence that people are *conservative* in this sense (Peterson and Beach 1967; Edwards 1982), though this may be due to a failure to believe the experimental setup (Corner, Harris, and Hahn 2010; Hahn and Harris 2014)—a source of ambiguity.

[35] Preregistration: https://aspredicted.org/8jg3e.pdf.

[36] One-sided *t*-test: $t(101) = 7.98$, $p < 0.001$, $d = 1.577$; the bootstrapped 95% confidence interval for the *difference* in posterior confidence between the two groups was $[16.02, 26.82]$.

[37] A 2 × 2 ANOVA indicated a main effect of valence ($F(1, 224) = 68.99$, $p < 0.001$, $\eta^2 = 0.217$), a main effect of ambiguity ($F(1, 224) = 6.39$, $p = 0.012$, $\eta^2 = 0.020$), and an interaction effect between the two ($F(1, 224) = 21.63$, $p < 0.001$, $\eta^2 = 0.068$). A bootstrapped 95% confidence interval for the difference of differences, that is, for (A-Headsers − A-Tailsers) − (U-Headsers − U-Tailsers), was $[7.19, 22.59]$, indicating that the former was larger.

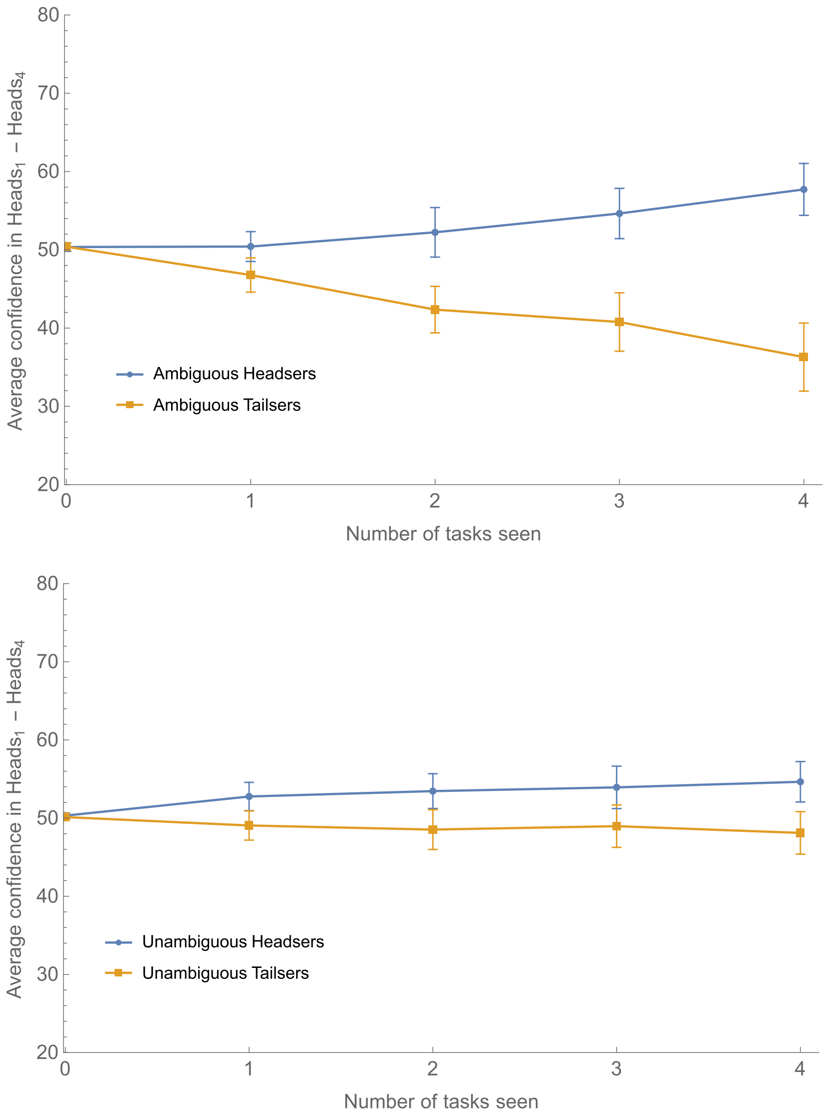


Figure 4. Means of participants' average confidence in *Heads*$_1$,. . . *Heads*$_4$ as they saw more tasks, in ambiguous (left) and unambiguous (right) conditions. Error bars represent 95% confidence intervals.

Yes and no. 'No' because full Value forbids it. 'Yes' because there is a weakening of Value—which we already knew we would have to make—that allows it.

The basic idea is to iterate cognitive searches. In the model from section 4, Haley knows the coin is fair but rationally estimates that the rational posterior is around 58%. So *if* we can repeat this with many independent fair coins and searches, since she is initially confident that around half the coins will land heads, she predicts that her average credence in $Heads_1, \ldots, Heads_n$ should rise to around 58%. Since they are independent, she can predict that she should become confident that *around 58% landed heads* and very confident that *more than half landed heads*. Since she is initially 50:50 in the latter, that is predictable polarization.

But there's a hitch. *Can* we iterate cognitive searches, given Value? Suppose the rational opinions for Haley go from H^1 to H^2 to… to H^n. I have shown how an individual step could be valuable despite being expectably polarizing. But ignore the steps; focus on the beginning and end. Let h be the claim that *more than half the coins landed heads.* If we can iterate, then at the beginning Haley can predict with (say, 90%) confidence that she should wind up (say, 90%) confident of h: $H^1(H^n(h) \geq 0.9) \geq 0.9$. It follows immediately that her initial opinions (H^1) do *not* value her final opinions (H^n); since she is should initially be 50% confident of h, she must think that almost half the time, the final 90% confidence will be misplaced! Thus she expects H^n to be less accurate about h than her initial opinions.[38]

Though it is not obvious, this implies that for some i, H^i does not value H^{i+1}, for Theorem A.4 (appendix A.5) shows that Value is 'transitive': if H^1 values H^2 and H^2 values H^3, then H^1 values H^3. (If we tried to simply iterate our model from section 4, then H^2 would not value H^3.)[39] You might understandably get off the boat here, insisting that epistemic rationality requires full Value, allowing expectable but forbidding predictable polarization.

But should you? We already knew that people do not obey full Value; after all, they sometimes forget things. And here is an easy theorem: if Haley might forget something—*anything*—then she cannot value her future opinions.[40]

[38] Formally, H^1 values H^n only if $H^1(H^n(h) \geq t) \geq s \Rightarrow H^1(h) \geq t \cdot s$ (Dorst 2020: Fact 5.5). So if $H^1(H^n(h) \geq 0.9) \geq 0.9$, we must have $H^1(h) \geq 0.9 \cdot 0.9 = 0.81 > 0.5$.

[39] If the first update was valuable, why would another copy fail to be? Because ambiguity can compound problematically—see discussions of 'double-bad-cases' in Williamson 2019: section 4; Das 2023; Dorst 2020: section A.1.

[40] If $H(q) = 1$, then H values $\widetilde{H}$ only if $H(\widetilde{H}(q) = 1) = 1$.

Forgetting is never ideal. Is it also always irrational? Surely not. Some things—like Mom's birthday—are bad to forget. Others—like what you ate last Tuesday—are not. The former are questions whose answers you should care about getting right; the latter are not. As stated, Value ignores this distinction: H values $\widetilde{H}$ if and only if for *any* decision problem, it prefers to let $\widetilde{H}$ decide; if and only if for *every* question, it expects $\widetilde{H}$ to be more accurate than itself; if and only if there is *no* subject matter the update can be Dutch-booked on.

That is a high bar. Most forms of deference are *question relative.* You defer to the forecaster about whether it will rain but not about whether your poncho is stylish; you defer to your future-self about how busy you will be next month but not about what you had for breakfast this morning. Value can be question relative too (Dorst et al. 2021). A *question Q* is a partition of logical space (Hamblin 1976)—a division of possibilities into groups that agree on the answer to Q. (E.g., "Will it rain tomorrow?" = $\{Rain, \neg Rain\}$.) *H values $\widetilde{H}$ with respect to Q* if and only if, for any decision *whose outcomes are determined by the answer to Q*, it prefers to let $\widetilde{H}$ decide. This entails that the update cannot be Dutch-booked *using bets about Q*, and it entails that H expects $\widetilde{H}$'s opinions *about Q* to be more accurate. (See section 5.1.)

Let's lower the bar. Fix the most fine-grained Q you (should) care about. I propose that if you should value an update with respect to Q, then it is a rational update:

> ***Q*-valuable rationality:** $\langle P, \widetilde{P} \rangle$ is a potentially rational update iff P values $\widetilde{P}$ with respect to the most fine-grained question Q that you should care about.

After all, if you should not care about a question, why must you expect to become more accurate about it in order to update rationally? You might object that such updates are not *fully* 'rational'. Still, you probably assumed that requiring *each* update to be expected to increase accuracy about Q would lead to expected *long-run* increases in accuracy about Q, guarding against predictable polarization about Q. I will show that it does not.[41]

[41] Although a variety of models show how limited memory can lead to polarization (Wilson 2014; Dallmann 2017; Fryer, Harms, and Jackson 2019; Loh and Phelan 2019; Singer et al. 2019), they all require losing information about (hence require updates that are not valuable with respect to) the question you polarize on.

Here's why. H can value $\widetilde{H}$ with respect to Q even if $\widetilde{H}$ forgets some things, so long as that forgetting does not affect $\widetilde{H}$'s opinions about Q. This yields one way of iterating cognitive searches.[42] Let Q be the question of *how all the cognitive searches went*, including whether Haley found a word and whether the coin landed heads or tails. Suppose this—so any question answered by it, for example, whether more than *h*alf landed heads (h)—is what Haley should care about. Each time she is presented with a string, she updates as discussed in section 4 (figure 2). Such updates satisfy (full) Value. But she knows that, after each, she will forget the letter-string (the details of the evidence she received). This forgetting does not affect her opinions about how the cognitive search went, so is valuable with respect to Q. What it does is *consolidate* her ambiguity. When she initially does not find the completion, she is left wondering whether the string is word-like, and hence whether she should be 1/3 or 2/3 confident it is completable. But once she forgets the string, she knows she can no longer be sensitive to whether it is word-like, and so she knows the rational way to respond to her (now impoverished) evidence is simply to stick with the opinion she ended up with. This consolidation of her ambiguity makes it so that when the next cognitive search comes around, she can again update as in section 4 and satisfy (full) Value. Rinse and repeat.

The main theoretical result of this article is that each step in this process is expected to make Haley more accurate about Q despite the whole sequence predictably polarizing her:

Theorem 5.1 (Informal). *Haley can start out 50% confident of h, know that each update in a sequence is valuable with respect to how all the coins land (hence whether h), and yet predict with arbitrary confidence that the sequence will make her arbitrarily confident of h.*

This is an epistemic form of a diachronic tragedy (Hedden 2015): at each stage, she expects the *next* step to make her more accurate and *later* ones to make her less so—despite knowing that once she takes the next step, she will *then* expect the later ones to make her more accurate

[42] An alternative strategy is to allow the question Haley cares about to (predictably) change across times: at time i, Haley cares only about outcome of the ith word-search task, so rationally does each word search. This avoids any forgetting but has the downside that Haley does not care about the claim (h) that she is predictably polarizing on throughout the process (see appendix A.6).

and so will be willing to take them. This is the slippery slope to radicalization.

More is true. If Thomas goes through the Tailser version of this process, the resulting polarization is also *persistent*: when Haley and Thomas discover that they have shifted in opposite directions, their now-polarized opinions remain extreme (Corollary 5.3).

What, intuitively, is happening? Initially Haley wants to do the first search (since it will give her an inkling about Q), but does not want to do the first two—for doing so might generate too much ambiguity to be valuable. Suppose she does the first and does not find a word, so she is left with ambiguous evidence ("Should I be 1/3 or 2/3 confident that there is a word?"). At this stage, she does not want to do the second. Then she forgets the first string, maintaining her opinions about Q but consolidating her ambiguity ("Okay, *now* I should be 2/3"). She thus stops worrying that the second search will yield too much ambiguity, and since it will give her more of an inkling about Q, she prefers to do it. And so it goes. . .

Since the (fair) coins are all independent, initially Haley is 50:50 on whether more than half will land heads and is quite confident that roughly half of them will. As the process unfolds, there are tosses (say, $Heads_2$, $Heads_5$,. . .) that she becomes sure landed heads (she finds completions). For the rest, her evidence was ambiguous, so she tends to have middling degrees of confidence—some slightly below 50%, others slightly above it. Across trials, her average credence in the $Heads_i$ rises to roughly 58%. To maintain coherence, she must therefore come to think that it is very likely that more than half the coins landed heads.

Of course, she predicted this rise in confidence. But so what? She had no idea *which* $Heads_i$ her credence would rise or fall in. Using the only evidence she has (the word searches), her confidence has risen a lot in some, risen a bit in others, and fallen slightly in still others. She cannot conclude that the ones it has fallen in landed tails—that would require assuming she has been rational, which she cannot be confident of. Thus the fact that she *initially* predicted that half would land heads cannot be used as a basis to lower her credence; in fact, she becomes progressively less confident in that prediction as the process unfolds. Thus Haley finds herself confident that more than half landed heads, with no rational way to lower that confidence. (She should expect lowering her credence in any of the $Heads_i$ to *decrease* her accuracy.)

Peeking over her shoulder, she notices that Thomas is now extremely *doubtful* that more than half the coins landed heads. But so what? She predicted as much from the outset, so it does not provide much evidence. Her confidence persists.[43]

They *can* reduce (but not eliminate) their disagreement if they start sharing which completions they found. But that is an exacting exercise: it takes the patience to talk through—and the ability to recall—the individual reasons underlying their opinions about *h*. Since time and memory are limited, Haley and Thomas may be left disagreeing about high-level claims (*most of the coins landed heads*) while being unable to share all the (rational) reasons they have for their differing opinions.

The upshot: predictable polarization could indeed be rational.

What, abstractly, is the structure that generates it? We need a 'high-level' target claim (e.g., *most of the coins landed heads*). We need a large collection of individual facts that bear on the target claim (e.g., the outcomes of individual coin tosses). We need the evidence about each such fact to be asymmetrically ambiguous *in different directions* for two groups—one group (Headsers) must be better at recognizing when a fact points one way ($Heads_i$); the other (Tailsers) must be better at recognizing when it points the other way ($Tails_i$). We expect discussion of individual facts to lead to (rough) agreement about which way those facts point. However, the opposing groups' high-level opinions are shaped by many more facts than they can recall or discuss—thus their asymmetric sensitivities leave them strongly disagreeing about the high-level claim.

To me, this feels familiar. Let's tell a better story.

For $Heads_i$ and $Tails_i$ substitute bits of evidence for and against the claim that *guns increase safety*. Going to a liberal university made me a Tailser—made me better at recognizing evidence against that claim. Living in a conservative town made Dan a Headser—made him better at recognizing evidence favoring that claim. Neither of us became worse at assessing evidence; we became *better*, in asymmetric ways. When we discuss individual facts (a school shooting, a case of self-defense), we often agree on which way they point. Yet since time and memory are limited, we are left disagreeing about high-level claims (*guns increase safety*) while being unable to share all the (rational) reasons we have for our differing opinions.

If that were what happened, then both of us could have predicted polarization as the outcome—as we could. And neither of us should be

[43] If they commonly knew they had been rational and exactly what each of their opinions were, their disagreement would disappear (Aumann 1976; Lederman 2015). But they do not know that.

moved now, when we discuss our persistent disagreements—as we are not. Nevertheless, while we each think the other is incorrect, we need not think they are dumb, or foolish, or irrational to believe what they do—as we do not.

If that were what happened. I am going to argue that it may have. That the example of Haley and Thomas is far more realistic than it seems. That we engage in cognitive search and face asymmetrically ambiguous evidence *all the time*. And that this helps explain real-world polarization. For that argument, jump to section 6; for the formal details of this section, read on.

5.1. The Formalities

A question Q is a partition of possibilities; $Q(w)$ is the partition-cell of w. A proposition p is about Q if and only if every complete answer to Q settles whether p (iff $p = \bigcup_i q_i$ for $q_i \in Q$). A decision problem is about Q if and only if every answer to Q settles the value of every option. P *Q-values* $\widetilde{P}$ (values $\widetilde{P}$ with respect to Q) if and only if it prefers to let $\widetilde{P}$ decide for any decision about Q. A *fixed-option Q-book* against an update is a pair of decision problems about Q such that deciding rationally before and after the update guarantees a loss. Q-Value entails that there is no Q-book against the update (Theorem A.5) and that for any quantity X whose value is determined by the answer to Q, P expects $\widetilde{P}$'s estimate of X to be more accurate than its own (Dorst et al. 2021: Theorem 3.2; Levinstein 2022). See section A.5.

Suppose Haley sees a sequence of n independent word-search tasks. Let Q_i be the partition of how the ith task went: $Q_i = \{N_i, C_i, F_i\}$, where N_i is the set of worlds where it is not completable, C_i is where it is but she does not find a completion, and F_i is where she finds one. Let Q be the question of how all the tasks went: for any w, w', $Q(w) = Q(w')$ if and only if for all i $Q_i(w) = Q_i(w')$. When Haley forgets a string, this *consolidates* the ambiguity: she holds fixed her opinions in Q but becomes certain (via imaging, Lewis 1976) they are now rational. H^i is the rational probability function after doing the ith task, and $\overline{H^i}$ is its consolidation. The updates from H^i to $\overline{H^i}$ are valuable with respect to Q. Meanwhile the updates from $\overline{H^i}$ to H^{i+1} are fully valuable, following the update from section 4: they Jeffrey-shift (Jeffrey, 1990) her opinions in the Q_i partition in different ways in different worlds, as indicated by figure 2 (e.g., in worlds in C_i, she Jeffrey-shifts to become 1/3 in N_i and 2/3 in C_i).

This yields Q-valuable predictable polarization about Q:

Theorem 5.1. *There is a sequence of probability functions H^0, $\overline{H^0}$, H^1, $\overline{H^1}$. . .,H^n, $\overline{H^n}$, a partition Q, and a proposition $h = \bigcup_i q_i$ (for $q_i \in Q$) such that, as $n \to \infty$:*

- *H^0 is (correctly) certain that $\overline{H^i}$ values H^{i+1} for each i;*
- *H^0 is (correctly) certain that H^i values $\overline{H^i}$ with respect to Q for each i;*
- *the sequence is predictably polarizing about h: $H^0(h) \approx \frac{1}{2}$, yet $H^0(\overline{H^n}(h) \approx 1) \approx 1$.*

See appendix A.6 for proof. Adding a Tailser leads to *persistent* polarization (Corollary 5.3).

The crux is that H^1 can think $\overline{H^1}$ makes (at least as good or) better decisions about Q than itself, and $\overline{H^1}$ can think H^2 makes better decisions about Q than itself, while H^1 thinks H^2 makes (some) *worse* decisions about Q than itself. How is this possible? Is $\overline{H^1}$'s judgment that H^2 makes better decisions about Q *itself* not a decision about Q (hence one H^1 should worry about)? *No.* The consolidation breaks the connection between Q and the rational opinions: we can no longer tell what $\overline{H^1}$ is based purely on the answer to Q, since once Haley forgets the string, she is rational to maintain her credence even if it was originally irrational. This means the judgment that H^2 makes better decisions about Q than $\overline{H^1}$ is not itself a decision about Q. That is what allows Q-Value to fail to be transitive.

Given this, you might want rational updates to be valuable about the combined question: *What is the answer to Q, and what are the rational opinions about Q?* That, I conjecture, would block predictable polarization. Does this cast doubt on its rationality? I do not think so. It is still true that every step is expected to make you more accurate about Q; if what you care about is the answer to Q, how can you be faulted for taking any such step?

6. The Confirmation Bias

Dan and I were not polarized by word searches. We were polarized by who we talked to, what we lived through, and how those factors shaped our ways of thinking. Dan fell in with libertarians, experienced failures of educational and criminal institutions, and became skeptical of many types of authority. I fell in with liberals, experienced favors of educational and criminal institutions, and became skeptical of many claims about individual responsibility.

Can my story explain this? Yes. I will show how ambiguity asymmetries may arise in the empirical processes that drive polarization, and that polarization can be predictable even if (unlike my word-search example) people have a *choice* about what evidence to receive.

Psychologists have documented many processes that predictably polarize people. *Confirmation bias* comprises tendencies to seek and interpret evidence in ways that strengthen your prior beliefs (Nickerson 1998; Whittlestone 2017). *Motivated reasoning* is the related tendency to selectively scrutinize uncongenial information (Kunda 1990; Kahan et al. 2017). And the *group polarization effect* is the tendency for discussions with likeminded others to make you more extreme (Isenberg 1986; Sunstein 2009). People who are aware of these tendencies are still subject to them (Pronin 2008; Lilienfeld, Ammirati, and Landfield 2009); hence Theorem 3.1 implies that (i) if evidence is unambiguous then they must be irrational, and (ii) Standard Bayesian models (see fn. 9) cannot rationalize them. I will show that ambiguous models can.

Confirmation bias first. This effect has been widely cited as a core driver of polarization in both academic[44] and popular[45] writings. Nevertheless, many researchers have noted that we lack good normative standards for assessing its rationality.[46] I hope to provide them.

Confirmation bias divides into at least two types: (1) *selective exposure*, the tendency to seek evidence that you expect to confirm your preferred hypothesis (Frey 1986; Hart et al. 2009), and (2) *biased assimilation*, the tendency to interpret mixed evidence as supporting your preferred hypothesis (Lord, Ross, and Lepper 1979; Taber and Lodge 2006). Here I focus on the latter, returning to the former in section 7.

[44] See Nickerson 1998; Taber and Lodge 2006; Risen and Gilovich 2007; Lilienfeld, Ammirati, and Landfield 2009; Stangor and Walinga 2014; Kahan et al. 2017; Mercier 2017; Mercier and Sperber 2017; Lazer et al. 2018; Talisse 2019.

[45] See Gilovich 1991; Fine 2005; Sunstein 2009; Kahneman 2011; Klein 2014, 2020; Wolfers 2014; Carmichael 2017; Robson 2018; Koerth 2019; Rogers 2020; Stanovich 2020.

[46] See Lord, Ross, and Lepper 1979; Lord and Taylor 2009; Taber and Lodge 2006; Crupi, Tentori, and Lombardi 2009; Ross 2012; Mercier 2017; Whittlestone 2017; Kinney and Bright 2021.

Examples of biased assimilation go like this.[47] Take two people—say, Dan and I—who strongly disagree about whether guns increase safety (s). Present us with two studies: one that (on its face) supports the claim, the other of which (on its face) tells against it. Give us time to think about them. Since you have given us the same information, you might expect it to dampen our disagreement. Generally, it will not. Instead, people tend to conclude that the *congruent* study—the one whose face-value reading supports their prior beliefs—is a more convincing study than the incongruent one. Thus on average, across situations like this, Dan will tend to increase his confidence in s, and I will tend to decrease mine.

Why? We will not simply dismiss the evidence against our beliefs—we will likely spend *more* time looking at it. As we do, we will often find legitimate flaws in the methodology, gaps in the reasoning, or alternative explanations that could explain away the data. Biased assimilation is driven by *selective scrutiny*: people spend more time looking for flaws with incongruent evidence than congruent evidence—the same mechanism that drives motivated reasoning.[48]

Thomas Kelly (2008) argues that selective scrutiny is rational and that it may rationalize some types of polarization. It is reasonable to spend more of our limited cognitive resources on surprising findings. If I doubt that guns increase safety, then a study suggesting they do should surprise me, while a study suggesting the opposite should not. It makes sense for me to scrutinize the former and for Dan to scrutinize the latter. Notice that if we do, we will end up receiving *different* evidence: I know more about one study, and Dan knows more about the other. Thus selective scrutiny is a type of selective exposure: exposure to flaws with incongruent evidence (cf. Kunda 1990). And if we *are not aware* that we are being selective—all we come away with is, "I saw one congruent

[47] Lord, Ross, and Lepper 1979 is the classic study; see also Gilovich 1983; Lord, Lepper, and Preston 1984; Plous 1991; Ditto and Lopez 1992; Liberman and Chaiken 1992; Miller et al. 1993; McHoskey 1995; Schuette and Fazio 1995; Kuhn and Lao 1996; Klaczynski and Narasimham 1998; Lundgren and Prislin 1998; Munro and Ditto 1997; Taber and Lodge 2006; Lord and Taylor 2009; Taber, Cann, and Kucsova 2009; Corner, Whitmarsh, and Xenias 2012; Ross 2012; Kahan 2013; Jern, Chang, and Kemp 2014; Kahan et al. 2017; Cook and Lewandowsky 2016; Liu 2017; Anglin 2019; Benoît and Dubra 2019.

[48] See Kunda 1990; Ditto and Lopez 1992; Lundgren and Prislin 1998; Kahan et al. 2012, 2017; Kahan 2013.

study and one *flawed* incongruent one"—then the resulting polarization is rational.

But, says Kelly, this only works if we are not aware we're being selective. If we are, we should not be surprised to find a flaw in only the incongruent study (cf. McWilliams 2021). (Compare: if you are aware you are fishing with a big net, you should not be surprised to catch only big fish.) In fact, if we *fail* to find a flaw in the incongruent study, we should *lower* our credence in our prior belief, since this suggests the incongruent evidence is stronger than we thought (McKenzie 2004). This is an instance of the point from section 3 that, without ambiguity, no rational strategy can lead to expectable polarization (Theorem 3.1; see Salow 2018).

And this is where Kelly and I part ways. Many of us *do* realize we are engaging in selective scrutiny. Indeed, it is standard scientific practice: adopt a hypothesis and then spend your time trying to explain away problems with it (Kuhn 1962; Solomon 1992). We are all familiar with how choosing a school to attend or a project to pursue can have a predictable impact on how we think and thus on how our beliefs evolve (Cook 1987).

The question: how could *knowingly* searching for flaws predictably polarize people?

My answer: the same way that knowingly searching for *words* can. Both are forms of cognitive search. Both involve an ambiguity asymmetry: if you find what you are looking for (a flaw, a word), it is easier to know how to react to the evidence; if you do not, you should be (more) unsure what to think. As a result, both induce asymmetric accuracy increases: if there is a flaw (a word), your credence that there is should on average increase a lot; if there is not, your credence should decrease only a bit. And again: the average of 'increase a lot' and 'decrease a bit' is 'increase a bit'—the process is expectably polarizing.

Suppose scrutinizing a study leads to the same structure of evidence as searching for a word, so we can model it in the same way (see section 6.1 for details). Which way it is polarizing depends on how you scrutinize. When I scrutinize a study suggesting that guns increase safety (*s*), this expectably *lowers* the rational credence in *s*, since finding a flaw would lower my credence. When Dan scrutinizes a study suggesting the opposite, that expectably *raises* the rational credence in *s*. Thus if it is rational to selectively scrutinize, then *even if you are aware of it*, the resulting ambiguity asymmetries will rationalize expectable polarization.[49]

But, given this polarizing model, *is* it rational to selectively scrutinize? You might think it could not be. After all, repeated selective scrutiny will predictably polarize you—so would it not be better to scrutinize even-handedly? This is where the diachronic tragedy rears its head. Just as with Theorem 5.1, if you were deciding on a policy for your whole life, you would expect to be more accurate if you did not selectively scrutinize; nevertheless, in *each instance*, when faced with a pair of conflicting studies, you expect selective scrutiny to be the best thing you can do *in that instant* to get to the truth.

How to assess the rationality of the choice in each instant? Since scrutinizing either study is (fully) valuable, both are expected to improve accuracy (on everything). So even if pragmatic considerations influence your choice—as some literature suggests (Kunda 1990; Kahan et al. 2017)—the process is arguably epistemically rational.

But more is true. Why do *I* tend to scrutinize incongruent studies over congruent ones? Because I expect doing so to make me more accurate, since it is more likely that I will be able to find a flaw and avoid ambiguity. I may think it is more likely to *contain* a flaw—but even if I do not, I will be more likely to *find* any flaws it contains. After all, part of being convinced of a claim is learning how to rebut arguments against it. This very article illustrates the point: what convinced me of its conclusions was, largely, figuring out how to rebut objections—that rational polarization violated Bayesianism (section 3), that it was purely theoretical (section 4), that ambiguity wasn't the driving force (section 4.2), that it could not be predictable (section 5), and so on. More generally, there are both theoretical (Aronowitz 2021) and empirical (Evans, Barston, and Pollard 1983; Kahan et al. 2017) reasons to think that people are better at finding flaws with evidence that tells against their beliefs—an idea at the heart of the adversarial model of academia.

Granting this, will polarization result? Here is an analogy. Suppose I will see a series of *pairs* of word-search tasks—one following Headser rules and the other following Tailser rules. Headser tasks use British English; Tailser tasks use American English. At each stage I can choose

[49] As in section 4, these updates are fully valuable. If (as in section 5) we allow for consolidations of higher-order uncertainty that are valuable with respect to some question Q (for example, which direction all the bits of relevant evidence point), this polarization can be predictable and persistent.

which to look at. Being an American, I expect to be better at finding words in the latter task than the former. So if at each stage I am guided by my desire to form accurate beliefs, I will tend to do the Tailser tasks more often. And since doing so leads to predictable polarization, I will wind up confident that less than half the coins landed heads.

How to verify this intuitive reasoning? Simulation. Randomly generate models of cognitive searches for flaws in studies and examine (1) whether a preference for accuracy can lead to selective scrutiny of studies that you are better at finding flaws in and (2) whether this preference can indeed lead to predictable polarization.

To (1): I randomly generated models and compared $P(\mathit{Find}|\mathit{Flaw})$ to expected accuracy, finding a robust correlation (figure 5, top). I then generated pairs of models in which you are (on average) more likely to find flaws that exist in the *in*congruent than the congruent study; expected accuracy quite often warrants scrutinizing the former (figure 5, bottom).

To (2): two groups of agents face a series of choices about which of two random studies to scrutinize. They start out 50% confident in a claim q, and at each stage they scrutinize in the way they expect to make their beliefs most accurate. But one group (red) is better at recognizing flaws in studies that tell against q, and the other (blue) is better at recognizing flaws in those that tell in favor of q. The result is polarization (figure 6).

These results suggest that irrationalist interpretations of biased assimilation and motivated reasoning are too quick: rational people who care about the truth but face ambiguous evidence will exhibit them. In fact, this model fits with a variety of empirical findings. It is built on the idea that people are better at finding flaws in incongruent than congruent evidence. They are.[50] It predicts that instructions like “do not be biased” or “be accurate” will not prevent biased assimilation, but that instructions that get people to scrutinize both sides equally *will.* They do.[51] And it suggests that bias will be more extreme when people think *harder*—when they scrutinize more, rather than less. It is.[52]

The upshot is that, insofar as confirmation bias and motivated reasoning drove me and Dan apart, this may have been due to ratio-

[50] See Evans, Barston, and Pollard 1983; Petty and Wegener 1998; Mercier and Sperber 2011; Kahan et al. 2012, 2017.

[51] See Koriat, Lichtenstein, and Fischhoff 1980; Lord, Lepper, and Preston 1984; Schuette and Fazio 1995; Lundgren and Prislin 1998; Liu 2017.

[52] See Fitzpatrick and Eagly 1981; Kuhn and Lao 1996; Downing, Judd, and Brauer 1992; Tesser, Martin, and Mendolia 1995; Kahan 2013.

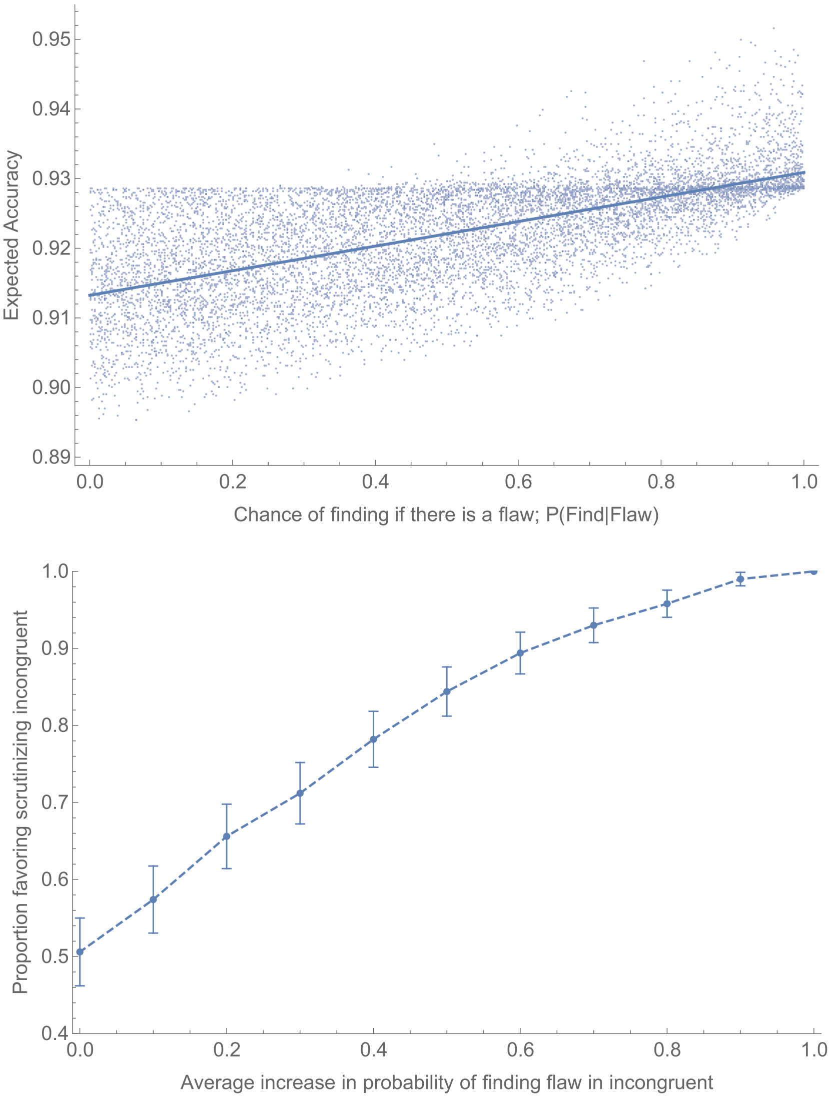


Figure 5. (Color online.) Top: Correlation between $P(\mathit{Find}|\mathit{Flaw})$ and the expected accuracy of scrutiny. Bottom: Rates of selective scrutiny based on expected accuracy (y-axis) grow as the average gap in $P(\mathit{Find}|\mathit{Flaw})$ between incongruent and congruent studies (x-axis) grows.

nal management of ambiguous evidence. Still, this model depends on differences in background knowledge and abilities to find flaws. How could such differences predictably *emerge*, simply from falling into differ-

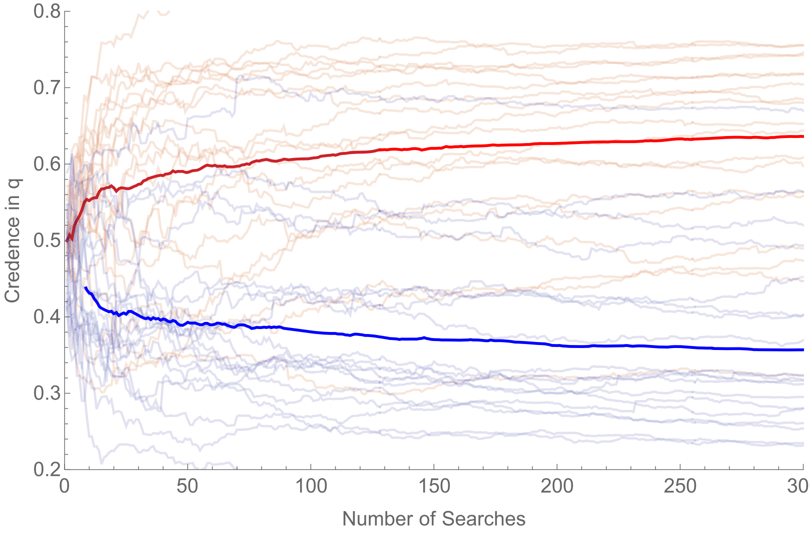
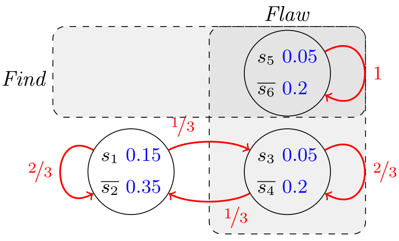


Figure 6. Agents faced with cognitive search choices, choosing via expected accuracy. Red agents are better at finding flaws in *q*-opposing studies; blue agents vice versa. Thin lines are individuals; thick lines are averages.

ent social circles? For the answer, skip to section 7; for the details from this section, read on.

6.1. *The Formalities*

Here I describe cognitive search models, which generalize the word-search model from figure 2 (see appendix C.1 for more details). They have the same structure (possibilities where you find a flaw, possibilities where you do not but there is one, etc.), but they multiply possibilities within each class to represent when the target proposition (s) is true or false, and they allow variation in priors and posteriors. Figure 7 is an example. I face a study favoring s, am currently 25% confident of s, and am scrutinizing for flaws. The s_i are where s is true; the $\overline{s_j}$ are where it is false. Prior probabilities (the blue numbers) are constant across worlds; posteriors are obtained by Jeffrey-shifting the prior P on the $\{Find\&Flaw, \neg Find\&Flaw, \neg Flaw\}$ partition as indicated by the labeled arrows (holding conditional probabilities like $P(\cdot|Find\&Flaw)$ fixed, but changing $P(Find\&Flaw)$). Thus the posterior probability for s is: if I find a flaw (s_5 and $\overline{s_6}$), $\frac{0.05}{0.05+0.20} = 0.2$; if there's a flaw that I do not find (s_3 and $\overline{s_4}$), $\frac{1}{3}(\frac{0.15}{0.5}) + \frac{2}{3}(\frac{0.05}{0.25}) = 0.2\overline{3}$; and if there is no flaw (s_1 and



$$Flaw = \{s_3, \overline{s_4}, s_5, \overline{s_6}\}$$
$$Find = \{s_5, \overline{s_6}\}$$
$$P = \begin{pmatrix} 0.15 & 0.35 & 0.05 & 0.2 & 0.05 & 0.2 \\ \dots \end{pmatrix}$$
$$\widetilde{P} = \begin{pmatrix} 0.2 & 0.4\overline{6} & 0.0\overline{6} & 0.2\overline{6} & 0 & 0 \\ 0.2 & 0.4\overline{6} & 0.0\overline{6} & 0.2\overline{6} & 0 & 0 \\ 0.1 & 0.2\overline{3} & 0.1\overline{3} & 0.5\overline{3} & 0 & 0 \\ 0.1 & 0.2\overline{3} & 0.1\overline{3} & 0.5\overline{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.2 & 0.8 \\ 0 & 0 & 0 & 0 & 0.2 & 0.8 \end{pmatrix}$$

Figure 7. (Color online.) Model of scrutinizing *s*-supporting evidence in Kripke model (left) and stochastic matrix (right). See figure 1 for interpretation.

$\overline{s_2}$), $\frac{2}{3}(\frac{0.15}{0.5}) + \frac{1}{3}(\frac{0.05}{0.25}) \approx 0.2\overline{6}$. If the study contains a flaw, *s* is 20% likely ($P(s|Flaw) = 0.2$); if it does not, *s* is 30% likely ($P(s|\neg Flaw) = 0.3$); and it is equally likely to contain a flaw as not ($P(Flaw) = 0.5 = P(\neg Flaw)$). But since evidence is less ambiguous when I find a flaw, the update is expectably polarizing.[53]

I measured accuracy with the Brier score (Brier 1950): the sum of squared distances between the probability of each possibility and its truth value, so the *in*accuracy of P at w is $B(P, w) := \sum_{x \in W}(\mathbb{1}_{\{x\}}(w) - P_w(x))^2$, and the accuracy $1 - B(P, w)$. For tractability, the simulations only tracked the agents' opinions in *s* and in the cognitive searches they were evaluating at a given time—it did not model their evolving opinions about *all* the searches. This is harmless, as a generalization of Theorem 5.1 (which I omit) shows that if we use a series of 'small-world' updates like this—which do not track past or future updates—we can stitch them together into a 'large-world' model that satisfies *Q*-Value.

7. The Group Polarization Effect

Once Dan and I had different background beliefs, selective scrutiny could pull us further apart. But our polarization became predictable when we fell into different social groups, *before* our beliefs had changed. How could ambiguity asymmetries start our divergence?

One answer is simple: different social groups incentivize different cognitive searches (Kahan et al. 2017). When Dan fell in with libertarians, that incentivized him to search for flaws in progovernment arguments, and vice versa for me.

[53] $\mathbb{E}_P(\widetilde{P}(Flaw)) \approx 0.583 > 0.5 = P(Flaw)$, so $\mathbb{E}_P(\widetilde{P}(s)) \approx 0.242 < 0.25 = P(s)$.

But clearly this is not the full explanation. Much polarization is due to the fact that group membership affects what information you receive. Libertarians discuss libertarian arguments; liberals discuss liberal ones. Both get their news from congenial sources; hence they diverge. This *group polarization effect* is widely documented (Myers and Lamm 1976; Isenberg 1986; Sunstein 2009; Talisse 2019). The mechanism driving it is unsurprising: people who believe a claim tend to share arguments that favor it (Toplak and Stanovich 2003; Wolfe and Britt 2008), and arguments for a claim tend—on average—to predictably persuade people of it (Vinokur and Burstein 1974; Burnstein and Vinokur 1977; Petty and Wegener 1998; Stafford 2015).[54] This is intuitive, so most explanations stop here.

They should not. A familiar point applies again: it is not just that *someone* can predict that we will be persuaded by arguments—it is that *we ourselves* can. If you are open minded (more on that caveat in a moment), you can expect that reading liberal arguments will make you more liberal. Theorem 3.1 again implies that if the evidence is unambiguous, rational Bayesians can expect no such thing (Salow 2018). Yet *we* can.

Everyone needs to explain this. Either we process arguments irrationally or they generate ambiguity asymmetries. I do not have a knockdown case for the latter, but here is the idea. Suppose you know you will be given an argument that guns increase safety (s). Given your background evidence, that argument will be either *good* (convincing) or *bad* (unconvincing): if it is good, it will warrant increasing your credence in s ("I hadn't thought of that"); if it is bad, it will warrant decreasing it ("That's the best they've got?"). You cannot be certain the argument will be good—if you were, you should have already raised your credence.[55] Nor will you be able to tell whether the argument was good after you have seen it: it is ambiguous, so you will rationally be unsure how you should interpret it. What you *can* expect is that the arguer will make it easier to recognize evidence favoring their position and harder to recognize evidence disfavoring it. There may even be a selection effect: good arguments tend to get repeated because they *are* good; bad arguments

[54] Some (e.g., Sunstein 2009) also point to *social comparison* (adopting your group's opinions so they like you). I set it aside because (1) arguments explain more of the effect (Isenberg 1986) and (2) every social comparison study I have seen fails to control for fact that others' opinions provide evidence (Elga 2007).

[55] If P values $\widetilde{P}$ with respect to $\{s, \neg s\}$ and $P(\widetilde{P}(s) \geq t) = 1$, then $P(s) \geq t$.

tend to get repeated because they *sound* good. Thus bad arguments will tend to be more ambiguous—harder to recognize as bad.

Here is an (overly) simple example. Suppose Jack was hurt, and someone is trying to convince you that he did not have a gun. Contrast two arguments:

> "Every weekend, Jack has a gun. But it was Monday, so he didn't have it."
> "Whenever Jack has a gun, it's a weekend. But it was Monday, so he didn't have it."

At a quick glance, or to the untrained eye, it is easier to recognize that the latter is valid than that the former is invalid. (Some fallacies are tempting!) Indeed, there is some evidence that people are worse at recognizing fallacies as fallacies than they are at recognizing validities as validities (Evans, Barston, and Pollard 1983; Cariani and Rips 2017: figure 1).

Generalizing, suppose that arguments are (on average) less ambiguous when they are good than when they are bad. Here is a *simple argument model.* When you see an argument, your credence that it is good should either increase or decrease. Value implies that it should increase when it is good and decrease when it is not, but it allows the *degree* to be asymmetric: the good-case increase is larger than the bad-case decrease. What follows? If two groups see randomly generated arguments—one (red) group sees arguments supporting *s*, while the other (blue) sees ones opposing *s*—then they predictably polarize (figure 8; see section 7.1 for details). The upshot is that being exposed to different arguments might have rationally, predictably polarized us.

But how does this simple argument model fit with my discussion of selective scrutiny (section 6)? If an argument is bad, shouldn't you be able to find a flaw and get *un*ambiguous evidence? Proposal: it depends on how you *engage.* If you engage passively (you do not scrutinize), the simple model makes sense—with just a quick glance, it is easier to recognize modus ponens as valid than affirming the consequent as invalid. But if you engage actively (you *do* scrutinize), the update becomes a cognitive search. This splits the *bad* possibilities into two: those in which you find a flaw and those in which you do not (see section 7.1 for details).

On this picture, whenever you see an argument you face a choice: scrutinize or not? Your choice affects how your rational opinions should expectably shift. To illustrate, imagine that two groups see arguments favoring *s*: one (red) group never scrutinizes; the other (blue) group always does. On natural parameterizations: if they know they *will not* find

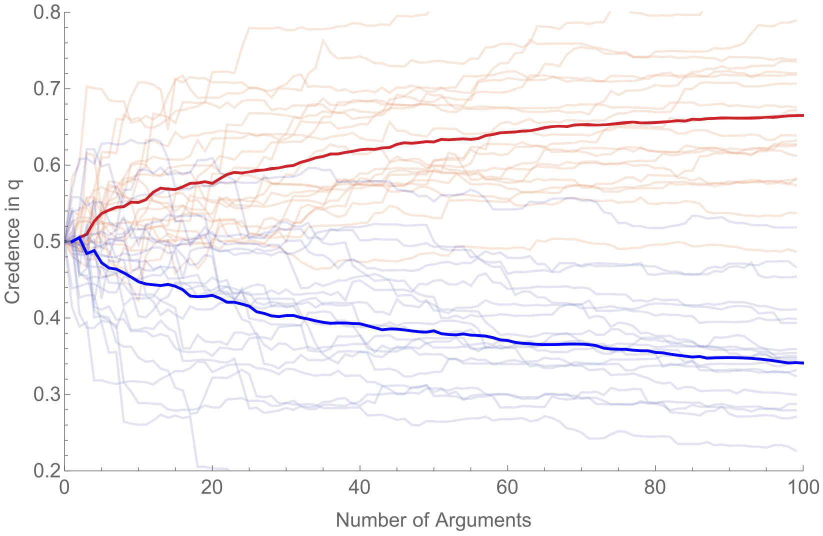


Figure 8. Red agents are presented with random argument models (from figure 10) favoring *s*, and blue agents are presented with models favoring ¬*s*. Thin lines are individuals; thick lines are averages.

a flaw even if there is one, scrutiny leaves the polarizing effects of the argument unchanged (figure 9, top left). If they know they *will* find a flaw if there is one, scrutiny removes all ambiguity—the update becomes a Standard Bayesian one with no expectable polarization (top right). And if there is a middling chance of finding a flaw, scrutiny dampens the polarizing effects of arguments (bottom left) and can even *reverse* the polarizing effects (bottom right).

The upshot is that if we always scrutinized arguments and had no self-doubt in our assessments, then our evidence would be unambiguous and predictable polarization would be irrational. But since we *cannot* scrutinize everything and we *should* have self-doubts, arguments can predictably polarize us despite being expected to improve accuracy.

Thus irrationalist interpretations of the group polarization effect are too quick. Indeed, when supplemented with the hypothesis that people selectively scrutinize *in*congruent arguments (section 6), this model fits with a variety of findings about persuasion. It predicts that there are two routes to engaging with arguments: a passive, low-effort one that predictably shifts opinions and an active, high-effort one for which the

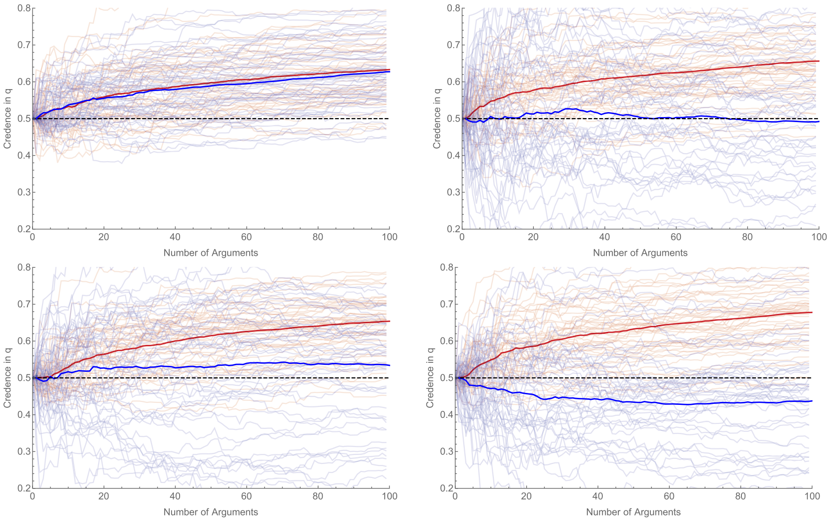


Figure 9. Two groups are presented with arguments favoring *q*; red group never scrutinizes, while blue group always does. Top left: 0% chance of finding flaw if there is one; full blue polarization. Top right: 100% chance of finding flaw if there is one; no blue polarization. Bottom: Middling chance of finding, with small (left) and large (right) amounts of ambiguity if they do not find; dampens (left) or reverses (right) blue polarization.

persuasive effects vary widely. There are.[56] It predicts that those who are (selective in scrutinizing but) *better* at finding flaws will end up with a more skewed assessment of the overall weight of evidence. They do.[57] It also predicts that manipulating how much people scrutinize will affect persuasion, with the biggest effects being on the evaluation of weak, congruent arguments (they will be surprised to find flaws) and strong, incongruent ones (they will be surprised *not* to find flaws). It does.[58]

Finally, this model may clarify the mixed findings on *selective exposure* (section 6)—the tendency to seek out congruent arguments over incongruent ones. Sometimes people do this (Fischer et al. 2005; Taber and Lodge 2006); other times they do not (Sears and Freedman 1967; Whittlestone 2017). Why? One throughline is that people are more inclined to engage in selective exposure when they expect the arguments to be of high quality (to not contain obvious flaws), less inclined

[56] See Petty 1994; Petty and Wegener 1998; Taber and Lodge 2006; Lundgren and Prislin 1998.

[57] See Kahan et al. 2012; Kahan 2013; Kahan et al. 2017; Bail et al. 2018.

[58] See Schuette and Fazio 1995; Petty and Wegener 1998; Liu 2017.

otherwise (Frey 1986; Hart et al. 2009). The model predicts this. When arguments are high quality, scrutiny is useless (you will not find a flaw even if there is one), so deciding which argument to see is just a comparison of simple-argument models. In that case, avoiding ambiguity will drive you to look at the argument you think is more likely to be good—generally, the one that supports your beliefs, leading to selective exposure. But when arguments are low quality, scrutiny makes a difference: avoiding ambiguity will spur you to look at the arguments you are most able to find a flaw in, that is, the *in*congruent arguments (contra the selective exposure effect).

Obviously this model is speculative; it needs to be refined and tested. But it shows that the group polarization effect is not necessarily a sign of irrationality.

7.1. *The Formalities*

Here I sketch the simple- and scrutinized-argument models (see appendices C.2 and C.3 for more details).

The simple-argument model partitions possibilities into those where the argument is good (G) and those where it is bad (B). The posteriors are obtained by Jeffrey-shifting on the $\{G, B\}$ partition, increasing credence in the true possibility, and hence satisfying (full) Value. But the *degree* of these shifts is asymmetric: since good arguments are easier to recognize, the shift is larger if G than if B (figure 10).[59]

What about scrutiny? Given an argument model, you choose whether to update in accordance with it or instead transform the update by splitting the *bad* possibilities into those where you do versus do not

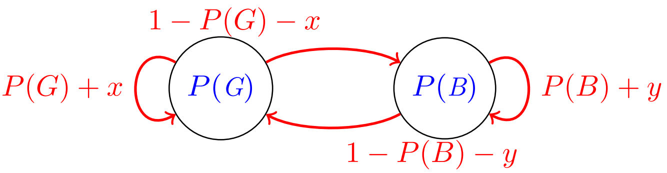


Figure 10. Schematic simple-argument model. If it is an argument for s, then $P(s|G) > P(s) > P(s|B)$; for $\neg s$, vice versa. Since bad arguments are more ambiguous than good ones, $y \leq x$.

[59] For example, if $P(s) = 0.5$, $P(G) = 0.5$, $P(s|G) = 0.6$, $P(s|B) = 0.4$, and $x = 0.4 > 0.1 = y$, then $\mathbb{E}_P(\widetilde{P}(G)) = 0.65 > 0.5 = P(G)$, so $\mathbb{E}_P(\widetilde{P}(s)) = 0.53 > 0.5 = P(s)$.

find a flaw, as diagrammed schematically in figure 14 (page 439). There are many ways to parameterize these models; see section C.3 for details.

8. A Better Story

Not long ago, I caught up with an old friend. Not Dan. A better friend. A friend who was with me that night we forgot something outside. A friend whose story is not mine to tell.

We talked about old times. About our lives. About politics. And about that damn bench. The details were stunning. But the outlines? Predictable. We were not surprised by each others' opinions; most of them, we could have guessed. That said, his *reasons* surprised me. I did not agree with them—with selective scrutiny, I concluded that some were misinformation, and many were missing the bigger picture. Nevertheless, given *his* networks of trust, *his* lived experience, and *his* background beliefs, they made perfect sense.

That conversation sticks with me. What should I think of him and his beliefs? He is bright and well meaning. He has had experiences—the failures of institutions, of communities, of friends—that I can only dimly imagine. The reasons he shares seem, given their context, perfectly sensible. Yet the overall picture seems radically distorted: the steps reasonable, but the destination wrong. How could that be?

For me, predictable polarization tends to induce this sort of double vision. I find myself unsurprised ("*Of course* you believe that"), but at the same time baffled ("*How* can you believe that?"). I am unsurprised, because I know the psychology: people glom onto the beliefs of their peers, confirm and entrench those beliefs, become extremely confident, and so on. I am baffled, because I often find that they are *not* just conforming, or pigheaded, or dogmatic. Yet if they are not, how do they end up where they do?

This double vision is starkest when I look inward. *I* am not just conforming, or pigheaded, or dogmatic. But the psychology works: if I told you my biography, you could tell me my beliefs.

This project is my attempt to square this circle. The mistake is to assume that we should *expect* individual steps toward the truth to lead to an accurate overall picture. If evidence were not ambiguous, we should expect this—but it is (section 3), so we should not. Instead, we face ambiguity asymmetries that make us better at recognizing evidence on one side than on the other (section 4). Wanting get to the truth, we take each individual step; by the end, the 'radically distorted' picture has become our own (section 5). This theoretical idea has both experimental support (section 4.2), and the potential to explain the mechanisms underlying confirmation bias (section 6) and the group polarization effect (section 7).

Obviously this does not show that real-world polarization is rational. What it suggests is that it *might* be—that it would not look terribly different if it were. And what it promises is a better way to think about our ideological opponents—and ourselves.

Assuming predictable polarization is irrational leaves me seeing my beliefs in double. It is incoherent to believe that "guns decrease safety, but I formed that belief irrationally." But how to avoid it? The evidence is overwhelming that guns *do* decrease safety. But the evidence is *also* overwhelming that my belief was formed by predictably-polarizing mechanisms.

Accepting the rationality of predictable polarization resolves the image. Yes, guns do decrease safety. Yes, the psychologists are right about why I believe as much. But no, I am not irrational for that. And no, my friends are not irrational for believing otherwise. Likewise for the religious beliefs we have formed through selective scrutiny, the political beliefs we have formed through selective exposure, and the philosophical beliefs we have formed through searching for evidence favoring our positions.

That is the promise of a story like this. It allows us to admit our own predictability without undermining our own deeply-held commitments—and without disparaging those of others.

Appendix A. Analytical Details

Appendix A gives all analytical details and proofs, including:

A.1 Higher-order probability models,
A.2 The Value-of-Evidence constraint,
A.3 Standard Bayesianism and the (im)possibility of valuable expectable polarization,
A.4 Word-search models,
A.5 Question-Relative Value, and
A.6 The predictable polarization theorem.

Appendix A.1. Higher-Order Probability Models

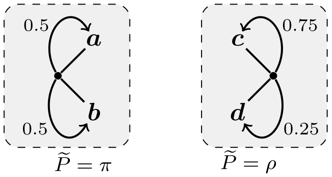
Following standard epistemic logic (Hintikka 1962; van Ditmarsch et al. 2015), we give a semantics for higher-order probability using a (finite) structure that can identify higher-order claims with events, that is, sets of worlds (i.e., propositions).[60] A *probability frame* $\langle W, \{P^i\}_{i \in N}\rangle$ is a (finite) set of worlds W and a set of functions P^i from worlds $w \in W$ to probability functions P^i_w defined over all subsets of W so that $P^i : W \to \Delta(W)$. Thus 'P^i' can be thought of as a *description* of a probability function—it picks out different functions in different worlds. In our case, it will be interpreted as "the rational credence function (for some particular agent) at time i." 'P^i_w' is a rigid designator that picks out the probability function that P^i associates with a given world w. When we are only concerned with one or two functions, I will drop indices, using P, P_w and $\widetilde{P}$, $\widetilde{P}_w$. I will also often enrich the structure with one or more (rigidly designated) probability functions, denoted $\pi, \delta, \eta, \ldots$.

W represents the propositions in the frame, so for any $p, q \subseteq W$, p is true at w if and only if $w \in p$, $\neg p = W \backslash p$, $p \wedge q = p \cap q$, $p \to q = \neg p \cup q$, and so on. All theorems are restricted to models with finite W—it is an open question how far they generalize.

We use P to identify facts about probabilities as sets of worlds in the frame, thus allowing us to 'unravel' higher-order probability claims into propositions. Thus for any $q \subseteq W$ and $t \in \mathbb{R}$ and $\pi \in \Delta(W)$: $[P(q) = t] := \{w \in W : P_w(q) = t\}$, $[P(q|r) \geq t] := \{w \in W : P_w(q \mid r) \geq t\}$, $[P = \pi] := \{w \in W : P_w = \pi\}$, and so on.

Since W is finite, we can think of a probability function as an assignment of nonnegative numbers to worlds that sum to 1, so we can diagram probability frames as we did in the main text using *Markov diagrams* (i.e., generalized Kripke frames): nodes represent worlds and an arrow labeled t from w to v says that $P_w(v) = t$. Equivalently, we can number the worlds $w_1, \ldots, w_n$ and write this information in a (square) *stochastic matrix* M in which $M_{ij} = P_{w_i}(w_j)$, that is, the probability that world i assigns to world j. A simple example of an (unambiguous) probability frame $\langle W, \widetilde{P}\rangle$ is given in figure 11.

[60] For explanations of such structures, see Williamson 2008 and Dorst 2019, forthcoming. For uses of them, see, for example, Gaifman 1988; Hild 1998; Samet 2000; Williamson 2000, 2014, 2019; Schervish, Seidenfeld, and Kadane 2004; Lasonen-Aarnio 2013, 2015; Campbell-Moore 2016; Salow 2018, 2019; Das 2022a, 2023; Dorst 2020; Dorst et al. 2021.



$$\begin{array}{c c} \pi \left\{ \begin{array}{c} \\ \\ \end{array} \right. & \multirow{4}{*}{$\begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.75 & 0.25 \\ 0 & 0 & 0.75 & 0.25 \end{pmatrix}$} \\ \rho \left\{ \begin{array}{c} \\ \\ \end{array} \right. & \end{array}$$

Figure 11. An unambiguous frame, in both Markov diagram and stochastic matrix notation. π assigns 0.5 to a and 0.5 to b; ρ assigns 0.75 to c and 0.25 to d.

Appendix A.2. The Value of Evidence

When is an update from prior P to posterior $\widetilde{P}$—updating from P_w to $\widetilde{P}_w$ in each world w—a potentially rational update? Following Dorst et al. 2021, I proposed that this is so when P prefers to outsource its decisions to $\widetilde{P}$, i.e. P *values* $\widetilde{P}$: it always expects $\widetilde{P}$ to make better decisions than itself. This is equivalent to saying that the update from P to $\widetilde{P}$ cannot be Dutch-booked, that it is always expected to increase accuracy, and that P obeys a particular ('Trust') deference principle toward $\widetilde{P}$. Let's formalize these in turn.

Consider a probability frame modeling the update, $\langle W, P, \widetilde{P} \rangle$, with W finite. An *option* O is a random variable: a function from worlds w to numbers $O(w) \in \mathbb{R}$ representing the utility that would be achieved by taking option O at world w. A *decision problem* is a finite set of options $\mathcal{O}$. A *strategy* S is a way of choosing options based on $\widetilde{P}$'s probabilities—that is, a function from w to $S_w \in \mathcal{O}$ such that $S_w = S_x$ whenever $\widetilde{P}_w = \widetilde{P}_x$. Abusing notation slightly, for any probability function π, let $\mathbb{E}_\pi(S)$ be π's expectation of following strategy S: $\mathbb{E}_\pi(S) := \sum_w \pi(w) S_w(w)$. $\widetilde{P}$ *recommends* a strategy S for $\mathcal{O}$ if and only if S always selects an option that maximizes expected value according to $\widetilde{P}$. For any probability function π, let $\mathbb{E}_\pi(O)$ be π's expectation of O: $\mathbb{E}_\pi(O) = \sum_t \pi(O = t) \cdot t = \sum_w \pi(w) O(w)$. Thus S is recommended by $\widetilde{P}$ if and only if, for all w and $O \in \mathcal{O}$, $\mathbb{E}_{\widetilde{P}_w}(S_w) \geq \mathbb{E}_{\widetilde{P}_w}(O)$.

Given this, we say a particular probability function π *values* $\widetilde{P}$ if and only if, for any decision problem, π expects following any strategy recommended by $\widetilde{P}$ to do at least as well as simply picking an option itself. Precisely: π values $\widetilde{P}$ if and only if for all $\mathcal{O}$, if $\widetilde{P}$ recommends S for $\mathcal{O}$, then for any $O \in \mathcal{O}$, $\mathbb{E}_\pi(S) \geq \mathbb{E}_\pi(O)$. We lift[61] this from a particular

[61] There is a subtlety here. As stated, P values $\widetilde{P}$ if and only if, at all worlds w, P_w prefers to let $\widetilde{P}$ (picked out descriptively) decide over *itself* (picked out rigidly), that is,

prior π to a description of the prior P by asserting that each at each world w, P_w values $\widetilde{P}$ in this sense:

> **Value:** P values $\widetilde{P}$ iff $\forall w$, $\mathcal{O}$, if $\widetilde{P}$ recommends S for $\mathcal{O}$, $\forall O \in \mathcal{O}$: $\mathbb{E}_{P_w}(S) \geq \mathbb{E}_{P_w}(O)$.
> P values $\widetilde{P}$ iff, for any decision problem, P prefers to let $\widetilde{P}$ decide on its behalf, rather than simply choose an option.

A *fixed-option Dutch book* is a pair of decision problems—both containing a 'no bet' option, one faced before and the other after the update—such that doing the rational thing at both times is guaranteed to result in a loss. Formally, given P_w and $\widetilde{P}$, it is a pair $\mathcal{O}_1$ and $\mathcal{O}_2$ (each including a constant $O_0 = 0$) such that $O \in \arg\max_{O' \in \mathcal{O}_1} \mathbb{E}_{P_w}(O')$ and S is recommended by $\widetilde{P}$ for $\mathcal{O}_2$, and yet $O(w) + S_w(w) < 0$ at every world w. A short but subtle proof shows that P_w values $\widetilde{P}$ if and only if there is no fixed-option Dutch book against updating from P_w to $\widetilde{P}$ (Dorst et al., 2021: fns. 21 and 22). Lifting this as before (cf. fn. 61), P values $\widetilde{P}$ if and only if there is no fixed-option Dutch book against updating from any of the P_w to $\widetilde{P}$.

An *estimate-accuracy measure* A_X for a random variable X takes an estimate $e \in \mathbb{R}$, a world w, and outputs the accuracy of e at w, $A_X(e, w)$—how 'close' e comes to $X(w)$. Writing $A_X(\pi)$ to abbreviate $A_X(\mathbb{E}_\pi(X))$, say that A_X is *strictly proper* if and only if any probability function expects its own estimate of X to be more accurate than any other (rigidly designated) estimate: for any π, $\mathbb{E}_\pi(A_X(\pi)) > \mathbb{E}_\pi(A_X(e))$ whenever $\mathbb{E}_\pi(X) \neq e$. Dorst et al. (2021, Theorems 3.2 and 5.1) show that P_w values $\widetilde{P}$ if and only if, for any quantity X and all strictly proper estimate-accuracy measures A_X, the expected accuracy of $\widetilde{P}$ is at least as great at that of P_w: $\mathbb{E}_{P_w}(A_X(\widetilde{P})) \geq \mathbb{E}_{P_w}(A_X(P_w))$. Once again lifting this to descriptions (cf. fn. 61), P values $\widetilde{P}$ if and only if each P_w expects $\widetilde{P}$ to have estimates at least as accurate as itself (P_w).

over P_w. When P has no higher-order uncertainty, P knows what P is, so 'letting P decide' is same as 'letting P_w decide', which is the same as choosing an option $O \in \mathcal{O}$—namely, the one that maximizes expectation according to P_w. But when P has higher-order uncertainty, it may be unsure what option it itself recommends. In that case, we might prefer to say that P values $\widetilde{P}$ when at each world w, P_w prefers to let $\widetilde{P}$ (picked out descriptively) decide rather than P (*also* picked out descriptively). These two formalizations are equivalent only if P is higher-order certain. I choose the former because it is the one used in Dorst et al. 2021 and whose formal properties are well understood. However, every update I use in this article is valuable (or, later on, valuable with respect to Q) in the latter sense as well, so the choice does not matter for our purposes.

Given a random variable X, let $[\widetilde{\mathbb{E}}(X) \geq t]$ be the proposition that $\widetilde{P}$'s expectation of X is at least t, so $[\widetilde{\mathbb{E}}(X) \geq t] := \{w \in W : \mathbb{E}_{\widetilde{P}_w}(X) \geq t\}$. The 'deference principle' that Value is equivalent to requires deferring to facts of this form:

> **Total Trust:** For any variable X and threshold t, $\mathbb{E}_\pi(X|\widetilde{\mathbb{E}}(X) \geq t) \geq t$.
> Given that $\widetilde{P}$'s estimate for X is at least t, have an estimate for X that is at least t.

Total Trust entails that $\mathbb{E}_\pi(X|\widetilde{\mathbb{E}}(X) \leq t) \leq t$, but it does *not* entail that $\mathbb{E}_\pi(X|\widetilde{\mathbb{E}}(X) = t) = t$; hence it is a weakening of standard 'Relection-style' deference principles like Function Reflection (section A.3 below; see Dorst et al. 2021 for discussion). Note that if we let X be the indicator function $\mathbb{1}_q$ for some proposition q, Total Trust implies that $\pi(q|\widetilde{P}(q) \geq t) \geq t$ and $\pi(q|\widetilde{P}(q) \leq t) \leq t$. Lifting this to descriptions (cf. fn. 61), P values $\widetilde{P}$ if and only if each P_w totally trusts $\widetilde{P}$.

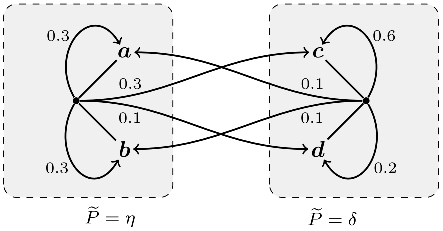
Appendix A.3. Ambiguity, Standard Bayesianism, and (Im)possibility Theorems

Recall the (often implicit) constraint implied by Standard Bayesianism:

> **No Ambiguity:** Rational opinions are always sure what the rational opinions are.
> Always, if $\widetilde{P} = \pi$, then $\widetilde{P}(\widetilde{P} = \pi) = 1$. That is, $\forall q, t$: if $\widetilde{P}(q) = t$, then $\widetilde{P}(\widetilde{P}(q) = t) = 1$.

No Ambiguity fails in any frame in which there are two worlds w and v such that $\widetilde{P}_w(v) > 0$ and yet $\widetilde{P}_w \neq \widetilde{P}_v$, for that means that $w \in [\widetilde{P} = \widetilde{P}_w]$ yet $v \notin [\widetilde{P} = \widetilde{P}_w]$, and hence that at w, $\widetilde{P} = \widetilde{P}_w$ but $\widetilde{P}(\widetilde{P} = \widetilde{P}_w) < 1$. Figure 12 represents an ambiguous frame wherein there are two possibly rational probability functions, $\widetilde{P}_a = \widetilde{P}_b = \eta$ and $\widetilde{P}_c = \widetilde{P}_d = \delta$, wherein η assigns 0.4 to δ being the rational function (and 0.6 to itself), while δ assigns 0.2 to η being the rational function (and 0.8 to itself). For more philosophical and technical background on such ambiguous probability frames, see Williamson 2008 and Dorst 2019, forthcoming.

Standard Bayesianism is a constraint on frames that captures the assumptions standardly built into Bayesian models. It holds if P has no higher-order uncertainty (the prior is known), and there is a partition whose cells represent the possible bits of evidence you could receive, such that $\widetilde{P}$ results from conditioning P on the true bit of evidence. Precisely:



$$\begin{array}{c} \eta \left\{ \begin{array}{c} \\ \\ \end{array} \right. \\ \delta \left\{ \begin{array}{c} \\ \\ \end{array} \right. \end{array} \begin{pmatrix} 0.3 & 0.3 & 0.3 & 0.1 \\ 0.3 & 0.3 & 0.3 & 0.1 \\ 0.1 & 0.1 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.6 & 0.2 \end{pmatrix}$$

Figure 12. An ambiguous frame. η assigns 0.3 to a, to b, and to c, and 0.1 to d; δ assigns 0.1 to a and to b, 0.6 to c, and 0.2 to d. Thus $\eta(\widetilde{P} = \eta) = 0.6$ and $\eta(\widetilde{P} = \delta) = 0.4$, while $\delta(\widetilde{P} = \eta) = 0.2$ and $\delta(\widetilde{P} = \delta) = 0.8$.

Definition. $\langle W, P, \widetilde{P} \rangle$ is *Standard Bayesian* if and only if there is a partition Π such that for each world w, $P_w(P = P_w) = 1$ and $\widetilde{P}_w(\cdot) = P_w(\cdot|\Pi(w))$, where $\Pi(w)$ is the partition cell of w.

This is (nearly) equivalent to the conjunction of Value and No Ambiguity.[62]

Theorem A.1. *If $\langle W, P, \widetilde{P} \rangle$ is Standard Bayesian, it validates No Ambiguity and Value. Conversely, if $\forall w$: $P_w(w) > 0$ (the prior is regular), No Ambiguity and Value are valid only if $\langle W, P, \widetilde{P} \rangle$ is Standard Bayesian.*

Proof. **(⇒:)** Suppose the update is Standard Bayesian. It is immediate that P satisfies No Ambiguity, since if $P = \pi = P_w$ at world w, then $P_w(P = \pi = P_w) = 1$. To show the same for $\widetilde{P}$, consider any $\widetilde{P}_w$. Since $\widetilde{P}_w = P_w(\cdot|\Pi(w))$, if $\widetilde{P}_w(x) > 0$, then $P_w(x) > 0$, and hence (since P satisfies No Ambiguity) $P_w = P_x$, that is, w and x share the same prior. Moreover, since $\widetilde{P}_w(\Pi(w)) = 1$, we know $x \in \Pi(w)$, so x and w are in the same partition cell: $\Pi(x) = \Pi(w)$, that is, w and x share the same evidence. It follows that $\widetilde{P}_x = P_x(\cdot|\Pi(x)) = P_w(\cdot|\Pi(w)) = \widetilde{P}_w$.

What remains is to show that P values $\widetilde{P}$. Consider any P_w and any decision problem $\mathcal{O}$ on W. Recall (section A.2) that a strategy S is a function from worlds v to options $S_v \in \mathcal{O}$ such that if $\widetilde{P}_v = \widetilde{P}_x$, then $S_v = S_x$. Also recall that S is recommended by $\widetilde{P}$ if and only if for each world v and any $O \in \mathcal{O}$, $\mathbb{E}_{\widetilde{P}_v}(S_v) \geq \mathbb{E}_{\widetilde{P}_v}(O)$. Notice that since $\widetilde{P}$ is not ambiguous

[62] Compare Samet 1999, who shows a similar result using a Reflection principle that is equivalent to No Ambiguity and Value, as shown in Dorst et al. 2021: fn. 17. See also Skyrms 1990 and Huttegger 2014, who show similar results assuming No Ambiguity.

it knows what option it recommends: for any v, $\widetilde{P}_v(\widetilde{P} = \widetilde{P}_v) = 1$ so that $\widetilde{P}_v(S = S_v) = 1$. Now we take an arbitrary option $O \in \mathcal{O}$ and show that $\mathbb{E}_{P_w}(S) \geq \mathbb{E}_{P_w}(O)$:

$$\begin{aligned}
\mathbb{E}_{P_w}(S) &= \sum_{\Pi(v)} P_w\big(\Pi(v)\big) \cdot \mathbb{E}_{P_w}\big(S|\Pi(v)\big) && \text{(total expectation)}\\
&= \sum_{\Pi(v)} P_w\big(\Pi(v)\big) \cdot \mathbb{E}_{\widetilde{P}_v}(S) && \text{(since } P_w\big(\cdot|\Pi(v)\big) = \widetilde{P}_v\text{)}\\
&= \sum_{\Pi(v)} P_w\big(\Pi(v)\big) \cdot \mathbb{E}_{\widetilde{P}_v}(S_v) && \text{(since } \widetilde{P}_v(S = S_v) = 1\text{)}\\
&\geq \sum_{\Pi(v)} P_w\big(\Pi(v)\big) \cdot \mathbb{E}_{\widetilde{P}_v}(O) && \text{(since } \mathbb{E}_{\widetilde{P}_v}(S_v) \geq \mathbb{E}_{\widetilde{P}_v}(O)\text{)}\\
&= \sum_{\Pi(v)} P_w\big(\Pi(v)\big) \cdot \mathbb{E}_{P_w}\big(O|\Pi(w)\big) = \mathbb{E}_{P_w}(O).
\end{aligned}$$

($\Leftarrow$:) Given $\langle W, P, \widetilde{P}\rangle$, suppose for all w, $P_w(w) > 0$ and the frame validates No Ambiguity and Value. No Ambiguity immediately implies that the prior is known: at each world w, $P_w(P = P_w) = 1$. Thus we need to find a partition Π such that $\widetilde{P}$ always results from conditioning P_w on the true member of Π.

Consider the possible posteriors (i.e., $\{\pi : \exists w : \widetilde{P}_w = \pi\}$), and label them $\pi_1, \ldots, \pi_n$. Notice that $\Pi := \{[\widetilde{P} = \pi_1], \ldots, [\widetilde{P} = \pi_n]\}$ partitions W and $\widetilde{P}$ is constant within each cell. Moreover, if $w \in [\widetilde{P} = \pi_i]$, then by No Ambiguity $\widetilde{P}_w(\widetilde{P} = \pi_i) = \widetilde{P}_w(\Pi(w)) = 1$, that is, $\widetilde{P}_w$ assigns probability 1 to its own partition cell.

Now suppose, for reductio, that there is a world w such that $\widetilde{P}_w \neq P_w(\cdot|\Pi(w))$. We know that $\widetilde{P}$ is constant within $\Pi(w)$, so there is a π such that for all $v \in \Pi(w)$, $\widetilde{P}_v = \pi$. Without loss of generality, suppose there is a q, t such that $\pi(q) > t > P_w(q|\Pi(w))$. We construct a decision problem that is a conditional bet on q given $\Pi(w)$ to show that P_w does not value $\widetilde{P}$. Let $\mathcal{O} = \{N, B\}$ where $N = 0$ everywhere, and

$$B(x) = \begin{cases} 1 - t & \text{if } x \in q \cap \Pi(w),\\ -t & \text{if } x \in \neg q \cap \Pi(w),\\ -1 & \text{if } x \notin \Pi(w). \end{cases}$$

What strategy is recommended by $\widetilde{P}$? Notice that for any $v \notin \Pi(w)$, by No Ambiguity $\widetilde{P}_v(\Pi(w)) = 0$, so $\widetilde{P}_v$ is certain that N pays out 0 while

B pays out -1; hence, $S_v = N$. Meanwhile, for any $x \in \Pi(w)$, we know that $\widetilde{P}_x(\Pi(w)) = 1$ and $\widetilde{P}_x(q) = \pi(q) > t$; hence, $\mathbb{E}_{\widetilde{P}_x}(B) > t(1-t) + (1-t)(-t) = 0 = \mathbb{E}_{\widetilde{P}_x}(N)$, and hence $S_x = B$. Thus the recommended strategy S is to take N at worlds not in $\Pi(w)$ and B at worlds inside it. But since P_w has a conditional credence in q given $\Pi(w)$ that is *below* t, it thinks this strategy is worse than simply taking N: $\mathbb{E}_{P_w}(S) = P_w(\neg\Pi(w)) \cdot 0 + P_w(\Pi(w)) \cdot \mathbb{E}_{P_w}(B|\Pi(w))$. Since $P_w(\Pi(w)) > 0$ (since $P_w(w) > 0$), this quantity is negative if and only if $\mathbb{E}_{P_w}(B|\Pi(w))$ is, and $\mathbb{E}_{P_w}(B|\Pi(w)) < t(1-t) + (1-t)(-t) = 0$; hence $\mathbb{E}_{P_w}(S) < 0 = \mathbb{E}_{P_w}(N)$. Value fails. □

Now turn to our impossibility result: given No Ambiguity, Value and Reflection are equivalent; if we assume Value as a constraint on rationality, Reflection failures (and expectable polarization) are possible only if evidence is ambiguous.

Theorem 3.1. *Given No Ambiguity, P values $\widetilde{P}$ iff P obeys Reflection toward $\widetilde{P}$.*

There are two steps. First we show that given No Ambiguity, Reflection is equivalent to an (otherwise stronger; see Dorst et al. 2021: fn. 18) 'Function Reflection' principle:

Function Reflection: $P_w(\cdot|\widetilde{P} = \pi) = \pi$ (whenever well defined).

Lemma 3.1.1. *Given No Ambiguity, Reflection holds if and only if Function Reflection holds.*

Proof. **(⇐:)** Notice that we can partition w into the possible posteriors $\widetilde{P}_1, \ldots, \widetilde{P}_n$; we have:

$$\begin{aligned}
&\mathbb{E}_{P_w}\big(\widetilde{P}(q)\big)\\
&\quad = \sum_{\widetilde{P}_i} P_w(\widetilde{P} = \widetilde{P}_i) \cdot \mathbb{E}_{P_w}\big(\widetilde{P}(q)|\widetilde{P} = \widetilde{P}_i\big) && \text{(total expectation)}\\
&\quad = \sum_{\widetilde{P}_i} P_w(\widetilde{P} = \widetilde{P}_i) \cdot \widetilde{P}_i(q)\\
&\quad = \sum_{\widetilde{P}_i} P_w(\widetilde{P} = \widetilde{P}_i) \cdot P_w(q|\widetilde{P} = \widetilde{P}_i) = P_w(q) && \text{(Function Reflection)}
\end{aligned}$$

(⇒:) For reductio, suppose there is a π such that $P_w(\cdot|\widetilde{P} = \pi) \neq \pi$. Without loss of generality, suppose $P_w(q|\widetilde{P} = \pi) > \pi(q)$. Consider

$q \wedge [\widetilde{P} = \pi]$. Since No Ambiguity is valid, at all worlds x, $\widetilde{P}_x(\widetilde{P} = \widetilde{P}_x) = 1$, so $\pi(q \wedge [\widetilde{P} = \pi]) = \pi(q)$; and if $\widetilde{P}_x \neq \pi$, then $\widetilde{P}_x(q \wedge [\widetilde{P} = \pi]) = 0$, so

$$\begin{aligned}
&\mathbb{E}_{P_w}(q \wedge \widetilde{P} = \pi)\\
&\quad = P_w(\widetilde{P} \neq \pi) \cdot 0 + P_w(\widetilde{P} = \pi) \cdot \pi\big(q \wedge [\widetilde{P} = \pi]\big)\\
&\quad = P_w(\widetilde{P} = \pi) \cdot \pi(q) \qquad\qquad (\text{since } \pi(\widetilde{P} = \pi) = 1)\\
&\quad < P_w(\widetilde{P} = \pi) \cdot P_w(q|\widetilde{P} = \pi)\\
&\quad = P_w\big(q \wedge [\widetilde{P} = \pi]\big).
\end{aligned}$$

So Reflection fails. □

Now we show that, given No Ambiguity, Function Reflection is equivalent to Value:

Lemma 3.1.2. *Given No Ambiguity, P_w values $\widetilde{P}$ iff it obeys Function Reflection.*

Proof. **(⇒:)** Suppose Function Reflection fails so there is a $\widetilde{P}_i$ and a w such that $P_w(\cdot|\widetilde{P} = \widetilde{P}_i) \neq \widetilde{P}_i$. Since this is well defined, we know that $P_w(\widetilde{P} = \widetilde{P}_i) > 0$. Without loss of generality, suppose $P_w(q|\widetilde{P} = \widetilde{P}_i) < t < \widetilde{P}_i(q)$. Let $\mathcal{O} = \{N, B\}$ where $N = 0$ everywhere and

$$B(x) = \begin{cases} 1 - t & \text{if } x \in q \cap [\widetilde{P} = \widetilde{P}_i],\\ -t & \text{if } x \in \neg q \cap [\widetilde{P} = \widetilde{P}_i],\\ -1 & \text{if } x \notin [\widetilde{P} = \widetilde{P}_i]. \end{cases}$$

What is recommended by $\widetilde{P}$? For any $v \notin \widetilde{P} = \widetilde{P}_i$, by No Ambiguity $\widetilde{P}_v(\widetilde{P} = \widetilde{P}_i) = 0$, so $S_v = N$. For any $x \in \widetilde{P} = \widetilde{P}_i$, we know that $\widetilde{P}_x(q) > t$ and by No Ambiguity $\widetilde{P}_x(\widetilde{P} = \widetilde{P}_i) = 1$, so $\mathbb{E}_{\widetilde{P}_x}(B) > 0 = \mathbb{E}_{\widetilde{P}_x}(N)$, so $S_x = B$. Thus the recommended strategy S takes N at $[\widetilde{P} \neq \widetilde{P}_i]$-worlds and B at $[\widetilde{P} = \widetilde{P}_i]$-worlds. So P_w's expectation of S is $\mathbb{E}_{P_w}(S) = P_w(\widetilde{P} \neq \widetilde{P}_i) \cdot 0 + P_w(\widetilde{P} = \widetilde{P}_i) \cdot \mathbb{E}_{P_w}(B|\widetilde{P} = \widetilde{P}_i)$. This is negative since $\mathbb{E}_{P_w}(B|\widetilde{P} = \widetilde{P}_i) < t \cdot (1 - t) + (1 - t)(-t) = 0$; hence $\mathbb{E}_{P_w}(S) < 0 = \mathbb{E}_{P_w}(N)$. Value fails.

(⇐:) Suppose P_w obeys Function Reflection. Taking an arbitrary $\mathcal{O}$ and recommended strategy S. Noting that that by No Ambiguity we have that $\widetilde{P}$ always knows what $\widetilde{P}$ is and hence what S recommends (so $\widetilde{P}_v(S = S_v) = 1$), we have:

$$\begin{aligned}
\mathbb{E}_{P_w}(S) &= \sum_{\widetilde{P}_i} P_w(\widetilde{P} = \widetilde{P}_i) \cdot \mathbb{E}_{P_w}(S|\widetilde{P} = \widetilde{P}_i) && \text{(total expectation)}\\
&= \sum_{\widetilde{P}_i} P_w(\widetilde{P} = \widetilde{P}_i) \cdot \mathbb{E}_{\widetilde{P}_i}(S) && \text{(Function Reflection)}\\
&= \sum_{\widetilde{P}_i} P_w(\widetilde{P} = \widetilde{P}_i) \cdot \mathbb{E}_{\widetilde{P}_i}(S_i) && (\widetilde{P}_i(S = S_i) = 1)\\
&\geq \sum_{\widetilde{P}_i} P_w(\widetilde{P} = \widetilde{P}_i) \cdot \mathbb{E}_{\widetilde{P}_i}(O) && (S \text{ is recommended})\\
&= \sum_{\widetilde{P}_i} P_w(\widetilde{P} = \widetilde{P}_i) \cdot \mathbb{E}_{P_w}(O|\widetilde{P} = \widetilde{P}_i) && \text{(Function Reflection)}\\
&= \mathbb{E}_{P_w}(O) && \text{(total expectation.)}
\end{aligned}$$

Thus Value holds. □

Theorem 3.1 is an immediate consequence of Lemmas 3.1.1 and 3.1.2.

Now turn to our possibility theorem (Theorem 3.2): whenever valuable evidence is ambiguous, it can be expectably polarizing. The easiest way to prove this is to appeal to the model-theoretic characterization of Value from Dorst et al. 2021. Given a function $\widetilde{P}_w$, we can consider its *informed* version $\widehat{\widetilde{P}}_w$, which removes its higher-order uncertainty (if it has any) by conditioning $\widetilde{P}_w$ on what the rational opinions were. Learning what the rational opinions were tells you how the rational opinions would respond to *that very information* (learning what $\widetilde{P}$ is tells you what all $\widetilde{P}$'s conditional opinions are as well), so $\widetilde{P}_w$ can then infer what new opinions are now rational upon learning what it learned (see Elga 2013; Stalnaker 2019; Dorst 2019). That is, let $\widehat{\widetilde{P}}_w := \widetilde{P}_w(\cdot|\widetilde{P} = \widetilde{P}_w)$. For example, informing η and δ from figure 12 (p. 405) would generate the frame in figure 11 (p. 402) since $\widehat{\eta} = \eta(\cdot|\widetilde{P} = \eta) = \eta(\cdot|\{a, b\}) = \pi$, and likewise $\widehat{\delta} = \rho$.

Now think of a probability function π over a set W of size $|W| = n$ as a point in Euclidean n-space, that is, a vector in which entry i is $\pi(w_i)$. The *convex hull* of a set of such points $\pi_1, \ldots \pi_n$ is the set of points obtainable by averaging them: $CH\{\pi_1, \ldots, \pi_n\} = \{\delta : \exists\lambda_i \geq 0$ and $\sum\lambda_i = 1$ such that $\delta = \sum\lambda_i\pi_i\}$. Given a probability function δ, let $\boldsymbol{C_\delta} := \{\pi : \delta(\widetilde{P} = \pi) > 0\}$ be the set of *C*andidates that δ thinks $\widetilde{P}$ might be. Let $C_\delta^- := C_\delta - \{\delta\}$ be the ones other than δ. Say that $\widetilde{P}_w$ is *modestly informed* if and only if it is an average of its informed self along with the other (uninformed) candidates, that is, if and only if $\widetilde{P}_w$ is in the convex hull of $\{\widehat{\widetilde{P}}_w\} \cup C^{-}_{\widetilde{P}_w}$. Then we have:

Theorem A.2 (Dorst et al. 2021, Theorem 4.1). *π values $\widetilde{P}$ iff each $\widetilde{P}_w$ in C_π is modestly informed, and π is in the convex hull of C_π.*

(A consequence is that if π values $\widetilde{P}$, then each $\widetilde{P}_w$ such that $\pi(w) > 0$ must also value $\widetilde{P}$.)

This allows us to prove that ambiguity suffices for valuable expectable polarization:

Theorem 3.2. *If $\widetilde{P}$ is valued by some π that assigns positive probability to it violating No Ambiguity, there are infinitely many P that value $\widetilde{P}$ and yet do not obey Reflection.*

Note that π assigns positive probability to $\widetilde{P}$ violating No Ambiguity if and only if $\pi(x) > 0$ with $\widetilde{P}_x(\widetilde{P} = \widetilde{P}_x) < 1$.

Proof. Let $\rho_1, \ldots, \rho_n$ be the potential realizations of $\widetilde{P}$, so $C_\pi = \{\rho_1, \ldots, \rho_n\}$. We know that each ρ_i is modestly informed and that π is in their convex hull.

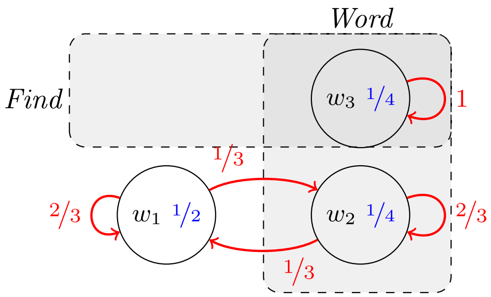
We begin by showing (following Samet 2000: Theorem 5) that, since one of the ρ_i is ambiguous, there is a $q \subseteq W$ and a ρ_i such that $\rho_i(q) \neq \mathbb{E}_{\rho_i}(\widetilde{P}(q))$. For reductio, suppose that for all ρ_i and q, $\rho_i(q) = \mathbb{E}_{\rho_i}(\widetilde{P}(q))$. Note that, formally, $\widetilde{P}$ is a finite Markov chain with W the state space and $\widetilde{P}_w(w')$ the probability of transitioning from w to w'. As such, we can partition W into its communicating classes $E_1, \ldots, E_k$, plus perhaps a set of transient states E_0. The claim that, for all q, $\rho_i(q) = \mathbb{E}_i(\widetilde{P}(q))$ is equivalent to the claim that ρ_i is a stationary distribution with respect to the Markov chain, that is, where M is the transition matrix and ρ_i is thought of as the (row) vector with $\rho_i(w_j)$ in column j, $\rho_i M = \rho_i$. By the Markov chain convergence theorem (e.g., Bertsekas and Tsitsiklis 2008, chap. 7), each $E_1, \ldots, E_k$ has a unique stationary distribution, and every stationary of M assigns 0 probability to E_0. These imply, first, that $\pi(E_0) = 0$, for otherwise π would not be in the convex hull of the (stationary) ρ_i. Since C_π includes all the ρ_i, this implies that E_0 is empty. Moreover, the fact that each E_i has a unique stationary, combined with our assumption that all $\rho_i(\cdot) = \mathbb{E}_{\rho_i}(\widetilde{P}(\cdot))$, implies that for any $w, w' \in E_i$, $\widetilde{P}_w = \widetilde{P}_{w'}$ since all $w \in E_i$ must equal that stationary. Since E_i is a communicating class, we also have that $\widetilde{P}_w(E_i) = 1$; hence $\widetilde{P}_w(\widetilde{P} = \widetilde{P}_w) = 1$.

Since this covers all the ρ_i, it implies that $\widetilde{P}$ is not ambiguous after all—contradiction.

Thus we reject our supposition: there are a ρ_i and q such that $\rho_i(q) \neq \mathbb{E}_{\rho_i}(\widetilde{P}(q))$. Without loss of generality, suppose $\rho_i(q) < \mathbb{E}_{\rho_i}(\widetilde{P}(q))$. Letting $\mathbb{1}_q$ be the indicator function of q (1 at $w \in q$, 0 elsewhere), $\rho_i(q) = \mathbb{E}_{\rho_i}(\mathbb{1}_q)$, so $\rho_i(q) < \mathbb{E}_{\rho_i}(\widetilde{P}(q))$ if and only if $0 < \mathbb{E}_{\rho_i}(\widetilde{P}(q)) - \mathbb{E}_{\rho_i}(\mathbb{1}_q)$ if and only if $\mathbb{E}_{\rho_i}(\widetilde{P}(q) - \mathbb{1}_q) > 0$. Thus it suffices to show that there are infinitely many δ such that δ values $\widetilde{P}$ and yet $\mathbb{E}_\delta(\widetilde{P}(q) - \mathbb{1}_q) > 0$. Pick some ρ_i that maximizes $\mathbb{E}_{\rho_i}(\widetilde{P}(q) - \mathbb{1}_q)$ within the frame (the frame is finite, so there is one), and pick any other $\rho_j \neq \rho_i$ (there must be at least one other since $\widetilde{P}$ is ambiguous). Now for any $\epsilon \in [0, 1]$, let $\eta_\epsilon := (1-\epsilon)\rho_i + \epsilon\rho_j$. Thinking of $\mathbb{E}_{\eta_\epsilon}(\widetilde{P}(q) - \mathbb{1}_q)$ as a function of ϵ, notice that this function is continuous and nonincreasing in ϵ, with maximum $\mathbb{E}_{\rho_i}(\widetilde{P}(q) - \mathbb{1}_q) > 0$ and minimum $\mathbb{E}_{\rho_j}(\widetilde{P}(q) - \mathbb{1}_q)$ (which may or may not be equal to $\mathbb{E}_{\rho_i}(\widetilde{P}(q) - \mathbb{1}_q)$). By the intermediate value theorem, this function must hit every value in between the two, meaning there are uncountably many values of ϵ such that $\mathbb{E}_{\eta_\epsilon}(\widetilde{P}(q) - \mathbb{1}_q) > 0$. Since each one of these η_ϵ are distinct (since $\rho_i \neq \rho_j$) and they are all in the convex hull of C_π (since $\rho_i, \rho_j \in C_\pi$), they all value $\widetilde{P}$ despite having $\eta_\epsilon(q) < \mathbb{E}_{\eta_\epsilon}(\widetilde{P}(q))$. So by picking various ϵ and then letting $P = \eta_\epsilon$ everywhere, we have infinitely many P that value $\widetilde{P}$ but do not obey Reflection toward it. □

Appendix A.4. Valuable Word Searches

Recall our simple word-search model, repeated from figure 2:



$$Word = \{w_2, w_3\}$$
$$Find = \{w_3\}$$
$$H = \begin{pmatrix} ^1/_2 & ^1/_4 & ^1/_4 \\ ^1/_2 & ^1/_4 & ^1/_4 \\ ^1/_2 & ^1/_4 & ^1/_4 \end{pmatrix}$$
$$\widetilde{H} = \begin{pmatrix} ^2/_3 & ^1/_3 & 0 \\ ^1/_3 & ^2/_3 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Intuitively H should value $\widetilde{H}$ since the latter is closer to the truth value of all propositions at all worlds. We can verify this using Theorem A.2 (p. 410). First, notice that $H_w = (^1/_2\ ^1/_4\ ^1/_4)$ is in the convex hull of the

$\widetilde{H}_i$ because $\frac{3}{4}H_{w_1} + \frac{1}{4}H_{w_3} = \frac{3}{4}(2/3\ 1/3\ 0) + \frac{1}{4}(0\ 0\ 1) = (2/4\ 1/4\ 1/4) = H_w$. Second, each $\widetilde{H}_i$ is modestly informed ($\widetilde{H}_{w_3}$ trivially so, as $\widetilde{H}_{w_3} = \widehat{\widetilde{H}}_{w_3}$). Note that $\widehat{\widetilde{H}}_{w_2} = (0\ 1\ 0)$ and $\widehat{\widetilde{H}}_{w_1} = (1\ 0\ 0)$. Thus $\widetilde{H}_{w_2} = (1/3\ 2/3\ 0) = \frac{1}{2}(2/3\ 1/3\ 0) + \frac{1}{2}(0\ 1\ 0) = \frac{1}{2}\widetilde{H}_{w_1} + \frac{1}{2}\widehat{\widetilde{H}}_{w_2}$, so $\widetilde{H}_{w_2}$ is modestly informed. Likewise, $\widetilde{H}_{w_1} = \frac{1}{2}\widehat{\widetilde{H}}_{w_1} + \frac{1}{2}H_{w_2}$.

Notably, recalling footnote 61, since H knows what H is, it thereby not only values $\widetilde{H}$ but also prefers $\widetilde{H}$ (whatever it is) to H (whatever *it* is) for all decision problems. This holds despite the fact that Reflection fails: $\mathbb{E}_{H_w}(\widetilde{H}(\mathit{Word})) \approx 0.583 > 0.5 = H_w(\mathit{Word})$.

This feature—that word searches are valuable but expectably polarizing—holds generally: a wide class of models of this structure are expected to increase your credence in *Word*, despite being valuable. Let a *word-search model* be as follows. There are three classes of worlds, $\{N, C, F\}$, where N is the set of worlds where there is no word, C is the set where there is one but you do not find it, and F is the set where you find it. $\mathit{Word} = C \cup F$ is the proposition that there is a word. The posterior always knows whether you found one: if $x \in F$, $\widetilde{H}_x(F) = 1$ and if $x \notin F$, $\widetilde{H}_x(F) = 0$. The prior H assigns positive probability to all worlds; let it be constant across worlds so that the prior has no higher-order uncertainty. Say the search is *bounded by conditioning* if and only if $\min_{n\in N} \widetilde{H}_n(\mathit{Word}) = H_w(\mathit{Word}|\neg F)$. Say that a search is *possibly ambiguous* if and only if you might be unsure of the rational posterior in *Word*, that is, if and only if there is an x such that for all t, $\widetilde{H}_x(\widetilde{H}(\mathit{Word}) = t) < 1$. Then:

Fact A.3. If H values $\widetilde{H}$ in a word-search model $\langle W, H, \widetilde{H}\rangle$ that is bounded by conditioning and possibly ambiguous, then $\mathbb{E}_H(\widetilde{H}(\mathit{Word})) > H(\mathit{Word})$.

Proof. Since $\widetilde{H}$ is possibly ambiguous, there is a $v \in W$ such that $\widetilde{H}_v(\widetilde{H}(\mathit{Word}) = t) < 1$ for all t. This v cannot be in F. Since H values $\widetilde{H}$, each $\widetilde{H}_w$ must value $\widetilde{H}$ as well. This implies that they must totally trust H (section A.2). Since for any $f \in F$, $\widetilde{H}_f(\mathit{Word}) = 1$ and $\widetilde{H}_f(\mathit{Word}|\widetilde{H}(\mathit{Word}) \leq t) \leq t$, we must have that $\widetilde{H}_f(\widetilde{H}(\mathit{Word}) \leq t) = 0$ for all $t < 1$; in other words, $\widetilde{H}_f(\widetilde{H}(\mathit{Word}) = 1) = 1$. Thus v must be in $N \cup C$.

Since any $v \in N \cup C = \neg F$ has $\widetilde{H}_v(N \cup C) = 1$, this implies that there must be at least two values of $\widetilde{H}(\mathit{Word})$ in $N \cup C$, so $\exists x \in N \cup C$ such that $\widetilde{H}_x(\mathit{Word}) \neq H_w(\mathit{Word}|\neg F)$. We know that $\forall n \in N$: $\widetilde{H}_n(\mathit{Word}) \geq H_w(\mathit{Word}|\neg F)$. Suppose for reductio that there is some $y \in C$

with $\widetilde{H}_y(Word) = t < H_w(Word|\neg F)$. Since this is lower than any $n \in N$, we have $[\widetilde{H}(Word) \leq t] \subseteq Word$; hence $H_w(Word|\widetilde{H}(Word \leq t) = 1 > t$, and hence H_w does not obey Total Trust toward $\widetilde{H}$—contradiction.

Thus for all y, $\widetilde{H}_y(Word) \geq H_w(Word|\neg F)$. Since there are at least two values of $\widetilde{H}(Word)$ in $N \cup C = \neg F$, there must be some $x \in \neg F$ such that $\widetilde{H}_x(Word) > H_w(Word|\neg F)$. Since H_w assigns positive probability to all worlds, this implies that $\mathbb{E}_{H_w}(\widetilde{H}(Word)|\neg F) > H_w(Word|\neg F)$. And from here we can infer that

$$\begin{aligned}\mathbb{E}_{H_w}\big(\widetilde{H}(Word)\big) &= H_w(F)\cdot 1 + H_w(\neg F)\cdot \mathbb{E}_{H_w}\big(\widetilde{H}(Word)|\neg F\big)\\ &> H_w(F)\cdot H_w(Word|F) + H_w(\neg F)\cdot H_w(Word|\neg F)\\ &= H_w(Word). \qquad \square\end{aligned}$$

Appendix A.5. Question-Relative Value

First we show that full Value is 'transitive', as discussed in section 5:

Fact A.4. If P^1 values P^2 and P^2 values P^3, then P^1 values P^3.

Proof. Consider any P^1_w, and let $C^3_w = \{P^3_v : P^1_w(v) > 0\}$ be the set of candidates P^1_w thinks P^3 might be. By Theorem A.2, it suffices to show that each $P^3_v \in C^3_w$ is modestly informed and that P^1_w is in their convex hull. Take an arbitrary P^3_v in C^3_w. There must be an x such that $P^1_w(x) > 0$ and $P^2_x(P^3 = P^3_v) > 0$—if not, then $P^1_w(P^3 = P^3_v|P^2(P^3 = P^3_v) \leq 0) = P^1(P^3 = P^3_v) > 0$, violating Total Trust and (so) the assumption that P^1_w values P^2. Since P^2 values P^3 and $P^2_x(P^3 = P^3_v) > 0$, this means P^3_v is modestly informed.

Now we show that P^1_w is in the convex hull of C^3_w. Let $C^2_w = \{P^2_x : P^1_w(x) > 0\}$, and take an arbitrary $P^2_x \in C^2_w$. If $P^2_x(P^3 = \pi) > 0$ for $\pi \notin C^3_w$, then $P^1_w(P^3 = \pi|P^2(P^3 = \pi) > 0) = 0$, violating Total Trust and hence the assumption that P^1_w values P^2. Thus $P^2_x(P^3 = \pi) > 0$ only if $\pi \in C^3_w$. Since P^2_x values P^3, this means that P^2_x is in the convex hull of C^3_w. Since P^2_x was arbitrary, this means *all* members of C^2_w are in the convex hull of C^3_w, so $CH(C^2_w) \subseteq CH(C^3_w)$. Since P^1_w values P^2, P^1_w is inside the former and so also inside the latter. □

Now turn to question-relative value. A question Q is a partition of W; let $Q(w)$ be the partition cell of w. A proposition $p \subseteq W$ is *about* Q if and only if $p = \bigcup_i q_i$ for $q_i \in Q$, that is if and only if p is a partial answer to the question Q. Recall that a decision problem $\mathcal{O}$ is any set of options (i.e., functions from worlds to numbers) on W. Say that an option O is *Q-measurable* if and only if Q settles the value of O, that is, for all w, w', if $Q(w) = Q(w')$, then $O(w) = O(w')$. Say that $\mathcal{O}_Q$ is a *decision about* Q if and only if each of its options is Q-measurable. Then π *Q-values* $\widetilde{P}$ if and only if it prefers to let $\widetilde{P}$ make any decision *about* Q. Lifting this to P:

> **Q-Value:** P Q-values $\widetilde{P}$ iff for all w and every $\mathcal{O}_Q$ about Q, if $\widetilde{P}$ recommends S for $\mathcal{O}_Q$, then $\forall O \in \mathcal{O}_Q : \mathbb{E}_{P_w}(S) \geq \mathbb{E}_{P_w}(O)$.
> P Q-values $\widetilde{P}$ iff, for any decision about Q, it prefers to let $\widetilde{P}$ decide on its behalf rather than make the decision itself.

As mentioned in the main text, we can also question-relativize our definition of a Dutch book. A *fixed-option Q-book* is a pair of decisions *about* Q—both containing a 'no bet' option, one presented before and the other after the update—such that doing the rational thing before and after is guaranteed to result in a loss. Formally, given P_w and $\widetilde{P}$, it is a pair $\mathcal{O}^1_Q$ and $\mathcal{O}^2_Q$ of decision problems about Q that both include a constant $O_0 = 0$ option, where $O \in \arg\max_{O' \in \mathcal{O}^1_Q} \mathbb{E}_{P_w}(O')$ and S is recommended by $\widetilde{P}$ for $\mathcal{O}^2_Q$ and yet $O(w) + S_w(w) < 0$ at every world w. Q-Value entails that no such book can be constructed against the update:

Theorem A.5. *If P_w Q-values $\widetilde{P}$, then there is no fixed-option Q-book against $\langle P_w, \widetilde{P} \rangle$.*

Proof. Suppose P_w Q-values $\widetilde{P}$, take any $\mathcal{O}^1_Q$ and $\mathcal{O}^2_Q$ about Q that both contain an $O_0 = 0$ option, and suppose O maximizes expectation amongst $\mathcal{O}^1_Q$ relative to P_w and S is recommended for $\mathcal{O}^2_Q$ by $\widetilde{P}$. By definition, $\mathbb{E}_{P_w}(O) \geq \mathbb{E}_{P_w}(O_0) = 0$, and since P_w values $\widetilde{P}$ about Q, $\mathbb{E}_{P_w}(S) \geq \mathbb{E}_{P_w}(O_0) = 0$; hence $\mathbb{E}_{P_w}(O + S) = \mathbb{E}_{P_w}(O) + \mathbb{E}_{P_w}(S) \geq 0$. Thus $\mathcal{O}^1_Q$ and $\mathcal{O}^2_Q$ do not constitute a Q-book, for if they did, then $P_w(O + S < 0) = 1$. □

Finally, note that one decision about Q is to choose a set of opinions about Q to be scored for accuracy. Thus Q-Value entails that P_w expects $\widetilde{P}$ to be at least as accurate as itself on any proper measure of the accuracy of opinions about Q (see Dorst et al. 2021, section 3).

Appendix A.6. The Predictable Theorem

I now turn to proving that updates that are valuable with respect to Q can nevertheless lead to predictable, persistent polarization about Q. The proof is long and the method is unintuitive, so I should say something about why. As mentioned in footnote 42, it would be straightforward to generate predictable polarization by iterating word-search tasks if we allowed the question Q Haley cares about to (predictably) change over time—then we could simply say that at time i she cares (only) about the outcome of the ith word search, and since each of those is valuable with respect to *that* question, there would be no obstacle to iteration. Though a sensible (and perhaps realistic) route to polarization, this faces the concern that it is not too surprising that Haley polarizes about *how many coins landed heads* if her updates are not constrained to be valuable about *that* question. The point of the following construction is to show that she can at all times care about the same question Q (namely, how all the word searches went; hence how all the coins landed and whether more than half landed heads), and nonetheless Q-Value will not prevent her from predictably polarizing on that question. The method of the construction—using *consolidations* of higher-order uncertainty, as discussed in the main text—is, I admit, rather baroque. But it is a possibility proof. I conjecture that there are more intuitive ways to get the same result.

I proceed in stages. First, I specify a model that iterates word-search tasks and consolidates higher-order uncertainty along the way. I then prove that each word-search update is fully valuable, while each consolidation update is valuable with respect to Q. I then establish the long-run predictable behavior of the final rational credence $\overline{H^n}$ in this model, showing it predictably polarizes on the proposition $h =$ *more than half the coins landed heads.* Finally, I add a Tailser to show that the polarization is also persistent.

Here is our initial goal:

Theorem 5.1. *There is a sequence of probability functions H^0, $\overline{H^0}$, H^1, $\overline{H^1}$...,H^n, $\overline{H^n}$, a partition Q, and a proposition $h = \bigcup_i q_i$ (for $q_i \in Q$) such that, as $n \to \infty$:*

- *H^0 is (correctly) certain that $\overline{H^i}$ values H^{i+1}, for each i;*
- *H^0 is (correctly) certain that H^i Q-values $\overline{H^i}$, for each i;*
- *the sequence is predictably polarizing about h: $H^0(h) \approx \frac{1}{2}$, yet $H^0(\overline{H^n}(h) \approx 1) \approx 1$.*

Haley the Headser faces a sequence of n independent word-search tasks, each determined by the toss of a (new, independent) fair coin that she is 50% confident will land heads. Since we want to consolidate her higher-order uncertainty between each update, we must include additional possibilities, initially ignored, where the outcome of each task is the same but her rational credence function updates in different ways; consolidations will use these possibilities to hold fixed her opinions in how the tasks went but remove her higher-order uncertainty.

For each task $i = 1, \ldots, n$, let $X_i = \{n_i, n_i', c_i, c_i', f_i\}$ be the set of outcomes. f_i indicates that she finds the completion, c_i and c_i' are where it is completable but she does not find it, and n_i and n_i' are where it is not completable. (c_i' and n_i' are the 'weird' outcomes, initially ignored, where the rational credence function updates differently.) Let our set of worlds $W = X_1 \times \cdots \times X_n$ be the sequence of all possible outcomes. Let $U = \{w : \exists i : c_i' \in w \text{ or } n_i' \in w\}$ be the set of weird update sequences that contain at least one c_i' or n_i'.

Over W we lay some partitions. Let

$$\begin{aligned} N_i &= \{w \in W : n_i \in w \text{ or } n_i' \in w\}, \\ C_i &= \{w \in W : c_i \in w \text{ or } c_i' \in w\}, \\ F_i &= \{w \in W : f_i \in w\}. \end{aligned}$$

Now, let $Q_i = \{N_i, C_i, F_i\}$ be the question of how the ith task went (did she find one, was there a completable one she missed, or was it not completable?), ignoring the further question of how her rational opinions changed. Now let Q be the combination of all these partitions so that $Q(x) = Q(y)$ if and only if for all i, $Q_i(x) = Q_i(y)$. Notice that $Heads_i = F_i \cup C_i$ and thus that any proposition about how the coins landed—one definable by specifying a set of sequences of heads and tails—is about Q.

Finally let U_i be the question of how the rational credence updated at i, so $U^i = \{U_n^i, U_c^i, U_f^i\}$ where

$$\begin{aligned} U_n^i &= \{w \in W : n_i \in w \text{ or } c_i' \in w\}, \\ U_c^i &= \{w \in W : c_i \in w \text{ or } n_i' \in w\}, \\ U_f^i &= \{w \in W : f_i \in w\}. \end{aligned}$$

As we will see, U_n^i is the set of worlds where H^i updated *as if* there was no completion (as if n_i) and U_c^i is that where H^i updated as if there was one (as if c_i).

Here are a few more bits of notation. Given a probability function π, let $\pi[x,y,z]_k$ (with $x,y,z \geq 0$ and summing to 1) be the probability function that results from Jeffrey-shifting (Jeffrey 1990) π on the partition $Q_k = \{N_k, C_k, F_k\}$ such that the posterior assigns x to N_k, y to C_k, and z to F_k. Explicitly, for any $p \subseteq W$:

$$\pi[x,y,z]_k(p) := x \cdot \pi(p \mid N_k) + y \cdot \pi(p \mid C_k) + z \cdot \pi(p \mid F_k).$$

Higher-order consolidations will happen by *imaging* (Lewis 1976): intuitively, throwing all probability mass from a set of worlds onto their 'closest' neighbors in which a given claim is true. Thus we will need to define a corresponding selection function (Stalnaker 1968) telling us which these closest neighbors are. Let $\wp(W)$ be the power set of W, that is, the set of propositions. For each world $w \in W$, let $g_w : \wp(W) \to W$ be a selection function that, given a nonempty proposition $p \in \wp(W)$ ($p \neq \emptyset$), outputs a world $g_w(p) \in p$ that is the 'closest' one to w in which p is true. We assume g obeys:

> **Strong centering:** if $w \in p$, then $g_w(p) = w$.
>
> *Q***-respecting:** if possible, g_w selects a world that agrees with w about Q.
> If $\exists x \in p$ such that $Q(x) = Q(w)$, then $g_w(p) \in Q(w)$.
>
> **Sequence-respecting:** g_w selects a world that agrees with w in as much of its final sequence as possible.
> If there are two worlds $x = \langle x_1, \ldots, x_n \rangle$ and $y = \langle y_1, \ldots, y_n \rangle$ that both are in p and have $Q(x) = Q(w) = Q(y)$ but y has a longer w-agreeing end-sequence ($x_n = w_n, \ldots$ but $x_{n-k} \neq w_{n-k}$, and $y_n = w_n, \ldots, y_{n-k} = w_{n-k}$), then $g_w(p) \neq x$.

Following Lewis 1976, for any probability function π, we let π imaged on p, $\pi(\cdot \| p)$, be the result of shifting all probability π assigns to $\neg p$-worlds to their closest p-world counterparts. Formally, for any world w:

$$\pi(w \| p) := \sum_{y \in W : g_y(p) = w} \pi(y).$$

Imaging shifts probability mass around but neither creates nor destroys it, so $\pi(\cdot \| p)$ is always a probability function. As a result, note that for any $r \subseteq W$:

$$\pi(r \| p) = \sum_{w \in r} \pi(w \| p)$$

$$= \sum_{w \in r} \sum_{y \in W : g_y(p) = w} \pi(y)$$
$$= \sum_{y \in W : g_y \in r} \pi(y).$$

Machinery in place, we can now define the series of probability functions $H^0, \overline{H^0}, H^1, \overline{H^1}, \ldots, H^n, \overline{H^n}$ that represent Haley's rational opinions over time. (H^i is that right after completing the ith word-search task, while $\overline{H^i}$ is some time after that when she has forgotten the string and so consolidated her higher-order uncertainty.) Recall that H^i is a description (so it picks out different probability functions at different worlds), whereas H^i_w is a rigid designator (that always picks out the function that H^i associates with w).

Recalling that $U = \{w : \exists i : c'_i \in w \text{ or } n'_i \in w\}$ is the set of worlds that contain a weird update, for any world $w \in W$, let H^0_w be such that $H^0_w(U) = 0$, and for each Q_j:

$H^0_w(N_i) = 1/2$;
$H^0_w(C_i) = 1/4$;
$H^0_w(F_i) = 1/4$.

Moreover assume H^0_w treats the Q_j as mutually independent; thus for any $q_{i_1}, \ldots, q_{i_k}$ in $Q_{j_1}, \ldots, Q_{j_k}$ respectively, $H^0_w(q_{i_1} \& \ldots \& q_{i_k}) = H^0_w(q_{i_1})H^0_w(q_{i_2}) \cdots H^0_w(q_{i_k})$. Since $H^0_w(U) = 0$, this pins down H^0_w uniquely over W; hence all worlds begin with the same prior.

Now define updates. For any world w and task i, the *consolidation* $\overline{H^i}$ comes by imaging on the proposition that the H^i equals the particular function H^i_w. Formally, for all w and i,

$$\overline{H^i_w} := H^i_w(\cdot \| H^i = H^i_w).$$

As we will see, these consolidation updates change her higher-order opinions (removing higher-order doubts) without changing her opinions about Q.

Finally, we define the regular (nonconsolidation) updates as Jeffrey shifts in the way indicated by the word-search model, except that c'_{i+1} and n'_{i+1} (the ones initially assigned 0 probability) update in the opposite way from what their word-search outcome would indicate. Thus for all w and $i < n$:

if $f_{i+1} \in w$, then $H^{i+1}_w = \overline{H^i_w}[0, 0, 1]_{i+1}$;

if $c_{i+1} \in w$ or $n'_{i+1} \in w$, then $H_w^{i+1} = \overline{H_w^i}[\frac{1}{3}, \frac{2}{3}, 0]_{i+1}$;
if $n_{i+1} \in w$ or $c'_{i+1} \in w$, then $H_w^{i+1} = \overline{H_w^i}[\frac{2}{3}, \frac{1}{3}, 0]_{i+1}$.

Having defined the iteration model, we now establish a variety of its features, including that its updates are (Q-)valuable, and the long-run behavior of H^n.

Lemma 5.1.1.

(1) *For each i and w, $\overline{H_w^i}$ is higher-order certain.*
(2) *Moreover, for $i > 1$, if $H_w^i(x) > 0$, then $\overline{H_w^{i-1}} = \overline{H_x^{i-1}}$.*

Proof. (1) Suppose $\overline{H_w^i}(x) > 0$; we show that $\overline{H_x^i} = \overline{H_w^i}$. By definition, $\overline{H_w^i}(x) = H_w^i(x \| H^i = H_w^i) > 0$. By the definition of imaging, $x \in [H^i = H_w^i]$, that is, $H_x^i = H_w^i$. Thus $\overline{H_x^i} = H_x^i(\cdot \| H^i = H_x^i) = H_w^i(\cdot \| H^i = H_w^i) = \overline{H_w^i}$. Since x was arbitrary, $\overline{H_w^i}(\overline{H^i} = \overline{H_w^i}) = 1$.

(2) By definition, H_w^i is obtained from $\overline{H_w^{i-1}}$ by Jeffrey-shifting in a way that preserves certainties; therefore if $H_w^i(x) > 0$, then $\overline{H_w^{i-1}}(x) > 0$, so by (1), $\overline{H_w^{i-1}} = \overline{H_x^{i-1}}$. □

Now we show that weird updates (n'_i and c'_i) are assigned probability 0 ahead of time:

Lemma 5.1.2. *For any w, x, $i < j$, if $n'_j \in x$ or $c'_j \in x$, then $H_w^i(x) = 0$ and $\overline{H_w^i}(x) = 0$.*

Proof. Proof is by induction. *Base case*: By construction, $H_w^0(U) = 0$, so $H_w^0(x) = 0$. Since $\overline{H_x^0} = H_x^0$, this is likewise so for $\overline{H_x^0}$. *Induction case*: Supposing it holds for all w with $k < i$, we show it holds for i. Since $H_w^i = \overline{H_w^{i-1}}[a_1, a_2, a_3]_i$ and this does not raise any probabilities from 0, since (by induction) $\overline{H_w^{i-1}}(x) = 0$, likewise $H_w^i(x) = 0$. Now suppose, for reductio, $\overline{H_w^i}(x) > 0$. Thus there must be a y such that $H_w^i(y) > 0$ and $g_y(H^i = H_w^i) = x$. But since H_w^i did not assign positive probability to any world with n'_j or c'_j in it, those are not in y and yet they are in x. If $H_y^i = H_w^i$, then (by strong centering) $g_y(H^i = H_w^i) = y$, so this is impossible; hence $H_y^i \neq H_w^i$. Since $H_w^i(y) > 0$, and if $w \in f_i$ then H_w^i would be higher-order certain, it must be that either (i) $w \in U_c^i$ and $y \in U_n^i$ or (ii) $w \in U_n^i$ and $y \in U_c^i$. Since we must have had $\overline{H_w^{i-1}}(y) > 0$, by the inductive hypothesis, we know either $c_i \in y$ or $n_i \in y$ (not $c'_i \in y$ or $n'_i \in y$). So if (i), then $y' = \langle y_1, \ldots, n'_i, \ldots, y_n \rangle$—which swaps out n'_i for n_i in y and is a world that is in the same Q-cell as y—updates the same as w, so $H^i_{y'} = H^i_w$. Since y' agrees with the end-sequence of y more than x does (since $n'_j \in x$ or $c'_j \in x$), by sequence-respecting, $g_y(H^i = H^i_w) \neq x$—contradiction. If (ii), parallel reasoning works by substituting c'_i into y, completing the proof. □

We now show that our consolidations never move probability mass from one Q-cell to another:

Lemma 5.1.3. *For any x, i, if $H^i_x(y) > 0$, then $g_y(H^i = H^i_x) \in Q(y)$.*

Proof. Suppose $H^i_x(y) > 0$. By Lemma 5.1.1, $\overline{H^{i-1}_x} = \overline{H^{i-1}_y}$. By Lemma 5.1.2 and the fact that H^i_x preserves $\overline{H^{i-1}_x}$'s certainties, neither $c'_i \in y$ nor $n'_i \in y$; hence either $f_i \in y$ or $c_i \in y$ or $n_i \in y$.

If $f_i \in x$, then of course $f_i \in y$ and so $H^i_y = H^i_x$, meaning that by strong centering $g_y(H^i = H^i_x) = y$, establishing the result.

If $c_i \in x$ or $n'_i \in x$, then $H^i_x = \overline{H^{i-1}_x}[\frac{1}{3}, \frac{2}{3}, 0]_i$. If $c_i \in y$, then $H^i_y = H^i_x$, so again we have the result. But suppose $n_i \in y$ instead. Then $y = \langle y_1, \ldots, y_{i-1}, n_i, y_{i+1}, \ldots, y_n \rangle$. Consider the possibility $y' = \langle y_1, \ldots, y_{i-1}, n'_i, y_{i+1}, \ldots, y_n \rangle$, which is the same as y except that it swaps n'_i for n_i. By construction, $Q(y') = Q(y)$ and $\overline{H^{i-1}_{y'}} = \overline{H^{i-1}_y} = \overline{H^{i-1}_x}$, so

$$\begin{aligned} H^i_{y'} &= \overline{H^{i-1}_{y'}}[\tfrac{1}{3}, \tfrac{2}{3}, 0]_i \\ &= \overline{H^{i-1}_x}[\tfrac{1}{3}, \tfrac{2}{3}, 0]_i = H^i_x. \end{aligned}$$

Thus there is a y' in $[H^i = H^i_x]$ such that $Q(y') = Q(y)$, so by Q-respecting, $g_y(H^i = H^i_x) \in Q(y)$, establishing the result.

If $n_i \in x$ or $c' \in x$, parallel reasoning (substituting c'_i for c_i) establishes the result. □

Since consolidations never move probability mass from one Q-cell to another, they do not change any opinions about Q:

Lemma 5.1.4. *For all x, i, and $q \in Q$, $\overline{H^i_x}(q) = H^i_x(q)$.*

Proof. By construction and the definition of imaging,

$$\begin{aligned} \overline{H^i_x}(q) &= H^i_x(q \| H^i = H^i_x) \\ &= \sum_{y \in W : g_y(H^i = H^i_x) \in q} H^i_x(y) \end{aligned}$$

$$= \sum_{y \in q : g_y(H^i = H^i_x) \in q} H^i_x(y) + \sum_{y \notin q : g_y(H^i = H^i_x) \in q} H^i_x(y).$$

By Lemma 5.1.3, all and only worlds in q map to worlds in q under $H^i = H^i_x$; thus $\{y \in q : g_y(H^i = H^i_x) \in q\} = \{y : y \in q\}$ and $\{y \notin q : g_y(H^i = H^i_x) \in q\} = \emptyset$. Therefore the right summand is 0, and the left summand equals $\sum_{y \in q} H^i_x(y) = H^i_x(q)$, as desired. □

Lemma 5.1.5. *For any w, $i < j$, $\overline{H^i_w}(F_j) = \overline{H^i_w}(C_j) = \frac{1}{4}$ and $\overline{H^i_w}(N_j) = \frac{1}{2}$ and $\overline{H^i_w}$ treats the Q_k as mutually independent.*

Proof. Proof is by induction. The *base case* is trivial by definition of H^0_w. *Induction step*: Suppose it holds for $k < i$. By definition, H^i_w is obtained by Jeffrey-shifting $\overline{H^{i-1}_w}$ on Q_i; since by the induction hypothesis $\overline{H^{i-1}_w}$ treats the Q_k as mutually independent and assigns $\frac{1}{4}$ to F_j and C_j, and $\frac{1}{2}$ to N_j, H^i_w does too. Now, by Lemma 5.1.4, $\overline{H^i_w}$ maintains the same distribution over Q as H^i_w has, establishing the result. □

Now we can establish that the Jeffrey-shift updates are fully valuable and that the consolidation updates are Q-valuable.

Lemma 5.1.6. *For all w and i, $\overline{H^i_w}$ values H^{i+1}_w.*

Proof. Letting $S^i_w := \{x \in W : \overline{H^i_w}(x) > 0\}$ be the support of H^i_w, by Theorem A.2, we must show that (1) for each $x \in S^i_w$, H^{i+1}_x is modestly informed and (2) $\overline{H^i_w}$ is in their convex hull.

(1) Taking an arbitrary $x \in S^i_w$, we show that H^{i+1}_x is modestly informed. By Lemma 5.1.1, note that since $\overline{H^i_w}(x)$ is higher-order certain, $\overline{H^i_x} = \overline{H^i_w}$. Now either (i) $f_{i+1} \in x$, or (ii) $c_{i+1} \in x$ or $n'_{i+1} \in x$, or (iii) $n_{i+1} \in x$ or $c'_{i+1} \in x$. Supposing (i), then $H^{i+1}_x = \overline{H^i_x}[0,0,1]_{i+1}$, meaning $H^{i+1}_x(F_i) = 1$ so that if $H^{i+1}_x(y) > 0$, then $f_{i+1} \in y$, and $H^{i+1}_y = H^{i+1}_x$. Hence $H^{i+1}_x(H^{i+1} = H^{i+1}_x) = 1$, so trivially H^{i+1} is modestly informed. On the other hand, if (ii) holds, then $H^{i+1}_x = \overline{H^i_x}[\frac{1}{3}, \frac{2}{3}, 0]_{i+1} = \overline{H^i_w}[\frac{1}{3}, \frac{2}{3}, 0]_{i+1}$—label this function π_c. If (iii) holds, then $H^{i+1}_x = \overline{H^i_x}[\frac{2}{3}, \frac{1}{3}, 0]_{i+1} = \overline{H^i_w}[\frac{2}{3}, \frac{1}{3}, 0]_{i+1}$—label this function π_n. Note that π_c and π_n both assign 1 to S^i_w and also assign 1 to $[H^{i+1} = \pi_c] \vee [H^{i+1} = \pi_n]$. Now, since by Lemma 5.1.2 we have that $\overline{H^i_w}$ assigns 0 to any world with n'_{i+1} or c'_{i+1} in it, it follows that π_c and π_n do too and hence that:

$$\widehat{\pi}_c = \pi_c(\cdot|H^{i+1} = \pi_c) = \overline{H^i_w}(\cdot|C_{i+1}),$$
$$\widehat{\pi}_n = \pi_n(\cdot|H^{i+1} = \pi_n) = \overline{H^i_w}(\cdot|N_{i+1}).$$

From this it follows that π_c (and, by parallel reasoning, π_n) is modestly informed, since

$$\begin{aligned}\frac{1}{2}\widehat{\pi}_c + \frac{1}{2}\pi_n &= \frac{1}{2}\overline{H^i_w}(\cdot|C_{i+1}) + \frac{1}{2}\Big(\frac{1}{3}\overline{H^i_w}(\cdot|C_{i+1}) + \frac{2}{3}\overline{H^i_w}(\cdot|N_{i+1})\Big)\\ &= \frac{1}{2}\overline{H^i_w}(\cdot|C_{i+1}) + \frac{1}{6}\overline{H^i_w}(\cdot|C_{i+1}) + \frac{1}{3}\overline{H^i_w}(\cdot|N_{i+1})\\ &= \frac{2}{3}\overline{H^i_w}(\cdot|C_{i+1}) + \frac{1}{3}\overline{H^i_w}(\cdot|N_{i+1})\\ &= \pi_c.\end{aligned}$$

Since π_c, π_n, and $\overline{H^i_w}(\cdot|F_{i+1})$ are the three realizations of H^{i+1} in S^i_w, this establishes (1).

(2) We now show that $\overline{H^i_w}$ is in their convex hull. Note that by Lemma 5.1.5 and total probability,

$$\overline{H^i_w} = \frac{1}{2}\overline{H^i_w}(\cdot|N_{i+1}) + \frac{1}{4}\overline{H^i_w}(\cdot|C_{i+1}) + \frac{1}{4}\overline{H^i_w}(\cdot|F_{i+1}).$$

Now notice that

$$\begin{aligned}&\frac{1}{4}\overline{H^i_w}(\cdot|F_{i+1}) + \frac{3}{4}\pi_n\\ &\quad= \frac{1}{4}\overline{H^i_w}(\cdot|F_{i+1}) + \frac{3}{4}\Big(\frac{1}{3}\overline{H^i_w}(\cdot|C_{i+1}) + \frac{2}{3}\overline{H^i_w}(\cdot|N_{i+1})\Big)\\ &\quad= \frac{1}{4}\overline{H^i_w}(\cdot|F_{i+1}) + \frac{1}{4}\overline{H^i_w}(\cdot|C_{i+1}) + \frac{1}{2}\overline{H^i_w}(\cdot|N_{i+1}) = \overline{H^i_w}.\end{aligned}$$

This establishes that $\overline{H^i_w}$ is in the convex hull of the realizations of H^{i+1} that it leaves open, completing the proof. □

Corollary 5.1.7. *For all w, i, H^i_w values H^i.*

Proof. For $i = 0$, this is trivial since H^0_w is higher-order certain. For $i > 0$, by construction, $H^i_w(x) > 0$ only if $\overline{H^{i-1}_w}(x) > 0$, and by Lemma 5.1.6, this implies that H^i_x is modestly informed. Since $H^i_w(H^i = H^i_w) > 0$, trivially H^i_w is in the convex hull of the realizations of H^i it leaves open. Thus by Theorem A.2, H^i_w values H^i. □

Since the consolidation updates do not shift credences in Q, the Q-Value step is quick:

Lemma 5.1.8. *For all x, i, H^i_x Q-values $\overline{H^i}$.*

Proof. By Lemma 5.1.4, for any $q \in Q$, $H^i_x(\overline{H^i}(q) = H^i(q)) = 1$. It follows that for any decision problem $\mathcal{O}_Q$ based on Q, H^i recommends strategy S for $\mathcal{O}_Q$ if and only if $\overline{H^i}$ recommends S for $\mathcal{O}_Q$. Since, by Corollary 5.1.7, H^i_x values H^i, it follows that H^i_x Q-values $\overline{H^i}$. □

Lemmas 5.1.6 and 5.1.8 establish the first two points of Theorem 5.1; we now focus on establishing the third.

Recall that h = *more than half the coins land heads* is a proposition about Q and that for each $Heads_i = F_i \cup C_i$, $H^0(Heads_i) = \frac{1}{2}$, mutually independently. Thus, letting $\#h$ be a random variable for the number of coins that land heads, $H^0(\#h = k)$ is a binomial distribution with parameters $\frac{1}{2}$ and n. The first part of the third point follows immediately: as $n \to \infty$ $H^0(h) \to \frac{1}{2}$.

To establish the second part of the third point, that $H^0(\overline{H^n(h)} \approx 1) \approx 1$, we establish the long-run behavior of H^n (which, by Lemma 5.1.4, establishes it for $\overline{H^n}$).

Lemma 5.1.9. *With $Heads_i = F_i \cup C_i$, we have that, for all w, i, H^0_w assigns probability 1 to:*

- $F_i \to [H^n(Heads_i) = 1]$*;*
- $C_i \to [H^n(Heads_i) = \frac{2}{3}]$*; and*
- $N_i \to [H^n(Heads_i) = \frac{1}{3}]$.

Proof. First focus on $H^i(Heads_i)$, returning to H^n in a moment. Combining Lemma 5.1.5 with the definition of the update, we know immediately that H^i_w's distribution over the partition $\langle N_i, C_i, F_i\rangle$ satisfies the following:

- if $f_i \in w$, then H^i_w's distribution over $\langle N_i, C_i, F_i\rangle$ is $(0, 0, 1)$;
- if $c_i \in w$ or $n'_i \in w$, then H^i_w's distribution over $\langle N_i, C_i, F_i\rangle$ is $(\frac{1}{3}, \frac{2}{3}, 0)$;
- if $n_i \in w$ or $c'_i \in w$, then H^i_w's distribution over $\langle N_i, C_i, F_i\rangle$ is $(\frac{2}{3}, \frac{1}{3}, 0)$.

Since $H^0(U) = 0$, H^0_w assigns 0 to any world with n'_i or c'_i in it; it suffices to show that $\overline{H^n}$ follow the same pattern as H^i. By Lemma 5.1.5, each $\overline{H^j}$ treats the Q_k as mutually independent, so by definition none of the later

Jeffrey shifts—for $j \geq i$, the update from $\overline{H^j}$ to H^{j+1}—change the probabilities in Q_j. By Lemma 5.1.4, none of the consolidations (from H^j to $\overline{H^j}$) do so either. Thus $\overline{H^n}$ follows the above pattern as well, establishing the result. □

From here, the law of large numbers quickly takes us to the desired conclusion:

Lemma 5.1.10. *For any* $\epsilon > 0$, *as* $n \to \infty$, $H^0(\overline{H^n}(h) \geq 1 - \epsilon) \to 1$.

Proof. By Lemma 5.1.4, it suffices to show the result for H^n.

Choosing an arbitrary $\epsilon > 0$, let $x \approx y$ mean that x is within ϵ of y. Sort the time indices into (random) groups by their outcomes, so $I_F := \{i : Q_i = F_i\}$, $I_C := \{i : Q_i = C_i\}$, and $I_N := \{i : Q_i = N_i\}$. Since H^0 treats the Q_i as independent and identically distributed (i.i.d.) with $H^0(F_i) = H^0(C_i) = \frac{1}{4}$, by the law of large numbers, as $n \to \infty$, $H^0(|I_F| \approx \frac{n}{4}\ \&\ |I_C| \approx \frac{n}{4}\ \&\ |I_N| \approx \frac{n}{2}) \to 1$. We want to show what follows if this obtains, so suppose it does: $|I_F| \approx \frac{n}{4}$ and $|I_C| \approx \frac{n}{4}$ and $|I_N| \approx \frac{n}{2}$. What is true of H^n? We have from Lemma 5.1.9 that H^n treats all the $Heads_i$ as mutually independent, is certain of $Heads_i$ if $i \in I_F$, is $\frac{2}{3}$ in it if $i \in I_C$, and is $\frac{1}{3}$ in it if $i \in I_N$:

for all $i \in I_F$, $H^n(Heads_i) = 1$;
for all $i \in I_C$, H^n treats $Heads_i$ as i.i.d. with $H^n(Heads_i) = \frac{2}{3}$; and
for all $i \in I_N$, H^n treats $Heads_i$ as i.i.d. with $H^n(Heads_i) = \frac{1}{3}$.

Thus by the weak law of large numbers, as $n \to \infty$, H^n becomes arbitrarily confident that the proportion of $Heads_i$ within each I_F, I_C, and I_N is close to 1, $\frac{2}{3}$, and $\frac{1}{3}$, respectively,

$$H^n\Big(\sum_{i\in I_F} \frac{\mathbb{1}_{Heads_i}}{|I_F|} = 1\Big) = 1, \tag{α}$$

$$H^n\Big(\sum_{i\in I_C} \frac{\mathbb{1}_{Heads_i}}{|I_C|} \approx \frac{2}{3}\Big) \to 1, \tag{β}$$

$$H^n\Big(\sum_{i\in I_N} \frac{\mathbb{1}_{Heads_i}}{|I_N|} \approx \frac{1}{3}\Big) \to 1. \tag{γ}$$

Note that that $\frac{|I_F|}{n}\sum_{i\in I_F} \frac{\mathbb{1}_{Heads_i}}{|I_F|} + \frac{|I_C|}{n}\sum_{i\in I_C} \frac{\mathbb{1}_{Heads_i}}{|I_C|} + \frac{|I_N|}{n}\sum_{i\in I_N} \frac{\mathbb{1}_{Heads_i}}{|I_N|} = \sum_{i=1}^{n} \frac{\mathbb{1}_{Heads_i}}{n}$ is the proportion of all flips that land heads. Combining the fact that $|I_F| \approx \frac{n}{4}$ and $|I_C| \approx \frac{n}{4}$ and $|I_N| \approx \frac{n}{2}$, with ($\alpha$), ($\beta$), and ($\gamma$), we have, as $n \to \infty$,

$$H^n\Big(\sum_{i=1}^{n} \frac{\mathbb{1}_{Heads_i}}{n} \approx \frac{1}{4}(1) + \frac{1}{4}\Big(\frac{2}{3}\Big) + \frac{1}{2}\Big(\frac{1}{3}\Big) = \frac{7}{12}\Big) \to 1.$$

Therefore, recalling that $h =$ *more than half the tosses land heads*:

$$H^n\Big(\sum_{i=1}^{n} \frac{\mathbb{1}_{Heads_i}}{n} > \frac{1}{2}\Big) = H^n(h) \approx 1.$$

Since this follows from $|I_F| \approx \frac{n}{4}$ and $|I_C| \approx \frac{n}{4}$ and $|I_N| \approx \frac{n}{2}$ and H^0 is arbitrarily confident of that conjunction, it follows that as $n \to \infty$, $H^0(H^n(h) \approx 1) \to 1$, completing the proof. □

This completes the proof of Theorem 5.1. Lemma 5.1.6 establishes the first point, Lemma 5.1.8 establishes the second, and the reasoning on page 423 combined with Lemma 5.1.10 establishes the third.

Finally, we can add Tailsers to this model to establish that such predictable, profound polarization is also *persistent*.

Corollary 5.3. *There are two sequences of probability functions H^0, $\overline{H^0}$,..., $\overline{H^n}$ and T^0, $\overline{T^0}$,...,$\overline{T^n}$, a partition Q, and a proposition $h = \bigcup_i q_i$ (for some $q_i \in Q$) such that, as $n \to \infty$:*

- *Both H^0 and T^0 are (correctly) certain that, for all i,*
    - *$\overline{H^i}$ values H^{i+1} and $\overline{T^i}$ values T^{i+1};*
    - *H^i Q-values $\overline{H^i}$, and T^i Q-values $\overline{T^i}$; and*
    - *$H^0 = T^0$, and in particular $H^0(h) = T^0(h) \approx \frac{1}{2}$.*
- *H^0 and T^0 are arbitrarily confident of $\overline{H^n}(h) \approx 1$ and $\overline{T^n}(h) \approx 0$ (predictability);*
- *H^0 and T^0 are arbitrarily confident of $\overline{H^n}(h|\overline{T^n}(h) \approx 0) \approx 1$ and $\overline{T^n}(h|\overline{H^n}(h) \approx 1) \approx 0$ (persistence).*

Proof. All but the final bullet point are straightforward generalizations of the proofs of Theorem 5.1, gotten by dividing possibilities further to track which updates T^i goes through, consolidating throughout the process in a way that maintains opinions about Q, and adding the partitions $Q_i^t = \{F_i^i, C_i^t, N_i^t\}$, where $F_i^t \cup C_i^t = N_i$ and $N_i^t = F_i \cup C_i$. By doing so, we create a model in which both H^0 and T^0 are (correctly) certain that:

- $F_i \& N_i^t \rightarrow (\overline{H^n}(Heads_i) = 1 \ \& \ \overline{T^n}(Heads_i) = \frac{2}{3})$,
- $C_i \& N_i^t \rightarrow (\overline{H^n}(Heads_i) = \frac{2}{3} \ \& \ \overline{T^n}(Heads_i) = \frac{2}{3})$,
- $N_i \& C_i^t \rightarrow (\overline{H^n}(Heads_i) = \frac{1}{3} \ \& \ \overline{T^n}(Heads_i) = \frac{1}{3})$, and
- $N_i \& F_i^t \rightarrow (\overline{H^n}(Heads_i) = \frac{1}{3} \ \& \ \overline{T^n}(Heads_i) = 0)$,

with $\overline{H^n}$ and $\overline{T^n}$ treating the $Heads_i$ as mutually independent. Moreover, $H^0 = T^0$, and both treat the Q_i as mutually independent, as well as the Q_i^t, assigning for example,

- $H^0(F_i) = H^0(C_i) = \frac{1}{4}$, while $H^0(N_i) = \frac{1}{2}$; and
- $H^0(F_i^t) = H^0(C_i^t) = \frac{1}{4}$, while $H^0(N_i^t) = \frac{1}{2}$.

By reasoning parallel to that in Lemma 5.1.10, as $n \rightarrow \infty$ both H^0 and T^0 become arbitrarily confident that

$$H^n\Big(\sum_{i=1}^{n} \frac{\mathbb{1}_{Heads_i}}{n} \approx \frac{7}{12}\Big) \approx 1, \quad \text{and so } H^n(h) \approx 1,$$

and that

$$T^n\Big(\sum_{i=1}^{n} \frac{\mathbb{1}_{Heads_i}}{n} \approx \frac{5}{12}\Big) \approx 1, \quad \text{and so } T^n(h) \approx 0.$$

To establish the final bullet point of persistent polarization, notice that by the weak law of large numbers, both H^0 and T^0 are arbitrarily confident that (where $I_{F^t} = \{i : Q_i^t = F_i^t\}$, etc.) $|I_F| \approx \frac{n}{4} \& |I_C| \approx \frac{n}{4} \& |I_{F^t}| \approx \frac{n}{4} \ \& \ |I_{C^t}| \approx \frac{n}{4}$. Supposing this conjunction obtains, we show that the resulting polarization is persistent for H^n and hence $\overline{H^n}$ (parallel reasoning works for $\overline{T^n}$)—which suffices to show that it is predictable and persistent.[63]

Note that, since H^n remains certain of the above four conditionals, we have:

(i) For all $i \in I_F$, since $H^n(F_i) = 1$, we have $H^n(T^n(Heads_i) = \frac{2}{3}) = 1$.
Therefore, $H^n(\sum_{i \in I_F} \frac{T^n(Heads_i)}{|I_F|} = \frac{2}{3}) = 1$.

(ii) For all $i \in I_C$, since $H^n(C_i) = \frac{2}{3}$ and $H^n(N_i) = \frac{1}{3}$, so $H^n(N_i \& F_t) = H^n(N_i \& C_t) = \frac{1}{6}$, we have $H^n(T^n(Heads_i) =$

[63] Strictly, we should use different bounds for the ≈ at different levels of nesting, but since all can be made arbitrarily small by making n large enough, I ignore this complication.

$\frac{2}{3}) = \frac{2}{3}$, $H^n(T^n(\mathit{Heads}_i) = 0) = \frac{1}{6}$, and $H^n(T^n(\mathit{Heads}_i) = \frac{1}{3}) = \frac{1}{6}$.

Therefore, if $\pi = H^n$, for all $i \in I_C$, $\mathbb{E}_\pi(T^n(\mathit{Heads}_i)) = \frac{2}{3}(\frac{2}{3}) + \frac{1}{6}(0) + \frac{1}{6}(\frac{1}{3}) = \frac{1}{2}$. Since H^n treats the $T^n(\mathit{Heads}_i)$ as independent, by the weak law of large numbers, as $n \to \infty$, $H^n(\sum_{i \in I_C} \frac{T^n(\mathit{Heads}_i)}{|I_C|} \approx \frac{1}{2}) \to 1$.

(iii) For all $i \in I_N$, since $H^n(C_i) = \frac{1}{3}$ and $H^n(N_i) = \frac{2}{3}$, so $H^n(N_i \& F_t) = H^n(N_i \& C_t) = \frac{1}{3}$, we have $H^n(T^n(\mathit{Heads}_i) = \frac{2}{3}) = \frac{1}{3}$, $H^n(T^n(\mathit{Heads}_i) = 0) = \frac{1}{3}$, and $H^n(T^n(\mathit{Heads}_i) = \frac{1}{3}) = \frac{1}{3}$.

Therefore, if $\pi = H^n$, for all $i \in I_N$, $\mathbb{E}_\pi(T^n(\mathit{Heads}_i)) = \frac{1}{3}(\frac{2}{3}) + \frac{1}{3}(\frac{1}{3}) = \frac{1}{3}$. Since H^n treats the $T^n(\mathit{Heads}_i)$ as independent, by the weak law of large numbers, as $n \to \infty$, $H^n(\sum_{i \in I_N} \frac{T^n(\mathit{Heads}_i)}{|I_N|} \approx \frac{1}{3}) \to 1$.

Since by hypothesis $|I_F| \approx \frac{n}{4} \approx |I_C|$ and $|I_N| \approx \frac{n}{2}$ and

$$\frac{|I_F|}{n} \sum_{i \in I_F} \frac{T^n(\mathit{Heads}_i)}{|I_F|} + \frac{|I_C|}{n} \sum_{i \in I_C} \frac{T^n(\mathit{Heads}_i)}{|I_C|} + \frac{|I_N|}{n} \sum_{i \in I_N} \frac{T^n(\mathit{Heads}_i)}{|I_N|}$$
$$= \sum_{i=1}^{n} \frac{T^n(\mathit{Heads}_i)}{n},$$

combining (i)–(iii) we have, as $n \to \infty$,

$$H^n\Big(\sum_{i=1}^{n} \frac{T^n(\mathit{Heads}_i)}{n} \approx \frac{1}{4}\Big(\frac{2}{3}\Big) + \frac{1}{4}\Big(\frac{1}{2}\Big) + \frac{1}{2}\Big(\frac{1}{3}\Big) = \frac{11}{24} \approx 0.458\Big) \to 1.$$

Therefore, H^n gets arbitrarily confident that T^n's average confidence in Heads_i is less than $\frac{1}{2}$: $H^n(\sum_{i=1}^{n} \frac{T^n(\mathit{Heads}_i)}{n} < \frac{1}{2}) \to 1$. And since H^n is certain that T^n treats the Heads_i independently, it follows that $H^n(T^n(\sum_{i=1}^{n} \frac{\mathbb{1}_{\mathit{Heads}_i}}{n} > \frac{1}{2}) \approx 0) \to 1$, that is, that $H^n(T^n(h) \approx 0) \to 1$. Thus it follows that as $n \to \infty$, $H^n(h | T^n(h) \approx 0) \to H^n(h) \to 1$. Since $\overline{H^n}(h) = H^n(h)$ and $\overline{T^n}(h) = T^n(h)$ and since H^0 is arbitrarily confident of this outcome, this establishes the desired result.

By parallel reasoning, it is likewise true that as $n \to \infty$, T^0 becomes arbitrarily confident that $\overline{T^n}(h | \overline{H^n}(h) \approx 1) \to \overline{T^n}(h) \to 0$, completing the proof. □

Appendix B. Experimental Details

Appendix B discusses the experiment from section 4.2.

Two hundred and fifty English speakers were recruited through Prolific (107 female, 139 male, 4 other; mean age = 27.06).[64] The hypothesis was that subjects would polarize more when given (potentially ambiguous) word searches than when given (unambiguous) draws from an urn. Subjects were randomly assigned to conditions in a 2 × 2 design that independently manipulated valence (Headsers vs. Tailsers) and ambiguity (ambiguous vs. unambiguous). I abbreviate the groups 'A-Hsers', 'A-Tsers', 'U-Hsers', and 'U-Tsers'. Each was told they would be given evidence about a series of four independent, fair coin tosses (in fact, the tosses were pseudorandomized to simulate two heads and two tails, in random orders). They were given standard instructions about how to use a 0–100% scale to rate their confidence in the answer to a yes/no question.

The A-group was told how word-search tasks work (section 4), and given three examples ('P_A_ET' [planet], 'CO_R_D', [uncompletable] and '_E_RT' [heart]). The A-Hsers were told they would see a completable string if the coin landed heads and an uncompletable one if it landed tails. (For A-Tsers 'heads' and 'tails' were reversed.) The U-group were told how the urn task worked (section 4.2). For U-Hsers, if the coin landed heads, then the urn contained one black marble and one non-black marble; if it landed tails, it contained two non-black marbles. (For U-Tsers, 'heads' and 'tails' were reversed.) The colors of the non-black marbles changed across trials to emphasize that they were different urns.

Both groups saw four tasks, each corresponding to a new coin flip, and were asked before and afterward how confident they were in that new flip's outcome.[65] The pretask question was an attention check, wherein they were instructed to move the slider to 50% since it was a new

[64] Preregistration: https://aspredicted.org/8jg3e.pdf. I made two mistakes at the preregistration phase: (1) failing to realize I had collected time-series data for individual participant's average confidence (which allowed me to increase statistical power over merely pooling all judgments) and (2) failing to plan both the ANOVA and difference-of-difference confidence intervals. The main text reported the results after correcting these mistakes; here I report the preregistered tests. The conclusions are the same.

[65] To minimize confusion in a somewhat complicated setup, for each task the A-group was asked how confident they were that "this string is completable"—this is equivalent to "this toss landed heads" for A-Hsers and "this toss landed tails" for A-Tsers. Since they know of these equivalences, I treated their answer for task i as (for Headsers) their credence in $Heads_i$ or (for Tailsers) their credence in $Tails_i$. Meanwhile, the U-Hsers

coin toss; as preregistered, I excluded (25 of 250) participants who failed two or more of these attention checks.

The order of the tasks was randomized. Each subject in the A-group saw two completable and two uncompletable strings. (The completable strings were randomly drawn from the list FO_E_T, ST_ _N, FR_ _L [forest/foment; stain/stern; frail/frill]; the uncompletable strings were drawn from the list TR_P_R, ST_ _RE, P_G_ER.) Each subject in the U-group saw three tasks in which a non-black marble was drawn, and one in which a black marble was, simulating the expected rate of drawing black marbles from a fair coin and urn.

From the responses of each individual to each question, I calculated their prior and posterior confidence that the coin landed heads in each toss (for Hsers, this was the number they reported as their confidence; for Tsers, it was obtained by subtracting this number from 100). I pooled such responses across participants and items to calculate the following statistics. (As discussed below, we obtain more statistical power if we group *by participant* and calculate their mean confidence as they view more tasks; those stronger statistics were reported in the main text in section 4.2, p. 377.)

I predicted (predictions 1–3) that the ambiguous evidence would lead to polarization and (predictions 4–6) that it would lead to *more* polarization than the unambiguous evidence:

1. The mean A-Hser posterior in heads would be higher than the prior (of 50%).
2. The mean A-Tser posterior in heads would be lower than the prior (of 50%).
3. The mean A-Hser posterior would be higher than the mean A-Tser posterior in heads.
4. The mean A-Hser posterior would be higher than the mean U-Hser posterior.
5. The mean A-Tser posterior would be lower than the mean U-Tser posterior.
6. The mean difference between A-Hser posteriors and A-Tser posteriors would be larger than that between the U-Hser posteriors and U-Tser posteriors.

were asked how confident they were that the toss landed heads, while the U-Tsers were asked how confident they were that the toss landed tails.

Table 1. Means and standard deviations for priors and posteriors in heads.

Group	Prior Mean (SD)	Posterior Mean (SD)
A-Hsers	50.35 (3.26)	57.71 (30.33)
A-Tsers	49.60 (2.90)	36.29 (31.04)
U-Hsers	50.31 (2.68)	54.56 (26.93)
U-Tsers	50.12 (2.33)	48.10 (28.47)

Table 1 gives the means and standard deviations of credences in heads for each group.

Predictions 1, 2, 3, 5, and 6 were confirmed with significant results; Prediction 4 had the divergence in the correct direction but it was not statistically significant. Precisely: one-sided paired t-test for Prediction 1 indicated that A-Hser priors were lower than A-Hser posteriors, with $t(219) = 3.58$, $p < 0.001$, and $d = 0.341$. One-sided paired t-test for Prediction 2 indicated that A-Tser posteriors were lower than A-Tser priors, with $t(191) = 5.90$, $p < 0.001$, and $d = 0.604$. One-sided independent samples t-test for Prediction 3 indicated that A-Hser posteriors were higher than A-Tser posteriors, with $t(410) = 7.07$, $p < 0.001$, and $d = 0.699$. One-sided independent samples t-test for Prediction 4 failed to indicate that A-Hser posteriors were higher than U-Hser posteriors, with $t(441) = 1.15$, $p = 0.125$, and $d = 0.107$. One-sided independent samples t-test for Prediction 5 indicated that A-Tser posteriors were below U-Tser posteriors, with $t(393) = 4.07$, $p < 0.001$, and $d = 0.398$.

Prediction 6 was (due to my oversight) handled poorly at preregistration—I only planned to calculate 95% confidence intervals for the differences between A-Hser and A-Tser posteriors as well as U-Hser and U-Tser posteriors, and compare them. This comparison went as predicted: the 95% confidence interval for the difference between A-Hsers and A-Tsers was [15.2, 27.2], while that for the difference between U-Hsers and U-Tsers was [1.8, 11.8]. Since the former dominates the latter, it indicates a larger difference.

What *should have* been planned was (a) a 2 × 2 ANOVA, and (b) a bootstrapped 95% confidence interval for the *difference* between the differences between A-Hsers/A-Tsers and U-Hsers/U-Tsers. (a) Analyzing the results using a 2 (valence: Hser vs. Tser) × 2 (ambiguity: A vs. U) ANOVA indicated that there was a main effect of valence ($F(1, 899) = 46.47$, $p < 0.001$, $\eta^2 = 0.048$), a marginally significant main effect of ambiguity ($F(1, 899) = 4.31$, $p = 0.038$, $\eta^2 = 0.005$), and an interaction effect between valence and ambiguity ($F(1, 899) = 14.57$,

$p < 0.001$, $\eta^2 = 0.015$), indicating that the divergence between Headsers and Tailsers was exacerbated by having ambiguous evidence. (b) Meanwhile, the empirically bootstrapped 95% confidence interval for the difference in differences between A-Hsers/A-Tsers and U-Hsers/U-Tsers was [7.2, 22.6], indicating that the Hsers and Tsers in the ambiguous condition diverged in opinion more than in the unambiguous condition. And while there *was* a significant difference between U-Hser posteriors (M = 54.64, SD = 26.93) and U-Tser posteriors (M = 48.10, SD = 28.47), with $t(486) = 2.61$ and (two-sided) $p = 0.009$, the effect size was smaller ($d = 0.236$) than for the difference between A-Hser and A-Tser posteriors (as mentioned, $d = 0.699$).

Another oversight at the preregistration was failing to use the time series data generated. Using the priors and posteriors for each participant, we can calculate their average confidence in heads after seeing n bits of evidence, for n ranging from 0 to 4.[66] (For Bayesians, this average confidence equals their estimate for the proportion of times the coin landed heads.) In other words, we can rerun the above statistics by pooling responses within subjects at each stage in their progression through the experiment. All the predicted results above hold true, with universally lower p-values and higher effect sizes, since the variance of the data has dropped. These are the statistics I reported in the main text (section 4.2, p. 377).

A supplemental prediction probed the hypothesis that (something like) the model in figure 2 is driving the effect. Within the ambiguous condition, I predicted that among those who *did not* find a completion, the average confidence that their string was completable would be higher if it *was* completable (bottom right possibility of figure 2) than if it was not (bottom left). This would indicate sensitivity to whether or not there was a word, over and above whether or not they found one. To test this, in addition to recording their confidence, the experiment explicitly asked subjects in the ambiguous condition whether they found a completion. We can then focus on those who said they did not, and compare the average confidence of those who were versus were not looking at a completable string. A one-sided independent samples t-test *failed* to indicate that the confidence of those who were not ($M = 39.00$, $SD = 19.90$) was lower than that of those who were ($M = 42.03$, $SD = 21.37$), with $t(243) = 1.11$ and $p = 0.13$ (one-sided). However,

[66] At stage 0, we average their priors for all tosses; at stage 1, we average their posterior for the first toss with their priors from the 3 remaining, and so on.

a substantial proportion of people who *claimed* to have found a word did not have the extreme confidence that they should have if so (39% of them were less than 95% confident there was a completion; 25% of them were less than 80%), suggesting that self-reports of 'finding' were unreliable. If we instead operationalize 'finding' as 'reporting 100% confidence there is a completion'—though, to be clear, this change was *not* preregistered—the prediction is confirmed: among those who were less than 100% confident there was a completion, a one-sided t-test indicated that the average confidence for those looking at *un*completable strings ($M = 44.60$, $SD = 25.15$) was below the average confidence for those looking at completable strings ($M = 52.26$, $SD = 22.98$), with $t(309) = 2.77$, $p = 0.003$, and $d = 0.32$.

Finally, two further (not preregistered—so take them with a grain of salt!) trends support the role of ambiguity. First, since ambiguity—uncertainty about how to react to evidence—should cause *variance* in people's opinions, we should expect the word-search condition to have more variance than the urn condition. It does. Restricting attention to those with weak (so potentially ambiguous) evidence—those who did not find a completion (A-group) or who did not see a black marble (U-group)—the variance in opinions was higher in the ambiguous condition than in the unambiguous one. This can be seen in the plots in figure 13 and is confirmed by tests for equality of variance.[67] (Notice that the there remains a nontrivial amount of variance even in the unambiguous condition; it may be that low levels of ambiguity—people being unsure how confident to be in response to a non-black marble—could be driving the slight polarization found in the unambiguous condition.)

Second, recall that the theory predicts that polarization will result from *asymmetric increases in accuracy*: Headsers will be better at recognizing heads cases, and Tailsers will be better at recognizing tails cases. As can be seen in table 2, this is what we find. When presented with uncompletable strings (*tails* cases for Headsers and *heads* cases for Tailsers), neither group's average posterior moved significantly from their priors of 50%. However, when they saw a completable string, it moved significantly in the direction of the truth. Hence asymmetric accuracy increases can drive polarization: the mean squared errors of Headsers average pri-

[67] A-Hsers' variance was 563.33, while U-Hsers' was 285.28 (Conover = 5.40, $p < 0.001$). A-Tsers' variance was 606.78, while U-Tsers' was 321.88 (Conover = 5.44, $p < 0.001$).

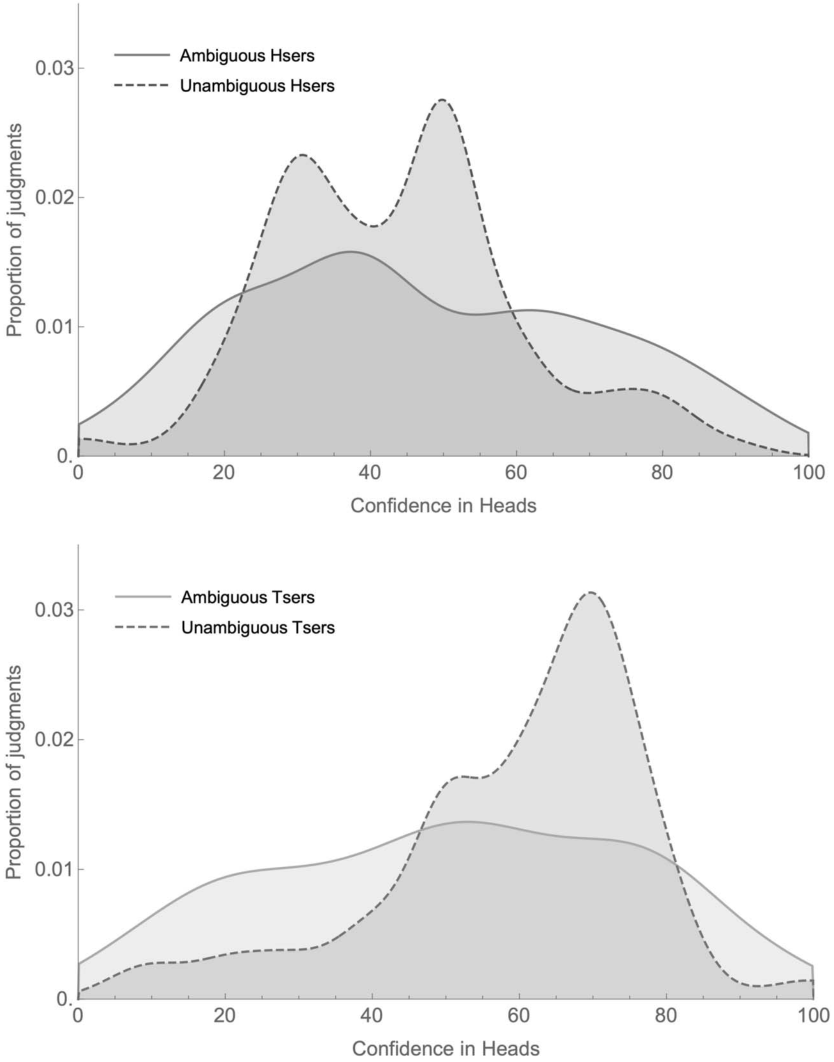


Figure 13. Density plots of confidence in *Heads*$_i$ when presented with weak evidence (uncompletable string or non-black marble).

ors versus posteriors is $0.5(1-0.5037)^2 + 0.5(0-0.5034)^2 = 0.250$ versus $0.5(1-0.6742)^2 + 0.5(0-0.4773)^2 = 0.167$. For Tailsers: 0.253 vs. 0.166.

Table 2. Ambiguous condition, mean prior and posterior confidence in *Heads*, by cases. * = *not* significantly different from 50%

	Headser Prior	Headser Posterior	Tailser Prior	Tailser Posterior
Heads cases:	50.37*	67.42	49.34*	48.00*
Tails cases:	50.34*	47.73*	49.86*	24.84
Overall:	50.35*	57.7	49.60*	36.29

Appendix C. Computational Details

Appendix C contains the details of the simulations used in sections 6 and 7. It can be read in tandem with the *Mathematica* notebook (https://github.com/kevindorst/RP_notebook), which contains a working version of all code.

Appendix C.1. Cognitive Search Models (Section 6)

This subsection explains the generalization of the word-search models that I call *cognitive search models.* Imagine an agent searching for flaws in a piece of evidence that bears on a proposition q. The general form of such a model starts with a known prior P and divides the worlds into three classes, depending on whether the agent finds a flaw (F), there is a flaw that they do not find (C; the search is 'Completable'), or there is no flaw (N). Within each class are (at least) two worlds that have the same posteriors but that differ on whether the target proposition q is true. Letting P_w be the known prior and $\widetilde{P}$ the posterior, a cognitive search model is any in which:

- $P_w(q|F) = P_w(q|C)$. (The existence of a flaw is what affects the probability of q, not whether you find it.)
- For any $n \in N$: $\widetilde{P}_n = P_w(\cdot|\neg F)$. (If there is no flaw, all you learn is that you did not find one.)
- For any $f \in F$: $\widetilde{P}_f = P_w(\cdot|F)$ (If you find a flaw, you learn exactly that.)
- For any $x, y \in C$: $\widetilde{P}_x = \widetilde{P}_y$; $\widetilde{P}_x(\neg F) = 1$; and $\widetilde{P}_x(C) \geq P_w(C|\neg F)$. (If there is a flaw that you do not find, that determines the rational credence; you learn that you did not find one, and you assign at least as much credence to there being a flaw you didn't find as you would if all you learned was that you didn't find one.)

Such models generalize the model of the word-search task from figure 2. For $x \in C$ and $y \in N$, we must have $\widetilde{P}_x(C) \geq \widetilde{P}_y(C)$ to satisfy the Value of Evidence. When $\widetilde{P}_x = \widetilde{P}_y$, the model is unambiguous and just consists in conditioning on whether or not you found a completion; but when $\widetilde{P}_x(C) > \widetilde{P}_y(C)$, the evidence is ambiguous (since $\widetilde{P}_y(x) > 0$ and $\widetilde{P}_x \neq \widetilde{P}_y$); this leads to expectable polarization.

The simplest cognitive search models consist of six worlds (two in each of F, C, and N) plus a prior over them. (In *Mathematica*, we represent this with a *seven*-world frame in which the first world encodes the prior and is assigned probability 0 by all worlds, including itself.) Such models can be parameterized in a variety of ways; the funtion `csModel` takes one such set of parameters and generates the resulting cognitive search model. The function `getCondCSModel` takes a prior in q, the degree to which finding a flaw would move it, and a probability of finding a flaw and outputs a cognitive search model by generating a random probability of there being a flaw (uniform from [0,1]) and then using that and the above to fix all the other parameters in a cognitive search model.

Given a cognitive search model and some posterior probability function $\widetilde{P}_w$, we can get the (Brier) *inaccuracy* of that function at w by taking the mean squared distance between its probability of each world x in the model and the truth value of $\{x\}$ at w. (We use this form of the Brier score—summing across worlds rather than across *propositions*—for computational tractability, since the number of propositions grows exponentially with the size of the model.) Thus `getGlobPartitionInAcc` takes a probability frame (specified using a stochastic matrix, where row i column j equals $\widetilde{P}_i(j)$) and a world w, and outputs the inaccuracy of $\widetilde{P}_w$ at w. By subtracting this number from 1, we get a measure of the *accuracy* of $\widetilde{P}_w$. And by taking the *expectation* of this value, according to our prior P, we get P's expected accuracy of the posterior rational credence function after the update.

We can then test the correlation between the probability of finding a flaw if there is one (i.e., $P(\mathit{Find}|\mathit{Flaw})$) and the expected accuracy of the update. There are a variety of ways to run such simulations. One issue is that when the `gBump` is large (i.e., the searches might shift your credence quite a bit) that introduces noise in the correlation. Thus I constrained such bumps to be small (as they will be in ensuing simulations), between 0 and 0.2. To minimize noise, I also fixed the prior in q at 0.5—but similar results are obtained by setting it to any other number. This simulation led to the plot on the left of figure 5 (p. 391).

Given this correlation, we can test what proportion of the time expected accuracy favors scrutinizing incongruent studies rather than congruent ones, as a function of how much more likely you are (on average) to find extant flaws in the former than the latter. The simulations I ran fixed a given prior in *q* and then generated pairs of cognitive search models (one would raise your credence in *q* if you found a flaw, while the other would lower it) such that the probability of finding a flaw was pulled from distributions with steadily higher means for the incongruent study and steadily lower means for the congruent one. As the gap grew, the proportion of pairs where expected accuracy favors scrutinizing the incongruent study grew as well. This led to the plot on the right of figure 5 (p. 391).

Finally, we can run a simulation of two groups of agents, presented with pairs of studies, but one group (red) is better at finding flaws with studies that tell against *q*, while the other group (blue) is better at finding flaws with those that tell in favor of *q*. At each stage, each agent chooses which study to scrutinize based on which one they expect to make them most accurate and then updates their credences with probability matching the various outcomes of that update model (i.e., their credences about how likely they are to undergo the various possible updates are calibrated with the objective chances).

There are a variety of choice points here; although variations on the theme will lead to the same results, here are the ones I made. Agents always have accurate beliefs about how likely they are to find a flaw in each study; this probability varies from a minimum of 0.1 to a maximum of 0.9. When scrutinizing *q*-detracting studies, red agents are pulling (uniformly) from $[0.1 + \texttt{findGap}, 0.9]$ and blue agents are pulling (uniformly) from $[0.1, 0.9 - \texttt{findGap}]$ (when scrutinizing *q*-supporting studies, vice versa). This parameter `findGap` can range from 0 (where there is no difference between the groups) to 0.8. The simulation displayed uses 0.5; generally the rate of polarization grows as `findgap` increases. The amount agents' credences would move if they found a flaw in the study was limited to an initial upper bound (of 0.125), which was steadily lowered as agents saw more studies and the 'weight' behind their credence in *q* was correspondingly increased. `hardenSpeed` is a parameter that controls how quickly agents harden in opinions; the smaller it is, the more polarization generally results but also the more chaotic their trajectories. The results of running the simulation with these parameters are displayed in figure 6 (p. 392).

Robustness. Fixing parameters, we can check for robustness by simulating 100 red ('pro') agents and 100 blue ('con') agents to get respective estimates for their posterior average credences at 0.603 (95% confidence interval = $[0.580, 0.626]$) and 0.387 (95% confidence interval = $[0.366, 0.409]$). These exact numbers depend on the parameters, so we can check for robustness by varying them. The end of the section 1 on cognitive search in the *Mathematica* notebook runs cross-variations on `findGap` and `hardenSpeed`, finding that as `findGap` grows and `hardenSpeed` shrinks, polarization becomes more extreme.

Appendix C.2. Argument Models

This subsection explains the simple argument models used in section 7 (without scrutiny). You know that you are about to be presented with an argument in favor of a given claim q. The model divides worlds into two classes depending on whether the argument is good (G) or bad (B). If the argument is good, it is rational to increase your confidence in q; if it is bad, it is rational to decrease it. For simplicity, we assume there are only two posteriors you could end up with. We assume the argument will be more ambiguous if it is bad. Letting P_w be the known prior and $\widetilde{P}$ be the posterior, a *simple argument (for q) model* is any in which $\{G, B\}$ is a partition and in which:

- $P_w(q|G) > P_w(q) > P_w(q|B)$ (If the argument is good, q is more likely to be true; if not, it is less.)
- For any x, y, if $x, y \in G$, $\widetilde{P}_x = \widetilde{P}_y$, and if $x, y \in B$, then $\widetilde{P}_x = \widetilde{P}_y$. (Whether the argument is good or bad determines the rational posterior.)
- $\exists \epsilon, \epsilon' > 0$, $\epsilon \geq \epsilon'$: if $g \in G$ and $b \in B$, $\widetilde{P}_g(G) = P_w(G) + \epsilon$ and $\widetilde{P}_b(B) = P_w(B) + \epsilon'$, and other probabilities are obtained by Jeffrey-shifting on these changes. (Whether good or bad, your credence should shift toward the truth, but since good arguments are easier to recognize, it should shift *more* if the former.)

Since $\widetilde{P}$ moves uniformly (though asymmetrically) toward the truth of $\{G, B\}$, P_w values $\widetilde{P}$. The simplest models consist of four worlds (two in each of G and B) plus a prior over them. (In Mathematica, we represent this with a *five*-world frame in which the first world encodes the prior and is assigned probability 0 by all worlds, including itself). Such models can be parameterized in a variety of ways; the function `getArgModel` does so using $P_w(q)$ (priorQ), $P_w(q|G)$ (gInf), $\widetilde{P}_g(G)$ for $g \in G$ (gConf), $P_w(q|B)$ (bInf), and $\widetilde{P}_b(B)$ for $b \in B$ (bConf).

An argument favors q if $P_w(q|G) > P_w(q)$; an argument disfavors q if it favors $\neg q$, that is, if $P_w(q|G) < P_w(q)$. getRandFavShiftArgModel and getRandDisShiftArgModel, respectively, generate random instances of such models. Given this, we can simulate presenting a group of (red) agents with (different) random arguments that favor q and a separate group of (blue) agents with (different) random arguments that disfavor q. Again, there are a variety of choice points in how to run such simulations. I assume agents always have accurate beliefs about how likely the arguments they are presented with are to be good or bad, and that all arguments are equally likely to be good—$P_w(G)$ was drawn uniformly from $[0, 1]$. Additionally, we can modify how much arguments could initially shift opinions and how quickly agent's opinions 'harden' (become less susceptible to change with new arguments). I simulated the result of 20 agents in each group, each witnessing 100 (different) random arguments, with an initial maximum potential shift (baseShift) of 0.2; the result is figure 8.

The code also allows for simulations to vary the rate at which each group of agents is presented with good arguments, using favGBound to lower-bound the probability that a red group member's argument is good ($P_w(G)$ drawn from [favGBound, 1]) and upper-bound the probability that a blue group member's is ($P_w(G)$ drawn from [0, 1 − favGBound]). The code in the *Mathematica* notebook runs simulations with 30 agents and 50 arguments, with the above parameters for possible shifts and hardening speed, with favGBound at 0, 0.25, 0.5, 0.75, and 0.95. The effects of varying this parameter are not straightforward—at low levels it does little, at middling levels it makes the groups' shifts more asymmetric, and at high levels it reduces the degree of belief change (I conjecture because agents are already quite confident about whether the argument is good or bad before seeing it, limiting its effects).

Robustness. Fixing parameters, I simulated 100 red (favorable argument) agents and 100 blue (disfavorable argument) agents being presented with 100 arguments each to get estimates for their mean posteriors of, respectively, 0.650 (with 95% confidence interval = $[0.630, 0.670]$) and 0.332 (with 95% confidence interval = $[0.311, 0.352]$). These exact numbers depend on the parameters, so we can check for robustness by varying them. The end of the section 2 on argument models in the *Mathematica* notebook finds that as baseShift

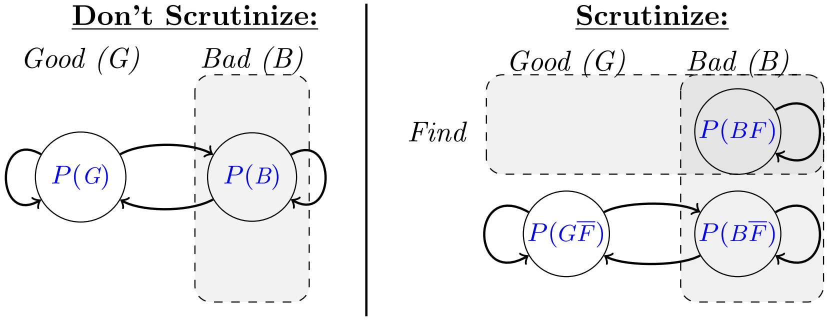


Figure 14. (Color online.) Schematic model of the choice of whether to scrutinize an argument.

grows and `hardenShift` shrinks, the amount and rate of polarization grows. All runs resulted in polarization.

Appendix C.3. Argument-Scrutiny Models

This subsection explains how to combine the simple argument models of section 7 with the cognitive search models in section 6 to yield *argument-scrutiny models*. As discussed in the main text, we begin with a simple argument model favoring some claim and then give the agent the choice to either scrutinize that argument or not. If she does not, the model remains the same and she updates as in section C.2; if she does scrutinize, the scenarios where the argument is bad (B) split into two, as in the right of figure 14. In one set of possibilities (F, top right), she finds a flaw with the argument; in another (C, bottom right), there is a flaw but she does not find it (the search is Completable). When the argument is good (G, left), there is no flaw ($N = G$).

Precisely, given an argument model as described in section C.2, with known prior P_w and posterior $\widetilde{P}$—realized as $\widetilde{P}_g$ if the argument is good and $\widetilde{P}_b$ if it is bad—scrutinizing it generates a cognitive search model with the partition $\{F, C, N\}$ fixing the posterior $\widetilde{P}$ as specified in section C.1 and the following constraints:

- $P_w(q|F) = P_w(q|C) = P_w(q|B)$. (Conditional on there being a flaw—whether or not you find it—the probability that q is true is the same as it would be if you learned the argument was bad.)
- $P_w(q|N) = P_w(q|G)$. (Conditional on there being no flaw, the probability that q is true is the same as it would be if you learned the argument was good.)

- If $x \in C$, then $\widetilde{P}_x(C) \geq \widetilde{P}_b(C|\neg F)$. (If there is a flaw that you do not find, your credence that there is should be at least as great as it should be if you did not scrutinize and updated your beliefs accordingly and then conditioned on the claim that you would not have found a flaw.)

The only subtle constraint is the third one. This ensures that, compared with the original argument model, not finding an extant flaw provides no more evidence against there being a flaw than simply conditioning on not finding one would. This is in keeping with our treatment of what happens in *N*-possibilities in cognitive search models. When $\widetilde{P}_x(C) = \widetilde{P}_b(C|\neg F)$, scrutiny adds no additional ambiguity over and above that already present in the argument model; when $\widetilde{P}_x(C) > \widetilde{P}_b(C|\neg F)$, the divergence between $\widetilde{P}_x$ (for $x \in C$) and $\widetilde{P}_y$ for $y \in N$ grows, increasing the ambiguity.

To generate such an argument-scrutiny model, we are given an argument model and must first extract its parameters—this is what `extractArgPars` does. The function `scrutArg` then uses this function to generate a cognitive search model meeting the above constraints. It takes three inputs: the original argument model (`frame`), the probability of finding a flaw in the argument if there is one (`pFind`), and the degree to which scrutiny increases ambiguity over and above the original argument, that is, the degree (if at all) to which $\widetilde{P}_x(C)$ approaches 1 over and above $\widetilde{P}_b(C|\neg F)$ (`jShift`, ranging from 0 to 1).

Given this, we can simulate what happens when both groups are presented with a series of (different) arguments favoring *q*, but one group (red) never scrutinizes them, while the other group (blue) always does. Again, there are a variety of choice points for how we model and constrain this. I used the same parameters for generating arguments that I used in section C.2 and ran four versions of the scrutiny simulation. Since scrutiny introduces more noise into the simulations, I used *50* agents and 100 arguments to see the trends.

In version (1), scrutinizing agents never find a flaw even if there is one (`pFind` = 0), and the scrutiny adds no ambiguity (`jShift` = 0). Such scrutiny does not change the original argument model, so agents who scrutinize polarize as much and in the same direction as those who do not—as seen in the top left of figure 9 (p. 397).

In version (2), scrutinizing agents *always* find a flaw if there is one (`pFind` =1), meaning that scrutiny removes all ambiguity. (The `jShift` parameter has no effect in this case.) Since scrutiny changes the model to an unambiguous one, by Theorem 3.1, scrutinizing agents do not expectedly polarize from their priors of 0.5—as seen in the top right of figure 9.

In version (3), scrutinizing agents *sometimes* find a flaw if there is one (`pFind` pulled uniformly from [0, 1]), and scrutiny introduces a small degree of ambiguity (`jShift` pulled uniformly from [0, 0.5]). The result is that scrutinizing agents polarize is the same direction as those that do not, but less so—as seen in the bottom left of figure 9.

In version (4), scrutinizing agents sometimes find a flaw if there is one (`pFind` pulled uniformly from [0, 1]), and scrutiny introduces *substantial* ambiguity (`jShift` pulled uniformly from [0, 1]). The result is that scrutinizing agents polarize in the *opposite* direction of those that do not—as seen in the bottom right of figure 9.

Robustness. Recall that pro agents in this simulation are identical to those from the main simulation of section C.2, meaning we have estimates for their mean posteriors with these parameters at 0.650 (95% confidence interval = [0.630, 0.670]). To check that the results in the above simulations (1)–(4) were robust, I ran the same parameters with 200 con agents and calculated estimates and confidence intervals for their posteriors. The results are as expected. In version (1), the mean posterior was 0.645 (95% confidence interval = [0.633, 0.658]), indicating that scrutinizing agents shift to a comparable degree to those who don't scrutinize. In version (2), the mean posterior was 0.503 (95% confidence interval = [0.474, 0.533]), indicating that agents do not predictably shift from their priors of 0.5. In version (3), the mean posterior was 0.551 (95% confidence interval = [0.530, 0.573]), confirming that such scrutiny dampens polarization. In version (4), the mean posterior was 0.463 (95% confidence interval = [0.442, 0.483]), confirming that such scrutiny reverses the direction of polarization.

References

Acemoglu, Daron, and Alexander Wolitzky. 2014. "Cycles of Conflict: An Economic Model." *American Economic Review* 104, no. 4: 1350–67.

Achen, Christopher H., and Larry M. Bartels. 2017. *Democracy for Realists: Why Elections Do Not Produce Responsive Government.* Princeton Studies in Political Behavior. Princeton: Princeton University Press.

Anderson, John R. 1990. *The Adaptive Character of Thought.* Mahwah, NJ: Erlbaum Associates.

Andreoni, James, and Tymofiy Mylovanov. 2012. "Diverging Opinions." *American Economic Journal: Microeconomics* 4, no. 1: 209–32.

Angere, Staffan, and Erik J. Olsson. 2017. "Publish Late, Publish Rarely!: Network Density and Group Performance in Scientific Communication." In *Scientific Collaboration and Collective Knowledge*, edited by Thomas Boyer-Kassem, Conor Mayo-Wilson, and Michael Weisberg, 34–62. Oxford: Oxford Unviersity Press.

Anglin, Stephanie M. 2019. "Do Beliefs Yield to Evidence? Examining Belief Perseverance Vs. Change in Response to Congruent Empirical Findings." *Journal of Experimental Social Psychology* 82, February: 176–99.

Ariely, Dan. 2008. *Predictably Irrational.* Harper Audio.

Aronowitz, Sara. 2021. "Exploring by Believing." *The Philosophical Review* 130, no. 3: 339–83.

Aumann, R. 1976. "Agreeing to Disagree." *The Annals of Statistics* 4, no. 6: 1236–39.

Austerweil, Joseph L., and Thomas L. Griffiths. 2011. "Seeking Confirmation Is Rational for Deterministic Hypotheses." *Cognitive Science* 35, no. 3: 499–526.

Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. "Exposure to Opposing Views on Social Media Can Increase Political Polarization." *Proceedings of the National Academy of Sciences* 115, no. 37: 9216–21.

Baliga, Sandeep, Eran Hanany, and Peter Klibanoff. 2013. "Polarization and Ambiguity." *American Economic Review* 103, no. 7: 3071–83.

Baron, Robert S., Sieg I. Hoppe, Chuan Feng Kao, Bethany Brunsman, Barbara Linneweh, and Diane Rogers. 1996. "Social Corroboration and Opinion Extremity." *Journal of Experimental Social Psychology* 32, no. 6: 537–60.

Baumgaertner, Bert O., Rebecca T. Tyson, and Stephen M. Krone. 2016. "Opinion Strength Influences the Spatial Dynamics of Opinion Formation." *Educational Research* 40, no. 4: 207–18.

Benoît, Jean Pierre, and Juan Dubra. 2019. "Apparent Bias: What Does Attitude Polarization Show?" *International Economic Review* 60, no. 4: 1675–703.

Bertsekas, Dmitri P., and John N. Tsitsiklis. 2008. *Introduction to Probability.* Second edition. Nashua, NH: Athena Scientific.

Blackwell, David. 1953. "Equivalent Comparisons of Experiments." *Annals of Mathematical Statistics* 24, no. 2: 265–72.

Bowen, T. Renee, Danil Dmitriev, and Simone Galperti. 2023. "Learning from Shared News: When Abundant Information Leads to Belief Polarization." *The Quarterly Journal of Economics* 138, no. 2: 955–1000.

Boxell, Levi, Matthew Gentzkow, and Jesse Shapiro. 2020. "Cross-Country Trends in Affective Polarization." NBER Working Paper No. w26669. Cambridge, MA: National Bureau of Economic Research. https://ssrn.com/abstract=3522318.

Bradley, Seamus, and Katie Steele. 2016. 'Can Free Evidence Be Bad? Value of Information for the Imprecise Probabilist'. *Philosophy of Science* 83, no. 1. https://doi.org/10.1086/684184.

Bregman, Rutger. 2017. *Utopia for Realists: And How We Can Get There*. London: Bloomsbury Publishing.

Brennan, Jason, 2016. *Against Democracy*. Princeton: Princeton University Press.

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78, no. 1: 1–3.

Briggs, Ray. 2009. "Distorted Reflection." *The Philosophical Review* 118, no. 1: 59–85.

Brown, Jacob R., and Ryan D. Enos. 2021. "The Measurement of Partisan Sorting for 180 Million Voters." *Nature Human Behaviour* 5: 998–1008.

Brownstein, Ronald, 2016. "How the Election Revealed the Divide Between City and Country." *The Atlantic*, November 17, 2016. https://www.theatlantic.com/politics/archive/2016/11/clinton-trump-city-country-divide/507902/.

Burnstein, Eugene, and Amiram Vinokur. 1977. "Persuasive Argumentation and Social Comparison as Determinants of Attitude Polarization." *Journal of Experimental Social Psychology* 13, no. 4: 315–32.

Callahan, Laura Frances. 2019. "Epistemic Existentialism." *Episteme* 18, no. 4: 539–54.

Camerer, Colin, and Martin Weber. 1992. "Recent Developments in Modeling Preferences: Uncertainty and Ambiguity." *Journal of Risk and Uncertainty* 5, no. 4: 325–70.

Campbell-Moore, Catrin. 2016. *Self-Referential Probability*. PhD diss, LMU Munich.

Cariani, Fabrizio, and Lance J. Rips. 2017. "Conditionals, Context, and the Suppression Effect." *Cognitive Science* 41, no. 3: 540–89.

Carmichael, Chloe. 2017. "Political Polarization Is A Psychology Problem." *HuffPost*, November 8, 2017. https://www.huffpost.com/entry/political-polarization-is-a-psychology-problem_b_5a01dd9ee4b07eb5118255e5.

Carr, Jennifer Rose. 2020. "Imprecise Evidence without Imprecise Credences." *Philosophical Studies* 177, no. 9: 2735–58.

Christensen, David. 2010. "Higher-Order Evidence." *Philosophy and Phenomenological Research* 81, no. 1: 185–215.

Cohen, G. A. 2000. *If You're an Egalitarian, How Come You're So Rich?* Cambridge, MA: Harvard University Press.

Cohen, Geoffrey L. 2003. "Party over Policy: The Dominating Impact of Group Influence on Political Beliefs." *Journal of Personality and Social Psychology* 85, no. 5: 808.

Cohen, L. Jonathan. 1981. "Can Human Irrationality Be Experimentally Demonstrated?" *Behavioral and Brain Sciences* 4, no. 3: 317–31.

Cook, J. Thomas. 1987. “Deciding to Believe without Self-Deception.” *Journal of Philosophy* 84, no. 8: 441–46.

Cook, John, and Stephan Lewandowsky. 2016. “Rational Irrationality: Modeling Climate Change Belief Polarization Using Bayesian Networks.” *Topics in Cognitive Science* 8, no. 1: 160–79.

Corner, Adam, Adam Harris, and Ulrike Hahn. 2010. “Conservatism in Belief Revision and Participant Skepticism.” In *Proceedings of the Annual Meeting of the Cognitive Science Society, 32.*

Corner, Adam, Lorraine Whitmarsh, and Dimitrios Xenias. 2012. “Uncertainty, Scepticism and Attitudes towards Climate Change: Biased Assimilation and Attitude Polarisation.” *Climatic Change* 114, no. 3: 463–78.

Crupi, Vincenzo, Katya Tentori, and Luigi Lombardi. 2009. “Pseudodiagnosticity Revisited.” *Psychological Review* 116, no. 4: 971–85.

Cushman, Fiery, 2020. “Rationalization Is Rational.” *Behavioral and Brain Sciences,* 43, e28: 1–59. doi:doi.org/10.1017/S0140525X19001730.

Dallmann, Justin. 2017. “When Obstinacy Is a Better Policy.” *Philosophers’ Imprint* 17, no. 24: 1–17.

Das, Nilanjan. 2019. “Accuracy and Ur-Prior Conditionalization.” *The Review of Symbolic Logic* 12, no. 1: 62–96.

Das, Nilanjan. 2022a. “Externalism and Exploitability.” *Philosophy and Phenomenological Research* 104, no. 1: 101–28.

Das, Nilanjan. 2022b. “Credal Imprecision and the Value of Evidence.” Preprint. *Noûs.*

Das, Nilanjan. 2023. “The Value of Biased Information.” *The British Journal for the Philosophy of Science* 74, no. 1: 25–55.

De Cruz, Helen. 2017. “Religious Disagreement: A Study among Academic Philosophers.” *Episteme* 14, no. 1: 71–87.

de Finetti, Bruno. 1977. “Probabilities of Probabilities: A Real Problem or a Misunderstanding.” In *New Developments in the Applications of Bayesian Methods,* edited by A. Aykac and C. Brumat, 1–10. Amsterdam: North-Holland.

DeMarzo, Peter M., Dimitri Vayanos, and Jeffrey Zwiebel. 2003. “Persuasion Bias, Social Influence, and Unidimensional Opinions.” *Quarterly Journal of Economics* 118, no. 3: 909–68.

Ditto, Peter H., and David F. Lopez. 1992. “Motivated Skepticism: Use of Differential Decision Criteria for Preferred and Nonpreferred Conclusions.” *Journal of Personality and Social Psychology,* 63, no. 4: 568.

Dixit, Avinash K., and Jörgen W. Weibull. 2007. “Political Polarization.” *Proceedings of the National Academy of Sciences of the United States of America,* 104, no. 2: 7351–56.

Dorst, Kevin. 2019. “Higher-Order Uncertainty.” In *Higher-Order Evidence: New Essays,* edited by Mattias Skipper Rasmussen and Asbjørn Steglich-Petersen, 35–61. Oxford: Oxford University Press.

Dorst, Kevin. 2020. “Evidence: A Guide for the Uncertain.” *Philosophy and Phenomenological Research* 100, no. 3: 586–632.

Dorst, Kevin. Forthcoming. “Higher-Order Evidence.” In *The Routledge Handbook for the Philosophy of Evidence*, edited by Maria Lasonen-Aarnio and Clayton Littlejohn. Philidelphia: Routledge.

Dorst, Kevin, Benjamin Levinstein, Bernhard Salow, Brooke E. Husic, and Branden Fitelson. 2021. “Deference Done Better.” *Philosophical Perspectives* 35, no. 1: 99–150.

Downing, James W., Charles M. Judd, and Markus Brauer. 1992. “Effects of Repeated Expressions on Attitude Extremity.” *Journal of Personality and Social Psychology* 63, no. 1: 17–29.

Easley, David, and Jon Kleinberg. 2010. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World.* Cambridge: Cambridge University Press.

Edwards, Ward. 1982. “Conservatism in Human Information Processing.” In *Judgment under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic, and Amos Tversky 359–69. Cambridge: Cambridge University Press.

Elga, Adam. 2007. “Reflection and Disagreement.” *Noûs* 41, no. 3: 478–502.

Elga, Adam. 2013. “The Puzzle of the Unmarked Clock and the New Rational Reflection Principle.” *Philosophical Studies* 164, no. 1: 127–39.

Elga, Adam, and Agustín Rayo. 2022. “Fragmentation and Logical Omniscience.” *Noûs* 56, no. 3: 716–41.

Ellsberg, Daniel. 1961. “Risk, Ambiguity, and the Savage Axioms.” *Quarterly Journal of Economics*, 75, no. 4: 643–69.

Evans, J. St B. T., Julie L. Barston, and Paul Pollard. 1983. “On the Conflict between Logic and Belief in Syllogistic Reasoning.” *Memory & Cognition* 11, no. 3: 295–306.

Feeney, Aidan, Jonathan St B. T. Evans, and John Clibbens. 2000. “Background Beliefs and Evidence Interpretation.” *Thinking & Reasoning* 6, no. 2: 97–124.

Fine, Cordelia. 2005. *A Mind of Its Own: How Your Brain Distorts and Deceives.* New York: W. W. Norton & Company.

Finkel, Eli J., Christopher A. Bail, Mina Cikara, Peter H. Ditto, Shanto Iyengar, Samara Klar, Lilliana Mason, Mary C. McGrath, Brendan Nyhan, David G. Rand, Linda J. Skitka, Joshua A. Tucker, Jay J. Van Bavel, Cynthia S. Wang, and James N. Druckman. 2020. “Political Sectarianism in America.” *Science* 370, no. 6516: 533–36.

Fischer, Peter, Eva Jonas, Dieter Frey, and Stefan Schulz-Hardt. 2005. “Selective Exposure to Information: The Impact of Information Limits.” *European Journal of Social Psychology* 35, no. 4: 469–92.

Fitzpatrick, Anne R., and Alice H. Eagly. 1981. “Anticipatory Belief Polarization as a Function of the Expertise of a Discussion Partner.” *Personality and Social Psychology Bulletin* 7, no. 4: 636–42.

Flache, Andreas, and Michael W. Macy. 2011. “Local Convergence and Global Diversity: From Interpersonal to Social Influence.” *Journal of Conflict Resolution* 55, no. 6: 970–95.

Fraser, Rachel. 2022. “Mushy Akrasia: Why Mushy Credences Are Rationally Permissible.” *Philosophy and Phenomenological Research* 105, no. 1: 79–106.

Frey, Dieter. 1986. “Recent Research on Selective Exposure to Information.” *Advances in Experimental Social Psychology* 19: 41–80.

Fryer, Roland G., Philipp Harms, and Matthew O. Jackson. 2019. “Updating Beliefs when Evidence is Open to Interpretation: Implications for Bias and Polarization.” *Journal of the European Economic Association* 17, no. 5: 1470–501.

Gaifman, Haim. 1988. “A Theory of Higher Order Probabilities.” In *Causation, Chance, and Credence, Volume 1*, edited by Brian Skyrms and William L. Harper, 191–219. Norwell, MA: Kluwer.

Gallow, J. Dmitri. 2021. “Updating for Externalists.” *Noûs* 55, no. 3: 487–516.

Geanakoplos, John. 1989. “Game Theory Without Partitions, and Applications to Speculation and Consensus.” *Cowles Foundation Discussion Papers*. 1158. https://elischolar.library.yale.edu/cowles-discussion-paper-series/1158.

Gershman, Samuel. 2021. *What Makes Us Smart: The Computational Logic of Human Cognition*. Princeton: Princeton University Press.

Gigerenzer, Gerd. 1991. “How to Make Cognitive Illusions Disappear: Beyond ‘Heuristics and Biases’.” *European Review of Social Psychology* 2, no. 1: 83–115.

Gigerenzer, Gerd. 2018. “The Bias Bias in Behavioral Economics.” *Review of Behavioral Economics* 5, no. 3-4: 303–36.

Gilovich, Thomas. 1983. “Biased Evaluation and Persistence in Gambling.” *Journal of Personality and Social Psychology* 44, no. 6: 1110–26.

Gilovich, Thomas. 1991. *How We Know What Isn’t So*. New York: Free Press.

Glaser, Markus, and Martin Weber. 2010. “Overconfidence.” In *Behavioral Finance: Investors, Corporations, and Markets*, edited by H. Kent Baker and John R. Nofsinger, 241–58. Hoboken, NJ: Wiley.

Good, I. J. 1967. “On the Principle of Total Evidence.” *The British Journal for the Philosophy of Science* 17, no. 4: 319–21.

Gopnik, Alison. 1996. “The Scientist as Child.” *Philosophy of Science* 63, December: 485–514.

Gopnik, Alison. 2012. “Scientific Thinking in Young Children: Theoretical Advances, Empirical Research, and Policy Implications.” *Science* 337, no. 6102: 1623–27.

Gopnik, Alison. 2020. “Childhood as a Solution to Explore–Exploit Tensions.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 375, no. 1803. http://dx.doi.org/10.1098/rstb.2019.0502.

Greaves, Hilary, and David Wallace. 2006. “Justifiying Conditonalization: Conditionalization Maximizes Expected Epistemic Utility.” *Mind* 115, no. 459: 607–32.

Griffiths, Thomas L., Nick Chater, Dennis Norris, and Alexandre Pouget. 2012. "How the Bayesians Got Their Beliefs (And What Those Beliefs Actually Are): Comment on Bowers and Davis (2012)." *Psychological Bulletin* 138, no. 3: 415–22.

Griffiths, Thomas L., Falk Lieder, and Noah D. Goodman. 2015. "Rational Use of Cognitive Resources: Levels of Analysis between the Computational and the Algorithmic." *Topics in Cognitive Science* 7, no. 2: 217–29.

Grönlund, Kimmo, Kaisa Herne, and Maija Setälä. 2015. "Does Enclave Deliberation Polarize Opinions?" *Political Behavior* 37, no. 4: 995–1020.

Hahn, Ulrike, and Adam J.L. Harris. 2014. "What Does It Mean to be Biased. Motivated Reasoning and Rationality." In *Psychology of Learning and Motivation*, edited by Brian Ross, Volume 61 of Advances in Research and Theory, 41–102. Amsterdam: Elsevier.

Haidt, Jonathan. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. New York: Knopf.

Halpern, Joseph Y. 2010. "I Don't Want to Think about It Now: Decision Theory with Costly Computation." In *Twelfth International Conference on the Principles of Knowledge Representation and Reasoning*, 182–90. Toronto: AAAI Press.

Hamblin, Charles L. 1976. "Questions in Montague English." In *Montague Grammar*, edited by Barbara H. Partee, 247–59. Amsterdam: Elsevier.

Hart, William, Dolores Albarracín, Alice H. Eagly, Inge Brechan, Matthew J. Lindberg, and Lisa Merrill. 2009. "Feeling Validated Versus Being Correct: A Meta-Analysis of Selective Exposure to Information." *Psychological Bulletin* 135, no. 4: 555–88.

Harvey, Nigel. 1997. "Confidence in Judgment." *Trends in Cognitive Sciences* 1, no. 2: 78–82.

Hastie, Reid, and Robyn M. Dawes. 2009. *Rational Choice in an Uncertain World: The Psychology of Judgment and Decision Making*. Thousand Oaks, CA: Sage Publications.

Hedden, Brian. 2015. "Options and Diachronic Tragedy." *Philosophy and Phenomenological Research* 90, no. 2: 423–51.

Hegselmann, Rainer, and Ulrich Krause. 2002. "Opinion Dynamics and Bounded Confidence: Models, Analysis and Simulation." *Journal of Artificial Societies and Social Simulation* 5, no. 3. https://www.jasss.org/5/3/2.html.

Henderson, Leah, and Alexander Gebharter. 2021. "The Role of Source Reliability in Belief Polarisation." *Synthese* 199, no. 3-4: 10253–76.

Hild, Matthias. 1998. "Auto-Epistemology and Updating." *Philosophical Studies* 92, no. 3: 321–61.

Hintikka, Jaako. 1962. *Knowledge and Belief*. Ithica, NY: Cornell University Press.

Horowitz, Sophie. 2014. "Epistemic Akrasia." *Noûs* 48, 4: 718–44.

Horowitz, Sophie. 2019. "The Truth Problem for Permissivism." *The Journal of Philosophy* 116, no. 5: 237–62.

Huttegger, Simon M. 2013. “In Defense of Reflection.” *Philosophy of Science* 80, no. 3: 413–33.

Huttegger, Simon M. 2014. “Learning Experiences and the Value of Knowledge.” *Philosophical Studies* 171, no. 2: 279–88.

Isaacs, Yoaav, and Jeffrey Sanford Russell. 2022. “Updating without Evidence.” Preprint. *Noûs*.

Isenberg, Daniel J. 1986. “Group Polarization: A Critical Review and Meta-Analysis.” *Journal of Personality and Social Psychology* 50, no. 6: 1141–51.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood. 2019. “The Origins and Consequences of Affective Polarization in the United States.” *Annual Review of Political Science* 22: 129–46.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. 2012. “Affect, Not Ideology: A Social Identity Perspective on Polarization.” *Public Opinion Quarterly* 76, no. 3: 405–31.

Jackson, Elizabeth. 2021. “A Defense of Intrapersonal Belief Permissivism.” *Episteme* 18, no. 2: 313–27.

Jamieson, Kathleen Hall, and Joseph N. Cappella. 2008. *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment.* Oxford: Oxford University Press.

Jeffrey, Richard C. 1990. *The Logic of Decision.* Chicago: University of Chicago press.

Jern, Alan, Kai Min K. Chang, and Charles Kemp. 2014. “Belief Polarization Is Not Always Irrational.” *Psychological Review* 121, no. 2: 206–24.

Johnson, Dominic D. P. 2009. *Overconfidence and War.* Cambridge, MA: Harvard University Press.

Jost, John T., Jack Glaser, Arie W. Kruglanski, and Frank J. Sulloway. 2003. “Political Conservatism as Motivated Social Cognition.” *Psychological Bulletin* 129, no. 3: 339–75.

Joyce, James M. 2010. “A Defense of Imprecise Credences in Inference and Decision Making.” *Philosophical Perspectives* 24, no. 1: 282–323.

Kadane, Joseph B., Mark Schervish, and Teddy Seidenfeld. 2008. “Is Ignorance Bliss?” *The Journal of Philosophy* 105, no. 1: 5–36.

Kadane, Joseph B., Mark J. Schervish, and Teddy Seidenfeld. 1996. “Reasoning to a Foregone Conclusion.” *Journal of the American Statistical Association* 91, no. 435: 1228–35.

Kahan, Dan M. 2013. “Ideology, Motivated Reasoning, and Cognitive Reflection.” *Judgment and Decision Making* 8, no. 4: 407–24.

Kahan, Dan M. 2018. “Why Smart People Are Vulnerable to Putting Tribe Before Truth.” *Scientific American*, December 3, 2018. https://blogs.scientificamerican.com/observations/why-smart-people-are-vulnerable-to-putting-tribe-before-truth/.

Kahan, Dan M., Ellen Peters, Erica Cantrell Dawson, and Paul Slovic. 2017. "Motivated Numeracy and Enlightened Self-Government." *Behavioural Public Policy* 1, no. 1: 54–86.

Kahan, Dan M., Ellen Peters, Maggie Wittlin, Paul Slovic, Lisa Larrimore Ouellette, Donald Braman, and Gregory Mandel. 2012. "The Polarizing Impact of Science Literacy and Numeracy on Perceived Climate Change Risks." *Nature Climate Change* 2, no. 10: 732–35.

Kahneman, Daniel. 2011. *Thinking Fast and Slow.* New York: Farrar, Straus, and Giroux.

Kahneman, Daniel, Paul Slovic, and Amos Tversky, eds. 1982. *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press.

Kahneman, Daniel, and Amos Tversky. 1996. "On the Reality of Cognitive Illusions." *Psychological Review* 103, no. 3: 582–91.

Kamenica, Emir, and Matthew Gentzkow. 2011. "Bayesian Persuasion." *American Economic Review* 101, no. 6: 2590–615.

Kelly, Thomas. 2008. "Disagreement, Dogmatism, and Belief Polarization." *The Journal of Philosophy* 105, no. 10: 611–33.

Kinney, David, and Liam Kofi Bright. 2021. "Risk Aversion and Elite-Group Ignorance." Preprint. *Philosophy and Phenomenological Research.*

Klaczynski, Paul A., and Gayathri Narasimham. 1998. "Development of Scientific Reasoning Biases: Cognitive Versus Ego-Protective Explanations." *Developmental Psychology* 34, no. 1: 175–87.

Klein, Ezra. 2014. "How Politics Makes Us Stupid." *Vox,* April 6, 2014. https://www.vox.com/2014/4/6/5556462/brain-dead-how-politics-makes-us-stupid.

Klein, Ezra. 2020. *Why We're Polarized.* London: Profile Books.

Klibanoff, Peter, Massimo Marinacci, and Sujoy Mukerji. 2005. "A Smooth Model of Decision Making under Ambiguity." *Econometrica* 73, no. 6: 1849–92.

Koerth, Maggie. 2019. "Why Partisans Look At The Same Evidence On Ukraine And See Wildly Different Things." *FiveThirtyEight,* October 3, 2019. https://fivethirtyeight.com/features/why-partisans-look-at-the-same-evidence-on-ukraine-and-see-wildly-different-things/.

Koriat, Asher, Sarah Lichtenstein, and Baruch Fischhoff, 1980. "Reasons for Confidence." *Journal of Experimental Psychology: Human Learning and Memory* 6, no. 2: 107–18.

Kossinets, Gueorgi, and Duncan J. Watts. 2009. "Origins of Homophily in an Evolving Social Network." *American Journal of Sociology* 115, no. 2: 405–50.

Krueger, Joachim I., and Adam L. Massey. 2009. "A Rational Reconstruction of Misbehavior." *Social Cognition* 27, no. 5: 786–812.

Kuhn, Deanna, and Joseph Lao. 1996. "Effects of Evidence on Atittudes: Is Polarization the Norm?" *Psycholohical Science* 7, no. 2: 115–20.

Kuhn, Thomas. 1962. *The Structure of Scientific Revolutions*. Chicago: The University of Chicago Press.

Kunda, Ziva. 1990. “The Case for Motivated Reasoning.” *Psychological Bulletin* 108, no. 3: 480–98.

Lakoff, George. 1997. *Moral Politics: What Conservatives Know that Liberals Don’t.* Chicago: University of Chicago Press.

Lasonen-Aarnio, Maria. 2013. “Disagreement and Evidential Attenuation.” *Noûs*, 47, no. 4: 767–94.

Lasonen-Aarnio, Maria. 2014. “Higher-Order Evidence and the Limits of Defeat.” *Philosophy and Phenomenological Research* 8, no. 2: 314–45.

Lasonen-Aarnio, Maria. 2015. “New Rational Reflection and Internalism about Rationality.” In *Oxford Studies in Epistemology, Volume 5*, edited by Tamar Szabó Gendler and John Hawthorne, 145–71. Oxford: Oxford University Press.

Lazer, David, Matthew Baum, Jochai Benkler, Adam Berinsky, Kelly Greenhill, Miriam Metzger, Brendan Nyhan, G. Pennycook, David Rothschild, Cass Sunstein, Emily Thorson, Duncan Watts, and Jonathan Zittrain. 2018. “The Science of Fake News.” *Science* 359, no. 6380: 1094–96.

Le Mens, Gaël, and Jerker Denrell. 2011. “Rational Learning and Information Sampling: On the ‘Naivety’ Assumption in Sampling Explanations of Judgment Biases.” *Psychological Review* 118, no. 2: 379–92.

Lederman, Harey. 2015. “People with Common Priors Can Agree to Disagree.” *The Review of Symbolic Logic* 8, no. 1: 11–45.

Levi, Isaac. 1974. “On Indeterminate Probabilities.” *The Journal of Philosophy* 71, no. 13: 391–418.

Levinstein, B. A. 2022. “Accuracy, Deference, and Chance.” Preprint. *The Philosophical Review.*

Levinstein, Benjamin, and Jack Spencer. 2022. “Bigger, Badder Bugs.” Working paper.

Lewis, David. 1976. “Probabilities of Conditionals and Conditional Probabilities.” *The Philosophical Review* 85, no. 3: 297–315.

Lewis, David. 1980. “A Subjectivist’s Guide to Objective Chance.” In *Studies in Inductive Logic and Probability, Volume 2*, edited by Richard C. Jeffrey, 263–93. Oakland: University of California Press.

Liberman, Akiva, and Shelly Chaiken. 1992. “Defensive Processing of Personally Relevant Health Messages.” *Personality and Social Psychology Bulletin* 18, no. 6: 669–79.

Lichtenstein, Sarah, Baruch Fischhoff, and Lawrence D. Phillips. 1982. “Calibration of Probabilities: The State of the Art to 1980.” In *Judgment under Uncertainty*, edited by Daniel Kahneman, Paul Slovic, and Amos Tversky, 306–34. Cambridge: Cambridge University Press.

Lieder, Falk, and Thomas L. Griffiths. 2019. “Resource-Rational Analysis: Understanding Human Cognition as the Optimal Use of Limited Computational Resources." *Behavioral and Brain Sciences* 43. https://doi.org/10.1017/s0140525x1900061x.

Lilienfeld, Scott O., Rachel Ammirati, and Kristin Landfield. 2009. "Giving Debiasing Away: Can Psychological Research on Correcting Cognitive Errors Promote Human Welfare?" *Perspectives on Psychological Science* 4, no. 4: 390–98.

Liu, Cheng-Hong. 2017. "Evaluating Arguments during Instigations of Defence Motivation and Accuracy Motivation." *British Journal of Psychology* 108, no. 2: 296–317.

Loh, Isaac, and Gregory Phelan. 2019. "Dimensionality and Disagreement: Asymptotic Belief Divergence in Response to Common Information." *International Economic Review* 60, no. 4: 1861–76.

Lord, Charles G., Mark R. Lepper, and Elizabeth Preston. 1984. "Considering the Opposite: A Corrective Strategy for Social Judgment." *Journal of Personality and Social Psychology* 47, no. 6: 1231–43.

Lord, Charles G., Lee Ross, and Mark R. Lepper. 1979. "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence." *Journal of Personality and Social Psychology* 37, no. 11: 2098–109.

Lord, Charles G., and Cheryl A. Taylor, 2009. "Biased Assimilation: Effects of Assumptions and Expectations on the Interpretation of New Evidence." *Social and Personality Psychology Compass* 3, no. 5: 827–41.

Lottes, Ilsa L., and Peter J. Kuriloff. 1994. "The Impact of College Experience on Political and Social Attitudes." *Sex Roles* 31, no. 1: 31–54.

Lundgren, Sharon R., and Radmila Prislin. 1998. "Motivated Cognitive Processing and Attitude Change." *Personality and Social Psychology Bulletin* 24, no. 7: 715–26.

Mandelbaum, Eric. 2018. "Troubles with Bayesianism: An Introduction to the Psychological Immune System." *Mind & Language* 34, no. 2: 141–57.

Mäs, Michael, and Andreas Flache. 2013. "Differentiation without Distancing: Explaining Bi-Polarization of Opinions without Negative Influence." *PLoS ONE*, 8, no. 11. https://doi.org/10.1371/journal.pone.0074516.

Mason, Lilliana. 2018. *Uncivil Agreement: How Politics Became Our Identity.* Chicago: The University of Chicago Press.

McHoskey, John W. 1995. "Case Closed? On the John F. Kennedy Assassination: Biased Assimilation of Evidence and Attitude Polarization." *Basic and Applied Social Psychology* 17, no. 3: 395–409.

McKenzie, Craig R. M. 2004. "Framing Effects in Inference Tasks—And Why They Are Normatively Defensible." *Memory & Cognition* 32, no. 6: 874–85.

Mcpherson, Miller, Lynn Smith-Lovin, and James M. Cook, 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27: 415–44.

McWilliams, Emily C. 2021. "Evidentialism and Belief Polarization." *Synthese* 198, no. 8: 7165–96.

Mercier, Hugo. 2017. "Confirmation Bias—Myside Bias." In *Cognitive Illusions: Intriguing Phenomena in Thinking, Judgment and Memory*, edited by R. F. Pohl, 99–114. Milton Park, Oxfordshire: Routledge.

Mercier, Hugo. 2020. *Not Born Yesterday*. Princeton, NJ: Princeton University Press.

Mercier, Hugo, and Dan Sperber. 2011. "Why Do Humans Reason? Arguments for an Argumentative Theory." *Behavioral and Brain Sciences* 34, no. 2: 57–74.

Mercier, Hugo, and Dan Sperber. 2017. *The Enigma of Reason*. Cambridge, MA: Harvard University Press.

Miller, Arthur G., John W. McHoskey, Cynthia M. Bane, and Timothy G. Dowd. 1993. "The Attitude Polarization Phenomenon: Role of Response Measure, Attitude Extremity, and Behavioral Consequences of Reported Attitude Change." *Journal of Personality and Social Psychology* 64, no. 4: 561–74.

Mills, Charles W. 2007. "White Ignorance." In *Race and Epistemologies of Ignorance*, edited by Shannon Sullivan and Nancy Tuana 13–38. Albany, NY: State University of New York Press.

Moore, Don A., Ashli B. Carter, and Heather H. J. Yang. 2015. "Wide of the Mark: Evidence on the Underlying Causes of Overprecision in Judgment." *Organizational Behavior and Human Decision Processes* 131: 110–20.

Moss, Sarah 2018. *Probabilistic Knowledge* Oxford: Oxford University Press.

Munro, Geoffrey D., and Peter H. Ditto. 1997. "Biased Assimilation, Attitude Polarization, and Affect in Reactions to Stereotype-Relevant Scientific Information." *Personality and Social Psychology Bulletin* 23, no. 6: 636–53.

Murray, Mark, 2018. "Poll: 58 Percent Say Gun Ownership Increases Safety." *NBC News*, March 23, 2018. https://www.nbcnews.com/politics/first-read/poll-58-percent-say-gun-ownership-increases-safety-n859231.

Myers, David G. 2012. *Social Psychology*. New York: McGraw-Hill Education.

Myers, David G., and Helmut Lamm. 1976. "The Group Polarization Phenomenon." *Psychological Bulletin* 83, no. 4: 602–27.

Nguyen, C. Thi. 2018. "Escape the Echo Chamber." *Aeon*, April 9, 2018. https://aeon.co/essays/why-its-as-hard-to-escape-an-echo-chamber-as-it-is-to-flee-a-cult.

Nickerson, Raymond S., 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises." *Review of General Psychology* 2, no. 2: 175–220.

Nielsen, Michael, and Rush T. Stewart. 2021. "Persistent Disagreement and Polarization in a Bayesian Setting." *British Journal for the Philosophy of Science* 72, no. 1: 51–78.

Nimark, Kristoffer P., and Savitar Sundaresan. 2019. "Inattention and Belief Polarization." *Journal of Economic Theory* 180: 203–28.

Oaksford, Mike, and Nick Chater. 1994. "A Rational Analysis of the Selection Task as Optimal Data Selection." *Psychological Review* 101, no. 4: 608–31.

Oaksford, Mike, and Nick Chater. 1998. *Rational Models of Cognition.* Oxford: Oxford University Press.

O'Connor, Cailin, and James Owen Weatherall. 2018. "Scientific Polarization." *European Journal for Philosophy of Science* 8, no. 3: 855–75.

Oddie, G. 1997. "Conditionalization, Cogency, and Cognitive Value." *The British Journal for the Philosophy of Science* 48, no. 4: 533–41.

Olsson, Erik J. 2013. "A Bayesian Simulation Model of Group Deliberation and Polarization." In *Bayesian Argumentation,* edited by Frank Zenker, 113–33. Berlin: Springer.

Ortoleva, Pietro, and Erik Snowberg. 2015. "Overconfidence in Political Behavior." *American Economic Review* 105, no. 2: 504–35.

Pallavicini, Josefine, Bjørn Hallsson, and Klemens Kappel. 2018. "Polarization in Groups of Bayesian Agents." *Synthese* 198: 1–55.

Pariser, Eli. 2012. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think.* London: Penguin Books.

Pascal, Blaise. 1660. "From Pensées." *Pensées,* 1.

Pennycook, Gordon, and David G. Rand. 2019. "Lazy, Not Biased: Susceptibility to Partisan Fake News is Better Explained by Lack of Reasoning Than by Motivated Reasoning." *Cognition* 188: 39–50.

Peterson, Cameron R., and Lee R. Beach. 1967. "Man As an Intuitive Statistician." *Psychological Bulletin* 68, no. 1: 29–46.

Pettigrew, Richard, and Michael G. Titelbaum. 2014. "Deference Done Right." *Philosopher's Imprint* 14, no. 35: 1–19.

Petty, R. E. 1994. "Two Routes to Persuasion: State of the Art." In *International Perspectives On Psychological Science, II: The State of the Art,* edited by Paul Bertelson, Paul Eelen, and Gery d'Ydewalle, 1–15. East Sussex, England: Psychology Press.

Petty, Richard E., and Duane T. Wegener. 1998. "Attitude Change: Multiple Roles for Persuasion Variables." In *The Handbook of Social Psychology,* edited by D. T. Gilbert, S. T. Fiske, and G. Lindzey, 323–90. New York: McGraw-Hill.

Plous, Scott. 1991. "Biases in the Assimilation of Technological Breakdowns: Do Accidents Make Us Safer?" *Journal of Applied Social Psychology* 21, no. 13: 1058–82.

Podgorski, Abelard. 2016. "Dynamic Permissivism." *Philosophical Studies* 173, no. 7: 1923–39.

Proietti, Carlo. 2017. "The Dynamics of Group Polarization." In *International Workshop on Logic, Rationality and Interaction, volume 10455,* edited by A. Baltag, J. Seligman, and T. Yamada, 195–208. Berlin: Springer.

Pronin, Emily. 2008. "How We See Ourselves and How We See Others." *Science* 320, no. 16: 1177–80.

Rabin, Matthew, and Joel Schrag. 1999. "First Impressions Matter: A Model of Confirmatory Bias." *Quarterly Journal of Economics* 114, no. 1: 37–82.

Ramsey, F. P. 1990. “Weight or the Value of Knowledge.” *British Journal for the Philosophy of Science* 41, no. 1: 1–4.

Risen, Jane, and Thomas Gilovich. 2007. “Informal Logical Fallacies.” In *Critical Thinking in Psychology*, edited by R. J. Sternberg, H. L. Roediger, and D. F. Halpern, 110–30. Cambridge: Cambridge University Press.

Rizzo, Mario J., and Glen Whitman. 2019. *Escaping Paternalism: Rationality, Behavioral Economics, and Public Policy*. Cambridge: Cambridge University Press.

Robson, David. 2018. “The Myth of the Online Echo Chamber.” *BBC*, April 16, 2018. https://www.bbc.com/future/article/20180416-the-myth-of-the-online-echo-chamber.

Rogers, Kayleigh. 2020. “Americans Were Primed To Believe The Current Onslaught Of Disinformation.” *FiveThirtyEight*, November 12, 2020. https://fivethirtyeight.com/features/americans-were-primed-to-believe-the-current-onslaught-of-disinformation/.

Ross, Lee. 2012. “Reflections on Biased Assimilation and Belief Polarization.” *Critical Review* 24, no. 2: 233–45.

Salow, Bernhard. 2018. “The Externalist’s Guide to Fishing for Compliments.” *Mind* 127, no. 507: 691–728.

Salow, Bernhard. 2019. “Elusive Externalism.” *Mind* 128, no. 510: 397–427.

Samet, Dov. 1999. “Bayesianism without Learning.” *Research in Economics* 53, no. 2: 227–42.

Samet, Dov. 2000. “Quantified Beliefs and Believed Quantities.” *Journal of Economic Theory* 95, no. 2: 169–85.

Savage, Leonard J. 1954. *The Foundations of Statistics*. New York: Wiley.

Schervish, M. J., T. Seidenfeld, and J. B. Kadane. 2004. “Stopping to Reflect.” *The Journal of Philosophy* 101, 6: 315–22.

Schoenfield, Miriam. 2012. “Chilling Out on Epistemic Rationality.” *Philosophical Studies* 158, no. 2: 197–219.

Schoenfield, Miriam. 2014. “Permission to Believe: Why Permissivism is True and What it Tells Us About Irrelevant Influences on Belief.” *Noûs* 48, no. 2: 193–218.

Schoenfield, Miriam. 2015. “A Dilemma for Calibrationism.” *Philosophy and Phenomenological Research* 91, no. 2: 425–55.

Schoenfield, Miriam. 2017. “Conditionalization Does Not (In General) Maximize Expected Accuracy.” *Mind* 126, no. 504: 1155–87.

Schoenfield, Miriam. 2018. “An Accuracy Based Approach to Higher Order Evidence.” *Philosophy and Phenomenological Research* 96, no. 3: 690–715.

Schuette, Robert A., and Russell H. Fazio. 1995. “Attitude Accessibility and Motivation as Determinants of Biased Processing: A Test of the MODE Model.” *Personality and Social Psychology Bulletin* 21, no. 7: 704–10.

Sears, David O., and Jonathan L. Freedman, 1967. “Selective Exposure to Information: A Critical Review.” *Public Opinion Quarterly* 31, no. 2: 194–213.

Seidenfeld, Teddy, and Larry Wasserman. 1993. “Dilation for Sets of Probabilities.” *The Annals of Statistics* 21, no. 3: 1139–54.

Siegel, Susanna. 2021. “The Problem of Culturally Normal Beliefs.” In *Ideology: New Essays*, edited by Robin Celikates, Sally Haslanger, and Jason Stanley, forthcoming. Oxford: Oxford University Press.

Simpson, Robert Mark. 2017. “Permissivism and the Arbitrariness Objection.” *Episteme* 14, no. 4: 519–38.

Singer, Daniel J., Aaron Bramson, Patrick Grim, Bennett Holman, Jiin Jung, Karen Kovaka, Anika Ranginani, and William J. Berger. 2019. “Rational Social and Political Polarization.” *Philosophical Studies* 176, no. 9: 2243–67.

Skyrms, Brian. 1990. “The Value of Knowledge.” *Minnesota Studies in the Philosophy of Science* 14: 245–66.

Sliwa, Paulina, and Sophie Horowitz. 2015. “Respecting *All* the Evidence.” *Philosophical Studies* 172, no. 11: 2835–58.

Solomon, Miriam. 1992. “Scientific Rationality and Human Reasoning.” *Philosophy of Science* 59, no. 3: 439–55.

Srinivasan, Amia. 2015. “Are We Luminous?” *Philosophy and Phenomenological Research* 90, no. 2: 294–319.

Stafford, Tom. 2015. *For Argument’s Sake: Evidence that Reason Can Change Minds.* Smashwords Edition.

Stafford, Tom. 2020. “Evidence for the Rationalisation Phenomenon Is Exaggerated.” *Behavioral and Brain Sciences*, 43: e48.

Stalnaker, Robert. 1968. “A Theory of Conditionals.” In *Studies in Logical Theory*, edited by Nicholas Rescher, 98–112. Oxford: Oxford University Press.

Stalnaker, Robert. 2019. “Rational Reflection, and the Notorious Unmarked Clock.” In *Knowledge and Conditionals: Essays on the Structure of Inquiry*, 99–112. Oxford: Oxford University Press.

Stangor, Charles, and Jennifer Walinga. 2014. *Introduction to Psychology.* BCcampus, BC Open Textbook Project.

Stanovich, Keith E. 2020. *The Bias That Divide Us: The Science and Politics of Myside Thinking.* Cambridge, MA: MIT Press.

Stone, Daniel F., 2019. ““Unmotivated Bias” and Partisan Hostility: Empirical Evidence.” *Journal of Behavioral and Experimental Economics* 79: 12–26.

Stone, Daniel F. 2020. “Just a Big Misunderstanding? Bias and Bayesian Affective Polarization.” *International Economic Review* 61, no. 1: 189–217.

Sunstein, C. 2009. *Going to Extremes: How Like Minds Unite and Divide.* Oxford: Oxford University Press.

Sunstein, Cass R. 2000. “Deliberative Trouble? Why Groups Go to Extremes.” *The Yale Law Journal* 110, no. 1: 71–119.

Sunstein, Cass R. 2017. *#Republic: Divided Democracy in the Age of Social Media.* Princeton: Princeton University Press.

Sutherland, Stuart. 1992. *Irrationality: The Enemy Within.* London: Pinker & Martin, Ltd..

Taber, Charles S., Damon Cann, and Simona Kucsova. 2009. "The Motivated Processing of Political Arguments." *Political Behavior* 31, no. 2: 137–55.

Taber, Charles S., and Milton Lodge. 2006. "Motivated Skepticism in the Evaluation of Political Beliefs." *American Journal of Political Science* 50, no. 3: 755–69.

Talisse, Robert B. 2019. *Overdoing Democracy: Why We Must Put Politics in Its Place.* Oxford: Oxford University Press.

Teller, Paul. 1973. "Conditionalization and Observation." *Synthese* 26, no. 2: 218–58.

Tenenbaum, Joshua B., and Thomas L. Griffiths. 2006. "Optimal Predictions in Everyday Cognition." *Psychological Science* 17, no. 9: 767–73.

Tenenbaum, Joshua B., Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. 2011. "How to Grow a Mind: Statistics, Structure, and Abstraction." *Science* 331, no. 6022: 1279–85.

Tesser, Abraham, Leonard Martin, and Marilyn Mendolia. 1995. "The Impact of Thought on Attitude Extremity and Attitude-Behavior Consistency." In *Attitude Strength: Antecedents and Consequences,* edited by R. E. Petty and J. A. Krosnick, 73–92. Mahwah, NJ: Lawrence Erlbaum.

Thaler, Richard H. 2015. *Misbehaving: The Making of Behavioural Economics.* New York: Penguin.

Todd, Peter M., Thomas T. Hills, Trevor W. Robbins, and Julia Lupp. 2012. *Cognitive Search: Evolution, Algorithms, and the Brain, Volume 9.* Cambridge, MA: MIT press.

Toplak, Maggie E., and Keith E. Stanovich. 2003. "Associations between Myside Bias on an Informal Reasoning Task and Amount of Post-Secondary Education." *Applied Cognitive Psychology* 17, no. 7: 851–60.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185, no. 4157: 1124–31.

van der Maas, Han L. J., Jonas Dalege, and Lourens Waldorp. 2020. "The Polarization Within and Across Individuals: The Hierarchical Ising Opinion Model." *Journal of Complex Networks* 8, no. 2. https://doi.org/10.1093/comnet/cnaa010.

van Ditmarsch, Hans, Joseph Y. Halpern, Wiebe van der Hoek, and Barteld Kooi. 2015. *Handbook of Epistemic Logic.* Rickmansworth: College Publications.

van Prooijen, Jan-Willem, and André P. M. Krouwel. 2019. "Psychological Features of Extreme Political Ideologies." *Current Directions in Psychological Science* 28, no. 2: 159–63.

Vavova, Katia. 2018. "Irrelevant Influences." *Philosophy and Phenomenological Research* 96, no. 1: 134–52.

Vinokur, Amiram, and Eugene Burstein. 1974. "Effects of Partially Shared Persuasive Arguments on Group-Induced Shifts: A Group-Problem-Solving Approach." *Journal of Personality and Social Psychology* 29, no. 3: 305–15.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. “The Spread of True and False News Online.” *Science* 359, no. 6380: 1146–51.

Weatherall, James Owen, and Cailin O’Connor. 2020. “Endogenous Epistemic Factionalization.” *Synthese* 198: 6179–200.

Weisberg, Jonathan. 2007. “Conditionalization, Reflection, and Self-Knowledge.” *Philosophical Studies* 135, no. 2: 179–97.

White, Roger. 2009. “Evidential Symmetry and Mushy Credence.” In *Oxford Studies in Epistemology, Volume 3*, edited by Tamar Szabó Gendler and John Hawthorne, 161–86. Oxford: Oxford University Press.

White, Roger. 2010. “You Just Believe that Because. . .” *Philosophical Perspectives* 24, no. 1: 573–615.

Whittlestone, Jess. 2017. “The Importance of Making Assumptions: Why Confirmation Is Not Necessarily a Bias.” PhD diss., University of Warwick.

Wilkinson, Will, 2018. “The Density Divide: Urbanization, Polarization, and Populist Backlash.” The Niskanen Center. https://www.niskanencenter.org/the-density-divide-urbanization-polarization-and-populist-backlash/.

Williams, Daniel. 2021. “Socially Adaptive Belief.” *Mind and Language,* 36, no. 3: 333–54.

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

Williamson, Timothy. 2008. “Why Epistemology Cannot be Operationalized.” In *Epistemology: New Essays,* edited by Quentin Smith, 277–300. Oxford: Oxford University Press.

Williamson, Timothy, 2014. “Very Improbable Knowing.” *Erkenntnis* 79, no. 5: 971–99.

Williamson, Timothy. 2019. “Evidence of Evidence in Epistemic Logic.” In *Higher-Order Evidence: New Essays,* edited by Mattias Skipper and Asbjørn Steglich-Petersen, 265–97. Oxford: Oxford University Press.

Wilson, Andrea. 2014. “Bounded Memory and Biases in Information Processing.” *Econometrica* 82, no. 6: 2257–94.

Wolfe, Christopher R., and M. Anne Britt. 2008. “The Locus of the Myside Bias in Written Argumentation.” *Thinking & Reasoning* 14, no. 1: 1–27.

Wolfers, Justin, 2014. “How Confirmation Bias Can Lead to a Spinning of Wheels.” *The New York Times,* October 31, 2014. https://www.nytimes.com/2014/11/01/upshot/how-confirmation-bias-can-lead-to-a-spinning-of-wheels.html.

Worsnip, Alex. 2019. “The obligation to diversify one’s sources: against epistemic partisanship in the consumption of news media.” In *Media Ethics, Free Speech, and the Requirements of Democracy,* edited by Carl Fox and Joe Saunders, 240–64. Milton Park, Oxfordshire: Routledge.

Ye, Ru. 2019. “The Arbitrariness Objection against Permissivism.” *Episteme* 18, no. 4: 654–73.

Zendejas Medina Pablo. 2022. "Just As Planned: Bayesianism, Externalism, and Plan Coherence." *Philosophers' Imprint.* Forthcoming.

Zhang, Snow, and Alexander Meehan, 2022. "Bayes Is Back." Working paper.

Zollman, Kevin J. S. 2021. "Network Epistemology: How Our Social Connections Shape Knowledge." *Episteme* 6, no. 2: 221–29.