Rational Polarization

Rational Polarization

Kevin Dorst
Massachusetts Institute of Technology

Forthcoming in *The Philosophical Review*
13,127 (or 11,369*) words

Abstract

Predictable polarization is everywhere: we can often predict how people's opinions—including our own—will shift over time. Extant theories either neglect the fact that we can predict our own polarization, or explain it through irrational mechanisms. They needn't. Empirical studies suggest that polarization is predictable when evidence is *ambiguous*, i.e. when the rational response is not obvious. I show how Bayesians should model such ambiguity, and then prove that—assuming rational updates are those which obey the value of evidence (Blackwell 1953; Good 1967)—ambiguity is necessary and sufficient for the rationality of predictable polarization. The main theoretical result is that there can be a series of such updates, each of which is individually expected to make you more accurate, but which together will predictably polarize you. Polarization results from *asymmetric increases in accuracy*. This mechanism is not only theoretically possible, but empirically plausible. I argue that *cognitive search*—searching a cognitively-accessible space for a particular item—often yields asymmetrically ambiguous evidence; I present an experiment supporting its polarizing effects; and I use simulations to show how it can explain two of the core causes of polarization: confirmation bias and the group polarization effect.

Keywords: Polarization, Ambiguous Evidence, Confirmation Bias, Value of Evidence, Reflection (Martingale) Principles, Bayesian Persuasion

Main Text:

Appendices:

*Informal track—formal subsections can be skipped without loss of continuity. **Abridged version** (under 10k words) is available at http://www.kevindorst.com/uploads/8/8/1/7/88177244/rp_abridged.pdf.

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 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 et al. 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 with certainty, but with some confidence—that I'd 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 (Iyengar et al. 2019; Boxell et al. 2020).

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

This requires rejecting standard Bayesian assumptions. Though it's often overlooked, they imply that predictable polarization must be irrational, regardless of varying evidential standards (Schoenfield 2014), background beliefs (Jern et al. 2014; Benoît and Dubra 2019), or distributions of trust (O'Connor and Weatherall 2018; Henderson and Gebharter 2021).

[1] Sutherland 1992; Lakoff 1997; Mills 2007; Lilienfeld et al. 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] Myers and Lamm 1976; Isenberg 1986; Baron et al. 1996; Sunstein 2000, 2009; Mcpherson et al. 2001; Cohen 2003; Pronin 2008; Iyengar et al. 2012; Mäs and Flache 2013, Myers 2012, Ch. 8, Baumgaertner et al. 2016; Brownstein 2016; Mason 2018; Wilkinson 2018; Talisse 2019; Siegel 2021; Williams 2021.

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

[4] Lichtenstein et al. 1982; Harvey 1997; Johnson 2009; Glaser and Weber 2010; Moore et al. 2015; Ortoleva and Snowberg 2015; van Prooijen and Krouwel 2019; Stone 2019.

[5] Jamieson and Cappella 2008; Pariser 2012; Nguyen 2018; Sunstein 2017; Vosoughi et al. 2018.

For they require your current opinion to always match your estimate of your future rational opinion, meaning you can't (rationally) do what we do all the time: predict the direction our actions will shift our opinions (§2).

But we *should* reject those assumptions, for they also imply that rational people can never be unsure whether they've been rational. Given *ambiguous evidence*—evidence that's 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 (§3). Indeed, common cognitive processes generate *asymmetric* ambiguities, making it easier to recognize evidence pointing in one direction than the other (§4). Each such update is expected to improve accuracy, despite the fact that a long series of them can predictably lead to profound polarization (§5). Moreover, this mechanism plausibly plays a role in the psychological processes that drive real-world polarization (§§6–7).

Although this story is built upon a series of technical results, the main ideas can be understood without them. Thus I've partitioned the paper: 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 means that if predictable polarization is due to irrationality, *my* beliefs are due to irrationality. Yet *I* can't admit that—at least not while maintaining my beliefs: it's incoherent ('akratic') to believe "Guns decrease safety, but it's irrational for me to believe that" (Horowitz 2014; Dorst 2020a). So if I'm not willing to give up my beliefs—as indeed I'm 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's also dubious. My friends were smarter (and quicker) than I was. My trajectory was more predictable than theirs were. Our divergence is due to our *circumstances*, not ourselves. A slight change in those, and I'd 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. Example: is the following word-search completable?

FR_ _L

If you find a completion, you know you should be 100% confident it's *c*ompletable (c). But if you don't 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% (§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's ambiguous, it leaves you where you were. So the (potentially ambiguous) evidence won't hurt, and might help.

Second: such *asymmetric accuracy-increases* can drive polarization. Iterate this with many claims $c_1, ..., c_n$ that you're 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'll 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 (§§3–5).

I'll 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 don't: 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'll 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 (§6); and arguments can predictably persuade you by making the evidence favoring their side less ambiguous than the evidence opposing it (§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's 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?

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'll polarize further. For example, Jern et al. 2014 and Henderson and Gebharter 2021 show that for two Bayesians who disagree about the likely causal paths or

[6] E.g. Tversky and Kahneman 1974; Kahneman et al. 1982; Kahneman and Tversky 1996; Fine 2005; Ariely 2008; Hastie and Dawes 2009; Kahneman 2011; Thaler 2015; Mandelbaum 2018.

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

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

[9] E.g. Feeney et al. 2000; Dixit and Weibull 2007; Austerweil and Griffiths 2011; Le Mens and Denrell 2011; Olsson 2013; Acemoglu and Wolitzky 2014; Jern et al. 2014; Cook and Lewandowsky 2016; Angere and Olsson 2017; Pallavicini et al. 2018; Benoît and Dubra 2019; Nimark and Sundaresan 2019; Nielsen and Stewart 2021; Bowen et al. 2021; Henderson and Gebharter 2021.

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 can't predict how they'll polarize: they can't know whether they'll learn E or $\neg E$, and learning the latter would push their opinions in the *other* direction ($P(q) > P(q|\neg E)$ and $P'(q|\neg E) > P'(q)$).

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 '$\widetilde{P}$' the rational one after the update. (More on my rationality-assumptions in §3.) Since you can be unsure what evidence you'll 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 can't estimate that it'll 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 can't 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, i.e. the true cell of a finite[10] partition (see §A.3). (Example: 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.
> 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're rational only if Standard Bayesianism is wrong. In this section I'll 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're 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 more generally.

Another example: biased sources. Make an estimate of the number of extreme weather events there'll be in the U.S. 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

[10] I'll restrict attention to finite models.

[11] See Kadane et al. 1996; Weisberg 2007; Briggs 2009; Salow 2018 for explanations. The 'Bayesian persuasion' literature (Kamenica and Gentzkow 2011) takes this constraint as axiomatic. As we'll see, it needn't.

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

in your climate-news consumption—say, reading only the most dire, doomsday climate-change reports? Obviously you expect this would increase your estimate! You're 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).

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

It's not just you. It's well-documented that people tend to shift their beliefs in the direction they're searching for evidence (e.g. Isenberg 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 et al. 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 you your initial opinion matches your initial estimate of the opinions you'll have at the end.[14] Yet as the epistemology of 'irrelevant influences' emphasizes, this defies common sense.[15] 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 wasn't, he could predict how his opinion would move given his choice. The choice *itself* was no evidence—so upon choosing Oxford, he still had no opinion, but could now predict that he'd 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'd parted ways—when our opinions hadn't moved, but our trajectories were clear. I'd started studying at an urban university; he'd started bartending in a rural town. Let P be my (rational) opinions then, and $\widetilde{P}$ be those it'd be rational to have 5 years later. Likewise for D and $\widetilde{D}$ for Dan. Let $\boldsymbol{s}$ be a partisan-coded claim, e.g. that *guns increase* $\boldsymbol{s}$*afety*. 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 didn't.[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'd 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'd be more doubtful. If this was rational, the following must be possible:

> **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'd receive radically different evidence,

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

[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] E.g. Cook 1987; Cohen 2000; White 2010; Schoenfield 2017b; Vavova 2018. For empirical work, see Mcpherson et al. 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).

Standard Bayesianism forbids our expectable polarization. (When Reflection fails in this way, I'll speak of a *single* person expectably polarizing.)

Some clarifications. First, I don't claim *politics* is predictable—it's 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 don't *predict* this, since you're 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 so is all we need to generate the epistemic problem of predictable polarization. In response, I'll show that updates that are expected to make you more accurate about every subject-matter and can't be Dutch-booked can nonetheless expectably polarize you (§§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 doesn't 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 (§5):

> **Predictable Polarization:** Dan and I could both *predict with confidence* that my credence in s should substantially drop, and his should substantially rise.
> $P\Big(\widetilde{P}(s) << P(s)\Big) \approx 1$ and $P\Big(\widetilde{D}(s) >> D(s)\Big) \approx 1$; 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 paper (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, let's 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 allows varying standards of reasoning across people (Schoenfield 2014) and times (Callahan 2019). Since what's rational to think (what you 'should' think) varies across worlds—with your evidence, priors, etc.—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, etc.[17]

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

How is it rational to *change* opinions? I won't 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: which updates $\langle P, \widetilde{P}\rangle$ represent *potentially-rational* updates, i.e. ones that 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 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** $\widetilde{P}$ iff the update $\langle P, \widetilde{P}\rangle$ satisfies these constraints (Appendix A.2)—iff, in other words, P prefers to give $\widetilde{P}$ power of attorney to make its decisions for it. I'll assume throughout—with a slight weakening in §5—that:

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

I'll assume that a *sequence* of updates $\langle P^1, P^2\rangle$, $\langle P^2, P^3\rangle$, ... is potentially-rational iff 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's commonly thought that Value (or the avoidance of Dutch books) on its own entails Reflection, and hence forbids expectable polarization. It doesn't:[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/blue) that $P_x(y) = t$ or (right/red) that $\widetilde{P}_x(y) = t$:
>
> 
>
>
> It's not hard to see that 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]

[18] You might add: "...and there's no available update preferable to $\widetilde{P}$." If so, what I'll 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 wasn't explicit until Dorst 2020a. Cf. Williamson 2000, Ch. 10.

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

How do Standard-Bayesians forbid this? In this model, at world g you learn that you're 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 the latter is an error: if sometimes you learn g, then when you don't learn g you learn something—namely, that *you didn't learn* g. In other words, they assume that rational updates are *introspective*: you always can be rationally sure of what you (did or didn't) 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's at b), it doesn't know what it learned! Thus it's 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 doesn't. 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's *rational* to have. When $\widetilde{P}$ is uncertain what $\widetilde{P}$ is, that means it's rational to be unsure what the rational opinions are—it's rational to have epistemic self-doubt.[22] Standard Bayesians assume that such self-doubts *couldn't* 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's hard to know what to make of it—when it's rational to be unsure what it's rational to think (Ellsberg 1961, 661). This higher-order model of ambiguity follows naturally from 'anti-luminous' epistemology, which argues that we often can't tell exactly what rationality requires of us (see Williamson 2000, chapters 4 and 10, and Srinivasan 2015). If you endorse anti-luminosity, you should permit ambiguity in this sense—and even if you have doubts about anti-luminosity in general, there's 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's 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 (§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-

[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 §A.1 and Samet 2000), and philosophically nontrivial on many interpretations (Lewis 1980; Williamson 2008; Pettigrew and Titelbaum 2014; Salow 2018; Dorst 2019, 2020a; Das 2020a,b; Levinstein 2022; Levinstein and Spencer 2022).

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

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

Bayesian—the result of conditioning a fixed prior on the true cell of a partition—iff 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 impose Reflection.

Meanwhile, extant models that *allow* expectable polarization do so using updates that violate Value, so are subject to Dutch books and are expected to make you less accurate.[25] What to make of this? If we allow *non*-valuable updates to be 'rational', the standard storytellers might fairly complain that we've 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 §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.

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's also a tight *empirical* connection between polarization and ambiguity. The intuitive cases of rational self-doubt—what I'm calling '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's 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 (Lord et al. e.g. 1979; Kunda e.g. 1990; Kahan et al. e.g. 2017; see §6). These effects are exacerbated by group discussions, where peer (dis)agreements have large effects on people's opinions (Isenberg 1986, e.g.; see §7). And when the evidence is made easier to interpret or discussion-norms are altered, biased processing often disappears (Lundgren and Prislin 1998; Grönlund et al. 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?

[25] E.g. Kadane et al. 1996; Rabin and Schrag 1999; Hegselmann and Krause 2002; DeMarzo et al. 2003; Halpern 2010; Flache and Macy 2011; Andreoni and Mylovanov 2012; Baliga et al. 2013; Wilson 2014; Baumgaertner et al. 2016; Proietti 2017; O'Connor and Weatherall 2018; Fryer et al. 2019; Loh and Phelan 2019; Singer et al. 2019; Stone 2020; van der Maas et al. 2020; Weatherall and O'Connor 2020; Zollman 2021.

[26] E.g. Schoenfield 2014; Podgorski 2016; Simpson 2017; Callahan 2019; Ye 2019; Jackson 2021.

[27] See the 'higher-order evidence' literature, e.g. Christensen 2010; Lasonen-Aarnio 2013, 2014, 2015; Horowitz 2014, 2019; Schoenfield 2015, 2018; Sliwa and Horowitz 2015; Fraser 2021; Dorst 2020b gives a summary.

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 2020). Given a string of letters and some blanks, you have a few seconds to figure out whether there's 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: 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've done this rationally, other times you don't. If you *find* a completion ('PLANET!'), you (often) know that it's rational to be certain there's a word (that $\widetilde{P}(\mathit{Word}) = 1$). But if you *don't* find a completion, you don't 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'll argue that this generates an ambiguity-asymmetry between completable and uncompletable searches, rationalizing expectable polarization. In §5 I'll turn to *predictable* polarization.

Meet Haley. She's wondering whether a fair coin landed heads. I'll show her a word-search determined by the outcome: if heads, it'll be completable; if tails, it'll be uncompletable. Thus her credence in heads equals her credence it's completable. She'll have 7 seconds, then she'll 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: $H(\mathit{Heads}) = 0.5$. But I claim her estimate for her posterior rational credence should be *higher* than 50%: $\mathbb{E}_H(\widetilde{H}(\mathit{Heads})) > 0.5$. Remember: estimates aren't predictions, so she needn't 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's 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'll say that we must find the most fine-grained question (partition) Q that Haley can always answer with certainty, and that she's rational iff she conditions on the true answer to Q. It's 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 ($\neg\mathit{Find}$); so $Q = \{\mathit{Find}, \neg\mathit{Find}\}$. (If she learns more, they'll insist there's 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's a word, she'll find it half the time ($H(\mathit{Find}|\,\mathit{Word}) = 1/2$);

and if there's not, she'll never find one ($H(\mathit{Find}|\neg\mathit{Word}) = 0$). Then $1/4$ of the time she'll learn *Find* ($1/2$ likely to be a word, and $1/2$ likely to find if so), making it rational to be sure there's a word: $\widetilde{H}(\mathit{Word}) = 1$. (And she'll *know* this is the rational reaction: $\widetilde{H}(\widetilde{H}(\mathit{Word}) = 1) = 1$.) The remaining $3/4$ of the time she'll learn only $\neg\mathit{Find}$, making it rational to slightly lower her credence: $\widetilde{H}(\mathit{Word}) = 1/3$.[28] (And since she'll know all she learned is that she didn't find one, she'll *know* that this is the rational reaction: $\widetilde{H}(\widetilde{H}(\mathit{Word}) = 1/3) = 1$.) Thus her expected future rational credence is $1/4 \cdot 1 + 3/4 \cdot 1/3 = 1/2$. No expectable polarization.

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

Reflect on your experience with another example:

_E_RT

When you haven't 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're missing something obvious. If you haven't found one when the 7-second timer goes off, your credence that there's a word may have gone down or gone up, but you won't (shouldn't!) be willing to bet the farm that it's moved in the rational direction. After all, sometimes it doesn't: 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're irrational, what about when you've *in fact* been rational to lower your credence? You should still wonder whether you've been *ir*rational. For example, if you don't find a completion to ST_ _RE and so drop your credence to $1/3$, you might still wonder if there's a word and (so) wonder if you should have a higher credence—even though in fact there isn't, so in fact you shouldn't. Rational people can doubt that they're rational, just as humble people can doubt that they're humble.

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

Here's a simple model (details in §4.1). Suppose, as before, it's $1/2$ likely there's no word (and so she doesn't find one), $1/4$ likely there's a word she finds, and $1/4$ likely there's a word she doesn't find. Moreover, suppose she knows the string will be word-like iff there's a word. If she finds a word, she's rational to become certain there's one: $\widetilde{H}(\mathit{Word}) = 1$. If

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

she doesn't find a word *and there is none*—so it's not word-like—she's rational to drop her confidence slightly: $\widetilde{H}(\textit{Word}) = 1/3$. So far this is just like the Standard-Bayesian model. Yet suppose that if she doesn't find a word *but there is one* (so the string is word-like), she's rational to raise her credence slightly—she should suspect it's word-like: $\widetilde{H}(\textit{Word}) = \frac{2}{3}$. This yields ambiguous evidence: if she doesn't find a word, she's rational to be unsure whether the rational posterior is $1/3$ or $2/3$: $\widetilde{H}(\widetilde{H}(\textit{Word}) = 1/3) > 0$ and $\widetilde{H}(\widetilde{H}(\textit{Word}) = 2/3) > 0$. (Which one it is depends on whether the string is word-like—but she's *also* rational to be unsure of that. There is no cognitive home; see 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's no word, both update to credence $1/3$; but if there is a word she doesn't 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 sometimes more accurate.[29] Thus neither is Dutch-bookable, and it's always rational to prefer the ambiguous one (§4.1).

Second, this update is expectably polarizing: Haley is initially 0.5 confident there's 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 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'd be a much wider range of possibly-rational posteriors. Appendix A.4 proves that a wide class of such models will lead to expectable polarization—so even if you object to the details, I hope you'll 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 §§6–7) 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: 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 to test the hypothesis that ambiguous evidence can polarize real people? Meet Thomas. Like Haley, he's about to see a word-search, determined by the (same) coin toss. But while she'll see a completable string iff heads, he'll see a completable string iff

[29] Is the comparison unfair, since the ambiguous posterior differs in more places than the Standard-Bayesian one? Insisting it's 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 2017a; 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 §A.3, Theorem A.2). So although this is a way to object, if you're onboard with ambiguous evidence you shouldn't worry about such 'unfairness'.

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

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

[32] If her actual opinion, $\widetilde{C}(\textit{Word})$, is an unbiased estimator of the rational opinion (meaning $\forall t : \mathbb{E}_H(\widetilde{C}(\textit{Word})|\widetilde{H}(\textit{Word}) = t) = t$), then it'll expectably polarize to the same degree. If it's a biased estimator, it may still polarize depending on the degree and direction of the bias.

tails. By parallel reasoning, Thomas's opinion should expectably polarize in the opposite direction: it'll 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'll see a completable string iff heads. The rest are *Tailsers:* like Thomas, they'll see a completable string iff tails. Headsers get evidence that's easier to assess when the coin lands heads; Tailsers get evidence that's 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; and 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'll 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 §4.2). This doesn't establish that the participants themselves could predict how they'd polarize, but it does support a necessary precondition of that—namely, that I could predict it.

More work is needed. But we've now seen—in principle, and perhaps in practice—how cognitive search could generate ambiguities that rationalize expectable polarization. What of *predictable* polarization? If you'd like to jump to that argument, skip to §5; for the technical (§4.1) or experimental (§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 doesn't 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: if $\widetilde{H}_i(j) > 0$, then $\widetilde{H}_i = \widetilde{H}_j$.

The ambiguous model (Figure 2 below) is identical except that in w_2 the rational posterior assigns higher credence to there being a word (the 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 doesn't 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's 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: $\mathbb{E}_H(\widetilde{H}(\textit{Word})) = 1/2 \cdot 1/3 + 1/4 \cdot 2/3 + 1/4 \cdot 1 = \frac{7}{12} \approx 0.58 > 0.5 = H(\textit{Word})$. 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've seen in the Example in §3.[33]

[33]If we assume all updates happen by conditioning, ambiguity occurs iff the possible propositions that

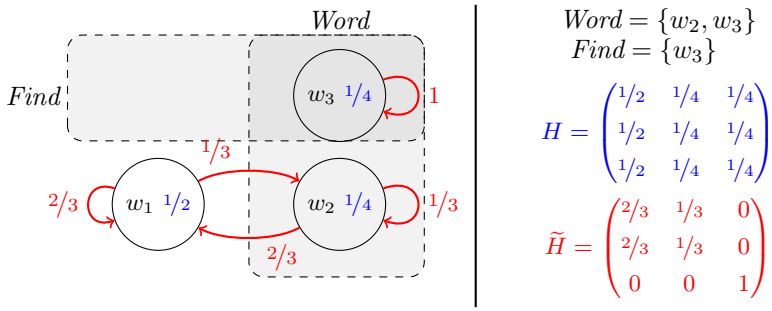


Figure 1: Standard-Bayesian model of a Headser's rational opinions. **Left:** Generalized-Kripke (Markov) diagram, in which blue numbers represent the prior probabilities of possibilities, and red arrows from circles represent the posterior probabilities *in* those possibilities. **Right:** The matrix H represents (constant) prior probabilities; the matrix $\widetilde{H}$ represents posteriors: row i column j is the probability, in world i, that it's 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's completable but she doesn't find one, etc.

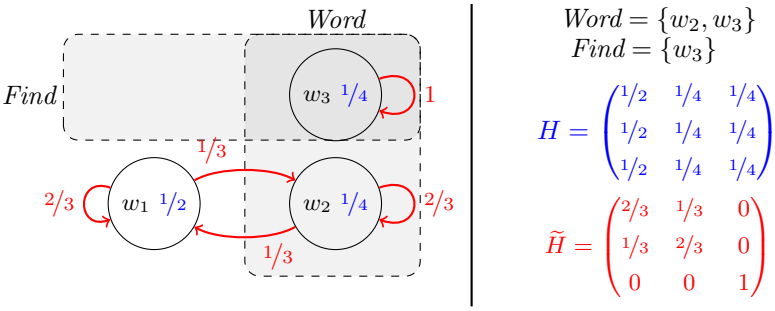


Figure 2: Ambiguous model of a Headser's rational opinions. See Figure 1 for interpretation.

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't she be *sure* there's a word in w_2? Because she can't 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 can't be *sure* she received 'word-like', how can she be sensitive to whether it's 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'll sketch an experiment that both suggests that cognitive search can cause people to polarize, and controls for a confound. (See Appendix B for more.)

might be rational to condition on don't form a partition. See e.g. Geanakoplos 1989; Williamson 2000, Ch. 10; Salow 2018; Dorst 2020a; Das 2019; Isaacs and Russell 2022; Zendejas Medina 2022.

The confound: ambiguous evidence is not simply *weak* evidence. Evidence is weak when it shouldn't move your opinions very much; evidence is ambiguous when *you shouldn't be sure how weak it is*. Highly-ambiguous evidence must be weak (Dorst 2020a, 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. I flip a coin, grabbing A if heads, B if tails. Then I draw a marble and show you. A black marble is strong evidence: you should be sure I'm holding A. A red marble is weak evidence: you should slightly boost your confidence that I'm holding B. Either way, it's unambiguous: if it's black, you should know that you should be sure I'm holding Urn A; if it's red, you should know that you should be $2/3$ confident I'm holding urn B. You should not have self-doubt.

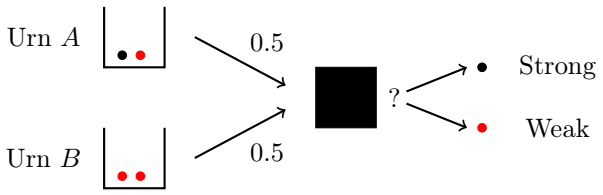


Figure 3: A case of unambiguous (but sometimes weak) evidence.

Upshot: 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's not, you get ambiguous evidence that it's not. But also: if the string is completable, you can get *strong* evidence that it is; if it's not, you get weak evidence that it's not. My theory predicts that the *ambiguity*-asymmetry drives polarization; but what if the weak/strong asymmetry does? What if people polarize because they under-react to weak evidence?[34]

The experiment tested this (pre-registration: https://aspredicted.org/8jg3e.pdf) 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 $Heads_1$,...,$Heads_4$ 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, etc. The Ambiguous condition polarized,[35] and did so significantly more than the Unambiguous one.[36] Appendix B reviews evidence

[34]There's 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 et al. 2010; Hahn and Harris 2014)—a source of ambiguity.

[35]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 $[16.02, 26.82]$.

[36]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

that ambiguity explains this effect—including the minor polarization in the 'Unambiguous' condition.

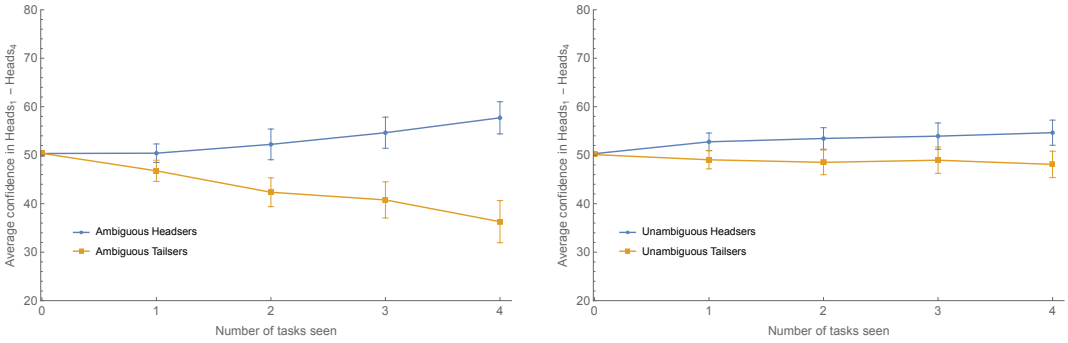


Figure 4: Means of participants' average confidence in $Heads_1$,...$Heads_4$ as they saw more tasks, in Ambiguous (left) and Unambiguous (right) conditions. Error bars represent 95% confidence intervals.

5 The Predictable Theorem

We've 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'd come to doubt that guns increase safety? Since estimates are not predictions, this doesn't follow from what we've seen so far. Can we go further?

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

The basic idea: iterate cognitive searches. In the model from §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's initially confident that around half the coins will land heads, she predicts that her average credence in $Heads_1, ..., Heads_n$ should rise to around 58%. Since they're 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's initially 50-50 in the latter, that's 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've 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* ***h****alf 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's 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.[37]

differences, i.e. for (A-Headsers − A-Tailsers) − (U-Headsers − U-Tailsers), was $[7.19, 22.59]$—indicating that the former was larger.

[37]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 2020a, Fact 5.5). So if $H^1(H^n(h) \geq 0.9) \geq 0.9$, we must have $H^1(h) \geq 0.9 \cdot 0.9 = 0.81 > 0.5$.

Though it's not obvious, this implies that for some i, H^i does not value H^{i+1}, for Theorem A.4 (§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 §4, then H^2 would not value H^3.[38]) 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 don't obey full Value—after all, they sometimes forget things. And here's an easy theorem: if Haley might forget something—*anything*—then she can't value her future opinions.[39]

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—aren't. The former are questions whose answers you should care about getting right; the latter are not. As stated, Value ignores this distinction: H values $\widetilde{H}$ iff for *any* decision problem, it prefers to let $\widetilde{H}$ decide; iff for *every* question, it expects $\widetilde{H}$ to be more accurate than itself; iff there is *no* subject-matter the update can be Dutch-booked on.

That's a high bar. Most forms of deference are *question-relative*. You defer to the forecaster about whether it'll rain, but not about whether your poncho is stylish; you defer to your future-self about how busy you'll 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*** iff, 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 §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's 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 aren't *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'll show that it doesn't.[40]

Here's why. H can value $\widetilde{H}$ with respect to Q even if $\widetilde{H}$ forgets some things, so long as that forgetting doesn't affect $\widetilde{H}$'s opinions about Q. This yields one way of iterating cognitive searches.[41] Let Q be the question of *how all the cognitive searches went*, including

[38] 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, §4; Das 2020b; Dorst 2020a, §A.1.

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

[40] Although a variety of models show how limited memory can lead to polarization (Wilson 2014; Dallmann 2017; Fryer et al. 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.

[41] 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, and so rationally does each word-search. This avoids any forgetting, but has the down-side that Haley does not care about the claim (h) she's predictably polarizing on throughout the process; see §A.6.

whether Haley found a word and whether the coin landed heads or tails. Suppose this—so any question answered by it, e.g. whether more than *h*alf landed heads (h)—is what Haley should care about. Each time she's presented with a string, she updates as discussed in §4 (Figure 2, page 15). Such updates satisfy (full) Value. But she knows that, after each, she'll forget the letter-string (the details of the evidence she received). This forgetting doesn't 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 doesn't find the completion, she's left wondering whether the string is word-like, and hence whether she should be 1/3 or 2/3 confident it's completable. But once she forgets the string, she knows she can no longer be sensitive to whether it's word-like, and so 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 §4 and satisfy (full) Value. Rinse and repeat.

The main theoretical result of this paper 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'll *then* expect the later ones to make her more accurate, 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've 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'll give her an inkling about Q), but doesn't want to do the first two—for doing so might generate too much ambiguity to be valuable. Suppose she does the first and doesn't find a word, so is left with ambiguous evidence ("Should I be 1/3 or 2/3 there's a word?"). At this stage she doesn't 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's 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 can't conclude that the ones it's fallen in landed tails—that would require assuming she's been rational, which she can't be confident of. Thus the fact that she *initially* predicted that half would land heads can't 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 doesn't provide much evidence. Her confidence persists.[42]

They *can* reduce (but not eliminate) their disagreement if they start sharing which completions they found. But that's 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.

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 will 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've predicted polarization as the outcome—as we could. And neither of us should be moved now, when we discuss our persistent disagreements—as we're not. Nevertheless, while we each think the other is incorrect, we needn't think they are dumb, or foolish, or irrational to believe what they do—as we don't.

If that were what happened. I'm going to argue that it may have. That the example of

[42] If they both knew they'd been rational and exactly what each of their opinions were, their disagreement would disappear (Aumann 1976; Lederman 2015). But they don't know that.

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 §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 iff 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 iff every answer to Q settles the value of every option. P ***Q-values*** $\widetilde{P}$ (values $\widetilde{P}$ with respect to Q) iff 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 and Levinstein 2022). See §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's not completable, C_i is where it is but she doesn't 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', $Q(w) = Q(w')$ iff 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're 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 §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? Isn't $\overline{H^1}$'s judgment that H^2 makes better decisions about Q *itself* 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's 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's what allows Q-Value to fail to be transitive.

Given this, you might want rational updates to be valuable about the combined question: *What's 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 don't think so. It's 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 weren't 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'll show how ambiguity-asymmetries may arise in the empirical processes that drive polarization, and that (unlike my word-search example) polarization can be predictable even if 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 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 et al. 2009), hence Theorem 3.1 implies that (i) if evidence is unambiguous then they must be irrational, and (ii) Standard-Bayesian models (see footnote 9) can't rationalize them. I'll show that ambiguous models can.

Confirmation bias first. This effect has been widely cited as a core driver of polarization in both academic[43] and popular[44] writings. Nevertheless, many researchers have noted that we lack good normative standards for assessing its rationality.[45] 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 et al. 1979; Taber and Lodge 2006). Here I'll focus on the latter, returning to the former in §7.

Examples of biased assimilation go like this.[46] Take two people—say, Dan and I—who

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

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

[45] Lord et al. 1979; Lord and Taylor 2009; Taber and Lodge 2006; Crupi et al. 2009; Ross 2012; Mercier 2017; Whittlestone 2017; Kinney and Bright 2021.

[46] Lord et al. 1979 is the classic study; see also Gilovich 1983; Lord et al. 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 et al. 2009; Corner et al. 2012; Ross 2012; Kahan 2013; Jern et al. 2014; Kahan et al. 2017; Cook and Lewandowsky 2016; Liu 2017; Anglin 2019; Benoît and Dubra 2019.

strongly disagree about whether guns increase *s*afety (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've given us the same information, you might expect it to dampen our disagreement. Generally, it won't. 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'll tend to decrease mine.

Why? We won't simply dismiss the evidence against our beliefs—we'll likely spend *more* time looking at it. As we do, we'll 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.[47]

Thomas Kelly (2008) argues that selective scrutiny is rational, and that it may rationalize some types of polarization. It's 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 shouldn't. It makes sense for me to scrutinize the former, and for Dan to scrutinize the latter. Notice that if we do, we'll end up receiving *different* evidence: I know more about one study, 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 *aren't aware* that we're being selective—all we come away with is, "I saw one congruent study, and one *flawed* incongruent one"—then the resulting polarization is rational.

But, says Kelly, this only works if we aren't aware we're being selective. If we are, we shouldn't be surprised to find a flaw in only the incongruent study (cf. McWilliams 2021). (Compare: if you're aware you're fishing with a big net, you shouldn't 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 §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're engaging in selective scrutiny. Indeed, it's standard scientific practice: adopt a hypothesis, and then spend your time trying to explain away problems with it (Kuhn 1962; Solomon 1992). We're 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're looking for (a flaw, a word), it's easier to know how to react to the evidence; if you don't, 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's 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.

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

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 §6.1 for details). Which way it's 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's rational to selectively scrutinize, then *even if you're aware of it*, the resulting ambiguity-asymmetries will rationalize expectable polarization.[48]

But, given this polarizing model, *is* it rational to selectively scrutinize? You might think it couldn't be. After all, repeated selective scrutiny will predictably polarize you—so wouldn't it 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'd expect to be more accurate if you didn't 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's more likely I'll be able to find a flaw, avoiding ambiguity. I may think it's more likely to *contain* a flaw—but even if I don't, I'll 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 paper illustrates the point: what convinced me of its conclusions was, largely, figuring out how to rebut objections—that rational polarization violated Bayesianism (§3), that it was purely theoretical (§4), that ambiguity wasn't the driving force (§4.2), that it couldn't be predictable (§5), and so on. More generally, there's both theoretical (Aronowitz 2020) and empirical (Evans et al. 1983; Kahan et al. 2017) reason 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's an analogy. Suppose I'll see a series of *pairs* of word-search tasks—one following Headser rules, the other following Tailser rules. Headser tasks use British English; Tailser tasks use American English. At each stage I can choose 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'm guided by my desire to form accurate beliefs, I'll tend to do the Tailser tasks more often. And since doing so leads to predictable polarization, I'll 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're better at finding flaws in, and (2) whether this preference can indeed lead to predictable polarization.

[48] As in §4, these updates are fully valuable. If (as in §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.

To (1): I randomly generated models and compared $P(Find|Flaw)$ to expected accuracy, finding a robust correlation (Figure 5, left). I then generated pairs of models in which you're (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, right).

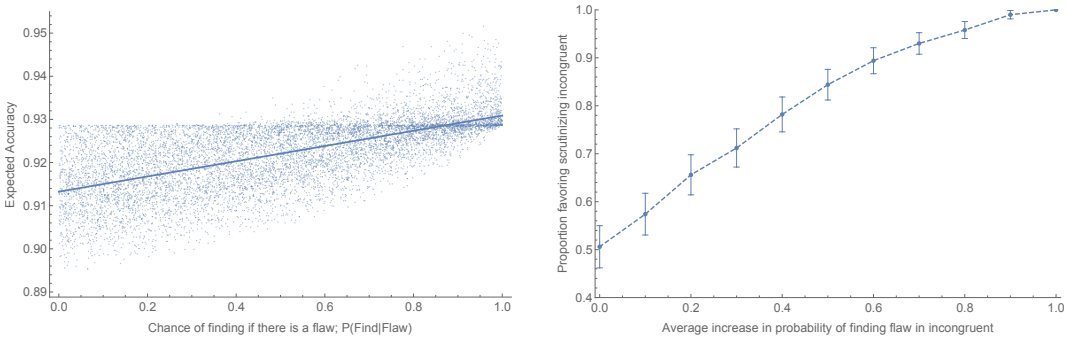


Figure 5: **Left:** Correlation between $P(Find|Flaw)$ and the expected accuracy of scrutiny. **Right:** Rates of selective scrutiny based on expected accuracy (y-axis) grow as the average gap in $P(Find|Flaw)$ between incongruent and congruent studies (x-axis) grows.

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).

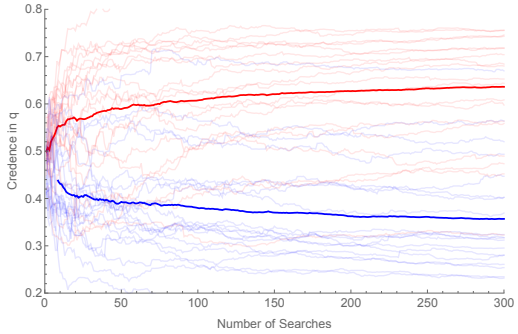


Figure 6: Agents faced with cognitive-search choices, choosing via expected accuracy. Red agents better at finding flaws in q-opposing studies; blue agents vice versa. Thin lines are individuals; thick lines are averages.

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's built on the idea that people are better at finding flaws in incongruent than congruent evidence. They are.[49] It predicts that instructions like "Don't be biased" or "Be accurate" won't prevent biased assimilation—but that instructions that get people to scrutinize both sides

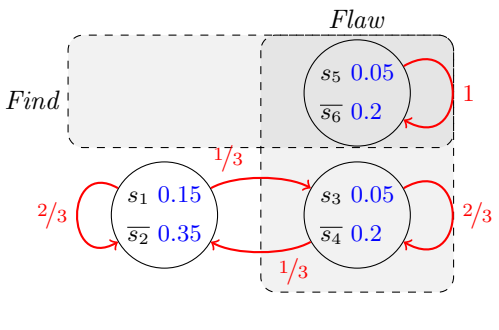
[49] Evans et al. 1983; Petty and Wegener 1998; Mercier and Sperber 2011; Kahan et al. 2012, 2017.

equally *will*. They do.[50] And it suggests that bias will be more extreme when people think *harder*—when they scrutinize more, rather than less. It is.[51]

Upshot: Insofar as confirmation bias and motivated reasoning drove me and Dan apart, this may have been due to rational 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 different social circles? For the answer, skip to §7; for the details from this section, read on.

6.1 The Formalities

Here I'll describe cognitive-search models, which generalize the word-search model from Figure 2 (see Appendix C.1 for more). They have the same structure—possibilities where you find a flaw, possibilities where you don't 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 $\overline{s_j}$ where it's false. Prior probabilities (the blue numbers) are constant across worlds; posteriors are obtained by Jeffrey-shifting the prior P on the $\{\mathit{Find}\&\mathit{Flaw}, \neg\mathit{Find}\&\mathit{Flaw}, \neg\mathit{Flaw}\}$ partition as indicated by the labeled arrows (holding conditional probabilities like $P(\cdot|\mathit{Find}\&\mathit{Flaw})$ fixed, but changing $P(\mathit{Find}\&\mathit{Flaw})$). 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 don't 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's no flaw (s_1 and $\overline{s_2}$), $\frac{2}{3}(\frac{0.15}{0.5}) + \frac{1}{3}(\frac{0.05}{0.25}) \approx 0.2\overline{6}$. If the study contains a flaw, s is 20% likely ($P(s|\mathit{Flaw}) = 0.2$); if it doesn't, s is 30% likely ($P(s|\neg\mathit{Flaw}) = 0.3$); and it's equally likely to contain a flaw as not ($P(\mathit{Flaw}) = 0.5 = P(\neg\mathit{Flaw})$). But since evidence is less ambiguous when I find a flaw, the update is expectably polarizing.[52]



$$\mathit{Flaw} = \{s_3, \overline{s_4}, s_5, \overline{s_6}\}$$
$$\mathit{Find} = \{s_5, \overline{s_6}\}$$

$$P = \begin{pmatrix} 0.15 & 0.35 & 0.05 & 0.2 & 0.05 & 0.2 \\ \dots \end{pmatrix}$$

$$\widetilde{P} = \begin{pmatrix} 0.2 & 0.4\overline{6} & 0.0\overline{6} & 0.2\overline{6} & 0 & 0 \\ 0.2 & 0.4\overline{6} & 0.0\overline{6} & 0.2\overline{6} & 0 & 0 \\ 0.1 & 0.2\overline{3} & 0.1\overline{3} & 0.5\overline{3} & 0 & 0 \\ 0.1 & 0.2\overline{3} & 0.1\overline{3} & 0.5\overline{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.2 & 0.8 \\ 0 & 0 & 0 & 0 & 0.2 & 0.8 \end{pmatrix}$$

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

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

[50] Koriat et al. 1980; Lord et al. 1984; Schuette and Fazio 1995; Lundgren and Prislin 1998; Liu 2017.
[51] Fitzpatrick and Eagly 1981; Kuhn and Lao 1996; Downing et al. 1992; Tesser et al. 1995; Kahan 2013.
[52] $\mathbb{E}_P(\widetilde{P}(\mathit{Flaw})) \approx 0.583 > 0.5 = P(\mathit{Flaw})$, so $\mathbb{E}_P(\widetilde{P}(s)) \approx 0.242 < 0.25 = P(s)$.

simulations only tracked the agents' opinions in s and in the cognitive-searches they were evaluating at a given time—it didn't 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 'small-world' updates like this—which don't 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 pro-government arguments; vice versa for me.

But clearly this isn't 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).[53] This is intuitive, so most explanations stop here.

They shouldn't. A familiar point applies again: it's not just that *someone* can predict that we'll be persuaded by arguments—it's that *we ourselves* can. If you're 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 don't have a knock-down case for the latter, but here's the idea. Suppose you know you'll 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's good, it'll warrant increasing your credence in s ("I hadn't thought of that"); if it's bad, it'll warrant decreasing it ("That's the best they've got?"). You can't be certain the argument will be good—if you were, you should've already raised your credence.[54] Nor will you be able to be sure whether the argument was good after you've seen it: it's ambiguous, so you'll 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

[53] 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've seen fails to control for fact that others' opinions provide evidence (Elga 2007).

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

because they *are* good; bad arguments tend to get repeated because they *sound* good. Thus bad arguments will tend to be more ambiguous, i.e. harder to recognize as bad.

Here's an (overly) simple example. Suppose Jack was hurt, and someone's is trying to convince you that he didn't 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's easier to recognize that the latter is valid than that the former is invalid. (Some fallacies are tempting!) Indeed, there's some evidence that people are worse at recognizing fallacies as fallacies than they are at recognizing validities as validities (Evans et al. 1983; Cariani and Rips 2017, Figure 1).

Suppose this generalizes: arguments are (on average) less ambiguous when they're good than when they're bad. Here's a *simple-argument model*. When you see an argument, your credence that it's good should either increase or decrease. Value implies that it should increase when it's good and decrease when it's not, but 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, the other (blue) sees ones opposing s—then they predictably polarize (Figure 8; see §7.1 for details). Upshot: being exposed to different arguments might've rationally, predictably polarized us.

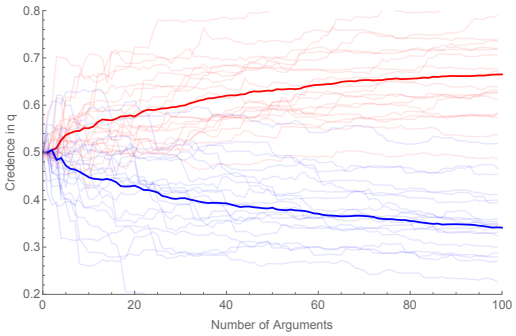


Figure 8: Red agents are presented with random argument models (from Figure 10) favoring s, and blue agents presented with models favoring $\neg s$. Thin lines are individuals; thick lines are averages.

But how does this simple-argument model fit with my discussion of selective scrutiny (§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 don't scrutinize), the simple model makes sense—with just a quick glance, it's 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 don't (see §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 *won't* find 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's 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).

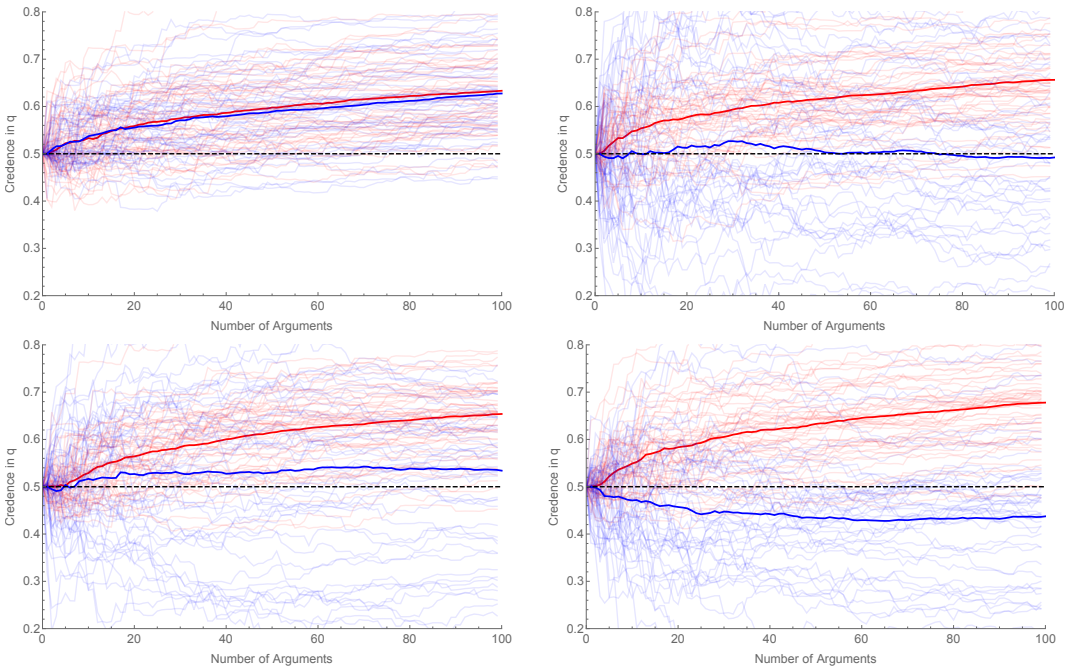


Figure 9: Two groups presented with arguments favoring q; red group never scrutinizes, blue group always does. **Top Left:** 0% chance of finding flaw if there is one; full blue polarization. **Top Right:** 100% chance of finding flaw if there is one; no blue polarization. **Bottom:** Middling chance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

Upshot: 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 *can't* 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 (§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 persuasive effects vary widely. There are.[55] 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.[56] And it predicts that manipulating how much people scrutinize will affect persuasion—with the biggest effects being on the evaluation of weak, congruent arguments (they'll be surprised to

[55]Petty 1994; Petty and Wegener 1998; Taber and Lodge 2006; Lundgren and Prislin 1998.
[56]Kahan et al. 2012; Kahan 2013; Kahan et al. 2017; Bail et al. 2018.

find flaws) and strong, incongruent ones (they'll be surprised *not* to find flaws). It does.[57]

Finally, this model may clarify the mixed findings on *selective exposure* (§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 don't (Sears and Freedman 1967; Whittlestone 2017). Why? One throughline: 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 otherwise (Frey 1986; Hart et al. 2009). The model predicts this. When arguments are high quality, scrutiny is useless (you won't 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're most able to find a flaw in—i.e. 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'll sketch the simple- and scrutinized-argument models (see Appendices C.2 and C.3).

The simple-argument model partitions possibilities into those where the argument is good (G) and those where it's bad (B). The posteriors are obtained by Jeffrey-shifting on the $\{G, B\}$ partition—increasing credence in the true possibility, 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).[58]

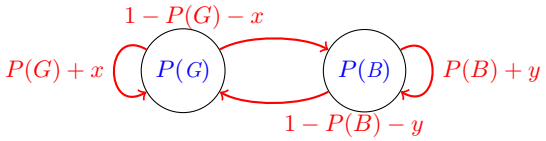


Figure 10: Schematic simple-argument model. If it's an argument for s, then $P(s|G) > P(s) > P(s|B)$; for $\neg s$, vice versa. Since bad arguments are more ambiguous than good ones, $y \leq x$.

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 vs. don't find a flaw, as diagrammed schematically in Figure 14 (page 60). There are many ways to parameterize these models; see §C.3 for details.

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

[58] E.g. 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 $\mathbb{E}_P(\widetilde{P}(G)) = 0.65 > 0.5 = P(G)$ and so $\mathbb{E}_P(\widetilde{P}(s)) = 0.53 > 0.5 = P(s)$.

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 weren't surprised by each others' opinions; most of them, we could've guessed. That said, his *reasons* surprised me. I didn't 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's bright and well-meaning. He's 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?") 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. Baffled, because I often find that they're *not* just conforming, or pigheaded, or dogmatic. Yet if they aren't, 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 weren't ambiguous, we should expect this (§2)—but it is, so we shouldn't (§3). Instead, we face ambiguity-asymmetries that make us better at recognizing evidence on one side than the other (§4). Wanting get to the truth, we take each individual step; by the end, the 'radically distorted' picture has become our own (§5). This theoretical idea has both experimental support (§4.2), and the potential to explain the mechanisms underlying confirmation bias (§6) and the group-polarization effect (§7).

Obviously this doesn't 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's incoherent to believe "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've formed through selective scrutiny, the political beliefs

we've formed through selective exposure, and the philosophical beliefs we've formed through searching for evidence favoring our positions.

That's the promise of this story. It allows us to admit our own predictability without undermining our own deeply-held commitments—and without disparaging those of others.[59]

[59]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, Branden Fitelson, Rachel Fraser, Jane Friedman, 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, Matthew 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.

Appendices

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.

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, i.e. 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'll 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're only concerned with one or two functions, I'll drop indices, using P, P_w and $\widetilde{P}$, $\widetilde{P}_w$. I'll also often enrich the structure with one or more (rigidly designated) probability functions, denoted π, δ, η,....

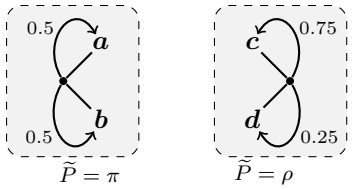
W represents the propositions in the frame, so for any $p, q \subseteq W$, p is true at w iff $w \in p$; $\neg p = W \setminus p$, $p \wedge q = p \cap q$, $p \to q = \neg p \cup q$ etc. All theorems are restricted to models with finite W—it's an open question how far they generalize.

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 $\pi \in \Delta(W)$: $[P(q) = t] := \{w \in W : P_w(q) = t\}$, $[P(q|r) \geq t] := \{w \in W : P_w(q|r) \geq t\}$, $[P = \pi] := \{w \in W : P_w = \pi\}$, etc.

Since W is finite, we can think of a probability function as an assignment of non-negative 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

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

the worlds $w_1, ..., w_n$ and write this information in a (square) *stochastic matrix* M in which $M_{ij} = P_{w_i}(w_j)$, i.e. 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.



$$\begin{matrix} \pi \\ \rho \end{matrix} \begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.75 & 0.25 \\ 0 & 0 & 0.75 & 0.25 \end{pmatrix}$$

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.

A.2 The Value of Evidence

When is an update from P to a 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's 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 simply a finite set of options $\mathcal{O}$. A **strategy** S is a way of choosing options based on $\widetilde{P}$'s probabilities, i.e. 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 $\mathcal{O}$ iff 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 $\widetilde{P}$ iff 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** $\widetilde{P}$ iff, 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: π values $\widetilde{P}$ iff 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 prior π to a description of the

[61] There's a subtlety here. As stated, P values $\widetilde{P}$ iff, at all worlds w, P_w prefers to let $\widetilde{P}$ (picked out descriptively) decide over *itself* (picked out rigidly), i.e. 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's the one used in Dorst et al. 2021, and whose formal properties are well-understood. However, every update I use in this paper is valuable (or, later on, valuable-with-respect-to-Q) in the latter sense as well, so the choice doesn't matter for our purposes.

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 it's 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 presented before and the other presented 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's a pair $\mathcal{O}_1$ and $\mathcal{O}_2$, 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 show that P_w values $\widetilde{P}$ iff 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. footnote 61), P values $\widetilde{P}$ iff 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 $X(w)$ (Schervish et al. 2014). Writing $A_X(\pi)$ to abbreviate $A_X(\mathbb{E}_\pi(X))$, say that A_X is **strictly proper** iff 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 $\widetilde{P}$ iff: 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. footnote 61), P values $\widetilde{P}$ iff each P_w expects $\widetilde{P}$ to have estimates at least as accurate as itself (P_w).

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's 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's a weakening of standard 'Relection-style' deference principles like Function Reflection (§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, it 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. footnote 61), P values $\widetilde{P}$ iff each P_w totally trusts $\widetilde{P}$.

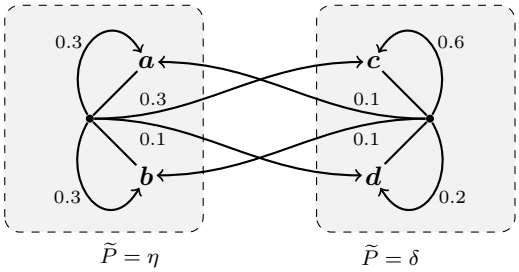
A.3 Ambiguity, Standard Bayesianism, (Im)possibility Theorems

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; Dorst 2019, 2020b.



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

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 $\delta(\widetilde{P} = \delta) = 0.8$.

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's 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:

Definition. $\langle W, P, \widetilde{P}\rangle$ is **Standard-Bayesian** iff 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's 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) $P_w = P_x$, i.e. 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: $\Pi(x) = \Pi(w)$, i.e. 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 (§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 $S_v = S_x$; and that S is recommended by $\widetilde{P}$ iff 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, 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

[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.

an arbitrary option $O \in \mathcal{O}$ and show that $\mathbb{E}_{P_w}(S) \geq \mathbb{E}_{P_w}(O)$:

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

(⇐:) 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 $\pi_1, ..., \pi_n$. Notice that $\Pi := \{[\widetilde{P} = \pi_1], ..., [\widetilde{P} = \pi_n]\}$ 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's 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 $v \in \Pi(w)$, $\widetilde{P}_v = \pi$. WLOG, suppose there is a q, t such that $\pi(q) > t > P_w(q \big| \Pi(w))$. We construct a decision problem that's a conditional bet on q given $\Pi(w)$ to show that P_w doesn't 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 $\widetilde{P}_x(\Pi(w)) = 1$ and $\widetilde{P}_x(q) = \pi(q) > t$, hence $\mathbb{E}_{\widetilde{P}_x}(B) > t(1-t) + (1-t)(-t) = 0 = \mathbb{E}_{\widetilde{P}_x}(N)$, 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's *below* t, it thinks this strategy is worse than simply taking N: $\mathbb{E}_{P_w}(S) = P_w(\neg\Pi(w)) \cdot 0 + P_w(\Pi(w)) \cdot \mathbb{E}_{P_w}(B \big| \Pi(w))$. Since $P_w(\Pi(w)) > 0$ (since $P_w(w) > 0$), this quantity is negative iff $\mathbb{E}_{P_w}(B \big| \Pi(w))$ is; and $\mathbb{E}_{P_w}(B \big| \Pi(w)) < t(1-t) + (1-t)(-t) = 0$, hence $\mathbb{E}_{P_w}(S) < 0 = \mathbb{E}_{P_w}(N)$. Value fails. □

Now our impossibility result: given No Ambiguity, Value and Reflection are equivalent;

thus, 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 $\widetilde{P}$.

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 iff Function Reflection holds.

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

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

(⇒:) For reductio suppose there's a π such that $P_w(\cdot|\widetilde{P} = \pi) \neq \pi$. WLOG, 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

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

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 $P_w(\widetilde{P} = \widetilde{P}_i) > 0$. WLOG, suppose $P_w(q|\widetilde{P} = \widetilde{P}_i) < t < \widetilde{P}_i(q)$. Let $\mathcal{O} = \{N, B\}$ were $N = 0$ 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, and 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 (page 36) would generate the frame in Figure 11 (page 34), 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, i.e. a vector in which entry i is $\pi(w_i)$. The **convex hull** of a set of such points $\pi_1, ...\pi_n$ is the set of points obtainable by averaging them: $CH\{\pi_1, ..., \pi_n\} = \{\delta : \exists \lambda_i \geq 0 \text{ and } \sum \lambda_i = 1 \text{ such that } \delta = \sum \lambda_i \pi_i\}$. Given a probability function δ, let $\boldsymbol{C_\delta} := \{\pi : \delta(\widetilde{P} = \pi) > 0\}$ be the set of Candidates that δ thinks $\widetilde{P}$ might be. Let $C^-_\delta := C_\delta - \{\delta\}$ the ones other than δ. Say that $\widetilde{P}_w$ is **modestly informed** iff it's an average of its informed self along with the other (uninformed) candidates, i.e. iff $\widetilde{P}_w$ 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 don't obey Reflection.

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

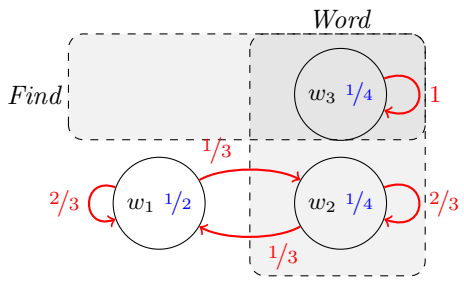
Proof. Let $\rho_1, ..., \rho_n$ be the potential realizations of $\widetilde{P}$, so $C_\pi = \{\rho_1, ..., \rho_n\}$. We know that each ρ_i is modestly informed, and that π is in their convex hull.

We begin by showing—following Samet 2000, Theorem 5—that, since once 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 $E_1, ..., E_k$, 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, i.e. where M is the transition matrix and ρ_i is thought of as the (row) vector with the $\rho_i(w_j)$ in column j, $\rho_i M = \rho_i$. By the Markov chain convergence theorem (e.g. Bertsekas and Tsitsiklis 2008, Ch. 7), each $E_1, ...E_k$ 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 is a ρ_i and q such that $\rho_i(q) \neq \mathbb{E}_{\rho_i}(\widetilde{P}(q))$. WLOG 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 $\rho_i(q) < \mathbb{E}_{\rho_i}(\widetilde{P}(q))$ iff $0 < \mathbb{E}_{\rho_i}(\widetilde{P}(q)) - \mathbb{E}_{\rho_i}(\mathbb{1}_q)$ iff $\mathbb{E}_{\rho_i}(\widetilde{P}(q) - \mathbb{1}_q) > 0$. 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 any other $\rho_j \neq \rho_i$ (there must be at least one other, since $\widetilde{P}$ is ambiguous). Now for any $\epsilon \in [0, 1]$, letting $\eta_\epsilon := (1 - \epsilon)\rho_i + \epsilon\rho_j$. Thinking of $\mathbb{E}_{\eta_\epsilon}(\widetilde{P}(q) - \mathbb{1}_q)$ as a function of ϵ, notice that this function is continuous and non-increasing 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. □

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 (page 39). 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: $\mathbb{E}_{H_w}(\widetilde{H}(Word)) \approx 0.583 > 0.5 = H_w(Word)$.

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 don't find it, and F the set where you find it. $Word = C \cup F$ is the proposition that there's 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 the prior has no higher-order uncertainty. Say the search is *bounded by conditioning* iff $\min_{n \in N} \widetilde{H}_n(Word) = H_w(Word|\neg F)$. Say that a search is *possibly ambiguous* iff you might be unsure of the rational posterior in *Word*, i.e. iff there is an x such that for all t: $\widetilde{H}_x(\widetilde{H}(Word) = t) < 1$. 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 $\mathbb{E}_H(\widetilde{H}(Word)) > H(Word)$.

Proof. Since $\widetilde{H}$ is possibly ambiguous, there is a $v \in W$ such that $\widetilde{H}_v(\widetilde{H}(Word) = t) < 1$ 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 (§A.2). Since for any $f \in F$, $\widetilde{H}_f(Word) = 1$, and also $\widetilde{H}_f(Word|\widetilde{H}(Word) \leq t) \leq t$, we must have that $\widetilde{H}_f(\widetilde{H}(Word) \leq t) = 0$ for all $t < 1$; in other words, $\widetilde{H}_f(\widetilde{H}(Word) = 1) = 1$. 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 there must be at least two values of $\widetilde{H}(Word)$ in $N \cup C$, so $\exists x \in N \cup C$ such that $\widetilde{H}_x(Word) \neq H_w(Word|\neg F)$. 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 $\widetilde{H}_y(\mathit{Word}) = t < H_w(\mathit{Word}|\neg F)$. Since this is lower than any $n \in N$, we have $[\widetilde{H}(\mathit{Word}) \leq t] \subseteq \mathit{Word}$, hence $H_w(\mathit{Word}|\widetilde{H}(\mathit{Word} \leq t) = 1 > t$, hence H_w doesn't obey Total Trust toward $\widetilde{H}$—contradiction.

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

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

□

A.5 Question-Relative Value

First we show that full Value is 'transitive', as discussed in §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 $P^2_x(P^3 = P^3_v) > 0$—for 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, 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 $\mathit{CH}(C^2_w) \subseteq \mathit{CH}(C^3_w)$. 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 iff $p = \bigcup_i q_i$ for $q_i \in Q$, i.e. iff 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* iff Q settles the value of O, i.e. for all w, w', if $Q(w) = Q(w')$, then $O(w) = O(w')$. Say that $\mathcal{O}_Q$ is a **decision about Q** iff each of its options is Q-measurable. Then: π *Q-values* $\widetilde{P}$ iff 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 decision problem $\mathcal{O}_Q$ 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's no fixed-option Q-book against $\langle P_w, \widetilde{P}\rangle$.

Proof. Suppose P_w Q-values $\widetilde{P}$, and 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, §3).

A.6 The Predictable Theorem

I'll 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 41, 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—for 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'd be no obstacle to iteration. Though a sensible (and perhaps realistic) route to polarization, this faces the concern that it's 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's a possibility proof. I conjecture that there are more-intuitive ways to get the same result.

I'll proceed in stages. First, I'll specify a model that iterates word-search tasks and consolidates higher-order uncertainty along the way. I'll then prove that each word-search update is fully valuable, while each consolidation update is valuable with respect to Q. I'll 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'll 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's 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 $i = 1, ..., n$, 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's completable but she doesn't find it, and n_i and n_i' are where it's 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 $W = X_1 \times ... \times X_n$ 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

$$N_i = \{w \in W : n_i \in w \text{ or } n_i' \in w\}$$
$$C_i = \{w \in W : c_i \in w \text{ or } c_i' \in w\}$$
$$F_i = \{w \in W : f_i \in w\};$$

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 $Q(x) = Q(y)$ iff 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 $U^i = \{U^i_n, U^i_c, U^i_f\}$ where

$$U^i_n = \{w \in W : n_i \in w \text{ or } c_i' \in w\}$$
$$U^i_c = \{w \in W : c_i \in w \text{ or } n_i' \in w\}$$
$$U^i_f = \{w \in W : f_i \in w\};$$

As we'll see, U^i_n is the set of worlds where H^i updated *as if* there was no completion (as if n_i) and U^i_c is that where H^i updated as if there was one (as if c_i).

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 $p \subseteq W$:

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

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'll 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, i.e. the set of propositions. For each world $w \in W$, let $g_w : \wp(W) \to W$ be a selection function which, 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 $x = \langle x_1, ..., x_n\rangle$ and $y = \langle y_1, ..., y_n\rangle$ which both are in p and have $Q(x) = Q(w) = Q(y)$, but y has a longer w-agreeing end-sequence ($x_n = w_n, ...$ but $x_{n-k} \neq w_{n-k}$, and $y_n = w_n, ..., y_{n-k} = w_{n-k}$), then $g_w(p) \neq x$.

Following Lewis 1976, for any probability function π, we let **π imaged on p**, $\boldsymbol{\pi(\cdot||p)}$, 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 not destroys it, so $\pi(\cdot||p)$ is always a probability function. As a result, note that for any $r \subseteq W$:

$$\begin{aligned}\pi(r||p) \;=\; \sum_{w \in r} \pi(w||p) \;&=\; \sum_{w \in r} \sum_{y \in W : g_y(p) = w} \pi(y) \\ &= \sum_{y \in W : g_y \in r} \pi(y)\end{aligned}$$

Machinery in place, we can now define the series of probability functions $H^0, \overline{H^0}, H^1, \overline{H^1}, ..., H^n, \overline{H^n}$ that represent Haleys 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's 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_i:

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

Moreover assume H^0_w treats the Q_i as mutually independent, thus for any $q_{i_1}, ..., q_{i_k}$ in $Q_{i_1}, ..., Q_{i_k}$ respectively, $H^0_w(q_{i_1}\&...\&q_{i_k}) = H^0_w(q_{i_1})H^0_w(q_{i_2}) \cdots H^0_w(q_{i_k})$. 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'll see, these consolidation-updates change her higher-order opinions (removing higher-order doubts) without changing her opinions about Q.

Finally, we define the regular (non-consolidation) 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 $H^{i+1}_w = \overline{H^i_w}[\frac{1}{3}, \frac{2}{3}, 0]_{i+1}$;
If $n_{i+1} \in w$ or $c'_{i+1} \in w$, then $H^{i+1}_w = \overline{H^i_w}[\frac{2}{3}, \frac{1}{3}, 0]_{i+1}$;

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

Lemma 5.1.1. (1) For each i and w: $\overline{H^i_w}$ is higher-order certain.
(2) Moreover, for $i > 1$, if $H^i_w(x) > 0$, then $\overline{H^{i-1}_w} = \overline{H^{i-1}_x}$.

Proof. (1) Suppose $\overline{H^i_w}(x) > 0$; we show that $\overline{H^i_x} = \overline{H^i_w}$. By definition, $\overline{H^i_w}(x) = H^i_w(x||H^i = H^i_w) > 0$. By the definition of imaging, $x \in [H^i = H^i_w]$, i.e. $H^i_x = H^i_w$. Thus $\overline{H^i_x} = H^i_x(\cdot||H^i = H^i_x) = H^i_w(\cdot||H^i = H^i_w) = \overline{H^i_w}$. Since x was arbitrary, $\overline{H^i_w}(\overline{H^i} = \overline{H^i_w}) = 1$.
(2) By definition H^i_w is obtained from $\overline{H^{i-1}_w}$ by Jeffrey-shifting in a way that preserves certainties, therefore if $H^i_w(x) > 0$ then $\overline{H^{i-1}_w}(x) > 0$, so by (1), $\overline{H^{i-1}_w} = \overline{H^{i-1}_x}$. □

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 $H^i_w(x) = 0$ and $\overline{H^i_w}(x) = 0$.

Proof. By induction. *Base case:* By construction, $H^0_w(U) = 0$, so $H^0_w(x) = 0$. Since $\overline{H^0_x} = H^0_x$, likewise for $\overline{H^0_x}$. *Induction case:* Supposing it holds for all w with $k < i$, we show it holds for i. Since $H^i_w = \overline{H^{i-1}_w}[a_1, a_2, a_3]_i$, and this doesn't raise any probabilities from 0, since (by induction) $\overline{H^{i-1}_w}(x) = 0$, likewise $H^i_w(x) = 0$. Now suppose, for reductio, $\overline{H^i_w}(x) > 0$. Thus there must be a y such that $H^i_w(y) > 0$ and $g_y(H^i = H^i_w) = x$. But since H^i_w didn't 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 $H^i_y = H^i_w$, then (by Strong Centering) $g_y(H^i = H^i_w) = y$, so this is impossible; hence $H^i_y \neq H^i_w$. Since $H^i_w(y) > 0$, and if $w \in f_i$ then H^i_w would be higher-order certain, it must be that either (i) $w \in U^i_c$ and $y \in U^i_n$, or (ii) $w \in U^i_n$ and $y \in U^i_c$. Since we must've had $\overline{H^{i-1}_w}(y) > 0$, by the inductive hypothesis we know either $c_i \in y$ or $n_i \in y$ (not $c'_i \in y$ nor $n'_i \in y$). So if (i), then $y' = \langle y_1, ..., n'_i ..., y_n\rangle$—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 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 $y = \langle y_1, ..., y_{i-1}, n_i, y_{i+1}..., y_n\rangle$. Consider the possibility $y' = \langle y_1, ..., y_{i-1}, n'_i, y_{i+1}, ..., y_n\rangle$, which is the same as y except that it swaps n'_i for n_i. By construction, $Q(y') = Q(y)$, and also $\overline{H^{i-1}_{y'}} = \overline{H^{i-1}_y} = \overline{H^{i-1}_x}$, 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 don't 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:

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

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. By induction. *Base case*: 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, so 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:

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

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

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

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,

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

Now notice that:

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

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 don't 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 $\mathcal{O}_Q$ iff $\overline{H^i}$ 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 bullet-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 $\mathit{Heads}_i = F_i \cup C_i$, $H^0(\mathit{Heads}_i) = \frac{1}{2}$, 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. Since each sequence of heads and tails is equally likely, and as $n \to \infty$ the proportion of sequences with more than half heads tends to $1/2$, the first part of the third bullet-point follows: $H^0(h) \approx \frac{1}{2}$.

To establish the second part of the third bullet-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 $\mathit{Heads}_i = F_i \cup C_i$, we have, for all w, i, H^0_w assigns probability 1 to:

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

Proof. First focus on $H^i(\mathit{Heads}_i)$, 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_i. And 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 i.i.d. with $H^0(F_i) = H^0(C_i) = \frac{1}{4}$, by the law of large numbers, as $n \to \infty$, $H^0\big(|I_F| \approx \frac{n}{4}\ \&\ |I_C| \approx \frac{n}{4}\ \&\ |I_N| \approx \frac{n}{2}\big) \to 1$. We want to show what follows if this obtains, so suppose it does: $|I_F| \approx \frac{n}{4}\ \&\ |I_C| \approx \frac{n}{4}\ \&\ |I_N| \approx \frac{n}{2}$. 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 $i \in I_F$, $H^n(\mathit{Heads}_i) = 1$;
For all $i \in I_C$, H^n treats Heads_i as i.i.d. with $H^n(\mathit{Heads}_i) = \frac{2}{3}$; and
For all $i \in I_N$, H^n treats Heads_i as i.i.d. with $H^n(\mathit{Heads}_i) = \frac{1}{3}$.

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:

$$H^n\big(\sum_{i\in I_F} \frac{\mathbb{1}_{\mathit{Heads}_i}}{|I_F|} = 1\big) = 1 \qquad (\alpha)$$

$$H^n\big(\sum_{i\in I_C} \frac{\mathbb{1}_{\mathit{Heads}_i}}{|I_C|} \approx \tfrac{2}{3}\big) \to 1 \qquad (\beta)$$

$$H^n\big(\sum_{i\in I_N} \frac{\mathbb{1}_{\mathit{Heads}_i}}{|I_N|} \approx \tfrac{1}{3}\big) \to 1 \qquad (\gamma)$$

Note that that $\frac{|I_F|}{n}\sum_{i\in I_F} \frac{\mathbb{1}_{\mathit{Heads}_i}}{|I_F|} + \frac{|I_C|}{n}\sum_{i\in I_C} \frac{\mathbb{1}_{\mathit{Heads}_i}}{|I_C|} + \frac{|I_N|}{n}\sum_{i\in I_N} \frac{\mathbb{1}_{\mathit{Heads}_i}}{|I_N|} = \sum_{i=1}^{n} \frac{\mathbb{1}_{\mathit{Heads}_i}}{n}$ is the proportion of all flips that land heads. Combining the fact that $|I_F| \approx \frac{n}{4}\ \&\ |I_C| \approx \frac{n}{4}\ \&\ |I_N| \approx \frac{n}{2}$, with (α), (β), and (γ), we have, as $n \to \infty$:

$$H^n\left(\sum_{i=1}^{n} \frac{\mathbb{1}_{\mathit{Heads}_i}}{n} \ \approx\ \tfrac{1}{4}(1) + \tfrac{1}{4}(\tfrac{2}{3}) + \tfrac{1}{2}(\tfrac{1}{3}) \ =\ \tfrac{7}{12}\right) \to 1$$

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

$$H^n\Big(\sum_{i=1}^{n} \frac{\mathbb{1}_{\mathit{Heads}_i}}{n} > \tfrac{1}{2}\Big) = H^n(h) \approx 1$$

Since this follows from $|I_F| \approx \frac{n}{4}\ \&\ |I_C| \approx \frac{n}{4}\ \&\ |I_N| \approx \frac{n}{2}$, 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 bullet-point, Lemma 5.1.8 establishes the second, and the reasoning on page 49 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 $\overline{H^n}\Big(h\Big|\overline{T^n}(h) \approx 0\Big) \approx 1$ and $\overline{T^n}\Big(h\Big|\overline{H^n}(h) \approx 1\Big) \approx 0$ (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 $Q^t_i = \{F^i_i, C^t_i, N^t_i\}$, where $F^t_i \cup C^t_i = N_i$ and $N^t_i = F_i \cup C_i$. By doing so, we create a model in which both H^0 and T^0 are (correctly) certain that:

- $F_i \& N^t_i \to \Big(\overline{H^n}(\mathit{Heads}_i) = 1 \;\&\; \overline{T^n}(\mathit{Heads}_i) = \frac{2}{3}\Big)$
- $C_i \& N^t_i \to \Big(\overline{H^n}(\mathit{Heads}_i) = \frac{2}{3} \;\&\; \overline{T^n}(\mathit{Heads}_i) = \frac{2}{3}\Big)$
- $N_i \& C^t_i \to \Big(\overline{H^n}(\mathit{Heads}_i) = \frac{1}{3} \;\&\; \overline{T^n}(\mathit{Heads}_i) = \frac{1}{3}\Big)$
- $N_i \& F^t_i \to \Big(\overline{H^n}(\mathit{Heads}_i) = \frac{1}{3} \;\&\; \overline{T^n}(\mathit{Heads}_i) = 0\Big)$

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^t_i, assigning e.g.:

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

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

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

and that

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

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 $I_{F^t} = \{i : Q^t_i = F^t_i\}$, 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 $H^n(T^n(\mathit{Heads}_i) = \frac{2}{3}) = 1$.
   Therefore, $H^n(\sum_{i \in I_F} \frac{T^n(\mathit{Heads}_i)}{|I_F|} = \frac{2}{3}) = 1$

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

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

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

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

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

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

$$H^n\Big(\sum_{i=1}^{n} \frac{T^n(Heads_i)}{n} \;\approx\; \tfrac{1}{4}(\tfrac{2}{3}) + \tfrac{1}{4}(\tfrac{1}{2}) + \tfrac{1}{2}(\tfrac{1}{3}) \;=\; \tfrac{11}{24} \;\approx\; 0.458\Big) \to 1$$

Therefore, H^n gets arbitrarily confident that T^n's average confidence in $Heads_i$ is less than $\frac{1}{2}$: $H^n(\sum_{i=1}^{n} \frac{T^n(Heads_i)}{n} < \frac{1}{2}) \to 1$. And since H^n is certain that T^n treats the $Heads_i$ independently, it follows that $H^n(T^n(\sum_{i=1}^{n} \frac{1_{Heads_i}}{n} > \frac{1}{2}) \approx 0) \to 1$, i.e. 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. □

B Experimental Details

Appendix B discusses the experiment from §4.2.

250 English-speakers were recruited through Prolific (107 F/139 M/4 Other; mean age = 27.06).[64] The hypothesis was that subjects would polarize more when given (potentially ambiguous) word-searches than (unambiguous) draws from an urn. Subjects were randomly assigned to conditions in a 2 × 2 design that independently manipulated valence (Headsers

[64]Pre-registration: https://aspredicted.org/8jg3e.pdf. I made two mistakes at the pre-registration 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 pre-registered tests. The conclusions are the same.

vs. Tailsers) and ambiguity (Ambiguous vs. Unambiguous). I'll abbreviate the groups '**A-Hsers**'; '**A-Tsers**'; '**U-Hsers**', and '**U-Tsers**'. Each was told they'd be given evidence about a series of four independent, fair coin tosses (in fact, the tosses were pseudo-randomized 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 (§4), and given three examples ('P_A_ET' [planet], 'CO_R_D', [uncompletable] and '_E_RT' [heart]). The A-Hsers were told they'd 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 (§4.2). For U-Hsers, if the coin landed heads then the urn contained 1 black marble and 1 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 pre-task question was an attention-check, wherein they were instructed to move the slider to 50% since it was a new 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 3 tasks in which a non-black marble was drawn, and 1 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. (*Note:* 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 §4.2, page 16.)

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.

[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 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.

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.

Here are the means and standard deviations of credence-in-heads for each group:

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)

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 $t(219) = 3.58$, $p < 0.001$, $d = 0.341$. One-sided paired t-test for Prediction 2 indicated that A-Tser posteriors were lower than A-Tser priors, with $t(191) = 5.90$, $p < 0.001$, $d = 0.604$. One-sided independent samples t-test for Prediction 3 indicated that A-Hser posteriors were higher than A-Tser posteriors, with $t(410) = 7.07$, $p < 0.001$, $d = 0.699$. One-sided independent samples t-test for Prediction 4 failed to indicate that A-Hser posteriors were higher than U-Hser posteriors, with $t(441) = 1.15$, $p = 0.125$, $d = 0.107$. One-sided independent samples t-test for Prediction 5 indicated that A-Tser posteriors were below U-Tser posteriors, with $t(393) = 4.07$, $p < 0.001$, $d = 0.398$.

Prediction 6 was (due to my oversight) handled poorly at pre-registration—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've* 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) by 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 pre-registration 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 re-run 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 (§4.2, page 16).

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 amongst those who *didn't* 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 wasn't (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 didn't, and then compare the average confidence of those who were vs. weren't looking at a completable string. A one-sided independent samples t-test *failed* to indicate that the confidence of those who weren't ($M = 39.00$, $SD = 19.90$) was lower than that of those who were ($M = 42.03$, $SD = 21.37$), with $t(243) = 1.11$, $p = 0.13$, one-sided. However, a substantial proportion of people who *claimed* to have found a word did not have the extreme confidence that they should've 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's a completion'—though, to be clear, this change was *not* pre-registered—the prediction is confirmed: amongst 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 $t(309) = 2.77$, $p = 0.003$, $d = 0.32$.

Finally, two further (not pre-registered—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 (A-group) who didn't find a completion, or (U-group) who didn't see a black marble—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; Tailsers will be

[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; etc.

[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, $p < 0.001$.

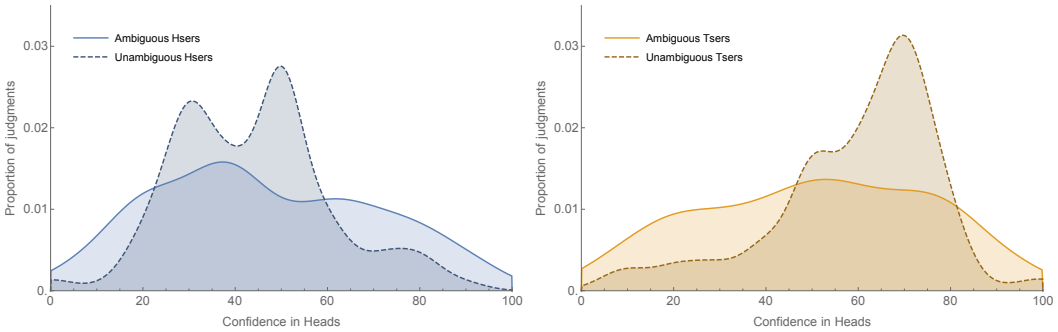


Figure 13: Density plots of confidence in Heads, given weak evidence.

better at recognizing tails-cases. As can be seen in Table 1, this is what we find. When presented with uncompletable strings (*Tails* cases for Headsers; *Heads* cases for Tailsers), neither group's average posterior moved significantly from their priors of 50%; but 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 their average priors vs. posteriors is: for Headsers, $0.5(1-0.5037)^2+0.5(0-0.5034)^2 = 0.250$ vs. $0.5(1-0.6742)^2+0.5(0-0.4773)^2 = 0.167$; for Tailsers, 0.253 vs. 0.166.

	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

Table 1: Ambiguous condition, mean prior and posterior confidence in *Heads*, by cases.
* = *not* significantly different from 50%.

C Computational Details

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

C.1 Cognitive-Search Models (§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 3 classes, depending on whether the agent finds a flaw (F), there is a flaw that they don't 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 which 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's no flaw, all you learn is that you didn't 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 don't find, that determines the rational credence; you learn that you didn't find one; and you assign at least as much credence to this possibility as you would if that were all you learned.)

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 6 worlds (two in each of F, C, and N) plus a prior over them. (In Mathematica, we represent this with a 7-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 it's 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. $P(\textit{Find}|\textit{Flaw})$) 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 (page 25).

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, 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 (page 25).

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 q-supporting studies, vice versa. This parameter `findGap` can range from 0 (where there's 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 (page 25).

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 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.

C.2 Argument Models (§7)

This subsection explains the simple-argument models used in §7 (without scrutiny). You know that you're 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's good, it's rational to increase your confidence in q; if it's bad, it's 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's bad. Letting P_w be the known prior and $\widetilde{P}$ is 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's 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 $\widetilde{P}_b(B) = P_w(B) + \epsilon'$.
  (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 4 worlds (two in each of G and B) plus a prior over them. (In Mathematica, we represent this with a *5*-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 $\neg q$, i.e. 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're 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 $[\texttt{favGBound}, 1]$) and upper-bound the probability that a blue-member's is ($P_w(G)$ drawn from $[0, 1 - \texttt{favGBound}]$). The code 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; 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 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` grows and `hardenShift` shrinks, the amount and rate of polarization grows. All runs resulted in polarization.

C.3 Argument-Scrutiny Models (§7)

This subsection explains how to combine the simple-argument models of §7 with the cognitive-search models in §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 doesn't, the model remains the same and she updates as in §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 doesn't find it (the search is Completable). When the argument is good (G, left), there is no flaw ($N = G$).

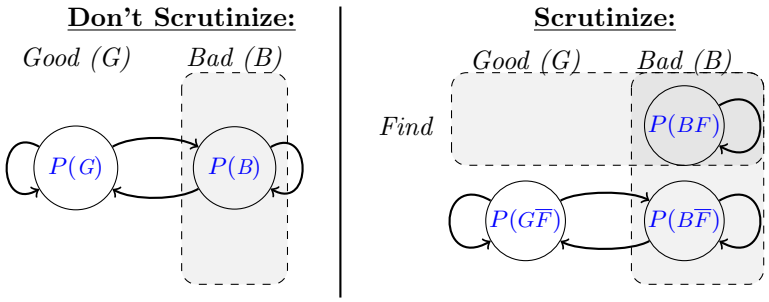


Figure 14: Schematic model of the choice of whether to scrutinize an argument.

Precisely, given an argument model as described in §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's bad—scrutinizing it generates a cognitive-search model with the partition $\{F, C, N\}$ fixing the posterior $\widetilde{P}$ as specified in §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's a flaw that you don't find, your credence that there is should be at least as great as it should be if you didn't scrutinize and updated your beliefs accordingly, and then conditioned on the claim that you wouldn't 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, i.e. 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 §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, and so agents who scrutinize polarize as much and in the same direction as those who don't—as seen in the top left of Figure 9 on page 29.

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 don't, 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 don't—as seen in the bottom right of Figure 9 on page 29.

Robustness. Recall that pro agents in this simulation are identical to those from the main simulation of §C.2, meaning we have estimates for their mean posteriors with these parameters at 0.650 with 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, with a 95% confidence interval of $[0.633, 0.658]$, indicating that scrutinizing agents shift to a comparable degree to those who don't scrutinize. In version (2), the mean posterior was 0.503, with a 95% confidence interval of $[0.474, 0.533]$, indicating that agents do not predictably shift from their priors of 0.5. In version (3), the mean posterior was 0.551, with a 95% confidence interval of $[0.530, 0.573]$,

confirming that such scrutiny dampens polarization. In version (4), the mean posterior was 0.463, with a 95% confidence interval of $[0.442, 0.483]$, confirming that such scrutiny reverses the direction of polarization.

References

Acemoglu, Daron and Wolitzky, Alexander, 2014. 'Cycles of conflict: An economic model'. *American Economic Review*, 104(4):1350–1367.

Achen, Christopher H and Bartels, Larry M, 2017. *Democracy for realists: Why elections do not produce responsive government*, volume 4. Princeton University Press.

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

Andreoni, James and Mylovanov, Tymofiy, 2012. 'Diverging opinions'. *American Economic Journal: Microeconomics*, 4(1):209–232.

Angere, Staffan and Olsson, Erik J, 2017. 'Publish late, publish rarely!: Network density and group performance in scientific communication'. *Scientific collaboration and collective knowledge*, 34–62.

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):176–199.

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

Aronowitz, Sara, 2020. 'Exploring by Believing'. *The Philosophical Review*, To Appear.

Aumann, R, 1976. 'Agreeing to Disagree'. *The Annals of Statistics*, 4:1236–1239.

Austerweil, Joseph L and Griffiths, Thomas L, 2011. 'Seeking Confirmation Is Rational for Deterministic Hypotheses'. *Cognitive Science*, 35:499–526.

Bail, Christopher A, Argyle, Lisa P, Brown, Taylor W, Bumpus, John P, Chen, Haohan, Hunzaker, M B Fallin, Lee, Jaemin, Mann, Marcus, Merhout, Friedolin, and Volfovsky, Alexander, 2018. 'Exposure to opposing views on social media can increase political polarization'. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221.

Baliga, Sandeep, Hanany, Eran, and Klibanoff, Peter, 2013. 'Polarization and Ambiguity †'. *The American Economic Review*, 103(2006264):3071–3083.

Baron, Robert S., Hoppe, Sieg I., Kao, Chuan Feng, Brunsman, Bethany, Linneweh, Barbara, and Rogers, Diane, 1996. 'Social corroboration and opinion extremity'. *Journal of Experimental Social Psychology*, 32(6):537–560.

Baumgaertner, Bert O., Tyson, Rebecca T., and Krone, Stephen M., 2016. 'Opinion strength influences the spatial dynamics of opinion formation'. *Educational Research*, 40(4):207–218.

Benoît, Jean Pierre and Dubra, Juan, 2019. 'Apparent Bias: What Does Attitude Polarization Show?' *International Economic Review*, 60(4):1675–1703.

Bertsekas, Dmitri P and Tsitsiklis, John N, 2008. *Introduction to Probability*. Athena Scientific, second edition.

Blackwell, David, 1953. 'Equivalent Comparisons of Experiments'. *The Annals of Mathematical Statistics*, 24(2):265–272.

Bowen, T. Renee, Dmitriev, Danil, and Galperti, Simone, 2021. 'Learning from Shared News: When Abundant Information Leads to Belief Polarization'. *SSRN Electronic Journal*, 1–112.

Boxell, Levi, Gentzkow, Matthew, and Shapiro, Jesse, 2020. 'Cross-Country Trends in Affective Polarization'. *National Bureau of Economic Research*, (June).

Bradley, Seamus and Steele, Katie, 2016. 'Can free evidence be bad? Value of information for the imprecise probabilist'. *Philosophy of Science*, 83(1):1–28.

Bregman, Rutger, 2017. *Utopia for realists: And how we can get there*. Bloomsbury Publishing.

Brennan, Jason, 2016. *Against democracy*. Princeton University Press.

Brier, Glenn W, 1950. 'Verification of forecasts expressed in terms of probability'. *Monthly weather review*, 78(1):1–3.

Briggs, R., 2009. 'Distorted Reflection'. *Philosophical Review*, 118(1):59–85.

Brown, Jacob R and Enos, Ryan D, 2021. 'The measurement of partisan sorting for 180 million voters'. *Nature Human Behaviour*, 1–11.

Brownstein, Ronald, 2016. 'How the Election Revealed the Divide Between City and Country'. *The Atlantic*.

Burnstein, Eugene and Vinokur, Amiram, 1977. 'Persuasive argumentation and social comparison as determinants of attitude polarization'. *Journal of Experimental Social Psychology*, 13(4):315–332.

Callahan, Laura Frances, 2019. 'Epistemic Existentialism'. *Episteme*, (2019):1–16.

Camerer, Colin and Weber, Martin, 1992. 'Recent developments in modeling preferences: Uncertainty and ambiguity'. *Journal of Risk and Uncertainty*, 5(4):325–370.

Campbell-Moore, Catrin, 2016. *Self-Referential Probability*. Ph.D. thesis.

Cariani, Fabrizio and Rips, Lance J., 2017. 'Conditionals, Context, and the Suppression Effect'. *Cognitive Science*, 41(3):540–589.

Carmichael, Chloe, 2017. 'Political Polarization Is A Psychology Problem'.

Carr, Jennifer Rose, 2020. 'Imprecise Evidence without Imprecise Credences'. *Philosophical Studies*, 177(9):2735–2758.

Christensen, David, 2010. 'Higher-Order Evidence'. *Philosophy and Phenomenological Research*, 81(1):185–215.

Cohen, G.A., 2000. *If You're an Egalitarian, How Come You're So Rich?* 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(5):808.

Cohen, L. Jonathan, 1981. 'Can human irrationality be experimentally demonstrated?' *Behavioral and Brain Sciences*, 4(3):317–331.

Cook, J. Thomas, 1987. 'Deciding to Believe without Self-Deception'. *Journal of Philosophy*, 84(8):441–446.

Cook, John and Lewandowsky, Stephan, 2016. 'Rational Irrationality: Modeling Climate Change Belief Polarization Using Bayesian Networks'. *Topics in Cognitive Science*, 8(1):160–179.

Corner, Adam, Harris, Adam, and Hahn, Ulrike, 2010. 'Conservatism in belief revision and participant skepticism'. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32.

Corner, Adam, Whitmarsh, Lorraine, and Xenias, Dimitrios, 2012. 'Uncertainty, scepticism and attitudes towards climate change: biased assimilation and attitude polarisation'. *Climatic change*, 114(3):463–478.

Crupi, Vincenzo, Tentori, Katya, and Lombardi, Luigi, 2009. 'Pseudodiagnosticity Revisited'. *Psychological Review*, 116(4):971–985.

Cushman, Fiery, 2020. 'Rationalization is rational'. *Behavioral and Brain Sciences*, 43.

Dallmann, Justin, 2017. 'When Obstinacy is a Better Policy'. *Philosophers' Imprint*, 17.

Das, Nilanjan, 2019. 'Accuracy and ur-prior conditionalization'. *The Review of Symbolic Logic*, 12(1):62–96.

———, 2020a. 'Externalism and Exploitability'. *Philosophy and Phenomenological Research*, To Appear.

———, 2020b. 'The Value of Biased Information'. *The British Journal for the Philosophy of Science*, To Appear.

———, 2022. 'Credal Imprecision and the Value of Evidence'. *Noûs*, To appear.

De Cruz, Helen, 2017. 'Religious disagreement: A study among academic philosophers'. *Episteme*, 14(1):71–87.

de Finetti, Bruno, 1977. 'Probabilities of probabilities: A real problem or a misunderstanding'. *New Developments in the Applications of Bayesian methods*, 1–10.

DeMarzo, Peter M, Vayanos, Dimitri, and Zwiebel, Jeffrey, 2003. 'Persuasion bias, social influence, and unidimensional opinions'. *The Quarterly journal of economics*, 118(3):909–968.

Ditto, Peter H and Lopez, David F, 1992. 'Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions.' *Journal of personality and social psychology*, 63(4):568.

Dixit, Avinash K and Weibull, Jörgen W, 2007. 'Political Polarization'. *Proceedings of the National Academy of Sciences of the United States of America*, 104(2):7351–7356.

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

———, 2020a. 'Evidence: A Guide for the Uncertain'. *Philosophy and Phenomenological Research*, 100(3):586–632.

———, 2020b. 'Higher-Order Evidence'. In Maria Lasonen-Aarnio and Clayton Littlejohn, eds., *The Routledge Handbook for the Philosophy of Evidence*. Routledge.

Dorst, Kevin, Levinstein, Benjamin, Salow, Bernhard, Husic, Brooke E., and Fitelson, Branden, 2021. 'Deference Done Better'. *Philosophical Perspectives*, To appear.

Downing, James W., Judd, Charles M., and Brauer, Markus, 1992. 'Effects of repeated expressions on attitude extremity.' *Journal of Personality and Social Psychology*, 63(1):17–29.

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

Edwards, Ward, 1982. 'Conservatism in Human Information Processing'. *Judgment under Uncertainty: Heuristics and Biases*, 359–369.

Elga, Adam, 2007. 'Reflection and Disagreement'. *Noûs*, 41(3):478–502.

———, 2013. 'The puzzle of the unmarked clock and the new rational reflection principle'. *Philosophical Studies*, 164(1):127–139.

Elga, Adam and Rayo, Agustín, 2020. 'Fragmentation and logical omniscience'. *Noûs*, To Appear.

Ellsberg, Daniel, 1961. 'Risk, Ambiguity, and the Savage Axioms'. *Quarterly Journal of Economics*, 75(4):643–669.

Evans, J. St B T, Barston, Julie L., and Pollard, Paul, 1983. 'On the conflict between logic and belief in syllogistic reasoning'. *Memory & Cognition*, 11(3):295–306.

Feeney, Aidan, Evans, Jonathan St B T, and Clibbens, John, 2000. 'Background beliefs and evidence interpretation'. *Thinking & reasoning*, 6(2):97–124.

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

Finkel, Eli J., Bail, Christopher A., Cikara, Mina, Ditto, Peter H., Iyengar, Shanto, Klar, Samara, Mason, Lilliana, McGrath, Mary C., Nyhan, Brendan, Rand, David G., Skitka, Linda J., Tucker, Joshua A., Van Bavel, Jay J., Wang, Cynthia S., and Druckman, James N., 2020. 'Political sectarianism in America'. *Science*, 370(6516):533–536.

Fischer, Peter, Jonas, Eva, Frey, Dieter, and Schulz-Hardt, Stefan, 2005. 'Selective exposure to information: The impact of information limits'. *European Journal of social psychology*, 35(4):469–492.

Fitzpatrick, Anne R and Eagly, Alice H, 1981. 'Anticipatory belief polarization as a function of the expertise of a discussion partner'. *Personality and Social Psychology Bulletin*, 7(4):636–642.

Flache, Andreas and Macy, Michael W, 2011. 'Local convergence and global diversity: From interpersonal to social influence'. *Journal of Conflict Resolution*, 55(6):970–995.

Fraser, Rachel, 2021. 'Mushy Akrasia'. *Philosophy and Phenomenological Research*, To Appear.

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

Fryer, Roland G., Harms, Philipp, and Jackson, Matthew O., 2019. 'Updating Beliefs when Evidence is Open to Interpretation: Implications for Bias and Polarization'. *Journal of the European Economic Association*, 17(5):1470–1501.

Gaifman, Haim, 1988. 'A Theory of Higher Order Probabilities'. In Brian Skyrms and William L Harper, eds., *Causation, Chance, and Credence*, volume 1, 191–219. Kluwer.

Gallow, J. Dmitri, 2021. 'Updating for Externalists'. *Noûs*, 55(3):487–516.

Geanakoplos, John, 1989. 'Game Theory Without Partitions, and Applications to Speculation and Consensus'. Cowles Fou.

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

Gigerenzer, Gerd, 1991. 'How to make cognitive illusions disappear: Beyond "heuristics and biases"'. *European review of social psychology*, 2(1):83–115.

———, 2018. 'The Bias Bias in Behavioral Economics'. *Review of Behavioral Economics*, 5(3-4):303–336.

Gilovich, Thomas, 1983. 'Biased evaluation and persistence in gambling.' *Journal of personality and social psychology*, 44(6):1110—-1126.

———, 1991. *How We Know What Isn't So*. Free Press.

Glaser, Markus and Weber, Martin, 2010. 'Overconfidence'. *Behavioral finance: Investors, corporations, and markets*, 241–258.

Good, I J, 1967. 'On the Principle of Total Evidence'. *The British Journal for the Philosophy of Science*, 17(4):319–321.

Gopnik, Alison, 1996. 'The Scientist as Child'. *Philosophy of Science*, 63(December):485–514.

———, 2012. 'Scientific thinking in young children: Theoretical advances, empirical research, and policy implications'. *Science*, 337(6102):1623–1627.

———, 2020. 'Childhood as a solution to explore–exploit tensions'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1803).

Greaves, Hilary and Wallace, David, 2006. 'Justifying conditionalisation : conditionalisation maximizes expected epistemic utility Introduction : Justifying conditionalisation'. 1–23.

Griffiths, Thomas L., Chater, Nick, Norris, Dennis, and Pouget, Alexandre, 2012. 'How the bayesians got their beliefs (and what those beliefs actually are): Comment on bowers and davis (2012)'. *Psychological Bulletin*, 138(3):415–422.

Griffiths, Thomas L., Lieder, Falk, and Goodman, Noah D., 2015. 'Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic'. *Topics in Cognitive Science*, 7(2):217–229.

Grönlund, Kimmo, Herne, Kaisa, and Setälä, Maija, 2015. 'Does enclave deliberation polarize opinions?' *Political Behavior*, 37(4):995–1020.

Hahn, Ulrike and Harris, Adam J.L., 2014. 'What Does It Mean to be Biased. Motivated Reasoning and Rationality.' In *Psychology of Learning and Motivation - Advances in Research and Theory*, volume 61, 41–102.

Haidt, Jonathan, 2012. *The righteous mind: Why good people are divided by politics and religion*. Vintage.

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*.

Hamblin, Charles L, 1976. 'Questions in montague english'. In *Montague grammar*, 247–259. Elsevier.

Hart, William, Albarracín, Dolores, Eagly, Alice H, Brechan, Inge, Lindberg, Matthew J, and Merrill, Lisa, 2009. 'Feeling validated versus being correct: a meta-analysis of selective exposure to information.' *Psychological bulletin*, 135(4):555.

Harvey, Nigel, 1997. 'Confidence in judgment'. *Trends in cognitive sciences*, 1(2):78–82.

Hastie, Reid and Dawes, Robyn M, 2009. *Rational choice in an uncertain world: The psychology of judgment and decision making*. Sage Publications.

Hedden, Brian, 2015. 'Options and Diachronic Tragedy'. *Philosophy and Phenomenological Research*, 90(2):423–451.

Hegselmann, Rainer and Krause, Ulrich, 2002. 'Opinion dynamics and bounded confidence: Models, analysis and simulation'. *Jasss*, 5(3).

Henderson, Leah and Gebharter, Alexander, 2021. 'The role of source reliability in belief polarisation'. *Synthese*, 1–23.

Hild, Matthias, 1998. 'Auto-epistemology and updating'. *Philosophical Studies*, 92(3):321–361.

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

Horowitz, Sophie, 2014. 'Epistemic Akrasia'. *Noûs*, 48(4):718–744.

———, 2019. 'Predictably Misleading Evidence'. In Mattias Skipper and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidence: New Essays*, 105–123. Oxford University Press.

Huttegger, Simon M, 2013. 'In defense of reflection'. *Philosophy of Science*, 80(3):413–433.

———, 2014. 'Learning experiences and the value of knowledge'. *Philosophical Studies*, 171(2):279–288.

Isaacs, Yoaav and Russell, Jeffrey Sanford, 2022. 'Updating Without Evidence'. *Noûs*, (March):1–33.

Isenberg, Daniel J., 1986. 'Group Polarization. A Critical Review and Meta-Analysis'. *Journal of Personality and Social Psychology*, 50(6):1141–1151.

Iyengar, Shanto, Lelkes, Yphtach, Levendusky, Matthew, Malhotra, Neil, and Westwood, Sean J., 2019. 'The origins and consequences of affective polarization in the United States'. *Annual Review of Political Science*, 22:129–146.

Iyengar, Shanto, Sood, Gaurav, and Lelkes, Yphtach, 2012. 'Affect, not ideology: A social identity perspective on polarization'. *Public Opinion Quarterly*, 76(3):405–431.

Jackson, Elizabeth, 2021. 'A Defense of Intrapersonal Belief Permissivism'. *Episteme*, 18(2):313–327.

Jamieson, Kathleen Hall and Cappella, Joseph N, 2008. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press.

Jeffrey, Richard C, 1990. *The logic of decision*. University of Chicago press.

Jern, Alan, Chang, Kai Min K., and Kemp, Charles, 2014. 'Belief polarization is not always irrational'. *Psychological Review*, 121(2):206–224.

Johnson, Dominic D P, 2009. *Overconfidence and war*. Harvard University Press.

Jost, John T., Glaser, Jack, Kruglanski, Arie W., and Sulloway, Frank J., 2003. 'Political Conservatism as Motivated Social Cognition'. *Psychological Bulletin*, 129(3):339–375.

Joyce, James M., 2010. 'A Defense of Imprecise Credences in Inference and Decision Making'. *Philosophical Perspectives*, 24:281–323.

Kadane, Joseph B, Schervish, Mark, and Seidenfeld, Teddy, 2008. 'Is ignorance bliss?' *The Journal of Philosophy*, 105(1):5–36.

Kadane, Joseph B., Schervish, Mark J., and Seidenfeld, Teddy, 1996. 'Reasoning to a foregone conclusion'. *Journal of the American Statistical Association*, 91(435):1228–1235.
Kahan, D M, 2018. 'Why smart people are vulnerable to putting tribe before truth'. *Scientific American*.
Kahan, Dan M., 2013. 'Ideology, motivated reasoning, and cognitive reflection'. *Judgment and Decision Making*, 8(4):407–424.
Kahan, Dan M., Peters, Ellen, Dawson, Erica Cantrell, and Slovic, Paul, 2017. 'Motivated numeracy and enlightened self-government'. *Behavioural Public Policy*, 1:54–86.
Kahan, Dan M., Peters, Ellen, Wittlin, Maggie, Slovic, Paul, Ouellette, Lisa Larrimore, Braman, Donald, and Mandel, Gregory, 2012. 'The polarizing impact of science literacy and numeracy on perceived climate change risks'. *Nature Climate Change*, 2(10):732–735.
Kahneman, Daniel, 2011. *Thinking Fast and Slow*. Farrar, Straus, and Giroux.
Kahneman, Daniel, Slovic, Paul, and Tversky, Amos, eds., 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press.
Kahneman, Daniel and Tversky, Amos, 1996. 'On the reality of cognitive illusions.'
Kamenica, Emir and Gentzkow, Matthew, 2011. 'Bayesian persuasion'. *American Economic Review*, 101(6):2590–2615.
Kelly, Thomas, 2008. 'Disagreement, Dogmatism, and Belief Polarization'. *The Journal of Philosophy*, 105(10):611–633.
Kinney, David and Bright, Liam Kofi, 2021. 'Risk aversion and elite-group ignorance'. *Philosophy and Phenomenological Research*, 1–23.
Klaczynski, Paul A and Narasimham, Gayathri, 1998. 'Development of scientific reasoning biases: Cognitive versus ego-protective explanations.' *Developmental Psychology*, 34(1):175.
Klein, Ezra, 2014. 'How politics makes us stupid'. *Vox*, 1–14.
———, 2020. *Why We're Polarized*. Profile Books.
Klibanoff, Peter, Marinacci, Massimo, and Mukerji, Sujoy, 2005. 'A smooth model of decision making under ambiguity'. *Econometrica*, 73(6):1849–1892.
Koerth, Maggie, 2019. 'Why Partisans Look At The Same Evidence On Ukraine And See Wildly Different Things'. *FiveThirtyEight*.
Koriat, Asher, Lichtenstein, Sarah, and Fischhoff, Baruch, 1980. 'Journal of Experimental Psychology : Human Learning and Memory Reasons for Confidence'. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2):107–118.
Kossinets, Gueorgi and Watts, Duncan J., 2009. 'Origins of Homophily in an Evolving Social Network'. *American Journal of Sociology*, 115(2):405–450.
Krueger, Joachim I and Massey, Adam L, 2009. 'A rational reconstruction of misbehavior'. *Social Cognition*, 27(5):786–812.
Kuhn, Deanna and Lao, Joseph, 1996. 'Effects of Evidence on Atittudes: is Polarization the Norm?' *Psycholohical Science*, 7(2):115–120.
Kuhn, Thomas, 1962. *The Structure of Scientific Revolutions*. The University of Chicago Press.
Kunda, Ziva, 1990. 'The case for motivated reasoning'. *Psychological Bulletin*, 108(3):480–498.
Lakoff, George, 1997. *Moral politics: What conservatives know that liberals don't*. University of Chicago Press.
Lasonen-Aarnio, Maria, 2013. 'Disagreement and Evidential Attenuation'. *Nous*, 47(4):767–794.
———, 2014. 'Higher-order evidence and the limits of defeat'. *Philosophy and Phenomenological Research*, 8(2):314–345.
———, 2015. 'New Rational Reflection and Internalism about Rationality'. In Tamar Szabó Gendler and John Hawthorne, eds., *Oxford Studies in Epistemology*, volume 5, 145–171. Oxford University Press.
Lazer, David, Baum, Matthew, Benkler, Jochai, Berinsky, Adam, Greenhill, Kelly, Metzger, Miriam, Nyhan, Brendan, Pennycook, G., Rothschild, David, Sunstein, Cass, Thorson, Emily, Watts, Duncan, and Zittrain, Jonathan, 2018. 'The science of fake news'. *Science*, 359(6380):1094–1096.
Le Mens, Gaël and Denrell, Jerker, 2011. 'Rational Learning and Information Sampling: On the " Naivety" Assumption in Sampling Explanations of Judgment Biases'. *Psychological Review*, 118(2):379–392.
Lederman, Harey, 2015. 'People with Common Priors Can Agree to Disagree'. *The Review of Symbolic Logic*, 8(1):11–45.
Levi, Isaac, 1974. 'On indeterminate probabilities'. *The Journal of Philosophy*, 71(13):391–418.
Levinstein, B. A., 2022. 'Accuracy, Deference, and Chance'. *Philosophical Review*, To appear.
Levinstein, Benjamin and Spencer, Jack, 2022. 'Bigger, Badder Bugs'.
Lewis, David, 1976. 'Probabilities of Conditionals and Conditional Probabilities'. *The Philosophical Review*, 85(3):297–315.
———, 1980. 'A subjectivist's guide to objective chance'. In Richard C Jeffrey, ed., *Studies in Inductive Logic and Probability*, volume 2, 263–293. University of California Press.
Liberman, Akiva and Chaiken, Shelly, 1992. 'Defensive processing of personally relevant health messages'. *Personality and Social Psychology Bulletin*, 18(6):669–679.
Lichtenstein, Sarah, Fischhoff, Baruch, and Phillips, Lawrence D., 1982. 'Calibration of probabilities: The state of the art to 1980'. In Daniel Kahneman, Paul Slovic, and Amos Tversky, eds., *Judgment under Uncertainty*, 306–334. Cambridge University Press.
Lieder, Falk and Griffiths, Thomas L., 2019. 'Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources'. *Behavioral and Brain Sciences*.
Lilienfeld, Scott O, Ammirati, Rachel, and Landfield, Kristin, 2009. 'Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare?' *Perspectives on psychological science*, 4(4):390–398.
Liu, Cheng-Hong, 2017. 'Evaluating arguments during instigations of defence motivation and accuracy motivation'. *British Journal of Psychology*, 108(2):296–317.

Loh, Isaac and Phelan, Gregory, 2019. 'Dimensionality and disagreement: Asymptotic belief divergence in response to common information'. *International Economic Review*, 60(4):1861–1876.
Lord, Charles G, Lepper, Mark R, and Preston, Elizabeth, 1984. 'Considering the opposite: a corrective strategy for social judgment.' *Journal of personality and social psychology*, 47(6):1231.
Lord, Charles G., Ross, Lee, and Lepper, Mark R., 1979. 'Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence'. *Journal of Personality and Social Psychology*, 37(11):2098–2109.
Lord, Charles G and Taylor, Cheryl A, 2009. 'Biased assimilation: Effects of assumptions and expectations on the interpretation of new evidence'. *Social and Personality Psychology Compass*, 3(5):827–841.
Lottes, Ilsa L and Kuriloff, Peter J, 1994. 'The impact of college experience on political and social attitudes'. *Sex Roles*, 31(1):31–54.
Lundgren, Sharon R and Prislin, Radmila, 1998. 'Motivated cognitive processing and attitude change'. *Personality and Social Psychology Bulletin*, 24(7):715–726.
Mandelbaum, Eric, 2018. 'Troubles with Bayesianism: An introduction to the psychological immune system'. *Mind & Language*, 1–17.
Mäs, Michael and Flache, Andreas, 2013. 'Differentiation without distancing. explaining bi-polarization of opinions without negative influence'. *PLoS ONE*, 8(11).
Mason, Lilliana, 2018. *Uncivil agreement: How politics became our identity*. 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(3):395–409.
McKenzie, Craig R M, 2004. 'Framing effects in inference tasks—and why they are normatively defensible'. *Memory & cognition*, 32(6):874–885.
Mcpherson, Miller, Smith-lovin, Lynn, and Cook, James M, 2001. 'Birds of a Feather: Homophily in Social Networks'. *Annual Review of Sociology*, 27:415–444.
McWilliams, Emily C., 2021. 'Evidentialism and belief polarization'. *Synthese*, 198(8):7165–7196.
Mercier, Hugo, 2017. 'Confirmation bias—Myside bias.' *Cognitive illusions: Intriguing phenomena in thinking, judgment and memory, 2nd ed.*, 99–114.
———, 2020. *Not born yesterday*. Princeton University Press.
Mercier, Hugo and Sperber, Dan, 2011. 'Why do humans reason ? Arguments for an argumentative theory'. 57–111.
———, 2017. *The enigma of reason*. Harvard University Press.
Miller, Arthur G., McHoskey, John W., Bane, Cynthia M., and Dowd, Timothy G., 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(4):561–574.
Mills, Charles W., 2007. 'White ignorance'. *Race and Epistemologies of Ignorance*, 13–38.
Moore, Don A, Carter, Ashli B, and Yang, Heather H J, 2015. 'Organizational Behavior and Human Decision Processes Wide of the mark : Evidence on the underlying causes of overprecision in judgment'. 131:110–120.
Moss, Sarah, 2018. *Probabilistic knowledge*. Oxford University Press.
Munro, Geoffrey D and Ditto, Peter H, 1997. 'Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information'. *Personality and Social Psychology Bulletin*, 23(6):636–653.
Murray, Mark, 2018. 'Poll: 58 percent say gun ownership increases safety'.
Myers, David G., 2012. *Social Psychology*. McGraw-Hill Education.
Myers, David G. and Lamm, Helmut, 1976. 'The group polarization phenomenon'. *Psychological Bulletin*, 83(4):602–627.
Nguyen, C. Thi, 2018. 'Escape the echo chamber'. *Aeon*.
Nickerson, Raymond S., 1998. 'Confirmation bias: A ubiquitous phenomenon in many guises.' *Review of General Psychology*, 2(2):175–220.
Nielsen, Michael and Stewart, Rush T, 2021. 'Persistent Disagreement and Polarization in a Bayesian Setting'. *British Journal for the Philosophy of Science*, 72(1):51–78.
Nimark, Kristoffer P. and Sundaresan, Savitar, 2019. 'Inattention and belief polarization'. *Journal of Economic Theory*, 180:203–228.
Oaksford, Mike and Chater, Nick, 1994. 'A Rational Analysis of the Selection Task as Optimal Data Selection'. *Psychological Review*, 101(4):608–631.
———, 1998. *Rational models of cognition*. Oxford University Press Oxford.
O'Connor, Cailin and Weatherall, James Owen, 2018. 'Scientific Polarization'. *European Journal for Philosophy of Science*, 8(3):855–875.
Oddie, Graham, 1997. 'Conditionalization, Cogency, and Cognitive Value'. *The British Journal for the Philosophy of Science*, 48(4):533–541.
Olsson, Erik J, 2013. 'A Bayesian simulation model of group deliberation and polarization'. In *Bayesian argumentation*, 113–133. Springer.
Ortoleva, Pietro and Snowberg, Erik, 2015. 'Overconfidence in political behavior'. *American Economic Review*, 105(2):504–535.
Pallavicini, Josefine, Hallsson, Bjørn, and Kappel, Klemens, 2018. *Polarization in groups of Bayesian agents*. Springer Netherlands.
Pariser, Eli, 2012. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin Books.
Pascal, Blaise, 1660. 'From Pensées'. *Pensées*, 1.
Pennycook, Gordon and Rand, David G., 2019. 'Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning'. *Cognition*, 188(September 2017):39–50.
Peterson, CAMERON R. and Beach, LEE R., 1967. 'Man As an Intuitive Statistician'. *Psychological Bulletin*, 68(1):29–46.
Pettigrew, Richard and Titelbaum, Michael G, 2014. 'Deference Done Right'. *Philosopher's Imprint*, 14(35):1–19.

Petty, RE, 1994. 'Two routes to persuasion: State of the art'. *International perspectives on psychological ...*, 2:1–15.

Petty, Richard E. and Wegener, Duane T., 1998. 'Attitude change: Multiple roles for persuasion variables'. *The handbook of social psychology*, 323–390.

Plous, Scott, 1991. 'Biases in the assimilation of technological breakdowns: Do accidents make us safer?' *Journal of Applied Social Psychology*, 21(13):1058–1082.

Podgorski, Abelard, 2016. 'Dynamic permissivism'. *Philosophical Studies*, 173(7):1923–1939.

Proietti, Carlo, 2017. 'The dynamics of group polarization'. In A. Baltag, J. Seligman, and T. Yamada, eds., *International Workshop on Logic, Rationality and Interaction*, volume 10455, 195–208.

Pronin, Emily, 2008. 'How We See Ourselves and How We See Others'. *Science*, 320(16):1177–1180.

Rabin, Matthew and Schrag, Joel, 1999. 'First impressions matter: a model of confirmatory bias'. *Quarterly Journal of Economics*, (February):37–82.

Ramsey, F. P., 1990. 'Weight or the value of knowledge'. *British Journal for the Philosophy of Science*, 41(1):1–4.

Risen, Jane and Gilovich, Thomas, 2007. 'Informal Logical Fallacies.' In R.J. Sternberg, H.L. Roediger, and D.F. Halpern, eds., *Critical thinking in psychology*, 110–130. Cambridge University Press.

Rizzo, Mario J and Whitman, Glen, 2019. *Escaping paternalism: Rationality, behavioral economics, and public policy*. Cambridge University Press.

Robson, David, 2018. 'The myth of the online echo chamber'.

Rogers, Kayleigh, 2020. 'Americans Were Primed To Believe The Current Onslaught Of Disinformation'.

Ross, Lee, 2012. 'Reflections on Biased Assimilation and Belief Polarization'. *Critical Review*, 24(2):233–245.

Salow, Bernhard, 2018. 'The Externalist's Guide to Fishing for Compliments'. *Mind*, 127(507):691–728.

———, 2019. 'Elusive externalism'. *Mind*, 128(510):397–427.

Samet, Dov, 1999. 'Bayesianism without learning'. *Research in Economics*, 53:227–242.

———, 2000. 'Quantified Beliefs and Believed Quantities'. *Journal of Economic Theory*, 95(2):169–185.

Savage, Leonard J, 1954. *The Foundations of Statistics*. Wiley Publications in Statistics.

Schervish, M. J., Seidenfeld, T., and Kadane, J.B., 2004. 'Stopping to Reflect'. *The Journal of Philosophy*, 101(6):315–322.

Schervish, M. J., Seidenfeld, Teddy, and Kadane, J. B., 2014. 'Dominating countably many forecasts'. *Annals of Statistics*, 42(2):728–756.

Schoenfield, Miriam, 2012. 'Chilling out on epistemic rationality'. *Philosophical Studies*, 158(2):197–219.

———, 2014. 'Permission to Believe: Why Permissivism is True and What it Tells Us About Irrelevant Influences on Belief'. *Nous*, 48(2):193–218.

———, 2015. 'A Dilemma for Calibrationism'. *Philosophy and Phenomenological Research*, 91(2):425–455.

———, 2017a. 'Conditionalization Does Not (In General) Maximize Expected Accuracy'. *Mind*, 126(504):1155–1187.

———, 2017b. 'Meditations on Beliefs Formed Arbitrarily'. *Oxford Studies in Epistemology*, To appear.

———, 2018. 'An Accuracy Based Approach to Higher Order Evidence'. *Philosophy and Phenomenological Research*, 96(3):690–715.

Schuette, Robert A and Fazio, Russell H, 1995. 'Attitude accessibility and motivation as determinants of biased processing: A test of the MODE model'. *Personality and Social Psychology Bulletin*, 21(7):704–710.

Sears, David O and Freedman, Jonathan L, 1967. 'Selective exposure to information: A critical review'. *Public Opinion Quarterly*, 31(2):194–213.

Seidenfeld, Teddy and Wasserman, Larry, 1993. 'Dilation for Sets of Probabilities'. *The Annals of Statistics*, 21(3):1139–1154.

Siegel, Susanna, 2021. 'The Problem of Culturally Normal Beliefs'. In Robin Celikates, Sally Haslanger, and Jason Stanley, eds., *Ideology: New Essays*, To Appear. Oxford University Press.

Simpson, Robert Mark, 2017. 'Permissivism and the arbitrariness objection'. *Episteme*, 14(4):519–538.

Singer, Daniel J, Bramson, Aaron, Grim, Patrick, Holman, Bennett, Jung, Jiin, Kovaka, Karen, Ranginani, Anika, and Berger, William J, 2019. 'Rational social and political polarization'. *Philosophical Studies*, 176(9):2243–2267.

Skyrms, Brian, 1990. 'The Value of Knowledge'. *Minnesota Studies in the Philosophy of Science*, 14:245–266.

Sliwa, Paulina and Horowitz, Sophie, 2015. 'Respecting *All* the Evidence'. *Philosophical Studies*, 172(11):2835–2858.

Solomon, Miriam, 1992. 'Scientific Rationality and Human Reasoning'. *Philosophy of Science*, 59(3):439–455.

Srinivasan, Amia, 2015. 'Are We Luminous?' *Philosophy and Phenomenological Research*, 90(2):294–319.

Stafford, Tom, 2015. *For argument's sake: evidence that reason can change minds*. Smashwords Edition.

———, 2020. 'Evidence for the rationalisation phenomenon is exaggerated'. *Behavioral and Brain Sciences*, 43.

Stalnaker, Robert, 1968. 'A Theory of Conditionals'. In Nicholas Rescher, ed., *Studies in Logical Theory*, 98–112. Oxford University Press.

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

Stangor, Charles and Walinga, Jennifer, 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*. MIT Press.

Stone, Daniel F., 2019. '"Unmotivated bias" and partisan hostility: Empirical evidence'. *Journal of Behavioral and Experimental Economics*, 79(August):12–26.

———, 2020. 'Just a Big Misunderstanding? Bias and Bayesian Affective Polarization'. *International Economic Review*, 61(1):189–217.

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

Sunstein, Cass R, 2000. 'Deliberative trouble? Why groups go to extremes'. *The Yale Law Journal*, 110(1).

———, 2017. *#Republic: Divided democracy in the age of social media*. Princeton University Press.

Sutherland, Stuart, 1992. *Irrationality: The enemy within*. Constable and Company.

Taber, Charles S, Cann, Damon, and Kucsova, Simona, 2009. 'The motivated processing of political arguments'. *Political Behavior*, 31(2):137–155.

Taber, Charles S and Lodge, Milton, 2006. 'Motivated Skepticism in the Evaluation of Political Beliefs'. *American Journal of Political Science*, 50(3):755–769.

Talisse, Robert B, 2019. *Overdoing democracy: Why we must put politics in its place*. Oxford University Press.

Teller, Paul, 1973. 'Conditionalization and observation'. *Synthese*, 26(2):218–258.

Tenenbaum, Joshua B and Griffiths, Thomas L, 2006. 'Optimal Predictions in Everyday Cognition'. *Psychological Science*, 17(9):767–773.

Tenenbaum, Joshua B, Kemp, Charles, Griffiths, Thomas L, and Goodman, Noah D, 2011. 'How to grow a mind: Statistics, structure, and abstraction'. *science*, 331(6022):1279–1285.

Tesser, Abraham, Martin, Leonard, and Mendolia, Marilyn, 1995. 'The impact of thought on attitude extremity and attitude-behavior consistency.'

Thaler, Richard H., 2015. *Misbehaving: The Making of Behavioural Economics*. Penguin.

Todd, Peter M, Hills, Thomas T, Robbins, Trevor W, and Lupp, Julia, 2012. *Cognitive search: Evolution, algorithms, and the brain*, volume 9. MIT press.

Toplak, Maggie E and Stanovich, Keith E, 2003. 'Associations between myside bias on an informal reasoning task and amount of post-secondary education'. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 17(7):851–860.

Tversky, Amos and Kahneman, Daniel, 1974. 'Judgment under uncertainty: Heuristics and biases'. *Science*, 185(4157):1124–1131.

van der Maas, Han L J, Dalege, Jonas, and Waldorp, Lourens, 2020. 'The polarization within and across individuals: the hierarchical Ising opinion model'. *Journal of Complex Networks*, 8(2).

van Ditmarsch, Hans, Halpern, Joseph Y, van der Hoek, Wiebe, and Kooi, Barteld, 2015. *Handbook of Epistemic Logic*. College Publications.

van Prooijen, Jan-Willem and Krouwel, André P M, 2019. 'Psychological Features of Extreme Political Ideologies'. *Current Directions in Psychological Science*, 28(2):159–163.

Vavova, Katia, 2018. 'Irrelevant Influences'. *Philosophy and Phenomenological Research*, 96(1):134–152.

Vinokur, Amiram and Burstein, Eugene, 1974. 'Effects of partially shared persuasive arguments on group-induced shifts: A group-problem-solving approach.' *Journal of Personality and Social Psychology*, 29(3):305–315.

Vosoughi, Soroush, Roy, Deb, and Aral, Sinan, 2018. 'The spread of true and false news online'. *Science*, 359(6380):1146–1151.

Weatherall, James Owen and O'Connor, Cailin, 2020. 'Endogenous epistemic factionalization'. *Synthese*, 1–23.

Weisberg, Jonathan, 2007. 'Conditionalization, reflection, and self-knowledge'. *Philosophical Studies*, 135(2):179–197.

White, Roger, 2009. 'On Treating Oneself and Others as Thermometers'. *Episteme*, 6(03):233–250.

———, 2010. 'You Just Believe that Because...' *Philosophical Perspectives*, 24:573–615.

Whittlestone, Jess, 2017. 'The importance of making assumptions : why confirmation is not necessarily a bias'. (July).

Wilkinson, Will, 2018. 'The Density Divide: Urbanization, Polarization, and Populist Backlash'. Technical report, The Niskanen Center.

Williams, Daniel, 2021. 'Socially adaptive belief'. *Mind and Language*, 36(3):333–354.

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

———, 2008. 'Why Epistemology Cannot be Operationalized'. In Quentin Smith, ed., *Epistemology: New Essays*, 277–300. Oxford University Press.

———, 2014. 'Very Improbable Knowing'. *Erkenntnis*, 79(5):971–999.

———, 2019. 'Evidence of Evidence in Epistemic Logic'. In Mattias Skipper and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidence: New Essays*, 265–297. Oxford University Press.

Wilson, Andrea, 2014. 'Bounded Memory and Biases in Information Processing'. *Econometrica*, 82(6):2257–2294.

Wolfe, Christopher R and Britt, M Anne, 2008. 'The locus of the myside bias in written argumentation'. *Thinking & reasoning*, 14(1):1–27.

Wolfers, Justin, 2014. 'How Confirmation Bias Can Lead to a Spinning of Wheels'.

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*, 240–264. Routledge.

Ye, Ru, 2019. 'The Arbitrariness Objection against Permissivism'. *Episteme*, (2019):1–20.

Zendejas Medina, Pablo, 2022. 'Just As Planned: Bayesianism, Externalism, and Plan Coherence'. *Philosophers' Imprint*, To Appear.

Zhang, Snow and Meehan, Alexander, 2022. 'Bayes Is Back'.

Zollman, Kevin J S, 2021. *Network Epistemology: How Our Social Connections Shape Knowledge*.