Rational Polarization

Kevin Dorst
kevindorst@pitt.edu
September 2021

Main Text: 14,351 (or 11,855*) words
Technical Appendices: 12,967 words

Abstract
Predictable polarization is everywhere. We can often predict the different directions that people’s opinions—including our own—will shift over time. Empirical studies suggest that this is so whenever evidence is ambiguous, a fact that’s often thought to demonstrate human bias or irrationality. It doesn’t. Bayesians will predictably polarize iff their evidence is ambiguous. And ours often is: the process of cognitive search—searching a cognitively-accessible space for an item of a particular profile—yields ambiguous evidence that can predictably polarize beliefs, despite being expected to make them more accurate. In principle, a series of such rational updates could lead to polarization that is predictable, profound, and persistent. Thus it’s theoretically possible that rational mechanisms drive predictable polarization. It’s also empirically plausible. I present a novel experiment confirming the polarizing effect of cognitive search, and then use models and simulations to show how such ambiguous evidence can help explain two of the core causes of polarization: confirmation bias and the group polarization effect.

Keywords: polarization, disagreement, irrelevant influences, ambiguous evidence, confirmation bias, group polarization effect, value of evidence, deference principles, diachronic tragedy

Main Text:
1. A Standard Story ...............................................................2
2. Predictable Polarization Requires Ambiguity .................................4
3. Cognitive Search Polarizes People .............................................10
4. Cognitive Search Explains Confirmation Bias ..............................21
5. Ambiguity Explains Group Polarization .....................................27
6. A Better Story .................................................................32

Appendices:
A. Analytical Details ..................................................................35
B. Experimental Details ............................................................48
C. Computational Details ..........................................................53
References ................................................................................58

*Informal track—the formal subsections can be skipped without loss of continuity.
1. A STANDARD STORY

1 A Standard Story

It was a bench.

"Why was this outside your window?", my parents asked pointedly.

I wasn’t quick on my feet. I had nothing. So I told them everything: “I’ve been sneaking
out”, I began...

That was when it stopped. That was one in a long line of lucky breaks—for me.

But not for them. 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 went off to universities in liberal cities, many
of them were stuck in our conservative hometown. While I was having my eyes opened by
higher-education, some of them were fighting for their lives.

I was a near thing, but they made it. Obviously so did I.

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 worries about far-right militias, and who about Black Lives Matter? 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.

There’s nothing surprising about that. Most societies display local conformity with global
disunity: people’s attitudes are predictable given their social group, despite varying widely
across such groups (Mcpherson et al. 2001; Brown and Enos 2021). Failed institutions cre-
ate libertarians, universities produce progressives, and the underside of Columbia, Missouri
shapes you in a different way than the academic centers of Cambridge, Massachusetts do.

The result is polarization: people who start out with similar beliefs but set out on dif-
ferent trajectories—like me and my old friends—often come to have disagreements that are
profound, persistent, and predictable (Cohen 2000; Sunstein 2009).

My question is why.

The standard story: predictable polarization is due to epistemic irrationality—the fact
that people’s beliefs are insufficiently constrained by accuracy or evidence.¹ Instead, people
glom onto the beliefs of their peers,² confirm and entrench those beliefs,³ become wildly
overconfident in them,⁴ and so on. When combined with the informational traps of the
modern internet,⁵ this story offers a simple explanation of both the ubiquity of polarization
generally, as well as its rise in recent decades (Iyengar et al. 2019; Boxell et al. 2020).

¹ E.g. Sutherland 1992; Lakoff 1997; Milks 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;

² Myers and Lammi 1976; Isenberg 1986; Baron et al. 1996; Sunstein 2000, 2009; Mcpherson et al. 2001;
Cohen 2003; Pronin 2008; Iyengar et al. 2012; Miß and Flache 2013, Myers 2012, Ch. 8, Baumgaertner et al.
2016; Brownstein 2016; Mason 2018; Wilkinson 2018; Talisse 2019; Siegel 2021.

³ Lord et al. 1979; Frey 1986; Kunda 1990; Nickerson 1998; Jost et al. 2003; Fine 2005; Taber and Lodge

⁴ Lichtenstein et al. 1982; Harvey 1997; Johnson 2009; Glaser and Weber 2010; Moore et al. 2015; Ortoleva

⁵ Van Heuvelen 2007; Jamieson and Cappella 2008; Nyhan and Reifler 2010; Pariser 2012; Nguyen 2018;
Sunstein 2017; Vosoughi et al. 2018.
But notice that this is a story in two parts. First is a series of empirical observations: people predictably conform to the beliefs of their peers, wind up very confident, etc. Second is a normative interpretation: if people were rationally constrained by the truth, they wouldn’t predictably conform to their peers, become so confident, etc. While the empirical observations are well-supported and widely-accepted, the normative interpretation is not. My claim is that it’s false. We can—we should—accept the empirical observations while denying the normative interpretation.

This isn’t easy. Standard Bayesian assumptions imply that the type of predictable polarization we observe empirically must be irrational (§2). In particular, variations in evidential standards (Schoenfield 2014; Callahan 2019), background beliefs (Jern et al. 2014; Benoît and Dubra 2019), or distributions of trust (O’Connor and Weatherall 2018; Pallavicini et al. 2018) do not rationalize the process. This is the best argument for the standard story.

My response will come in three steps. First, I’ll locate a flaw in the standard assumptions that opens the theoretical possibility of rational, predictable polarization (§2). Second, I’ll point to a process that shows that this possibility is not merely theoretical (§3). Third, I’ll argue that such processes play a central role in the psychological mechanisms that drive real-world polarization (§§4–5). In more detail:

The flaw is that the standard assumptions neglect the fact that some types of evidence are *more ambiguous*—harder to know how to interpret—than others (§2). As a result, it’s possible to face ambiguity-asymmetries that make it easier to recognize evidence pointing in one direction than another. In such situations, a bias is present—but it is a bias in the evidence, and thus is one that even rational people would be susceptible to.

The process I’ll point to is a *cognitive search* (Todd et al. 2012)—searching a cognitively-accessible space for an item of a particular profile. I’ll show how such processes generate ambiguity-asymmetries that can lead rational people who start out with the same beliefs to polarize in ways that are predictable, profound, and persistent (§3). This contradicts standard approaches to disagreement, which imply that at most one side could rationally stand their ground. Thus, in principle, my friends and I could start out agreeing and all predict—with arbitrarily high confidence—that rationally responding to our (differing, ambiguous) evidence would lead me to become liberal, them to become conservative, and none of us to budge upon realizing what had happened. This all despite the fact that at each stage in our journeys, we are each rational to expect that updating on our (differing, ambiguous) evidence will make our (Bayesian) beliefs more accurate. Predictable polarization amounts to an epistemic form of a *diachronic tragedy* (Hedden 2015): doing the best we can at each stage to get accurate beliefs leads, in the long run, to those beliefs predictably splitting apart.

The argument I’ll make is that cognitive search and ambiguous evidence play an important role in explaining real-world polarization. I’ll present a novel experiment confirming their polarizing effects (§3.2), and then use models and simulations to show how they can drive two core causes of polarization: confirmation bias (§4) and the group-polarization effect (§5). The upshot: if we always searched for flaws in arguments, and if we had no

---


2. PREDICTABLE POLARIZATION REQUIRES AMBIGUITY

self-doubt about our ability to do so, our evidence would be unambiguous and standard Bayesian models would be right. But since we can’t scrutinize everything—and since we should have self-doubts—our evidence is systematically ambiguous in a way that leads to predictable, rational polarization. That’ll complete my (rational) story.

Although this story is built on a series of technical results, those results aren’t needed to assess the story itself. Thus I’ve partitioned the paper: those interested in the story but not the technicalities may skip the formal subsections and footnotes without loss of continuity.

But why do I think we need a story like this? Because of what happened to my old friends—and to me. Because when I try to apply the standard story to us, it doesn’t fit. Long before we polarized, we could predict that we would: that I’d come to believe that (say) owning a gun makes you less safe, and they’d come to believe the opposite. Indeed, my higher-education-based polarization was more predictable than theirs. Perhaps back then I could blame my future, predictable change-of-mind on irrationality. But now? If I continue to do so, I’ll akratically believe “Owning a gun makes you less safe, but I believe that for irrational reasons.” That’s not a rational option (Horowitz 2014). Of course, I could instead give up my belief about guns—but, by parallel reasoning, I’d have to give up pretty much all of my political beliefs. That’s also not a rational option. It seems I can’t continue to think that my polarization was caused by irrationality; so if I continue to believe that irrationality is the cause, I must believe that they were the irrational ones. But that too seems wrong. My beliefs have changed even more predictably than theirs have. They were brighter (and quicker) than I was. 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.

My aim is to tell a story that does justice to all of these features of polarization—its predictability, its profundity, its persistence, its arbitrariness—without resorting to implausible claims about the rational superiority of those who, I think, have gotten things right.

2 Predictable Polarization Requires Ambiguity

This project began in an unlikely place. I was working on rational self-doubt—scenarios in which rational people are unsure whether they’re rational. If you have complicated evidence or peers who disagree with you about how to interpret it, you might be unsure whether you’re thinking rationally. My aim was to build well-behaved models of such scenarios.

Then Trump got elected.

Like many in liberal bubbles, I was shocked. And like some from conservative towns, I was embarrassed to realize that I knew no Trump voters. Evidently I had buried my roots deeper than I’d thought. What had happened to them? What had happened to me?

I became fascinated by polarization, so I set my previous project to one side to start (what I thought would be) an entirely new one. Pundits blamed polarization on irrationality. But my old friends hadn’t seemed irrational back then. And my beliefs—which had polarized more predictably than theirs—did not seem irrational now. So I wondered. To what extent

---

could predictable polarization—in politics, but also academia (Becher and Trowler 2001), science (Kuhn 1962), and elsewhere—result from epistemically rational causes, i.e. individuals trying their best to get to the truth?

That might seem silly. Hasn’t psychology and behavioral economics taught us that people are systematically irrational? Many think so. But many don’t. They critique the replicability and interpretations of such empirical work, and appeal to a burgeoning research program that uses rational models of cognition to explain the mind’s ability to routinely perform intractably-complex tasks that the most advanced computers cannot.

But perhaps this misses the point. Just look at what people believe! The world is full of climate-deniers and vaccine-skeptics (or, from a different reader: police-abolitionists and socialists). Surely such beliefs must be irrational, right?

Not obviously. Those people have incredibly different background beliefs, networks of trust, and life experiences (evidence) than you do—and it’s well-known that such differences can lead approximately- or even ideally-rational people to persistently disagree.

Take me and one of my old friends, Dan. Consider a moment soon after we’d parted ways—when our opinions had not yet moved substantially, but our trajectories were clear. I was a few months into studying political science at a liberal university; Dan was a few months into a job at a bar in rural Missouri. Let ‘s’ rigidly designate the particular probability function that I had then. Meanwhile, let ‘P’ be a description for the probability function (whatever it is) that it’d be rational for me to have in 2021. Likewise with ‘δ’ and ‘D’ for Dan. (I’ll be neutral on the interpretation of “rational” opinions like the opinions that, were I to start thinking about my evidence properly, I’d adopt.)

It’s straightforward to build rational models in which someone is in a position to expect that our different trajectories of (perhaps misleading) evidence will lead our rational opinions to polarize. For example, letting s be the claim that owning a gun makes you safer, they might expect it to be rational for my confidence in s to fall, and Dan’s to rise:

**Expected Polarization:** Someone’s expectation is that my confidence in s should drop, and Dan’s should rise.

For some specified probability function ρ: \( E_\rho (P(s) - \pi(s)) < 0 \) and \( E_\rho (D(s) - \delta(s)) > 0 \).

---


12I’ll focus on polarization in opinions, rather than feelings or behaviors. Though some contest whether people are polarizing in policy opinions (Fiorina 2016; Mason 2016), it’s clear that they’re polarizing in political opinions generally—e.g. on claims like owning a gun makes you safer (Murray 2018).

13‘P’ must be a description—picking out different probability functions in different possibilities—because it’s an object of uncertainty; I didn’t know what opinions it’d be rational for me to have (see Schervish et al. 2004; Williamson 2008; Dorst 2019; Dorst et al. 2021 for discussion). I’ll use lowercase Greek letters (\( \pi, \delta, ... \)) for particular probability functions, and upper-case Romans (\( P, D, ... \)) for ones that vary across possibilities.

14I’ll model the rational opinions for any given person and time with a unique (White 2005; Horowitz 2013; Schultheis 2018), precise (White 2009; Schoenfield 2012; Carr 2020) probability function. This allows both interpersonal (Schoenfield 2014) and intertemporal (Callahan 2019) permissivism—all it assumes is that at a given time, for a given person there’s only one permissible probability function to adopt. This is a modeling assumption I make both for tractability and because it makes it harder to demonstrate rational polarization.

15\( E_\rho \) is the expectation function for ρ, so for any random variable X: \( E_\rho (X) = \sum_t \rho(X = t) \cdot t \).
2. PREDICTABLE POLARIZATION REQUIRES AMBIGUITY

There are many models of polarization that show that this is possible—that from some suitable epistemic position (say, someone who knows the workings of the model, or the evidence Dan and I will each receive), it can be expected that our rational opinions will move apart.\(^\text{16}\) Does that settle it?

No. Real-world polarization is starker than that. It’s not just that someone could expect our polarization—it’s that we ourselves were in a position to expect it. After all, it’s no secret that failed institutions create libertarians, while universities produce progressives.

More generally, as work on arbitrary-influences on beliefs has made clear, you can often form clear expectations about the direction your choices and circumstances will push your beliefs.\(^\text{17}\) Thus rationalizing polarization requires (at least) showing the possibility of:

**Expectable Polarization:** Dan and I could both expect that my confidence in \(s\) should rise, and his should drop.

\[
E_{\pi}(P(s) - \pi(s)) < 0 \text{ and } E_{\pi}(D(s) - \delta(s)) > 0; \text{ likewise for } E_{\delta}.
\]

Indeed, arguably we must be able to not merely expect such rational shifts (on average), but to predict them with confidence:

**Predictable Polarization:** I and Dan could both predict with confidence that my confidence in \(s\) should substantially drop, and his should substantially rise.

\[
\pi(P(s) \ll \pi(s)) \approx 1 \text{ and } \pi(D(s) \gg \delta(s)) \approx 1; \text{ likewise for } \delta.
\]

And since our disagreements persist upon learning of each others’ opinion, such predictable polarization would also need to be (rationally) **persistent**:

**Predictable, Persistent Polarization:** I and Dan could both predict with confidence that my confidence in \(s\) should substantially drop—and his should substantially rise—**even upon learning of each other’s new opinions**.

\[
\pi\left(P(s|D(s) \gg \delta(s)) \ll \pi(s)\right) \approx 1 \text{ and } \pi\left(D(s|P(s) \ll \pi(s)) \gg \delta(s)\right) \approx 1; \text{ likewise for } \delta.
\]

This catalogue of types of polarization is ordered by strength—predictable polarization implies expectable polarization (but not vice versa), etc.

Here’s the problem: no extant theory of rational polarization can explain how polarization is even expectable—let alone predictable or persistent. The reason is simple. Since expectable polarization requires me to expect that my future, rational,\(^\text{16}\) E.g. Hegselmann and Krause 2002; Jern et al. 2014; Wilson 2014; O’Connor and Weatherall 2018, 2019; Pallavicini et al. 2018; Benoît and Dubra 2019; Singer et al. 2019; Nielsen and Stewart 2021; Zollman 2021.

\(^\text{17}\) E.g. Cook 1987; Cohen 2000; White 2010; Schoenfield 2017; Vavova 2018. For empirical work, see Mcpherson et al. 2001; Kossinets and Watts 2009; Sunstein 2009; Easley and Kleinberg 2010; De Cruz 2017.

\(^\text{18}\) See Kadane et al. 1996; Weisberg 2007; Briggs 2009; Salow 2018 for explanations of this result. Expectable polarization implies \(E_{\pi}(P(s) - \pi(s)) < 0\), hence \(E_{\pi}(P(s)) < E_{\pi}(\pi(s)) = \pi(s)\), violating Reflection.

In any standard Bayesian model—with \(P\) always obtained by conditioning \(\pi\) on the true cell of a finite partition—Reflection is a theorem, and therefore polarization cannot be expectable.\(^\text{18}\) This is intuitive. If at an initial time I could expect that my future, rational,
more-informed self will be less confident of $s$, shouldn’t I now lower my confidence in $s$? (Shouldn’t I defer to my future-self?) If so, there cannot be a rational divergence between what I expect my (rational) future-self to believe, and what I now believe. Yet such a divergence is paradigmatic of stories like mine and Dan’s: a few months into my degree, I had not yet become less confident that owning a gun makes you safer, but I could fully expect that being at a liberal university would make me so. (In cases like this, I’ll speak of a single person expectably polarizing.)

Crucial point: it doesn’t matter what I think of the reliability of higher-education. The fact that my polarization was predictable is enough to show that it cannot have been rational—at least according to standard Bayesian models, since they imply Reflection.

This isn’t an accident. If rational opinions are modeled with (precise) probabilities, Reflection follows from two core assumptions about rationality.

The first is what’s known as the Value of Evidence.\footnote{Ramsey 1990; Blackwell 1953; Savage 1954; Good 1967; see Salow 2020; Dorst et al. 2021 for discussion.} This is the idea that epistemic rationality must be a guide to belief and action: responding rationally to evidence had better be expected to lead you to have more accurate beliefs, and to make more fruitful decisions. (If it didn’t, then you should prefer not to rationally respond to new evidence!) This can be formalized in a variety of ways: as a ban on the possibility of taking bets that lead to a sure loss (a ban on “Dutch books”); as a requirement to expect being rational to increase your accuracy; or as a requirement to expect being rational to lead to better decisions. All three are equivalent (§2.1 and Appendix A). Say that $\pi$ values $P$ iff the transition from $\pi$ to $P$ meets this Value-of-Evidence constraint. The first assumption is that directly updating your beliefs from $\pi$ to $P$ is (epistemically) rational only if $\pi$ values $P$.

The second assumption—usually implicit, but entailed by the formalism used—is No Self-Doubt: if you’ve responded rationally to your evidence, you must be certain that you’ve done so. Where ‘$P$’ is a description for the probability function it’s rational to have, this is the assumption that, for any claim $q$, you should be sure how confident you should be in $q$: for all $q$ there is a $t$ such that $P(P(q) = t) = 1$.\footnote{Any formalism that treats ‘$P$’ as a rigid designator for a particular probability function encodes this assumption, since then $P$ can’t vary across possibilities and so can’t be an object of uncertainty. Similarly, if $P$ is assumed to be obtained by conditioning a fixed prior on a partition, then this assumption follows. See Aumann 1999; Samet 1999; Williamson 2008; Elga 2013; Lederman 2015; Dorst 2019; Dorst et al. 2021.}

Combining the Value of Evidence with No Self-Doubt entails Reflection (Theorem 2.1, §2.1), and hence entails that expectable polarization cannot be rational. This is an impossibility result: any theory of rational polarization must either reject one of these assumptions, or fail to account for the type of predictable polarization we observe empirically.

I know of no account of rational polarization that rejects No Self-Doubt.\footnote{Salow 2018—which I take much inspiration from—shows that self-doubt can lead to expectable polarization (cf. Williamson 2000, Ch. 10); but he uses this as an argument for No Self-Doubt.} Those that use standard Bayesian models accept the Value of Evidence, and so can’t account for the fact that we can expect our own polarization.\footnote{This includes 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; Benoît and Dubra 2019; Nimark and Sundaresan 2019; ?; Henderson 2021, and perhaps Pallavicini et al. 2018.} On the other hand, those that do allow for agents
2. PREDICTABLE POLARIZATION REQUIRES AMBIGUITY

to expect their own polarization do so, to the best of my knowledge, by using belief-updates that violate the Value of Evidence and therefore can be subject to Dutch books.\textsuperscript{23}

What to make of this fact? I think the Value of Evidence is true. I will assume throughout—with a modification to be introduced in §3—that updating your beliefs directly from $\pi$ to $P$ is rational \textit{if and only if} $\pi$ values $P$. (And I’ll assume that a \textit{sequence} of belief-updates is rational \textit{iff} each individual update in the sequence is valuable.) I assume this because it provides a bright line between rationality and irrationality: rational updates are those that you should expect to make you more accurate; irrational ones are those that you should not. If we cross this line—allowing disvaluable updates to be “rational”—those inclined towards the standard, irrationalist story of polarization might fairly complain that we’ve moved the goalposts. For example, several philosophers have argued that allowing evidence to be \textit{permissive}—open to multiple rational interpretations—can nullify the force about arbitrary, predictable-polarizing influences on our beliefs.\textsuperscript{24} The above result sheds doubt on this: even if evidence is permissive, expectable shifts between permissible standards will be expected to make you less accurate. In what sense, then, do they differ from clearly-irrational forms of (say) confirmation bias or motivated reasoning?

This isn’t conclusive. But it is suggestive. The Value of Evidence is well-motivated; No Self Doubt is not. In fact, it’s \textit{ unmotivated}. Peruse the psychological literature on polarization. You’ll find that people engage in biased processing—seeing evidence in a way that favors their prior beliefs—when evidence is \textit{ambiguous}: mixed, complex, or hard to know how to interpret.\textsuperscript{25} You’ll find that these effects are often exacerbated by discussing the evidence in group settings, where peer (dis)agreements in interpretation have large effects on people’s subsequent opinions.\textsuperscript{26} And you’ll find that when the evidence is made easier to interpret or group norms are altered to encourage constructive disagreement, biased processing and polarization often disappear (Lundgren and Prislin 1998; Grönlund et al. 2015; Anglin 2019).

In other words: real-world polarization tends to involve ambiguous evidence and peers who (dis)agree about it. Those are exactly the cases that motivate the literature on rational self-doubt—the literature that Trump’s election had torn me away from.

What if we take that connection seriously? \textit{Hypothesis:} evidence is ambiguous when No Self Doubt is false—that is, when it’s rational to be unsure what beliefs it’s rational to have in response to that evidence. (For some $q$ and all $t$, $P(P(q) = t) < 1$; see §2.1.) More colloquially: your evidence is ambiguous when it doesn’t wear its verdicts on its sleeve; thus, in response to it, you should be unsure what you should think.

So-defined, ambiguous evidence can satisfy the Value of Evidence (Geanakoplos 1989;


\textsuperscript{24}E.g. Schoenfield 2014; Podgorski 2016; Simpson 2017; Callahan 2019; Ye 2019; Jackson 2021.


2. PREDICTABLE POLARIZATION REQUIRES AMBIGUITY

Dorst et al. 2021). And Theorem 2.2 (§2.1) implies that whenever valuable evidence is ambiguous in this sense, it can be expectably polarizing. Why? Because whenever evidence is ambiguous, there is a perspective from which it is asymmetrically ambiguous—it’s easier to recognize when the evidence favors \( q \) than when it tells against it. As a result, it can be expected that, on average, getting the evidence will warrant more confidence in \( q \).

Combining our theorems, the Value of Evidence implies that ambiguity is necessary and sufficient for rational expectable polarization. Thus we have both an impossibility and a possibility proof. Without ambiguous evidence, expectable polarization must be irrational—hence most accounts of rational polarization fail. But if evidence can be ambiguous—i.e. self-doubt can be rational—then expectable polarization could be rational.

This is the flaw in the standard assumptions. The rest of the paper will try to make good on it: I’ll argue that there is a process by which valuable-but-ambiguous evidence can polarize people in a way that can be predictable and persistent (§3), and that this process plausibly plays a role in two of the core drivers of real-world polarization (§§4–5).

If you want to get on with that argument, skip to §3; if you want more details, read on.

2.1 The formalities

I’ll consign full formal definitions and proofs to Appendix A.1—but here are the basics.

We can think of a decision problem as a set of options that, if taken, would each lead to different values at different worlds. \( \pi \) values \( P \) iff, for any such decision problem, \( \pi \)'s expected value of letting \( P \) choose an option on its behalf is (weakly) higher than that of bypassing \( P \) and choosing for itself. (\( \pi \) wants to give \( P \) power of attorney.) As Dorst et al. 2021 show, this is equivalent to (1) the claim that \( \pi \) expects \( P \)'s estimates to be more accurate than its own on every standard (strictly proper) way of measuring accuracy, and to (2) the claim that no Dutch book can be made against transitioning from \( \pi \) to \( P \) (§A.1).

\( P \) is ambiguous iff there is a \( q \) such that for all \( t \): \( P(P(q) = t) < 1 \). So-defined, modeling evidential ambiguity involves modeling higher-order probabilities—§A.1 explains how.\(^{27}\) The important thing to remember is that \( P \) captures the opinions you should have, not the ones you do have. Thus having ambiguous evidence is fully compatible with being rational and knowing what your own opinions are—it simply implies that you don’t know whether such opinions are rational. You may well know that \( \pi \) captures your actual opinions, in fact be rational (\( P = \pi \)), but be (rationally) unsure whether your opinions are rational: \( \pi(P = \pi) < 1 \). (Recall that since ‘\( P \)’ is a description, it picks out different functions at different worlds; so it may be that \( P = \pi \) at \( w \) while at some \( w' \), \( P \neq \pi \). Compare: it may be that Paws is the fastest dog in \( w \) while at some \( w' \), Paws is not the fastest dog.)

There are other models of ambiguous evidence. Those that use precise probabilities fall foul of our impossibility theorem (see footnotes 22 and 23), but some use imprecise probabilities.\(^{28}\) I do not claim ownership of the term—my claim is simply that rational self-doubt is one sensible model of ambiguity. Moreover, note that (1) standard models of imprecise

\(^{27}\)Folk wisdom notwithstanding (Savage 1954; de Finetti 1977), coherent, finite, nontrivial models of higher-order probability are simple generalizations of the standard models of epistemic logic; see e.g. Aumann 1976; Gaifman 1988; Samet 2000; Williamson 2000, 2008; Salow 2018; Dorst 2019, 2020a; Das 2020a,b.

\(^{28}\)E.g. Levi 1974; Seidenfeld and Wasserman 1993; Joyce 2010; Schoenfield 2012—but see Carr 2020.
probability fail to respect the value of evidence (Kadane et al. 2008; Bradley and Steele 2016), and (2) imprecise probability on its own will not lead to expectable polarization.\footnote{In standard models, when imprecise probabilities lead to “dilation”, the future upper- and lower-probabilities are bounded by the current ones—meaning, for instance, the expected future lower-probability cannot be higher than the current one. See Seidenfeld 1981; Seidenfeld and Wasserman 1993; Good 1974.}

Given this, here are our (im)possibility results. (All results restricted to finite models; see §A.1.) First: given the value of evidence, ambiguity is necessary for expectable polarization:

**Theorem 2.1.** If $P$ is always unambiguous and $\pi$ values $P$, then for all $q$: $\pi(q) = \mathbb{E}_\pi(P(q))$.

An immediate consequence: if at each stage in some sequence of updates it’s valuable for me to get some new evidence, and that evidence is never ambiguous, then no matter what evidence I get, expectable polarization could not be rational even in the long run.

Next, our possibility result: given the value of evidence, ambiguity is sufficient for expectable polarization. Say that $P$ is \textit{valuable} iff there is some $\pi$ that assigns positive probability to all its realizations and yet values it. Then:

**Theorem 2.2.** If $P$ is valuable and possibly ambiguous, there are infinitely many $\pi$ such that $\pi$ values $P$ and yet there is a $q$ for which $\mathbb{E}_\pi(P(q)) > \pi(q)$.

This implies that expectable polarization can be rational even in the short run. And as I’ll argue, iterating such expectably-polarizing processes can generate a series of rational steps that lead to long run polarization that is predictable and persistent.

3 Cognitive Search Polarizes People

In principle, ambiguous evidence could lead to rational expectable polarization—but this possibility remains entirely abstract. Are there realistic cognitive processes that lead to it? And can they engender predictable, persistent polarization?

There are—and they can. Consider a word-search task (cf. Elga and Rayo 2020). You’re given a string of letters and some blanks, and you have a handful of seconds to figure out whether there’s an (English) word that completes the string. For example:

\[
\text{P\_A\_ET}
\]

And the answer is... yes, there is a completion. (Hint: what is Mars?) Another:

\[
\text{P\_G\_ER}
\]

And the answer is... no, there is no (legitimate) completion.

To solve a word-search task, you need to do a cognitive search (Todd et al. 2012). This is a process in which you search an accessible cognitive-space for an item of a particular profile—think of searching your memory for an example, searching your reasoning for a flaw, or searching your lexicon for a word. The crucial feature: it’s easier to recognize that \textit{there} is a completion to your search than to recognize that there’s none. To recognize that there is a completion, all you have to do is find it; to recognize that there’s none, you have to
rule out the possibility that there’s an accessible item that you’ve missed. The former is an existential task (find an instance); the latter is a universal task (show that there is none).

My claim is that this leads to an asymmetry in how ambiguous your evidence is—and thus that it can lead to (rational) expectable polarization.

Meet Haley. She’s wondering whether a fair coin, which I just flipped, landed heads. To give her a chance, I’ll present her with a word-search task, determined by the coin flip: if the coin landed heads, it’ll be a completable task; if it landed tails, it’ll be uncompletable. Thus her confidence that the coin lands heads should equal her confidence that the string is completable. I’ll give her 7 seconds to look at the string, and then she’ll write down how confident she is that the coin landed heads. She knows all of this.

She’s currently 50% confident it’ll land heads. Nevertheless, I claim that our (and her!) current estimate for her posterior rational confidence should be higher than 50%—she should be expectably polarized by the word-search. This doesn’t mean we should be confident she’ll write down a higher number: sometimes her confidence will go up, other times it’ll go down. Rather, it means that across many trials like this, we should be confident that the average of the numbers she should write down will be higher than 50%.

Why? The idea is simple. It’s easier for her to recognize that the string is completable than to recognize that it’s not. So if the coin lands heads, her confidence that it did should (on average) increase a lot; if it doesn’t, her confidence that it did should (on average) decrease a bit—and the average of “increase a lot” and “decrease a bit” is “increase a bit”.

Now in more detail. Say that a string is obvious iff, upon seeing it for 7 seconds, she should be sure it’s completable. (If she’s unsure, she’s failed to properly use her evidence.) Some strings are clearly obvious. Witness:

\[
\text{C\_T}
\]

Others are clearly not:

\[
\_\_XT\_\_M\_\_M
\]

Though the string is completable, she’s not irrational for failing to complete it.

If what she learned when she looked at the string was simply whether or not it’s obvious, her evidence would be unambiguous and so would not be expectably polarizing. But that’s not what she learns. Since it’s easier to tell there’s an obvious word than to tell that there’s none, she can’t always tell whether it’s obvious. If it is obvious, then (by definition) she’s in a position to be sure that it is—she should find a completion. But if it’s not obvious, she may well be unable to be sure it’s not. Staring at _EA_ K, she’s unsure—she hasn’t found one, but maybe, like _EA_ T, she’s about to realize it’s obvious (‘HEART’!). In fact, she’s not, since _EA_ K has no (obvious) completions. But she’s not in a position to be sure of that after only 7 seconds. All she knows is that she didn’t find a completion within 7 seconds—some evidence that it’s not obvious, but inconclusive: she should still wonder whether her failure to find one was because she was rational (it wasn’t obvious), or she was

---

\[^{30}\] If \( P \) is updated by conditioning \( \pi \) on the true member of \{ Obv, \neg Obv \}, then \( \mathbb{E}_\pi(P(q)) = \pi(Obv) \cdot \mathbb{E}_\pi(P(q)|Obv) + \pi(\neg Obv) \cdot \mathbb{E}_\pi(P(q)|\neg Obv) = \pi(Obv) \cdot \pi(q|Obv) + \pi(\neg Obv) \cdot \pi(q|\neg Obv) = \pi(q) \).
irrational (she missed an obvious one). That is: when the string is not obvious, she has ambiguous evidence—she should be unsure whether she should be sure there’s a completion.

Ambiguous-evidence structures like these are sometimes called “negative access failures” (see Williamson 2000, Ch. 8; Lasonen-Aarnio 2015; Salow 2018, 2019). Though it’s controversial whether they are possible, I think cognitive-search makes a strong case that they are—§3.1 provides both an argument and some simple models. Nevertheless, since my main aim is to show what happens if such ambiguous-evidence structures are possible, I’ll move on to showing why the update is expectably polarizing.

Start with a simple case: suppose she knows that if there’s a completion, it’s obvious. That means that if the coin lands heads (there’s a completion), she should be able to tell that it did (she should find one). But if it lands tails (there’s no completion), she can’t be sure that it did—after all, perhaps there’s an obvious completion that she’s missed. In other words, it’s easier for her to be rationally confident when the coin lands heads than when the coin lands tails. Precisely: there’s initially a 50% chance her confidence should rise all the way to 100%, and a 50% chance it should drop to something greater than 0% (say, to 20%). Thus the average posterior rational credence that the coin landed heads is greater than 50% (say, 60%)—higher than her initial confidence. That’s expectable polarization.

Nevertheless, the update is valuable. For notice: the rational opinions are always moving in the direction of truth. She starts out 50% confident that the coin’ll land heads. When it does, she should become more confident (100%) that it did; when it doesn’t, she should become less confident (20%) that it did. The evidence is valuable because the rational opinions shift toward the truth; yet polarization is expectable because these shifts are asymmetric.

The same reasoning applies if she doesn’t know that completable strings will be obvious. When they are obvious, she should know that they are; but when they aren’t, she should still (often) leave open that they are; thus she should expect that, on average, the rational confidence will rise. And since her opinion about whether the string is obvious should always be shifting (albeit asymmetrically) in the direction of truth, the update is still valuable. (See §3.1 for models and proofs.)

This explains why the rational opinions for her to have will expectably polarize. But—as I’ve been at pains to emphasize—her actual opinions may well be irrational. (Whenever she has self-doubt, she thinks they might be!) Why should we expect her actual opinions—the numbers she actually writes down—to polarize as well? Because if she’s approximately rational, the natural hypothesis is that her actual opinions will be some noisy indicator of the rational ones. Since the rational credences are expectably polarized, her actual opinions thus will be too. (For simplicity, I’ll continue to only explicitly model the rational opinions.)

That’s the theory. How can we test it?

Simple. Meet Thomas. Like Haley, he’s about to see a word-search task. Whether it’s completable is also determined by the (same) coin toss. But whereas Haley will see a completable string iff the coin lands heads, Thomas will see a completable string iff the coin

---

31They are failures of the axiom \( P(q) < 1 \rightarrow [P(P(q) < 1) = 1] \); see §3.1.

32Other hypotheses are possible. Given a model of the rational opinions like those in §3.1, any hypothesis that meets two conditions will engender expectable polarization on her actual opinions: (1) all she’s certain of is whether or not she’s found a completion and (2) her actual confidence is always within the span of what she’s certain the rational confidence might be, but is not always on the extreme edge of that span.
lands *tails*. By exactly parallel reasoning, Thomas’s opinion will expectably polarize in the opposite direction: he’ll be better at recognizing when the coin lands tails than when it lands heads; therefore his posterior rational credence in heads will, on average, be lower than 50%.

In fact, it’s time you met the whole room. Half of them are *Headsers*: like Haley, they’ll see a completable string iff the coin lands heads. The other half are *Tailsers*: like Thomas, they’ll see a completable string iff the coin lands tails.

Headsers should be better at recognizing cases where the coin lands heads; Tailsers should be better at recognizing cases where the coin lands tails. Thus if the coin lands heads, the average Headser will be quite confident it did, while the average Tailser will be only somewhat confident it did. And if it doesn’t land heads, the average Headser will be somewhat doubtful it did, and the average Tailser will be quite doubtful that it did. Since both groups start out 50% confident in *heads*, that means everyone can predict that their average opinions will split apart.

Well—do they?

I’ve tested this in two ways. The fun way: I’ve presented this scenario to my audiences during a handful of talks. (Indeed, Haley and Thomas were in the first such audience.) In 6 of 7 of them, the average Headser posterior was higher than the average Tailser one.

Second, the rigorous way: I’ve run this experiment properly, finding a statistically significant (and large) difference in the two groups’ mean posterior degrees of confidence that the coin landed heads (Headsers: 57.7%; Tailsers: 36.3%, \( p < 0.001, d = 1.57 \)). The experiment also controlled for a confound: it provided evidence that it is the ambiguity of an uncompletable string that drives the polarization, rather than the simple fact that an uncompletable string provides weaker evidence than a completable one. For details, see §3.2.

Upshot: cognitive search engenders (valuable) ambiguity-asymmetries that can lead real people to expectably polarize.

**What about predictable, persistent polarization?** The basic idea is simple: iterate this process. In the models given in §3.1, the prior estimate for Haley’s posterior rational credence is about 60%. So if we just repeat with a bunch of independent coins and word-searches, it seems that we should be able to predict that—since probably about half the coins will land heads—her average confidence will rise to around 60%. Since the coins are independent, that implies that she should become confident that around 60% of the coins landed heads, and hence that she should become very confident that more than half of the coins landed heads.

But there’s a hitch. Two, actually. The first: *Can* we just iterate this process, while respecting the Value of Evidence? Suppose Haley starts out with probability function \( \eta \), and then over time the rational opinions for her become \( H^1 \), then \( H^2 \), ..., then \( H^n \). I’ve been focusing on showing that the individual updates—from \( \eta \) to \( H^1 \), from \( H^1 \) to \( H^2 \), etc.—could be valuable despite being expectably polarizing. But ignore the individual updates for a second; focus on the beginning and the end. Let \( h \) be the claim that *more than half of the coins landed heads*. If indeed we can rationally iterate this process, then at the beginning Haley can predict with confidence that the final rational credence function for her will be confident of \( h \)—for example, \( \eta(H^n(h) \geq 0.9) \geq 0.9 \). It follows that her initial opinions (\( \eta \)) do
not value her final opinions $(H^n)$. After all, she’s initially about 50% confident of $h$; so if she’s confident that her final opinions will be confident of $h$, she must expect that almost half the time, that confidence will be misplaced!\footnote{Formally, $\eta$ values $H^n$ only if $\eta(H^n(q) \geq t) \geq s$ implies $\eta(q) \geq t \cdot s$ (Dorst 2020a, Fact 5.5). So if $\eta(H^n(h) \geq 0.9) \geq 0.9$, we must have $\eta(h) \geq 0.81$—which we don’t.} Thus $\eta$ doesn’t value $H^n$. Crucial question: does it follow that some individual update in this sequence must be irrational—that there is an $i$ such that $H^i$ doesn’t value $H^{i+1}$?

I don’t know. If the evidence were unambiguous, then indeed it would follow. For ambiguous evidence, that’s an open question. But it turns out we don’t need to settle it. For there’s a second hitch—and solving the second will also solve the first.

The second hitch: people forget things! In particular, if I actually presented Haley with a series of 100 word-search tasks, by the end she would start to forget what the beginning strings were. And here’s an easy theorem: $\pi$ values $P$ only if it’s certain that $P$ is certain of everything that $\pi$ is; if $\pi(q) = 1$, then $\pi$ values $P$ only if $\pi(P(q) = 1) = 1$. So if Haley thinks she might forget something—anything—then she doesn’t value her future opinions. Does that mean the possibility of rational polarization is sunk?

No. Some things—like Mom’s birthday—are bad to forget; others—like what you ate for breakfast—are not. The former are questions that you care about being accurate about, and whose answers are likely to affect what decisions you should make; the latter aren’t.

As stated, the Value of Evidence ignores this distinction: $\pi$ values $P$ iff for any decision problem, it prefers to let $P$ make the decision; iff for every question, it expects $P$ to be more accurate than itself. But that’s a high bar. Most forms of deference are local—you defer about some questions, but not others (Dorst et al. 2021). You defer to the forecaster about whether it’ll rain tomorrow, but not about whether your poncho looks cool; you defer to your future-self (next month) about where you should go on vacation, but not about what you had for breakfast this morning.

We can weaken the Value of Evidence to make it question-relative in this way. A question $Q$ is a partition (any partition) of logical space (Hamblin 1976; Roberts 2012)—a way of dividing possibilities into those that agree on the answer to $Q$. For example, “Will it rain tomorrow?” = \{Rain, ¬Rain\}. $\pi$ values $P$ with respect to $Q$ iff, for any decision whose outcomes are determined by the answer $Q$, it prefers to let $P$ decide for it. If so, $\pi$ also expects $P$ to be more accurate than itself about any proposition whose truth-value is determined by $Q$. (See §3.1 for details.)

What happens if we lower the bar to question-relative Value? It solves our two problems. Fix the most fine-grained question $Q$ you (should) care about. I claim that if an update is valuable with respect to $Q$, then it is a rational update. This permits (some) forgetting. In particular, if what you forget doesn’t affect your opinions about the question $Q$ you care about, then it is valuable with respect to $Q$, and hence is rational.\footnote{Precisely: If $\pi(p) = 1$, $P(p) < 1$, but for all $q \in Q$: $\pi(q) = P(q)$, then $\pi$ values $P$ with respect to $Q.$} Of course, real people do sometimes forget things that matter, so this is still an idealized model. But what it allows us to show is that forgetting information about $Q$ is not essential to predictable polarization about $Q$. Although there are a variety of models that show how limited memory can lead to polarization (Wilson 2014; Dallmann 2017; Fryer et al. 2019; Loh and Phelan 2019; Singer et al. 2019), so far as I know none of them allow you to polarize about a given question without losing information about that very question.
3. COGNITIVE SEARCH POLARIZES PEOPLE

And this in turn gives us a way to rationally iterate the word-search tasks. Let \( Q \) be the question of how each of the coins land; imagine this is what Haley should care about. Each time she is presented with a string, she updates in the way discussed above—this update is valuable with respect to every question. Then some time later she forgets the string used in that completion. This forgetting doesn’t affect her opinions about how the coins land, and so is valuable with respect to \( Q \), and hence is rationally permissible. What it does, however, is consolidate her self-doubt: although she previously wondered whether a completion was cognitively accessible, once she forgets the string she knows that she’s no longer in a position to complete it. Hence although she often has self-doubt upon first seeing each task, that self-doubt is consolidated by the time the next task comes around—she knows the rational thing to do (now) is to stick with the opinion she was left with when she forgot what the string was. Then she’s presented with the next word-search task...

If this is how her memory works, then predictable polarization can be rational. Let \( h \) be the claim that more than half the coins land heads. She starts out about 50% confident of \( h \). Theorem 3.1 (§3.1) shows that each update in the sequence is valuable with respect to \( Q = \text{how each of the coins landed} \), and yet that—as the number of word-searchs grows—she can predict with arbitrarily high confidence that, by the end, she should be near-certain that more than half the coins landed heads: \( \eta(h) \approx \frac{1}{2} \), yet \( \eta(H^n(h)) \approx 1 \). This amounts to an epistemic form of a diachronic tragedy (Hedden 2015): it is a series of updates, each of which you should expect to make you more accurate about \( Q \), but which in the long run you should (initially\(^{36}\) expect to make you less so.

Moreover, if Thomas goes through the Tailser-version of this process in parallel, the resulting predictable polarization is also persistent: both of them can predict that the other’s opinions should move in the opposite direction, and hence when they find out what has happened, their newly-polarized opinions remain extreme (Corollary 3.2, §3.1).

What, intuitively, is happening? Since the (fair) coins are all independent, initially they’re both 50% confident of \( h \), and are quite confident that roughly half the coins will land heads. As the process unfolds, there are many tosses (say, Heads\(_1\), Heads\(_3\), Heads\(_6\),...) for which Haley becomes sure that they landed heads (she finds completions); for the others, the letter-string was ambiguous, so she has middling degrees of confidence. As a result, her average confidence in Heads\(_i\) rises to around 60%. Since she knows the coins are independent, she must therefore (by the law of large numbers) come to think that it’s very likely that more than half the coins landed heads.

Of course, she was confident from the outset that her confidence in this would rise. But so what? She had no idea which Heads\(_i\) her confidence would rise or fall in; using the only relevant evidence she has (the word-searches), her confidence has risen a lot in some and fallen a little in others; as a result, her average confidence has risen. The fact that she initially expected half the coins to land heads, or that she predicted that her confidence would shift in something like this way, provides no evidence against Heads\(_1\), Heads\(_3\), etc. (which she’s sure

\(^{35}\)A way, not the way. This section contains no impossibility proofs. There may be ways to iterate the process without any forgetting—for example, if the salient question changes throughout the process. Indeed, for all I know there is some way to get predictable polarization with a series of fully valuable updates.

\(^{36}\)Unlike a diachronic tragedy in Hedden’s sense, you don’t always prefer to not take the full sequence of updates—by the end, you’ll be convinced that your posterior is more accurate than your prior.
3. COGNITIVE SEARCH POLARIZES PEOPLE

Thus she finds herself confident that more than half the coins landed heads, with no way to lower that confidence: she should expect that lowering her confidence in any of the $Heads_i$ would make her less accurate—both about how that coin landed, and overall.

Peeking over her shoulder, she also notices that Thomas is now extremely doubtful that more than half the coins landed heads. But so what? She predicted that from the outset; it doesn’t provide substantial evidence against any of the $Heads_i$. So her confidence persists.

Two caveats. If Haley and Thomas knew that they’d both been rational, and knew exactly what each others’ opinions were, then their disagreement would disappear (Aumann 1976, cf. Lederman 2015). But they don’t know that—Haley doesn’t even know that her opinion in $Heads_2$ was rational (though it was), never mind that all her opinions were.

Second, there are ways that they can ameliorate their disagreements. If they start sharing what they did and didn’t find (“I was sure there was a completion on the first toss, was unsure on the second,...”), this would reduce (though often not eliminate) their disagreement. But that is an exacting exercise: it takes the patience to talk through (and the ability to recall) all of the individual bases for their opinion about $h$. Thus it may well be that Haley and Thomas are 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: cognitive search provides a mechanism by which predictable, profound, persistent polarization could indeed be rational.

That means, in principle, we could tell a better story about polarization. For $Heads_i$ and $Tails_i$, substitute bits of evidence for and against the claim that owning a gun makes you safer. Going off to liberal universities made me a Tailser—made me better at recognizing bits of evidence against that claim. Staying in a conservative town made Dan a Headser—made him better at recognizing bits of evidence favoring that claim. Neither of us were made worse at assessing evidence—we were made better, in asymmetric ways. But as a result, it may well be that we are left disagreeing about high-level claims (“guns make people safer”), while being unable to share all the (rational) reasons we have for those differing opinions.

If that were what happened, then both of us could’ve predicted polarization as the outcome—as we did. And neither of us should be moved now, when we discuss our profound disagreements—as we’re not. Nevertheless, we could both acknowledge that while we think the other side is wrong, we needn’t think that 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 Haley and Thomas polarizing as a result of word-search tasks is far more realistic than it seems. We engage in cognitive search and face ambiguous evidence all the time—and these facts helps to explain the drivers of real-world polarization. If you’d like to jump to that argument, skip to §4; if you’d like to see the formal (§3.1) or experimental (§3.2) details of this section’s argument, read on.

37If she could be certain that roughly half the coins would land heads, then she could use this certainty as a basis for lowering her confidence in the coins she’s not sure of. But she can’t be certain of that—and indeed, she becomes progressively less confident as she spots more $Heads_i$. 

16
3. COGNITIVE SEARCH POLARIZES PEOPLE

3.1 The formalities

Here I’ll (1) argue that the word-search engenders an ambiguity-asymmetry, (2) offer valuable-but-polarizing models of the word-search task, and (3) explain question-relative Value and the predictable-polarization result.

(1) Cognitive search engenders ambiguity. Let η be Haley’s prior and H the posterior rational credence function upon looking at a word-search task for 7 seconds. Recall that the string is obvious iff she should be sure it’s completable: where c is the claim that it’s completable, Obv ↔ [H(c) = 1]. I claimed that if the string is obvious, she can be sure that it is; but if it’s not, she should still leave open that it is. The first claim is straightforward; the second is what requires argument.

The claim is that if the string is not obvious (H(c) < 1), she should still leave open that it is (H(H(c) = 1) > 0). So it suffices to show that H(H(c) = 1) > 0.

**Premise 1:** After 7 seconds, she should not be sure there’s no completion: H(c) > 0.

**Premise 2:** After 7 seconds, conditional on there being a completion, she should leave open that she’ll find it in another moment. Where $C^+$ is the probability function she will have 8 seconds after being shown the string, $H(C^+(c) = 1|c) > 0$.

**Premise 3:** After 7 seconds, conditional on there being a completion and her in fact finding it after 8 seconds, she should leave open that she should’ve found it in 7 (what she does do in 8 seconds she, perhaps, can do in 7): $H(H(c) = 1|c\&C^+(c) = 1) > 0$.

It follows that she should leave open that she should be sure there’s a word, i.e. leave open that it’s obvious: $H(H(c) = 1) > 0$, i.e. $H(Obv) > 0$.38 Whenever $H(c) < 1$, that’s a negative access failure.

(2) Models of the word-search. Figure 1 presents a simple model of the various possible rational opinions for Haley after looking at her string, in two forms: the left is a generalized Kripke (or Markov) diagram; the right is stochastic-matrix notation. (See Appendix A for the formal semantics, and Dorst 2020b; Dorst et al. 2021 for background.)

As indicated by the blue numbers, her prior η assigns 1/2 to the possibility that there’s no completion, 1/8 to the possibility that there’s a non-obvious completion, and so on. The posterior rational opinions, H, vary from world to world—they are obtained by conditioning η on either ~Find, Obv&~Find, or Find.39 In particular, if she finds a completion within 7 seconds, she should be certain that she did. If she doesn’t find a completion but there is an obvious one, she should be certain (she should realize) that she didn’t find one but that in fact there’s an obvious one.40 And if there’s no completion or it’s not obvious, she should realize she didn’t find one and be 2/3 confident there’s no completion—but also be 1/6 confident that there’s an obvious completion that she missed. In these (red) possibilities,

---

38 $H(H(c) = 1) \geq H(c\&C^+(c) = 1|H(c) = 1) = H(c) \cdot H(C^+(c) = 1|c) \cdot H(H(c) = 1|c\&C^+(c) = 1) > 0$.

39 Thus the update is not partitional—a common feature of ambiguous evidence. See Geanakoplos 1989; Williamson 2000, Ch. 10; Williamson 2019; Salow 2018; Dorst 2020a.

40 This highlights why we cannot interpret “H” as “the opinions she would’ve had, if she were rational”, but must instead interpret it as “the opinions her evidence warrants adopting” (Williamson 2000, Ch. 10). If there’s an obvious completion, then (in some sense) if she had been rational, she would’ve found it, and so would know that she had. Nevertheless, if she does not in fact find it, then that fact is part of her evidence—so her evidence warrants being sure of it. Thus the opinions her evidence warrants adopting are “Arg! Although I didn’t find it within 7 seconds, there was an obvious one.”
she has ambiguous evidence (and a negative access failure), since $H(c) < 1$, but $H(\text{Obv}) > 0$, hence $H(H(c) = 1) > 0$.

Any model with this structure is expectably polarizing on the claim that the string is obvious.41 (So even if you object to the details, I hope you’ll agree with my point that updates like this are possible.) In this case, Haley initially is 50% confident there’s a completion, but her expectation for the future rational confidence is $7/12 \approx 0.58$.42

For simplicity, going forward I’ll coarse-grain models like this by collapsing possibilities within the “completable, but don’t find” region,43 yielding the model in Figure 2. Think of this as representing the average rational opinions for Haley to have within the three remaining classes of possibilities. This model captures the same type of ambiguity-asymmetry (if it’s completable, she can get unambiguous evidence; if it’s not she can’t), is equally expectably polarizing, etc.

Why are these updates valuable, despite being expectably-polarizing? The reason is easy to see in the coarse-grained model: $H$ is more accurate than $\eta$ on every proposition, in every world. For example: if the string is not completable, $\eta$ is $1/2$ confident of this while $H$ is $2/3$; if the string is completable but she won’t find it, $\eta$ is $1/3$ confident of this while $H$ is $2/3$, etc.

Accuracy always increases—it just increases asymmetrically: it increases more if the task is completable than if it’s not. (Similar reasoning applies to Figure 1—see Appendix A.)

(3) Question-relativity and predictable polarization. Let a question $Q$ be a partition of logical space; given a world $w$, $Q(w)$ is its partition-cell. A proposition $p$ is about $Q$ iff every complete answer to $Q$ settles the truth-value of $p$ (iff $p = \bigcup_i q_i$ for $q_i \in Q$). A

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{example_diagram.png}
\caption{Model of a Header’s rational opinions. Left: Generalized Kripke (Markov) diagram, in which blue numbers inside circles represent the prior probabilities of those possibilities, and labeled arrows from circles represent the posterior probabilities in those possibilities. Right: The vector $\eta$ represents prior probabilities; the matrix $H$ represents posteriors: row $i$ column $j$ is the probability that, in world $i$, it’s rational to assign to being in world $j$; thus the top row of $H$ says what Haley’s probabilities should be if the string isn’t completable; the second row says what they should be if it’s completable but not obvious, etc.}
\end{figure}

\footnotesize
41 Suppose $\text{Obv}$ implies $H(\text{Obv}) = 1$, but $\neg\text{Obv}$ is consistent with $H(\text{Obv}) > 0$ and so $\mathbb{E}_\eta(H(\text{Obv})|\neg\text{Obv}) > 0$.

42 $\mathbb{E}_\eta(H(c)) = 1/2(1/3) + 1/3(4/3) + 1/3(1) + 1/3(1) = 7/12$.

43 Thinking of probability functions as vectors, $\mathbb{E}_\eta(P|\neg\text{Find}) = \frac{1}{2}(2/3, 1/6, 1/6, 0) + \frac{1}{2}(0, 0, 1, 0) = (1/3, 1/6, 1/2, 0)$, which—collapsing the middle two possibilities—coarsens to $(1/3, 2/3, 0)$. 
Figure 2: Coarse-grained model of a Header’s rational opinions, in both (left:) generalized Kripke-model and (right:) stochastic-matrix notation. See Figure 1 for how to read the diagrams.

decision-problem is about $Q$ iff every answer to $Q$ settles the values of every option in the decision-problem. $\pi$ values $P$ with respect to $Q$ iff for every decision problem about $Q$, $\pi$ expects $P$’s decision to be better than its own. (See Appendix A.2 for more details.)

Suppose we present Haley with a sequence of $n$ mutually-independent word-search tasks, using the model from Figure 2. Let $Q_i$ be the partition of how the $i$th task went: $Q_i = \{N, C, F\}$, where $N$ is the set of worlds where it’s not completable, $C$ the set where she doesn’t find a completion, and $F$ the set where she finds a completion. Let $Q$ be the combination of all the $Q_i$, so that $Q$ settles how each task went (and, in particular, whether each coin landed heads): for any worlds $w, w'$, $Q(w) = Q(w')$ iff for all $i$, $Q_i(w) = Q_i(w')$.

After being presented with each task and forming the resulting rational credence, Haley then forgets the string. This has the effect of consolidating her higher-order uncertainty: it holds fixed her opinions in $Q$, but makes her certain that those opinions are (now) rational.

Her rational opinions at each stage are $H^0, \overline{H^0}, H^1, \overline{H^1}, ..., H^n, \overline{H^n}$. The transitions from $H^i$ to $\overline{H^i}$ consolidate her higher-order uncertainty while holding her opinions about $Q$ fixed—they are valuable with respect to $Q$. The transitions from $\overline{H^i}$ to $H^{i+1}$ look exactly like the update in Figure 2, Jeffrey-shifting (Jeffrey 1990) her opinions on the $Q_i$ partition in different ways in different worlds, in the way specified by that model. These transitions are valuable with respect to every question (not just $Q$).

Letting $h$ be the claim that more than half the coins land heads, this process leads to predictable polarization (see Appendix A.2 for proof):

**Theorem 3.1.** 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 some $q_i \in Q$) 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 $i$;
- $H^0(h) \approx \frac{1}{2}$ and $H^0(\overline{H^0}(h) \approx \frac{1}{2}) = 1$; and yet
- $H^0(\overline{H^n}(h) \approx 1) \approx 1$.

Moreover, adding Thomas the Tailser to the model results in polarization that is also persistent—see Corollary 3.2, Appendix A.2.
3. COGNITIVE SEARCH POLARIZES PEOPLE

3.2 The experiment

Here I’ll sketch an experiment that both confirmed the that cognitive-search can cause polarization, and controlled for a confound in the ambiguity-based explanation of this effect. (See Appendix B for more details.)

The confound is this. Ambiguous evidence is not the same thing as 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. Evidence can be weak without being ambiguous.

Figure 3 gives an example. Urn A contains one black and one red marble; Urn B contains two red marbles. I flip a coin, grab Urn A if heads, and Urn B if tails; then I draw a single marble and reveal it. If I draw a black marble, that’s strong evidence: you should be sure I’m holding urn A. If I draw a red marble, that’s weak evidence: you should slightly boost your confidence that I’m holding Urn B. But either way, it’s unambiguous evidence. If you see a black marble, you should know that you should be sure I’m holding Urn A; if you see a red marble, you should know that you should be $2/3$ confident I’m holding urn B. In neither case should you have self-doubt.

![Figure 3](image)

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

So the strong/weak asymmetry is not the same as the unambiguous/ambiguous asymmetry. In the word-search task, both are present. As I’ve said: if the string is completable, you can get unambiguous evidence that it is; if it’s not, you can only 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 can only get weak evidence that it’s not.

My theory predicts that it’s the ambiguity-asymmetry that’s driving the polarization; but what if it’s the weak/strong asymmetry? Perhaps the reason people polarize is that they under-react when they see weak evidence: when they don’t find a word, they don’t move their opinions as much as they ought to.\(^{44}\)

The experiment tested this.\(^{45}\) Participants were divided into Headsers and Tailsers. Headsers sometimes got strong evidence when a coin landed heads, and always got weak evidence when it landed tails; Tailsers vice versa. There were two conditions. The unambiguous condition saw draws from the Urn task above; the ambiguous condition saw word-search tasks. My prediction was that the ambiguous condition would polarize, and that it would do so more than the unambiguous condition.

---

\(^{44}\)There is indeed evidence that (in certain settings) people are conservative Bayesians, meaning they under-react to some types of evidence (Peterson and Beach 1967; Edwards 1982). But many of these effects are due to failure to fully believe or understand the experimental setup (Corner et al. 2010; Hahn and Harris 2014)—arguably, a source of ambiguity.

\(^{45}\)Detailed report in Appendix B; pre-registration here: [https://aspredicted.org/8jg3e.pdf](https://aspredicted.org/8jg3e.pdf).
4. COGNITIVE SEARCH EXPLAINS CONFIRMATION BIAS

It worked. The plots in Figure 4 show how the average individual’s mean confidence in heads evolved as they were presented with more tasks. As can easily be seen, the ambiguous condition polarized, and did so significantly more than the unambiguous condition. In addition, Appendix B reviews evidence that it is ambiguity that explains this effect.

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

4 Cognitive Search Explains 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 the libertarian crowd, experienced the unjust failures of educational and criminal institutions, and became independent and skeptical of many types of authority. I fell out with that crowd and in with the liberal nerds, experienced the unearned favors of educational and criminal institutions, and became skeptical of many claims of individual responsibility and autonomy.

Can my rational story explain this? I’m going to make the case that it can: that it is both theoretically possible and empirically plausible that ambiguous evidence plays a role in predictable polarization like ours.

Psychologists have documented many mechanisms that predictably polarize people. Confirmation bias comprises a variety of tendencies to seek and interpret evidence in a way that predictably strengthens your prior beliefs (Nickerson 1998; Whittlestone 2017). Motivated reasoning is the closely-related tendency to subject uncongenial information to especially high scrutiny (Kunda 1990; Kahan et al. 2017). And the group polarization effect is the process by which talking with likeminded others (or listening to like-minded sources) tends to lead you to become more extreme in your opinions (Isenberg 1986; Sunstein 2009).

46 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].

47 A 2 (Headser vs. Tailser) by 2 (ambiguous vs. unambiguous) ANOVA indicated a main effect of being a Headser/Tailser (F(1, 224) = 68.99, p < 0.001, η² = 0.217), a main effect of ambiguity (F(1, 224) = 6.39, p = 0.012, η² = 0.020), and—crucially—an interaction effect between the two (F(1, 224) = 21.63, p < 0.001, η² = 0.068), indicating that the divergence was exacerbated by ambiguity. Moreover, an empirically bootstrapped 95% confidence interval for the difference of differences—i.e. (A-Headsers − A-Tailsers) − (U-Headsers − U-Tailsers)—was positive ([7.19, 22.59]), indicating that the former difference was larger.
4. COGNITIVE SEARCH EXPLAINS CONFIRMATION BIAS

People who are aware of these predictably-polarizing mechanisms are still subject to them (Pronin 2008; Lilienfeld et al. 2009). Theorem 2.1 therefore implies that if evidence is unambiguous, they must be irrational.\textsuperscript{48} But in this section and the next I’ll argue that ambiguous evidence is plausibly present in a way that may indeed make them rational.

Confirmation bias first. This effect has been widely cited as a core driver of polarization in both academic\textsuperscript{49} and popular\textsuperscript{50} writings. Nevertheless, many empirical researchers have noted that we lack good normative standards for assessing its rationality.\textsuperscript{51} My goal is to provide those standards.

Confirmation bias is standardly divided into (at least) two tendencies: (1) \textit{selective exposure}, the tendency to seek evidence that you expect to confirm your preferred hypothesis (Frey 1986; Hart et al. 2009), and (2) \textit{biased assimilation}, the tendency to interpret mixed evidence as supporting your preferred hypothesis (Lord et al. 1979; Taber and Lodge 2006). I’ll focus on the latter in this section, and come back to the former at the end of §5.

The classic examples of biased assimilation go like this.\textsuperscript{52} Take two people who strongly disagree about some claim—say, Dan and I, who disagree about whether owning a gun makes you safer. Present both of us with two studies: one that (on its face) supports this claim, the other of which (on its face) tells against it. Give us time to think about them. Since you’ve given us with the same information, you might hope this will 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 many situations like this—Dan will tend to increase his confidence that guns make people safer, and I will tend to decrease mine.

Why? It’s not because we’ll simply dismiss or ignore the evidence against our beliefs. To the contrary: we’ll likely spend more time looking at the details of the incongruent study. We’ll search for flaws in the methodology, gaps in the reasoning, and alternative explanations that could explain away the data. And we’ll likely come up with legitimate flaws.\textsuperscript{53}

Thus the mechanism that drives biased assimilation is \textit{selective scrutiny}: people spend more time looking for flaws with incongruent evidence than with congruent evidence. And thus biased assimilation results from the same mechanism that drives \textit{motivated reasoning}\textsuperscript{54}—explaining one will explain both.

Thomas Kelly (2008) argues that selective scrutiny may be rational, and that it may

\textsuperscript{48}And thus that standard Bayesian models—such as Feeney et al. 2000; Austerweil and Griffiths 2011; Le Mens and Denrell 2011; Jern et al. 2014; Benoît and Dubra 2019—can’t fully rationalize them.

\textsuperscript{49}E.g. 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.


\textsuperscript{51}Lord et al. 1979; Lord and Taylor 2009; Taber and Lodge 2006; Crupi et al. 2009; Ross 2012; Mercier 2017; Whittlestone 2017.


rationalize some types of polarization. After all, it makes sense to spend more of our (limited) cognitive resources trying to understand surprising findings. If I believe that guns don’t make people safer, then seeing a study suggesting that they do should surprise me, while seeing a study suggesting the opposite shouldn’t. Thus it makes sense for me to more thoroughly scrutinize the former. (And for Dan to more thoroughly scrutinize the latter.)

Now notice: if Dan and I scrutinize different studies, we end up receiving different total evidence: I know more about one, Dan knows more about the other (Kelly 2008). Thus selective scrutiny (and therefore biased assimilation and motivated reasoning) seems to reduce to a type of selective exposure, namely selective exposure to flaws with incongruent evidence (cf. Kunda 1990). And if Dan and I aren’t aware that we’re selectively scrutinizing—all we come away with is, “I saw a congruent study in which I didn’t see a flaw, and an incongruent one in which I did”—then the resulting polarization may be rational.

But, says Kelly, this only works if we aren’t aware that we’re selectively scrutinizing. If I know that I’m doing so, then it shouldn’t be too surprising to find a flaw in the incongruent but not the congruent study (cf. McWilliams 2019). And in fact if I fail to find a flaw in the incongruent study I should lower my confidence in my prior belief, since this suggests that the evidence against it is stronger than I thought (McKenzie 2004). This is an instance of the general point that, without ambiguity, no strategy of evaluating evidence can lead to predictable polarization (Theorem 2.1; see Salow 2018).

And this is where Kelly and I part ways. After all, many of us do realize we’re engaging in selective scrutiny. This often touted as standard scientific practice: scientists adopt a hypothesis and then spend most of their time trying to explain away problems with it (Kuhn 1962; Solomon 1992). Indeed, we are all familiar with the way in which choosing a particular academic track—whether it’s a graduate school to attend, or a project to pursue—has a predictable impact on where we will place our argumentative focus, and thus how our beliefs will evolve (Cook 1987; Cohen 2000). Certainly knowing about selective scrutiny hasn’t stopped me from predictably polarizing as I’ve carried out my research projects—and I suspect it hasn’t stopped you from doing the same as you’ve carried out yours.

So our question: How could knowingly searching for flaws lead rational people to predictably polarize?

My answer: the same way that knowingly searching for words can lead them to do so.

Searching for flaws in a study is a form of cognitive search. It’s easier to recognize that there is a flaw than to recognize that there’s none—the former is an existential task, the latter is a universal one. Indeed, spotting a flaw can happen at any point, just as spotting a word—“Aha, ‘EXTRÆMISM!’”—can. Thus if you are not yet sure there’s a flaw, you should often be unsure whether you should be. And thus if there is a flaw, your confidence that there is should (on average) increase a lot; but if there is no flaw, your confidence that there is one should (on average) decrease only a bit.

In short, my claim is that scrutinizing a bit of evidence leads to the same sort of epistemic structure as our word-search task does (and indeed, that the argument that cognitive search engenders ambiguity in §3.1 applies equally-well to it). As a result, such scrutiny provides evidence that is both valuable and yet expectably polarizing. Which way it’s polarizing depends on what sort of study you’re scrutinizing. If I scrutinize a study suggesting that
owning a gun makes you safer (s), this expectably lowers the rational confidence in s, since becoming sure that it contains a flaw would lower my confidence. If Dan scrutinizes a study suggesting the opposite, that expectably raises the rational confidence in s. (See §4.1 and Appendix C.1 for details.)

Upshot: if—as Kelly suggests—it can be rational to selectively scrutinize incongruent studies, then even if you’re aware you’re doing this, the resulting ambiguity-asymmetries will lead to rational expectable polarization. (As in §3 and §3.1, these updates are fully valuable; and if we allow for consolidations of higher-order uncertainty that are valuable with respect to some question Q (for example, whether guns make you safer, or which direction all the bits of relevant evidence point), this polarization can be predictable and persistent.)

But is it rational to selectively scrutinize incongruent evidence? Wouldn’t it be better to scrutinize the two sides evenly? If I were deciding for my whole epistemic life, then yes: since selectively scrutinizing leads to predictable polarization, a strategy that’s (initially) expected to make me more accurate about s in the long-run would be to evenly scrutinize both sides. But when I’m presented with a pair of conflicting studies, that’s not what I’m deciding. I’m not deciding how to invest my cognitive resources over months or years; I’m deciding how to invest them now (Hedden 2015). Given that, how is it rational for me to decide?

Since both updates are valuable, either way I should expect the evidence it provides to warrant more accurate opinions (about everything, including whether owning a gun makes you safer). Because of this, even if I allow blatantly pragmatic considerations to guide my choice—as some of the literature on motivated reasoning purports to suggest (Kunda 1990; Kahan et al. 2017; Klein 2020)—the process is still arguably epistemically rational.

But I think often more is true. Why am I often more inclined to scrutinize the incongruent study? Because I often think it’s more likely that I’ll be able to find a flaw in the incongruent study—and if I do, I’ll know what to think about it (the evidence becomes unambiguous). I think this in part because I may think it’s more likely to contain a flaw—but even when I don’t think that, I still often think I’ll be more likely to find any flaws that may exist.

Why? Because part of being convinced of a claim is learning how to rebut arguments against it. This very paper can be used to illustrate the point: what put me in a position to write it was figuring out how to rebut objections to the rationality of polarization—that it violated Bayesianism (§2), that it was purely theoretical (§3), that ambiguity wasn’t the driving force (§3.2), and so on. Indeed, 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 arguments that tell against their beliefs than those that support them.

Granting this, will polarization result? Here’s an analogy. Suppose I’ll be presented with a series of pairs of word-search tasks—one following Headser rules, the other following Tailser rules. At each stage I can choose which task to look at. The twist: the Headser tasks use British English, while the Tailser ones use American English. Being an American, I expect to be better at finding words in the latter task than the former. As a result, if at each stage I’m guided by my desire to form accurate beliefs about how that particular coin landed, I’ll tend to do the Tailser tasks more often. And since doing so leads to predictable polarization, I’ll predictably wind up confident that less than half the coins landed heads.

This reasoning is intuitive—but how can we verify it? Simulation. We can randomly
4. COGNITIVE SEARCH EXPLAINS CONFIRMATION BIAS

generate models of cognitive searches for flaws in studies, and examine (1) whether a preference for accurate beliefs could lead people to selectively scrutinize studies in which they think they’re more likely to find any flaws that exist, and (2) whether this preference can indeed lead to predictable polarization. (See §4.1 and Appendix C for details.)

To answer the first question, I randomly generated such models and then measured the correlation between \( \pi(Find|Flaw) \) and the expected accuracy of the update, finding a robust correlation (left side of Figure 5). I then generated pairs of such models in which you’re (on average) more likely to find flaws that exist in the incongruent study than the congruent study, and recorded what proportion of the time expected accuracy favors scrutinizing the former. As the right side of Figure 5 shows, as you get better at scrutinizing incongruent (compared to congruent) studies, the rate of rational selective scrutiny rises quickly.

Figure 5: Accuracy and finding flaws. Left: There’s a robust correlation between \( \pi(Find|Flaw) \) and the expected accuracy of scrutinizing a study. Right: This leads rates of selective scrutiny (y-axis) to grow as the (average) gap in \( \pi(Find|Flaw) \) between incongruent and congruent studies (x-axis) grows.

To answer the second question, I took two groups of agents who each face a series of (randomly-generated) choices of which of two studies to scrutinize. They all start out 50% confident in a target proposition \( q \), and each stage they each scrutinize in the way they expect will make their beliefs more accurate. The only difference is that one group (red) is better at recognizing flaws in studies that tell against \( q \), and the other (blue) is better at recognizing flaws in those that tell in favor of \( q \). The result is polarization (Figure 6).

These results show that the irrationalist interpretations of biased assimilation and motivated reasoning are too quick: rational people who care about the truth but must manage ambiguous evidence will exhibit them. In fact, the hypothesized mechanism of these models fits well with data about a variety of intensifiers and moderators of these effects. It’s built on the idea that people are better at finding flaws in incongruent than congruent arguments. They are.\(^{55}\) It predicts that instructions like “don’t be biases” or “try to be accurate” will not prevent biased assimilation—but that ones that get people to scrutinize both sides equally will. They do.\(^{56}\) And it suggests that biased assimilation will be more extreme when people think harder—when they scrutinize more, rather than less. It is.\(^{57}\)

4. COGNITIVE SEARCH EXPLAINS CONFIRMATION BIAS

Upshot: Insofar as confirmation bias and motivated reasoning helped drive me and Dan apart, these results suggest that rational management of ambiguous evidence may have played an important role.

Nevertheless, these models are built on a difference in our background beliefs and abilities to find flaws. How could such differences predictably emerge, simply by us falling into different social circles? If you'd like to move on to that question, skip to §5; if you'd like to see the details of the argument from this section, read on.

4.1 The formalities

Here I’ll describe some of the details of cognitive-search models; see Appendix C.1 for more.

The model of cognitive search used in these simulations is a generalization of the coarse-grained model of the word-search task from Figure 2. They have the same structure—a class of worlds where you find a flaw, a class where you don’t but there is one, and a class where there is none—but they multiply possibilities within each class to represent situations where the target proposition (s, that owning a gun makes you safer) is either true or false.

Figure 7 presents an example: I’ve been presented with a study favoring s, currently am 25% confident of it, and am scrutinizing that study for flaws. The s_i are possibilities where s is true; the s_i are ones where it’s false. The prior probabilities, π, are captures by the blue numbers next to each possibility, while the posteriors are obtained by Jeffrey-shifting π on the {Find&Flaw, ¬Find&Flaw, ¬Flaw} partition in the way indicated by the labeled arrows coming from the circles. Thus the posterior probability for s is: if I find a flaw (top right), \[ \frac{0.05}{0.05+0.20} = 0.2; \] if there’s a flaw that I don’t find (bottom right), \[ \frac{1}{3} \left( \frac{0.15}{0.05} + \frac{2}{3} \left( \frac{0.05}{0.20} \right) \right) \approx 0.233; \] and if there’s no flaw (bottom left), \[ \frac{2}{3} \left( \frac{0.15}{0.05} \right) + \frac{1}{3} \left( \frac{0.05}{0.20} \right) \approx 0.266. \]

In this model, if the study contains a flaw it’s 20% likely that s is true (\( \pi(s|\text{Flaw}) = 0.2 \)); if it doesn’t, it’s 30% likely (\( \pi(s|\neg\text{Flaw}) = 0.3 \)); and it’s equally likely to contain a flaw as to not (\( \pi(\text{Flaw}) = 0.5 = \pi(\neg\text{Flaw}) \)). But since it’s easier to recognize that there is a flaw than
that there’s not, this update leads to expectable polarization.\textsuperscript{58} For the full details about cognitive search models, see Appendix C.1.

Two final notes. First, to measure accuracy I used the partitional version of the well-known Brier score (Brier 1950; Joyce 2009), which takes the model’s partition of possibilities and calculates the sum of squared distances between the probability of each partition-cell and its truth-value: the inaccuracy of $P$ at $w$ is $B(P, w) = \sum_{x \in W} (1 \{x\}(w) - P_w(x))^2$. This is a measure of inaccuracy, so $1 - B(P, w)$ is our measure of accuracy.

Second, for tractability, the simulations only tracked the agents opinion in $s$ and in the cognitive-searches they were choosing between at any given time—it did not model their evolving opinions about all the cognitive-searches they had done. This is a harmless simplification because a generalization of Theorem 3.1 (which I omit for brevity) guarantees that when we simulate a series of updates using “small-world” models (like Figure 7)—which don’t represent propositions about past or future updates—we can always stitch such models together into a “grand-world” model that captures the (question-relative) value of evidence.

5 Ambiguity Explains Group Polarization

Once Dan and I had different abilities to find flaws in arguments, the cognitive-search mechanism described in the last section could predictably pull us apart. But how could ambiguous evidence start our divergence? The step that made our polarization predictable was when we fell into different social groups. How could this fact lead rational people who have very similar beliefs to predictably diverge?

One mechanism is straightforward. Being in different social groups can incentivize doing different cognitive searches (Kahan et al. 2017; Klein 2020)—when Dan fell in with a crowd of libertarians, that may have incentivized him to search for flaws in pro-government arguments. As we’ve seen, this could lead to predictable polarization despite him reasonably expecting each step in the process to make his beliefs more accurate.

But clearly this isn’t the full explanation. Much of what leads us to polarize is simply the fact that belonging to different groups affects what arguments and information we see.

\textsuperscript{58}$E_w(P(Flaw)) \approx 0.35 > 0.5 = \pi(Flaw)$, and therefore $E_w(P(s)) \approx 0.242 < 0.25 = \pi(s)$. Nevertheless, the update is valuable (with respect to every question), just as with our word-search models.
Groups of libertarians tend to discuss libertarian arguments; groups of liberals tend to discuss liberal ones; both tend to get their news from sources that are sympathetic to their beliefs; hence they diverge. This *group polarization effect* has been widely documented: discussion amongst likeminded people tends to make them more extreme in their opinions (Myers and Lamm 1976; Isenber 1986; Sunstein 2009; Talisse 2019). The central mechanism that drives this effect is unsurprising: people who believe a claim tend to share more arguments that favor it than that disfavor it (Toplak and Stanovich 2003; Wolfe and Britt 2008), and arguments for a given claim tend—that is to predictably persuade people of it (Vinokur and Burstein 1974; Burnstein and Vinokur 1977; Petty and Wegener 1998; Stafford 2015).

This is intuitive enough, so most explanations of the effect stop here. They shouldn’t. The point from §2 applies once again: it’s not just that *someone* can predict that we’ll be persuaded by arguments—it’s that *we ourselves* can predict it. If you are open-minded (more on that in a moment), then if you start reading libertarian arguments every day, you can predict that over the long-run doing so will boost your confidence in libertarian positions. As we know, Theorem 2.1 implies that if the evidence is unambiguous, then rational Bayesians could expect no such thing (Salow 2018). Yet we can.

Everyone needs an explanation of this. One possibility is that we expect to be irrational. But there’s another: perhaps arguments are another source of ambiguity-asymmetries.

Here’s the idea. Suppose you know that someone is about to present you with an argument for $s$: the claim that owning a gun makes you safer. Given your background evidence, that argument will be either *good* or *bad*: if it’s good, it’ll warrant increasing your confidence in $s$ (“Huh, I hadn’t thought of that before”); if it’s bad, it’ll warrant decreasing it (“That’s the best they can come up with?”). You can’t be certain that the argument will be good—for if you were, you should already raise your confidence. Nor will you necessarily be able to be certain whether the argument was good or bad after you’ve been presented with it: the evidence is ambiguous; you can rationally be unsure how you should interpret it. What you can expect is that the arguer will try to make it easier to recognize the reasons favoring their position, and harder to recognize the ones disfavoring it. More generally, there’s a selection effect on arguments: good arguments tend to be made because they are *good*; bad arguments tend to be made because they *sound* good. Since bad arguments don’t tend to wear their badness on their sleeves, they tend to be more ambiguous, i.e. harder to recognize as bad.

Here’s an (overly) simple example. Suppose Kris was hurt, and your libertarian friend is trying to convince you that he didn’t have his gun with him at the time. Contrast two arguments for this claim:

“Every time he went to the North Side, Kris carried a gun. But he got hurt south of town—so he didn’t have it with him.”

“Every time Kris carried a gun, he went to the North Side. But he got hurt south of town—so he didn’t have it with him.”

---

59 Some (e.g. Sunstein 2009) suggest that “social comparison”—adopting the opinions of those in our group, simply because they’re in our group—is also a factor. I set this aside because (1) most agree that arguments explain more of the effect (Isenber 1986), and (2) every study I’ve seen supporting social comparison fails to control for a confound: the opinions of others are a source of evidence (Elga 2007).

60 Although the Value of Evidence allows failures of Reflection, if $\pi$ values $P$ and is *certain* that $P(s) \geq t$, then $\pi(s) \geq t$ (Dorst 2020a).
The latter is valid, the former is not (it’s a version of affirming the consequent). Yet—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. (That’s part of what makes tempting fallacies tempting.) Indeed, there’s empirical 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 is right, and that it generalizes: at least in many contexts, arguments tend to be less ambiguous when they’re good than when they’re bad. Here’s a simple model. When you’re presented with an argument, your confidence that it’s good should (on average) either increase or decrease. To respect the Value of Evidence, it should (on average) increase when it’s good, and (on average) decrease when it’s bad. But the degree to which it should do so is asymmetric: since it’s easier to recognize good arguments than bad ones, your confidence should increase more when it’s good than it should decrease when it’s bad.

What follows? We can simulate it. If we present two groups of people with different arguments like this—one (red) group with arguments supporting $s$, and the other (blue) with arguments opposing $s$—then, as Figure 8 shows, they predictably polarize. (See §5.1 and Appendix C.2 for details.) Thus simply placing me and Dan in different social groups—and thereby presenting us with different arguments—could lead to our predictable polarization.

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 individual agents; thick lines are group averages.

While this all may seem plausible enough, you may be left wondering how it fits with my discussion of biased assimilation. After all, if an argument is bad—it contains a flaw—shouldn’t you be able to find the flaw, and therefore get unambiguous evidence?

This is a good objection; I was stuck on it for an embarrassingly long time. But the answer is simple: it depends on how you engage with the argument. If you engage passively (you don’t scrutinize it), the above model makes sense—with nothing but 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 it), this transforms your epistemic situation into a cognitive-search update. On a natural model, this has the effect of breaking the bad-argument possibilities into two: those in which you find a flaw with it (so get unambiguous evidence), and those in which you fail to find a flaw (and so still have ambiguous evidence).
Thus whenever you are presented with an argument, you face a choice: scrutinize it, or not? Which one you choose affects which direction and to what degree your opinions will predictably shift. To get a sense for how, imagine that two groups of agents are presented with arguments favoring $s$—but one (red) group never scrutinizes them, while the other (blue) group always does. Although there are a variety of ways to parameterize these models (see Appendix C.3), natural settings yield the following results (Figure 9). If blue agents know that they’re not going to find a flaw even if there is one, scrutiny leaves the original model unchanged and so doesn’t affect the polarizing effects of the arguments (top left). If they know they are going to find a flaw if there is one, scrutiny removes all ambiguity, turning the model into a standard-Bayesian one with no expected shifts in opinion (top right). And if there’s a middling chance of finding a flaw, scrutiny dampens the polarization effect (bottom left), and can even lead to polarization in the opposite direction (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 polarization. **Top Right:** 100% chance of finding flaw, if there is one; no polarization. **Bottom:** Middling chance of finding flaw if there is one, with small (left) and large (right) amounts of ambiguity if not—can dampen (left) or reverse (right) polarization.

Upshot: if we always scrutinized arguments and had no self-doubt about our ability to find their flaws, then our evidence would be unambiguous and we would not be predictably persuaded—standard Bayesian models would be right; predictable polarization would be irrational. But since we can’t scrutinize everything—and even if we could, we should still have self-doubts—arguments can rationally, predictably polarize us.

These results show that irrationalist interpretations of the group polarization effect are too quick: people who rationally expect the arguments they’re seeing to make their beliefs
more accurate can nevertheless, over the long run, be predictably polarized by them.

Indeed, this model of how this happens—especially when supplemented with the hypothesis (§4) that people are more likely to scrutinize incongruent than congruent arguments—fits well with a variety of findings from the literature on persuasion. It predicts that there are two routes to engaging with arguments: a passive, low-effort one that moves opinions in a predictable direction; and an active, high-effort one in which the persuasive effects depend massively on the details. There are.\textsuperscript{61} It predicts that people presented with arguments that oppose their beliefs will react differently depending on how they engage with them—and that those who are better at finding flaws may well end up with a more biased assessment of the overall weight of evidence. They do.\textsuperscript{62} And it predicts that manipulating how much scrutiny people give to arguments will have large effects on persuasion—with the biggest effects being on the evaluation of weak, congruent arguments (they’ll be surprised to find flaws) and of strong, incongruent arguments (they’ll be surprised not to find flaws). It does.\textsuperscript{63}

Finally, this model might help make sense of the mixed findings surrounding selective exposure (§4)—the tendency to prefer to look at 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 through-line seems to be that people are more inclined to engage in selective exposure when they expect the arguments to be of high quality (i.e. not to contain obvious flaws), and less inclined otherwise (Frey 1986; Hart et al. 2009). The above model predicts this. When arguments are strong, scrutiny is useless (you don’t think you’ll be able to find a flaw even if there is one); so deciding which argument to look at 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 aren’t strong, scrutiny matters (you expect to find flaws); thus avoiding ambiguity will spur you to look at the arguments you think you’re most likely to find a flaw in—often the incongruent arguments, contra the selective exposure effect.

Much remains to be done. These speculative empirical remarks should be backed up with quantitative studies; these simple theoretical models should be further refined. Nevertheless, what we’ve seen so far this suffices to show that many of the empirically-observed predictably-polarizing effects—including the ones that drove me and Dan apart—are to be expected from rational people who care about the truth but face ambiguous evidence.

\textbf{5.1 The formalities}

Here I’ll sketch the models of arguments and scrutiny; see Appendices C.2 and C.3 for more.

The simple model of arguments divides possibilities into those in which the argument is good ($G$) and those in which it’s bad ($B$). You start out with some prior degree of confidence in each of these. Whichever it is, the posterior rational credence is obtained by Jeffrey-shifting on the \{G, B\} partition—increasing confidence in the true possibility, thus increasing accuracy and satisfying the Value of Evidence. However the degree of these shifts

\textsuperscript{61}Petty 1994; Petty and Wegener 1998; Taber and Lodge 2006; Lundgren and Prislin 1998.


is asymmetric: since good arguments are easier to recognize, the shift is larger if $G$ than if $B$. This model is summarized in Figure 10. (For example, if $\pi(s) = 0.5$, $\pi(G) = 0.5$, $\pi(s|G) = 0.6$, $\pi(s|B) = 0.4$, and $x = 0.4 > 0.1 = y$, then $E_\pi(P(G)) = 0.65 > 0.5 = \pi(G)$ and so $E_\pi(P(s)) = 0.53 > 0.5 = \pi(s)$.)

\[
\begin{array}{cc}
\text{Good (G)} & \text{Bad (B)} \\
\pi(G) + x & \pi(B) + y \\
1 - \pi(G) - x & 1 - \pi(B) - y
\end{array}
\]

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

What about scrutiny? Given an argument-model, the agent chooses whether to update in accordance with it, or instead transform the update by splitting the Bad possibilities into those in which they do vs. do not find a flaw in the argument, as diagrammed schematically (without labeled arrows) in Figure 11. There are a variety choices that must be made about the parameters of these models; see §C.3 for details.

**Figure 11:** Schematic model of the choice of whether to scrutinize an argument.

6 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 left something outside my window. A friend whose story is his own.

We talked about old times. Alec and Kris; history class and lifeguarding; those nights by the lake—and that god-damned bench.

We talked about our lives. How mine has disappeared into academia—finding my passion so quickly, but losing my roots so thoroughly. How his had nearly fallen apart—but then how he’d gotten himself out, gotten it together, gotten it back.

We talked about politics. About Trump and Biden; about Covid and vaccines; about race and politics; about police and guns.
6. A BETTER STORY

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. But what they made clear was how—with his networks of trust, his lived experience, and his (ambiguous) evidence—a person just as bright and well-meaning as I could be led, predictably, to such a different place.

That night, I realized something: I finally believed my own story. You see, the progression of this paper has mirrored the progression of this project. It began as a “flashy” research program: abstract models of rationality made a prediction from pure theory that was borne out by empirical psychology—predictable polarization could result from ambiguous evidence (§2). Though interesting, this was purely theoretical. I eventually managed to turn this result into a concrete example—a word-search task—that both polarized real people, and demonstrated a possibility that many thought impossible: a series of locally optimal steps toward the truth could lead, predictably, to profound and persistent shifts away from it (§3).

It started to look like a promising direction—but there were obstacles. If it were correct, the crucial feature had to be asymmetries in ambiguity (rather than strength) of evidence; so I did an experiment, and found that indeed it was (§3.2). Yet if this was to be an empirically-relevant framework, this had to move outside the lab—the real-world causes of polarization would have to be explicable in terms of evidential ambiguity. So I dove into the literatures on confirmation bias (§4) and group polarization (§5)—and I found that indeed they were.

Then—finally—I got ahold of that friend. And—finally—I believed it. This was no longer just a flashy research program or an interesting theoretical idea or a promising research direction. It was, maybe, true: in the face of ambiguous evidence and self-doubt, concern for the truth might predictably lead rational people apart. And if it were true, it mattered: for how we think about our political opponents, and ourselves; for understanding society’s polarized mix of local conformity with global disunity; and—maybe—for trying to fix it.

There’s much more to the story of how I became convinced—blind alleys I won’t bother sharing, and promising directions that, here, I can’t. And I know that you probably aren’t yet convinced. You shouldn’t be! (Much remains to be done.) But my hope is that by walking you down this road I’ve been on, I’ve at least gotten you to where I began: seeing the ambiguous-evidence theory of rational polarization as an interesting theoretical idea, a promising research direction, and the sort of framework that might bring a useful normative yardstick to the vast empirical literature on predictable polarization.

Of course, you may be skeptical of Dan’s rationality—or of mine. Indeed, I just told you a less-than-fully-objective story about how I predictably became convinced of the rationality of my old friends.

But hold on. Why should that make you skeptical? Four years ago, with an open mind I began a project arguing that predictable polarization could be rational. Predictably, I’ve been convinced that it can be. The processes that convinced me were the ones that drive polarization generally: discussions with those sympathetic to my position; scrutiny of objections from the opposing side; and a search for evidence that confirmed my hypothesis. So if you’re skeptical of my rationality—or of Dan’s—presumably that’s because you buy the standard story that such processes are irrational.
Should you? Notice that there’s an asymmetry between my story and the alternative. The position I end up with is coherent: “Predictable polarization is rational, and I myself came to believe that through rational polarizing mechanisms”—in short, “p, and I’m rational to believe that p”. In contrast, imagine if my story had the opposite conclusion: I had started with an open mind on a project arguing for the irrationality of polarization, and had predictably become convinced. Then my position would be incoherent: “Predictable polarization is irrational, and I myself came to believe that through irrational polarizing mechanisms”—in short, “p, but I’m irrational to believe that p”. This in an epistemically akratic state—one that both common sense (Horowitz 2014) and the Value of Evidence (Dorst 2020a) imply must be irrational. (If you rationally believe that you’re not rational to believe p, you can’t also rationally believe that p.) Upshot: I couldn’t coherently tell the same story with the opposite conclusion.

Now notice: the exact same asymmetry applies to you. Take any one of your beliefs—about religion, politics, philosophy, etc.—that, in hindsight, you recognize as formed by the predictably-polarizing mechanisms we’ve discussed. Perhaps you were raised in a religious household, or went to a liberal university, or worked for a long time on a project arguing for a particular conclusion. Regardless: if you buy the standard story that predictably-polarizing processes are irrational, you’re in trouble—for if you don’t give up your own belief, you’ll incoherently believe “p, but I’ve come to believe that through irrational polarizing mechanisms”. In contrast: if you buy my story, you can acknowledge all the features of your belief—its predictability, its profundity, its persistence, its arbitrariness—while still coherently maintaining it.

That, presumably, is something you’d like.

Well, it’s yours for the taking. Read my arguments with an open mind; engage in selective scrutiny on my behalf; discuss these ideas with those who are sympathetic to them; and you too can become predictably—and rationally!—convinced of rational polarization.64

---

64 Far too many people have helped me with this project for me to properly thank all of them. I received helpful 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, and the University of Lisbon. Bernhard Salow, Jack Spencer, Dmitri Gallow, Roger White, Sally Haslanger, Caspar Hare, Kieran Setiya, and Bob Stalnaker each played a formative role at the start of this project. I received helpful feedback along the way from Martina Calderisi, Agnes Callard, Chris Dorst, Rachel Fraser, Peter Gerdes, Dan Greco, Brian Hedden, Hay Hodges, Michael Hannon, Harvey Lederman, Amnia Loets, Matt Mandelkern, Travis McKenna, Teddy Seidenfeld, Tom Stafford, Mason Westfall, Kevin Zollman, and many others—including many anonymous blog- and social-media commenters. Special thanks to Liam Kofi Bright, Thomas Byrne, Miriam Schoenfield, Ginger Schultheis, and Quinn White for helping me pull the project together—and to some old friends, for reminding me what about it mattered.
A. ANALYTICAL DETAILS

Appendices

A.1 Higher-order probability, Value, and Theorems 2.1 and 2.2

Following the practice of standard epistemic logic (Hintikka 1962; van Ditmarsch et al. 2015), we give a semantics for higher-order probability by using a (finite) structure that can identify higher-order claims (of any order) with events, i.e. sets of worlds, i.e. propositions. A probability frame \( \langle W, \{P^i\}_{i \in \mathbb{N}} \rangle \) consists of 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 \( P^i : W \rightarrow \Delta(W) \). Thus \( 'P^i' \) can be thought of as a description of a probability function—it picks out different such functions in different worlds. In our case, it’ll always be interpreted as ‘the rational credence function (for some particular agent) at some particular time \( i \)’.

Note that \( 'P^i_w' \) is a rigid designator that picks out the (unique) probability function that \( P^i \) associates with a given world \( w \). In cases where we’re only concerned about one description of a probability function, I’ll drop the index and use \( P, P_w, \) etc. In addition, I’ll often enrich the structure of a probability frame with one or more (rigidly designated) probability functions, denoted \( \pi, \delta, \rho, \eta, \) ....

\( W \) is used to represent the propositions in the frame, so for any \( p, q \subseteq W \), \( p \) is true at \( w \) iff \( w \in p; \neg p = W \setminus p, p \land q = p \cap q, p \rightarrow q = \neg p \cup q \) etc. All theorems are restricted to models with finite \( W \)—it’s an open and interesting 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 to be sets of worlds. 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.

Say that \( P \) is possibly ambiguous iff at some world it has higher-order uncertainty, iff there is a world \( w \) and proposition \( q \) such that for all \( t \): \( P_w(P(q) = t) < 1 \); iff there are two probability functions \( \pi \neq \rho \) such that \( P_w(P = \pi) > 0 \) and \( P_w(P = \rho) > 0 \). Since \( W \) is finite, we can think of a probability function just as an assignment of non-negative numbers that sum to 1 to the various worlds. Thus we can diagram probability frames as we did in the main text using Markov diagrams: nodes represent states, and an arrow labeled \( t \) from node \( x \) to node \( y \) represents that \( P_x(y) = t \). For example, Figure 12 represents an unambiguous frame, since the two classes of probability functions (left and right) in the frame do not assign any probability to each other. Meanwhile, Figure 13 represents an ambiguous frame, wherein the left class assigns 0.4 to the right one, and the right one assigns 0.2 to the left one.

\footnote{For more uses of (structures like) probability frames, see Gaifman 1988; Samet 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. See Williamson 2008 and Dorst 2019, 2020b for summaries.}
When is a transition from an initial probability function $\pi$, to a posterior $P$—updating to different probabilities $P_w$ in different worlds $w$—a rational update? If you should expect it to lead you to make better decisions, no matter what decision you face, then I claimed it is. We can formalize this as follows (following Dorst et al. 2021). Given a probability frame $⟨W, P⟩$, let an option $O$ be a function from worlds $w$ to numbers $O(w)$ (so options are random variables), representing the utility that would be achieved by taking option $O$ at $w$. Let a decision problem be a finite set of options $O$. Let a strategy $S$ be a way of choosing options based on $P$’s probabilities, i.e. a function from $w$ to $S_w ∈ O$ such that $S_w = S_x$ whenever $P_w = P_x$. $P$ recommends a strategy $S$ for $O$ if $S$ always selects an option that maximizes expected value according to $P$. Recalling that $E_\pi(O) = \sum_{t ∈ \mathbb{R}} \pi(O = t) \cdot t$ is the expected value of option $O$ according to $\pi$, a strategy is recommended by $P$ iff for all $w$ and $O ∈ O$: $E_{P_w}(S_w) ≥ E_{P_x}(O)$. Then $\pi$ values $P$ if, for every decision problem, the expected value of following a strategy recommended by $P$ is at least as high as taking the option that maximizes expected value according to $\pi$. Abusing notation slightly so that $E_\pi(S)$ is the expected value of following a strategy $S$, i.e. $E_\pi(S) := \sum_w \pi(w)S_w(w)$, that is:

**Value:** $\pi$ values $P$ iff $\forall O$: if $P$ recommends $S$ for $O$, then $\forall O ∈ O$: $E_\pi(S) ≥ E_\pi(O)$.

$\pi$ values $P$ if, for any decision problem, $\pi$ prefers to let $P$ decide on $\pi$’s behalf, rather than make the decision itself.

Dorst et al. 2021 argue that Value is what it takes for $\pi$ to defer to $P$, and give a characterization of what this amounts to (Theorem 5.2). For our purposes, the relevant points are that $\pi$ values $P$ if no (fixed-option\textsuperscript{66}) Dutch book can be constructed for updating from $\pi$ to $P$ (i.e. the transition cannot be made to lead to a sure loss), iff on every generally

---

\textsuperscript{66}A fixed-option Dutch book is one in which options are random variables as above, and thus a situation in which you have the same options at every world. See Dorst et al. 2021, §1.
strictly proper\textsuperscript{67} way of measuring the accuracy of estimates (Schervish 1989; Levinstein 2020; Campbell-Moore 2020), \( \pi \) expects \( P \) to be more accurate than itself.

It’ll expedite the results below to appeal to the details of one part of their characterization. Given a function \( P_w \) that perhaps has higher-order uncertainty (so that \( P_w(P = P_w) < 1 \)), we can considered it’s informed version \( \hat{P}_w \) which removes this higher-order uncertainty, i.e. \( \hat{P}_w := P_w(\cdot | P = P_w) \) (Elga 2013; Stalmaker 2019; Dorst 2019). For example, informing \( \eta \) and \( \delta \) in Figure 13 would generate the frame in Figure 12, since \( \hat{\eta} = \eta(\cdot | P = \eta) = \eta(\cdot | \{a, b\}) = \pi \) and likewise \( \hat{\delta} = \rho \).

Now think of a probability function \( \pi \) 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 probability functions \( \pi_1, \ldots, \pi_n \) is the set of points obtainable by averaging them; formally, \( CH\{\pi_1, \ldots, \pi_n\} = \{\delta : \exists \lambda_i \geq 0 \text{ and } \sum \lambda_i = 1 \text{ such that } \delta = \sum \lambda_i \pi_i\} \). Let \( C_w = \{\rho : P_w(\cdot | P = \rho) > 0\} \) be the set of candidate probability functions that \( P_w \) thinks might be \( P \), and \( C_{w}^- = C_w - \{P_w\} \) be the set of such functions other than \( P_w \). Then say that \( P_w \) is modestly informed \textsuperscript{i}ff it is an average of its informed self \( \hat{P}_w \) along with the other candidates \( C_{w}^- \), i.e. iff \( P_w \) is in the convex hull of \( \{\hat{P}_w\} \cup C_{w}^- \). Finally let \( C_\pi = \{P_w : \pi(P = P_w) > 0\} \) be the set of candidates that \( \pi \) leaves open might be \( P \). Then we have:

**Theorem A.1.** \( \pi \) values \( P \) iff each \( P_w \) in \( C_\pi \) is modestly informed, and \( \pi \) is in the convex hull of \( C_\pi \).

With this result we can easily prove our Theorems 2.1 and 2.2 that ambiguity is necessary and sufficient for expectable polarization.

**Theorem 2.1.** If \( P \) is always unambiguous and \( \pi \) values \( P \), then for all \( q: \pi(q) = \mathbb{E}_\pi(P(q)). \)

**Proof.** If \( \pi \) values \( P \), then each \( P_w \in C_\pi \) is modestly informed and \( \pi \) is in their convex hull. This implies that for any \( P_w \in C_\pi \), \( P_w(\cdot | P = P_w) > 0 \) (Dorst et al. 2021, Lemma 7.2.5). If \( P \) is unambiguous, then each \( P_w \) is certain of what \( P \) is, and since \( P_w(\cdot | P = P_w) > 0 \), it follows that \( P_w(\cdot | P = P_w) = 1 \). Since \( \pi \) is in the convex hull of the \( P_w \), it follows that for each \( P_w \), \( \pi(\cdot | P = P_w) = P_w \). (Taking any \( q \), we know \( \pi(q|P = P_w) = \frac{\pi(q|P = P_w)}{\pi(P = P_w)} = \frac{\lambda_w P_w(q|P = P_w)}{\lambda_w P_w(P = P_w)} = \frac{P_w(q|P = P_w)}{P_w(P = P_w)} = P_w(q|P = P_w) = P_w(q) \), where the third and final equalities comes from the fact that for all \( i \), \( P_i(\cdot | P_i) = 1 \).) It follows immediately that, for any \( q \):

\[
\pi(q) = \sum_{P_w} \pi(\cdot | P = P_w) \pi(q | P = P_w) \quad \text{(Total probability)}
\]

\[
= \sum_{P_w} \pi(\cdot | P = P_w) P_w(q) \quad \text{(Total expectation)}
\]

\[
= \sum_{P_w} \pi(\cdot | P = P_w) E_\pi(P(q) | P = P_w) = E_\pi(P(q)) \quad \text{(Total expectation)}
\]

\textsuperscript{67}A proper scoring rule is one such that every (rigidly designated) probability function expects itself to be at least as accurate as any other. They are the standard ways of measuring (in)accuracy—see e.g. Schervish 1989; Pettigrew 2016; Campbell-Moore 2020; Campbell-Moore and Levinstein 2020; Levinstein 2020.
Recall that we say that $P$ is valuable iff there is some $\rho$ that values it while also leaving open all its potential realizations, i.e. for all $w \in W$: $\rho(P = P^w) > 0$. Then:

**Theorem 2.2.** If $P$ is valuable and possibly ambiguous, there are infinitely many $\pi$ such that $\pi$ values $P$ and yet there is a $q$ for which $E_\pi(P(q)) > \pi(q)$.

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

We begin by showing that there is a $q \subseteq W$ and a $\rho_i$ such that $\rho_i(q) \neq E_{\rho_i}(P(q))$, following Samet 2000, Theorem 5. For reductio, suppose that for all $\rho_i$ and $q$, $\rho_i(q) = E_{\rho_i}(P(q))$. Note that $P$ can be viewed as a finite Markov chain with $W$ the state space and $P_w(w')$ the probability of transitioning from $w$ to $w'$. As such, we can partition $W$ into its communicating classes $E_1, \ldots, E_k$, plus perhaps a set of transient states $E_0$. The claim that, for all $q$, $\rho_i(q) = E_i(P(q))$ is equivalent to the claim that $\rho_i$ is a stationary distribution with respect to the Markov chain, i.e. where $M$ is the transition matrix and $\rho_i$ is thought of as the (row) vector with the $\rho_i(w)$ in each column, $\rho_i M = \rho_i$. By the Markov chain convergence theorem, each $E_1, \ldots, E_k$ has a unique stationary distribution, and every stationary of $M$ assigns 0 probability to $E_0$. These imply, first, that $\pi(E_0) = 0$, for otherwise $\pi$ would not be in the convex hull of the (stationary) $\rho_i$. Since $C_\pi$ includes all realizations of $P$, this implies that $E_0$ is empty. Moreover, the fact that each $E_i$ has a unique stationary, combined with our assumption that all $\rho_i(q) = E_{\rho_i}(P(q))$ implies that for any $w, w' \in E_i$, $P_w = P_{w'}$ and (since $E_i$ is a communicating class) $P_w(E_i) = 1$, which implies that $P$ is not ambiguous after all—contradiction.

Thus we know that there is a $\rho_i$ and $q$ such that $\rho_i(q) \neq E_{\rho_i}(P(q))$. Since $\rho_i(q) < E_{\rho_i}(P(q))$ iff $\pi(-q) > E_{\rho_i}(P(-q))$, WLOG assume $\rho_i(q) < E_{\rho_i}(P(q))$. Equivalently, where $1_q$ is the indicator function of $q$ (1 at $w \in q$, 0 elsewhere), $E_{\rho_i}(P(q) - 1_q) > 0$. We want to show that there are uncountably many $\delta$ such that $\delta$ values $P$ and yet $E_\delta(P(q) - 1_q) > 0$. Pick some $\rho_i$ that maximizes $E_{\rho_i}(P(q) - 1_q)$ 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 $P$ is ambiguous).

Now for any $\epsilon \in [0, 1]$, letting $\eta := (1 - \epsilon)\rho_i + \epsilon\rho_j$, and thinking of $E_{\eta_i}(P(q) - 1_q)$ as a function of $\epsilon$, notice that this function is continuous and non-increasing in $\epsilon$, with maximum $E_{\rho_i}(P(q) - 1_q) > 0$ and minimum $E_{\rho_j}(P(q) - 1_q)$. By the intermediate value theorem, this function must hit every value in between the two—meaning there are uncountably many values of $\epsilon$ such that $E_{\eta_i}(P(q) - 1_q) > 0$. Since each one of these $\eta_i$ are distinct (since $\rho_i \neq \rho_j$), and they are all in the convex hull of $C_\pi$ (since $\rho_i, \rho_j \in C_\pi$), they all value $P$ despite having $\eta_i(q) < E_{\eta_i}(P(q))$.

**A.2 Word-searches, question-relativity, and iteration**

In this subsection, I’ll first show that our two simple word-search models (from §3) are valuable, then introduce question-relative value and show how by iterating such updates we can get predictable, profound, and persistent polarization (Theorem 3.1 and Corollary 3.2).

We first show that the (coarse-grained) model of word-search tasks in Figure 2 (page 19) is such that the prior $\eta$ values the posterior $H$. Let $W = \{n, c, f\}$, where $f = \text{Find}, c =$
A. ANALYTICAL DETAILS

~\text{Find}&\text{Completable}, and n = \neg\text{Completable}. Recall that \eta(f) = \eta(c) = \frac{1}{3} and \eta(n) = \frac{1}{2}, so that (thinking of probability functions as vectors over \text{W}), \eta = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}), while \text{H}_f = (0, 0, 1), \text{H}_c = (\frac{1}{3}, \frac{2}{3}, 0), \text{H}_n = (\frac{2}{3}, \frac{1}{3}, 0). \text{By Theorem A.1, we must show that each } \text{H}_i \text{ is modestly informed and that } \eta \text{ is in their convex hull.}

Note that \text{H}_f(H - \text{H}_f) = 1, so \text{H}_f = \tilde{\text{H}}_f and hence is modestly informed. Meanwhile, \tilde{\text{H}}_c = (0, 1, 0) and \tilde{\text{H}}_n = (1, 0, 0). Thus \frac{1}{2} \tilde{\text{H}}_c + \frac{1}{2} \text{H}_n = \frac{1}{2}(0, 1, 0) + \frac{1}{2}(\frac{2}{3}, \frac{1}{3}, 0) = (0, \frac{3}{6}, 0) + (\frac{1}{3}, \frac{1}{6}, 0) = (\frac{1}{3}, \frac{2}{3}, 0) = \text{H}_c, \text{ so } \text{H}_c \text{ is modestly informed. Similarly, } \frac{1}{2} \text{H}_n + \frac{1}{2} \text{H}_c = \text{H}_n, \text{ so } \text{H}_n \text{ is also modestly informed. Finally note that } \frac{1}{2}(0, 0, 1) + 0(\frac{1}{3}, \frac{2}{3}, 0) + \frac{1}{2}(\frac{2}{3}, \frac{1}{3}, 0) = (0, 0, \frac{1}{2}) + (\frac{2}{3}, \frac{1}{3}, 0) = (\frac{1}{3}, \frac{1}{2}, \frac{1}{2}) = \eta, \text{ so } \eta \text{ is in the convex hull of } \{\text{H}_f, \text{H}_c, \text{H}_n\}. \text{By Theorem A.1, this shows that } \eta \text{ values } \text{H}.

Next turn to the model of the word-search task in Figure 1 (page 18). Recall that there are four possibilities, label them \text{W} = \{n, c, o, f\}, for not \text{Completable}, \text{Completable}, \neg\text{Find}, \text{Obvious}, \text{\& Find}, \text{and Find}, \text{respectively. Recall that } \eta = (\frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}), \text{while:}

\begin{align*}
\text{H}_f &= (0, 0, 0, 1); \\
\text{H}_o &= (0, 0, 1, 0); \text{ and} \\
\text{H}_c &= \text{H}_n = (\frac{2}{3}, \frac{1}{3}, \frac{1}{6}, 0).
\end{align*}

Clearly \text{H}_f \text{ and } \text{H}_o \text{ are modestly informed. Note that } \tilde{\text{H}}_c = \tilde{\text{H}}_n = (\frac{2}{3}, \frac{1}{3}, 0, 0), \text{ and that } \\
\frac{2}{6} \tilde{\text{H}}_c + \frac{1}{6} \text{H}_o = \frac{2}{6}(\frac{1}{3}, \frac{1}{3}, 0, 0) + \frac{1}{6}(0, 0, 1, 0) = \frac{4}{6} \frac{1}{6}, 0, 0) + (0, 0, \frac{1}{6}, 0) = (\frac{2}{3}, \frac{1}{6}, \frac{1}{6}, 0) = \text{H}_c, \text{ so } \text{H}_c \text{ (and likewise } \text{H}_n) \text{ is modestly informed. Finally, notice that } \frac{1}{2} \text{H}_f + \frac{1}{2} \text{H}_n = \frac{1}{4}(0, 0, 0, 1) + \\
\frac{1}{3}(\frac{2}{3}, \frac{1}{2}, \frac{1}{2}, 0) = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}) = \eta, \text{ so } \eta \text{ is in the convex hull of } \{\text{H}_n, \text{H}_c, \text{H}_o, \text{H}_f\}. \text{By Theorem A.1, this shows that } \eta \text{ values } \text{H}.

I now turn to question-relative value, and how this can be combined with higher-order consolidations to iterate the (coarse-grained) Headser/Tailser models to lead to predictable and persistent polarization.

Let a question \text{Q} be a partition of \text{W}, \text{and let } \text{Q}(w) = \text{the partition-cell of } w \in \text{W}. \text{A proposition } p \subseteq \text{W} \text{ is about } \text{Q} \text{ iff } p = \bigcup_i q_i \text{ for } q_i \in \text{Q}, \text{i.e. iff } p \text{ is a partial answer to the question } \text{Q} \text{(Hamblin 1976; Roberts 2012)}. \text{Recall that a decision problem } \text{O} \text{ is any set of options (i.e. random variables—functions from worlds to numbers) on } \text{W}. \text{Say that an option } \text{O} \text{ is } \text{Q-measurable} \text{ iff } \text{Q} \text{ settles the value of } \text{O}, \text{i.e. for all } w, w', \text{if } \text{Q}(w) = \text{Q}(w'), \text{then } \text{O}(w) = \text{O}(w'). \text{Say that } \text{O}_\text{Q} \text{ is a decision about } \text{Q} \text{ iff each of its options is } \text{Q-measurable. Then: } \pi \text{ values } \text{P} \text{ with respect to } \text{Q} \text{ iff it prefers to let } \text{P} \text{ decide any decision about } \text{Q}:

Value wrt \text{Q}: \pi \text{ values } \text{P} \text{ with respect to } \text{Q} \text{ iff: for every decision problem } \text{O}_\text{Q} \text{ about } \text{Q}, \text{if } \text{P} \text{ recommends } \text{S} \text{ for } \text{O}_\text{Q}, \text{then } \forall \text{O} \in \text{O}_\text{Q} : \text{E}_\pi(S) \geq \text{E}_\pi(O).

\pi \text{ values } \text{P} \text{ with respect to } \text{Q} \text{ iff, for any decision about } \text{Q}, \text{it prefers to let } \text{P} \text{ decide on } \pi's \text{ behalf, rather than make the decision itself.}

Note that one decision about \text{Q} is to choose a credence in any proposition \text{p} \text{ about } \text{Q}, and then have that credence scored using any proper scoring rule (Dorst et al. 2021, §3); therefore to for \pi to value \text{P} with respect to \text{Q}, it must expect \text{P} to be more accurate than itself on any (set of) opinion(s) about \text{Q}.

Given this, I’ll now turn to proving that updates that are valuable with respect to \text{Q} can nevertheless lead to predictable, profound polarization about \text{Q}:
A. ANALYTICAL DETAILS

Theorem 3.1. 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 some $q_i \in Q$) such that, as $n \to \infty$:

- $H^0$ is (correctly) certain that $\overline{H}$ values $H^{i+1}$, for each $i$;
- $H^0$ is (correctly) certain that $H^i$ values $\overline{H}$ with respect to $Q$, for each $i$;
- $H^0(h) \approx \frac{1}{2}$ and $H^0(H^n(h)) \approx \frac{1}{2}$; and
- $H^0(\overline{H}^n(h) \approx 1)$ \approx 1.

I’ll proceed in stages. First, I’ll specify a model that iterates (coarse-grained) word-search tasks and consolidates higher-order uncertainty along the way. I’ll then prove that each substantive update is valuable period, while each consolidation update is valuable with respect to each. I’ll then show the long-run predictable behavior of the final rational credence $H^n$ in this model, and show that the proposition $h = \text{more than half the coins landed heads}$ is one on which $H^n$ is predictably polarized. Afterwards, we will add a Tailser to the model and show that such polarization is also persistent.

**Definition of iteration model.** Consider Haley the Headser, who 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 doubts.

For each task $i = 1, \ldots, n$, let $X_i = \{n_i, n'_i, c_i, c'_i, f_i\}$ be the set of outcomes. $f_i$ indicates that she finds the completion, $c_i$ and $c'_i$ are where it’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 $W = X_1 \times \ldots \times X_n$ 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 $i$th 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 the sequences of heads and tails—is about $Q$. Finally let $U_i$ be the question of how the rational credence updated at $i$, so $U^t = \{U^t_n, U^t_c, U^t_f\}$ where

\[
U^t_n = \{w \in W : n_i \in w \text{ or } c'_i \in w\}
\]

\[
U^t_c = \{w \in W : c_i \in w \text{ or } n'_i \in w\}
\]

\[
U^t_f = \{w \in W : f_i \in w\};
\]
A. ANALYTICAL DETAILS

A few more bits of notation. Given a probability function $\pi$, 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 $\pi$ 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 $p \subseteq W$:

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

Finally, higher-order consolidations will happen not by conditioning but by imaging (Lewis 1976), so we’ll need to define a corresponding selection function (Stalnaker 1968). For each world $w \in W$, let $g_w : \wp(W) \to W$ be a selection function which is given a proposition $p$ and outputs a world $g_w(p) \in p$ (whenever $p \neq \emptyset$) that is the ‘closest’ one to $w$ where $p$ is true. We assume $g$ obeys three constraints:

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

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

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

Machinery in place, we can define the series of probability functions $H^0, \overline{H}^0, H^1, \overline{H}^1, \ldots, H^n, \overline{H}^n$ that represent Haleys rational opinions over time. ($H^i$ is that right after completing the $i$th 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 defined so 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 $q_{i_1}, \ldots, q_{i_k}$ in $Q_{i_1}, \ldots, Q_{i_k}$ respectively, $H^0_w(q_{i_1}, \ldots, q_{i_k}) = H^0_w(q_{i_1})H^0_w(q_{i_2}) \cdots H^0_w(q_{i_k})$. Since $H^0_w(U) = 0$, this pins down $H^0_w$ uniquely over $W$, hence all worlds begin with the same prior.

Now we define the updates as follows. For any world $w$ and task $i$, the consolidation $\overline{H}^i_w$ is obtained by imaging on the proposition that the $H^i$ equals the particular function $H^i_w$. 

41
A. ANALYTICAL DETAILS

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 first-order opinions about \( Q \).

Finally, we define the regular (non-consolidation) updates as Jeffrey-shifts in exactly the way indicated by the (coarse-grained) word-search model, except that \( c'_i+1 \) and \( n'_i+1 \) (the ones initially assigned 0 probability) update in a way 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 \( H^{i+1}_w = \overline{H}^i_w[\frac{1}{3}, 2, 0]_{i+1} \);
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 and what the long-run behavior of \( H^n \) is.

**Lemma 3.1.1.** (1) For each \( i \) and \( w \): \( \overline{H}^i_w(x) \) is higher-order certain. (2) Moreover, for \( i > 1 \), if \( H^i_w(x) > 0 \), then \( \overline{H}^{i-1}_w = \overline{H}^i_w \).

**Proof.** (1) Suppose \( \overline{H}^i_w(x) > 0 \). 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 \( \overline{H}^i_w(x|H^i = H^i_w) = H^i_w(x|H^i = H^i_w) = \overline{H}^i_w \). Since \( x \) was arbitrary, \( \overline{H}^i_w(\overline{H}^i_w) = 1 \).

(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), \( \overline{H}^{i-1}_w = \overline{H}^i_w \). 

Now we show that weird updates are always assigned probability 0 ahead of time:

**Lemma 3.1.2.** 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 \( \overline{H}^i_w \). **Induction:** Supposing it holds for all \( w \) with \( k < i \), we show it holds for \( i \). Since \( H^i_w = \overline{H}^{i-1}_w[1, 0] \), 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 \( y \in U^i_y \), or (ii) \( w \in U^i_y \) 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 \( y' = (y_1, ..., y'_n) \)—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:
A. ANALYTICAL DETAILS

Lemma 3.1.3. For any \( x, i \): if \( H_x^i(y) > 0 \), then \( g_y(H^i = H_x^i) \in Q(y) \).

Proof. By Lemma 3.1.1 (2), \( \overline{H_x^{i-1}} = \overline{H_y^{i-1}} \). By Lemma 3.1.2 and the fact that \( H_x^i \) preserves \( H_y^{i-1} \)'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 \( H_y^i = H_x^i \), meaning that by Strong Centering \( g_y(H^i = H_x^i) = y \), establishing the result.

If \( c_i \in x \) or \( n_i' \in x \), then \( H_x^i = \overline{H_y^{i-1}}[\frac{1}{3}, \frac{2}{3}, 0]_i \). If \( c_i \in y \), then \( H_y^i = H_x^i \), 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, of course \( Q(y') = Q(y) \), and also \( \overline{H_y^{i-1}} = \overline{H_y^i} = \overline{H_x^{i-1}} \), so

\[
H_y^i = \overline{H_y^{i-1}}[\frac{1}{3}, \frac{2}{3}, 0]_i = \overline{H_x^{i-1}}[\frac{1}{3}, \frac{2}{3}, 0]_i = H_x^i.
\]

Thus there is a \( y' \) in \( \{H^i = H_x^i\} \) such that \( Q(y') = Q(y) \), so by \( Q \)-Respecting \( g_y(H^i = H_x^i) \in Q(y) \), establishing the result.

If \( n_i \in x \) or \( c_i' \in x \), parallel reasoning (substituting \( c_i' \) for \( c_i \)) establishes the result. \( \square \)

Lemma 3.1.4. For all \( x, i \) and \( q \in Q, \overline{H_x^i}(q) = H_x^i(q) \).

Proof. By construction:

\[
\overline{H_x^i}(q) = H_x^i(q|H^i = H_x^i) = \sum_{y \in q} H_x^i(y|H^i = H_x^i) = \sum_{y \in q} \sum_{z \in W: g_z(H^i = H_x^i) = y} H_x^i(z).
\]

But by Lemma 3.1.4, all and only worlds in \( q \) are mapped to worlds in \( y \) by imaging on \( H^i = H_x^i \), hence the above sum equals \( \sum_{y \in q} H_x^i(y) = H_x^i(q) \), as desired. \( \square \)

Lemma 3.1.5. For any \( w, i < j \), \( \overline{H_w^i}(F_j) = \overline{H_w^i}(C_j) = \frac{1}{4} \) and \( \overline{H_w^i}(N_j) = \frac{1}{2} \) and \( \overline{H_w^i} \) treats the \( Q_i \) as mutually independent.

Proof. By induction. Base case: trivial by definition of \( H_w^0 \). Induction step: Suppose it holds for \( k < i \). By definition, \( H_w^i \) is obtained by Jeffrey-shifting \( \overline{H_w^{i-1}} \) on \( Q_i \), so since by the induction hypothesis \( \overline{H_w^{i-1}} \) treats the \( Q_i \) as mutually independent and assigns \( \frac{1}{4} \) to \( F_j \) and \( C_j \), and \( \frac{1}{2} \) to \( N_j \), \( H_w^i \) does too. Now by Lemma 3.1.4, \( \overline{H_w^i} \) maintains the same distribution over \( Q \) as \( H_w^i \) has, establishing the result. \( \square \)

Lemma 3.1.6. For all \( w \) and \( i \), \( \overline{H_w^i} \) values \( H_w^{i+1} \).

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

(1) Taking an arbitrary $x \in S^i_w$, we show that $H^{i+1}_x$ is modestly informed. By Lemma 3.1.1 (1), note that since $H^{i+1}_x(\pi)$ is higher-order certain, $H^{i+1}_x = H^{i+1}_w$. 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 $H^{i+1}_x = H^{i+1}_x(\xi) = 1$ 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 $H^{i+1}_x = H^{i+1}_w(\xi) = H^{i+1}_w(0,1,i+1)$—label this function $\pi_c$. If (iii) holds, then $H^{i+1}_x = H^{i+1}_w(0,1,i+1)$—label this function $\pi_n$. Note that $\pi_c$ and $\pi_n$ both assign 1 to $S^i_w$, and also assign 1 to $|H^{i+1} = \pi_c| \lor |H^{i+1} = \pi_n|$. Now, since by Lemma 3.1.2 we have that $H^{i+1}_w$ assigns 0 to any world with $n_{i+1}'$ or $c_{i+1}'$ in it, it follows that $\pi_c$ and $\pi_n$ do too, and hence that:

$$\hat{\pi}_c = \pi_c(\pi^{i+1} = \pi_c) = H^{i+1}_w(\pi^{i+1} = \pi_c)$$

$$\hat{\pi}_n = \pi_n(\pi^{i+1} = \pi_n) = H^{i+1}_w(\pi^{i+1} = \pi_n)$$

From this it follows that $\pi_c$ (and, by parallel reasoning, $\pi_n$) is modestly informed, since:

$$\frac{1}{2} \hat{\pi}_c + \frac{1}{2} \hat{\pi}_n = \frac{1}{2} H^{i+1}_w(\pi^{i+1} = \pi_c) + \frac{1}{2} \left( \frac{1}{3} H^{i+1}_w(\pi^{i+1} = \pi_c) + \frac{2}{3} H^{i+1}_w(\pi^{i+1} = \pi_n) \right)$$

$$= \frac{1}{2} H^{i+1}_w(\pi^{i+1} = \pi_c) + \frac{1}{2} H^{i+1}_w(\pi^{i+1} = \pi_n) + \frac{1}{2} H^{i+1}_w(\pi^{i+1} = \pi_n)$$

$$= \frac{2}{3} H^{i+1}_w(\pi^{i+1} = \pi_c) + \frac{1}{3} H^{i+1}_w(\pi^{i+1} = \pi_n)$$

Since $\pi_c$, $\pi_n$, and $H^{i+1}_w(\pi^{i+1} = \pi)$ are the three realizations of $H^{i+1}$ in $S^i_w$, this establishes (1).

(2) We now show that $H^{i+1}_w$ is in their convex hull. Note that by Lemma 3.1.5 and total probability,

$$H^{i+1}_w = \frac{1}{2} H^{i+1}_w(\pi^{i+1} = \pi_c) + \frac{1}{2} H^{i+1}_w(\pi^{i+1} = \pi_n) + \frac{1}{2} H^{i+1}_w(\pi^{i+1} = \pi_n)$$

Now notice that:

$$\frac{1}{4} H^{i+1}_w(\pi^{i+1} = \pi_c) + \frac{3}{4} \pi_n = \frac{1}{4} H^{i+1}_w(\pi^{i+1} = \pi_c) + \frac{3}{4} \left( \frac{1}{3} H^{i+1}_w(\pi^{i+1} = \pi_c) + \frac{2}{3} H^{i+1}_w(\pi^{i+1} = \pi_n) \right)$$

$$= \frac{1}{4} H^{i+1}_w(\pi^{i+1} = \pi_c) + \frac{1}{4} H^{i+1}_w(\pi^{i+1} = \pi_n) + \frac{1}{4} H^{i+1}_w(\pi^{i+1} = \pi_n) = H^{i+1}_w$$

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

Corollary 3.1.7. For all $w, i$: $H^i_w$ values $H^i$.

Proof. For $i = 0$, this is trivial. For $i > 0$, by construction, $H^i_w(x) > 0$ only if $H^{i-1}_w(x) > 0$, and by Lemma 3.1.6, this implies that $H^i_w$ is modestly informed. Since $H^i_w(\pi^i = H^i_w) > 0$, trivially $H^i_w$ is in the convex hull of the realizations of $H^i$ it leaves open. Thus by Theorem A.1, the result holds.

Lemma 3.1.8. For all $x, i$: $H^i_x$ values $H^i$ with respect to $Q$. 

44
A. ANALYTICAL DETAILS

Proof. By Lemma 3.1.4, for any \( q \in Q \): \( H'_x(\overline{H}(q)) = H'(q) = 1 \). It follows that for any decision-problem \( O_Q \) based on \( Q \), \( H' \) recommends strategy \( S \) for \( O_Q \) iff \( \overline{H}' \) recommends \( S \) for \( O_Q \). Since, by Corollary 3.1.7, \( H'_x \) values \( H' \), it follows immediately that \( H'_x \) values \( \overline{H}' \) with respect to \( Q \).

Lemmas 3.1.6 and 3.1.8 establish the first two bullet-points of Theorem 3.1; we now focus on establishing the second two.

Recall that \( h = more than half the coins land heads \) is a proposition about \( Q \), and that for each \( Heads_i = F_i \cup C_i \), \( H'(Heads) = \frac{1}{2} \), 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 \( \frac{1}{2} \), the third bullet-point follows: \( H^0(h) \approx \frac{1}{2} \), and since \( H^0 \) is higher-order certain, \( H^0(\overline{H}(\# h) = 1) \approx 1 \).

To establish the final bullet-point, that \( H^0(\overline{H}(\# h) = 1) \approx 1 \), we establish the long-run behavior of \( H^n \) (which, by Lemma 3.1.4, establishes it for \( \overline{H^n} \)).

**Lemma 3.1.9.** With \( Heads_i = F_i \cup C_i \), we have, for all \( w, i \), \( H^0_w \) assigns probability 1 to:

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

Proof. Combining Lemma 3.1.5 with the definition of the update, we know immediately that \( H^0_w \)'s distribution over the partition \( \langle F_i, C_i, N_i \rangle \) follows the above pattern:

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

Since \( H^0(U) = 0 \), so \( H^0_w \) 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. By Lemma 3.1.5, each \( \overline{H} \) treats the \( Q_k \) as mutually independent, so by definition none of the later Jeffrey-shifts—for \( j \geq i \), the update from \( \overline{H} \) to \( \overline{H}^{j+1} \) change the probabilities in \( Q_k \). And by Lemma 3.1.4, none of the consolidations (from \( \overline{H} \) to \( \overline{H}^i \)) 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 3.1.10.** For any \( \epsilon > 0 \), as \( n \to \infty \), \( H^0(H^n(h) \geq 1 - \epsilon \to 1 \).

Proof. Choosing an arbitrary \( \epsilon > 0 \), let \( x \approx y \) mean that \( x \) is within \( \epsilon \) of \( y \). Sort the time indices into 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 that the \( Q_i \) as i.i.d. with \( H^0(F_i) = H^0(C_i) = \frac{1}{2} \), by the law of large numbers, as \( n \to \infty \), \( H^0(|I_F| \approx \frac{n}{4} & |I_C| \approx \frac{n}{4} \) \& \( |I_N| \approx \frac{n}{2} \)) \to 1. We want to show what follows if this obtains, so suppose it does: \( |I_F| \approx \frac{n}{4} \) \& \( |I_C| \approx \frac{n}{4} \) \& \( |I_N| \approx \frac{n}{2} \). What is true of \( H^n? \) We have from Lemma 3.1.9 that:

- For all \( i \in I_F \), \( H^n(Heads) = 1; \)
- For all \( i \in I_C \), \( H^n \) treats \( Heads_i \) as i.i.d. with \( H^n(Heads) = \frac{2}{3}; \) and
A. Analytical Details

For all \( i \in I_N \), \( H^\alpha \) treats Heads\(_i\) as i.i.d. with \( H^\alpha(\text{Heads}_i) = \frac{1}{3} \). Thus by the weak law of large numbers, as \( n \to \infty \) we have:

\[
H^n\left( \sum_{i \in I_F} \frac{\mathbb{1}_{\text{Heads}_i}}{|I_F|} = 1 \right) = 1 \quad (\alpha)
\]

\[
H^n\left( \sum_{i \in I_C} \frac{\mathbb{1}_{\text{Heads}_i}}{|I_C|} \approx \frac{2}{3} \right) \to 1 \quad (\beta)
\]

\[
H^n\left( \sum_{i \in I_N} \frac{\mathbb{1}_{\text{Heads}_i}}{|I_N|} \approx \frac{1}{3} \right) \to 1 \quad (\gamma)
\]

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

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

And therefore, recalling that \( h = \text{more than half the tosses land heads} \):

\[
H^n\left( \sum_{i=1}^n \frac{\mathbb{1}_{\text{Heads}_i}}{n} > \frac{1}{2} \right) = H^n(h) \approx 1
\]

Since this follows from \( |I_F| \approx \frac{n}{4} \) \& \( |I_C| \approx \frac{n}{4} \) \& \( |I_N| \approx \frac{n}{4} \), 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. \( \square \)

This completes the proof of Theorem 3.1: Lemma 3.1.6 establishes the first bullet-point, Lemma 3.1.8 establishes the second, the reasoning on page 45 establishes the third, and Lemma 3.1.10 establishes the fourth.

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

**Corollary 3.2.** There are two sequences of probability functions \( H^0, \overline{H}^0, \ldots, \overline{H}^\alpha \) and \( T^0, \overline{T}^0, \ldots, \overline{T}^\alpha \), 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 \( i \):
  - \( H^\alpha \) values \( H^{i+1} \) and \( T^{i+1} \);
  - \( H^0 \) values \( H^{\bar{i}} \) with respect to \( Q \), and \( T^{\bar{i}} \) values \( T^{\bar{i}} \) with respect to \( Q \); and
  - \( H^0 = T^0 \), and in particular \( H^0(h) = T^0(h) \approx \frac{1}{2} \); yet
  - \( H^0 \) and \( T^0 \) are both arbitrarily confident of \( \overline{H}^\alpha(h) \approx 1 \) and \( \overline{T}^\alpha(h) \approx 0 \); and
  - \( H^0 \) and \( T^0 \) are arbitrarily confident of \( \overline{H}^\alpha(h \mid \overline{H}^\alpha(h) \approx 0) \approx 1 \) and \( \overline{T}^\alpha(h \mid \overline{T}^\alpha(h) \approx 1) \approx 0 \).

**Proof.** All but the final bullet-point are straightforward generalizations of the proofs of Theorem 3.1, gotten by dividing possibilities further to track which updates \( T^\alpha \) goes through, consolidating throughout the process in a way that maintains opinions about \( Q \), and adding the partitions \( Q_i^t = \{ F_i^t, C_i^t, N_i^t \} \), where \( F_i^t \cup C_i^t = N_i^t \) and \( N_i^t = F_i \cup C_i \). By doing so, we create a model in which both \( H^\alpha \) and \( T^\alpha \) are (correctly) certain that:
A. ANALYTICAL DETAILS

\[ F_i \& N_i^n \rightarrow \left( H^n(\text{Heads}_i) = 1 \& T^n(\text{Heads}_i) = \frac{2}{3} \right) \]
\[ C_i \& N_i^n \rightarrow \left( H^n(\text{Heads}_i) = \frac{2}{3} \& T^n(\text{Heads}_i) = \frac{2}{3} \right) \]
\[ N_i \& C_i^n \rightarrow \left( H^n(\text{Heads}_i) = \frac{1}{4} \& T^n(\text{Heads}_i) = \frac{3}{4} \right) \]
\[ N_i \& F_i^n \rightarrow \left( H^n(\text{Heads}_i) = \frac{1}{4} \& T^n(\text{Heads}_i) = 0 \right) \]

with \( H^n \) and \( T^n \) treating the \( \text{Heads}_i \) as mutually independent. Moreover, \( H^0 = T^0 \), and both treat the \( Q_i \) as mutually independent, as well as the \( Q'_i \), assigning e.g.:

\[ H^0(F_i) = H^0(C_i) = \frac{1}{4}, \text{ while } H^0(N_i) = \frac{3}{4} ; \text{ and } \]
\[ H^0(F_i') = H^0(C_i') = \frac{1}{4}, \text{ while } H^0(N_i') = \frac{1}{2}. \]

By reasoning parallel to that in Lemma 3.1.10, as \( n \to \infty \) both \( H^0 \) and \( T^0 \) become arbitrarily confident that

\[ H^n \left( \sum_{i=1}^{n} \frac{\text{Heads}_i}{n} \right) \approx \frac{7}{12} \]
and so \( H^n(h) \approx 1 \),

and that

\[ T^n \left( \sum_{i=1}^{n} \frac{\text{Heads}_i}{n} \right) \approx \frac{5}{12} \]
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 \( I_F = \{ i : Q_i = F_i \} \), etc.) \( |I_F| \approx \frac{n}{2} \& |I_C| \approx \frac{n}{4} \& |I_{F'}| \approx \frac{n}{4} \& |I_{C'}| \approx \frac{n}{8} \). Supposing this obtains, we show that the resulting polarization on \( H^n \) (and, by parallel reasoning, \( T^n \)) is persistent—which suffices to show that it is predictable, profound, and persistent.\(^{68}\)

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

i) For all \( i \in I_F \), since \( H^n(F_i) = 1 \), that \( H^n(T^n(\text{Heads}_i) = \frac{2}{3}) = 1 \).

Therefore, \( H^n \left( \sum_{i \in I_F} \frac{T^n(\text{Heads}_i)}{|I_F|} \right) = \frac{2}{3} \) = 1

ii) For all \( i \in I_C \), since \( H^n(C_i) = \frac{2}{3} \) and \( H^n(N_i) = \frac{1}{4} \), so \( H^n(N_i \& F_i) = H^n(N_i \& C_i) = \frac{1}{6} \), we have: \( H^n(T^n(\text{Heads}_i) = \frac{2}{3}) = \frac{2}{3}, H^n(T^n(\text{Heads}_i) = 0) = \frac{1}{6} \), and \( H^n(T^n(\text{Heads}_i) = \frac{1}{3}) = \frac{1}{6} \).

Therefore, if \( \pi = H^n \), for all \( i \in I_C \), \( \mathbb{E}_\pi(T^n(\text{Heads}_i)) = \frac{2}{3} \left( \frac{2}{3} \right) + \frac{1}{4} \left( \frac{1}{6} \right) = \frac{1}{2} \). Since \( H^n \) treats the \( T^n(\text{Heads}_i) \) as independent, by the weak law of large numbers, as \( n \to \infty \), \( H^n \left( \sum_{i \in I_C} \frac{T^n(\text{Heads}_i)}{|I_C|} \right) \approx \frac{1}{2} \) → 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 \( H^n(N_i \& F_i) = H^n(N_i \& C_i) = \frac{1}{3} \), we have: \( H^n(T^n(\text{Heads}_i) = \frac{2}{3}) = \frac{1}{3}, H^n(T^n(\text{Heads}_i) = 0) = \frac{1}{3}, \text{ and } H^n(T^n(\text{Heads}_i) = \frac{1}{3}) = \frac{1}{3}. \)

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

\(^{68}\)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.
B. EXPERIMENTAL DETAILS

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

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

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

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

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 $H^n(T^n(\sum_{i=1}^{n} \frac{1_{\text{Heads}_i}}{n} > \frac{1}{2}) \approx 0) \to 1$, 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 $\overline{T}(h) = H^n(h)$ and $\overline{T}(h) = T^n(h)$, 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 $\overline{T}(h|\overline{T}(h) \approx 1) \to \overline{T}(h) \to 0$, completing the proof. \hfill \square

B  Experimental Details

This appendix contains the details of the experiment discussed in §3.2.

250 participants were recruited through Prolific (107 F/139 M/4 Other; mean age = 27.06; pre-registration here: https://aspredicted.org/8jg3e.pdf). Subjects were (pseudo)randomly divided into Ambiguous (A) and Unambiguous (U) conditions. Within each condition, they were further (pseudorandomly) divided into “Headsers” and “Tailsers”. I will abbreviate the groups “A-Hsers”; “A-Tsers”; “U-Hsers”, and “U-Tsers”. Each group was told they’d be given evidence about a series of independent, fair coin tosses. Both groups were given standard instructions about how to use a 0–100% scale to rate their confidence in the answer to a yes/no question, and then given specialized instructions.

The A group was informed about how word-search tasks work, and given three examples (‘P_A_ET’ [planet], ‘CO_R_D’, [uncompletable] and ‘E_RT’ [heart]). The A-Hsers were instructed that they’d see a completable string if the coin landed heads, and an uncompletable if it landed tails. The A-Tsers were instructed vice versa.

The U group was given instructions about how the urn task worked. 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.

I made several 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 participants’ judgments—and (2) failing to plan both the ANOVA and difference-of-difference confidence intervals that could further confirm my second prediction. The main text reported the results after correcting this; here I report both pre-registered and post-hoc tests (the upshots are the same).
B. EXPERIMENTAL DETAILS

Both groups saw four independent tasks, and were asked before and afterwards how confident they were in the outcome.\textsuperscript{70} The pre-task question was an attention-check, wherein they were instructed to move the slider to 50%; it was pre-registered that I would exclude participants who failed two or more of these attention-checks. In total, 25 of 250 participants were excluded in this way.

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 group 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 all 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 what was reported in the main text in §3.2, page 21.)

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

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

Predictions 1, 2, 3, 5, and 6 were confirmed with statistically significant results; Prediction 4 had the divergence in the correct direction but it was not statistically significant. Plots of prior and posterior mean confidences in each group, along with 95\% confidence intervals, displayed in Figure 14:

In more detail: one-sided paired t-test for Prediction 1 indicated that A-Hser priors (M = 50.35, SD = 3.26) were lower than A-Hser posteriors (M = 57.71, SD = 30.33) with \( t(219) = 3.58, p < 0.001, d = 0.341 \). One-sided paired t-test for Prediction 2 indicated that A-Tser posteriors (M = 36.29, SD = 31.04) were lower than A-Tser priors (M = 49.60, SD = 2.90), with \( t(191) = 5.90, p < 0.001, d = 0.604 \). And one-sided independent samples t-test for Prediction 3 indicated that A-Hser posteriors (M = 57.71, SD = 30.33) were higher than A-Tser posteriors (M = 36.29, SD = 31.04), with \( t(410) = 7.07, p < 0.001, d = 0.699 \).

Meanwhile, one-sided independent samples t-test for Prediction 4 failed to indicate that A-Hser posteriors (M = 57.71, SD = 30.33) were higher than U-Hser posteriors (M = 54.64, 70The A-group was asked how confident they were that “that the string is completable”—equivalent to “coin landed heads” for A-Hsers, and “coin landed tails” for A-Tsers. The U-group was asked how confident they were that the coin landed heads (U-Hsers) or tails (U-Tsers).
B. EXPERIMENTAL DETAILS

Figure 14: Mean prior and posterior confidence in heads in A (left) and U (right) conditions. Bars represent 95% confidence intervals.

SD = 26.93), with \( t(441) = 1.15, p = 0.125, d = 0.107 \). But one-sided independent samples t-test for Prediction 5 indicated that A-Tser posteriors (M = 36.29, SD = 31.04) were below U-Tser posteriors (M = 48.10, SD = 28.47), with \( t(393) = 4.07, p < 0.001, d = 0.398 \).

Prediction 6 was (due to my oversight) handled poorly at the pre-registration stage—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, I later realized, was to do (a) a 2 × 2 ANOVA, and (b) an empirically bootstrapped 95% confidence interval for the difference between the differences between A-Hsers/A-Tsers and U-Hsers/U-Tsers.

(a) Let valence be the variable for whether the subject was a Headser (= 1) or Tailser (= 0), and ambiguity be the variable for whether the subject was in the A (= 1) or U (= 0) group. 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 (as should’ve been predicted) 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 Tailers was exacerbated by having ambiguous evidence.

(b) Meanwhile, the empirically bootstrapped 95% confidence interval for the difference between 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, \( d = 0.699 \)).

A further oversight on my part at the pre-registration phase was that I only realized after the fact that I actually had access to time-series data about how the participants’ confidence evolves over time. In particular, using their priors and posteriors for each of the four coin
tosses, I could calculate their average confidence in heads after seeing \( n \) bits of evidence, for \( n \) ranging from 0 to 4.\(^{71}\) (For Bayesians, this average confidence equals their estimate for the proportion of times the coin landed heads.\(^{72}\)

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 this way of carving up the data (with universally lower p-values and higher effect sizes, since the variance of the data has dropped; Prediction 5 is still the only non-significant effect). These are the statistics I reported in the main text (§3.2, page 21).

A supplemental prediction was intended to probe 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. This data failed to confirm the supplemental prediction (\( t(243) = 1.11, p = 0.13, \text{one-sided} \)). However, I noticed that 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, and 25% of them were less than 80%), indicating that self-reports of ‘finding’ might’ve been unreliable. If we rerun the data by operationalizing ‘finding’ as ‘reporting 100% confidence there’s a completion’, the prediction is confirmed.\(^{73}\)

Finally, post-hoc analyses show two further trends that support the role of ambiguity.

First, since it’s natural to expect ambiguity—uncertainty about how to react to evidence—to cause variance in people’s opinions, we should expect the word-search condition to have more variance than the urn condition. This is what we find. Restricting attention to those with weak (so, potentially ambiguous) evidence—those who did not find a completion, or who did not see a black marble—the variance in opinions was much higher in the ambiguous condition than in the unambiguous one. This can be seen in the plots in Figure 15, and is confirmed by tests for equality of variance.\(^{74}\) (Notice that the there remains a nontrivial amount of variance even in the unambiguous condition. Thus it may be that low levels of ambiguity—people being unsure how confident to be in response to a red marble—could be driving the small degree of polarization found in the unambiguous condition.)

Second, recall that the above (§3) model predicts that the mechanism that will drive po-

\(^{71}\) I.e. at stage 0 average their priors for all tosses; at stage 1, average their posterior for the first toss with their priors from the 3 remaining; at stage 2, average their posteriors for the first two tosses with their priors from the remaining 2, etc.

\(^{72}\) Where \( P \) if their probabilistic credence function and \( I_{h_i} \) is the indicator variable for Heads \( i \) (1 if heads, 0 if tails), \( \sum_{i=1}^{4} \frac{P(h_i)}{\sum_{i=1}^{4} I_{h_i}} = \sum_{i=1}^{4} \frac{E(I_{h_i})}{E[\sum_{i=1}^{4} I_{h_i}]} = E[\text{proportion of heads}] \).

\(^{73}\) Amongst those who were less than 100% confident there was a completion, a one-sided t-test found that the average confidence for those looking at uncompletable strings (\( M = 44.60, SD = 25.15 \)) was significantly below the average confidence for those looking at completable strings (\( M = 52.26, SD = 22.98 \)), with \( t(309) = 2.77, p = 0.003, d = 0.32 \).

\(^{74}\) Word-search Headsers’ variance was 563.33, while urn Headsers’ was 285.28, Conover = 5.40, \( p < 0.001 \); and word-search Tailsers’ was 606.78, while urn Tailsers’ was 321.88, Conover = 5.44, \( p < 0.001 \).
Figure 16: Ambiguous condition, mean prior and posterior confidence in Heads, by cases.

* = not significantly different from 50%.

75Interestingly, recalling the hypothesis that the average of people’s actual opinions will equal the rational opinions (§3), these data suggest that an even simpler (value-validating) ambiguous-evidence model may better capture this asymmetry by ignoring the Find/¬Find distinction, e.g.:

Completable, (c)
C. COMPUTATIONAL DETAILS

C Computational Details

This appendix contains the details of the simulations used in §§4–5. It can be read on its own, or 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 (§4)

This subsection explains the generalizations of the (course-grained) word-search-task models that I call cognitive search models. We are to imagine an agent doing a cognitive search for flaws in a piece of evidence that purports to support or tell against a given proposition $q$.

The general form of such a model 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 $\pi$ be the prior and $P$ the posterior (with $P_w$ its realization at world $w$), a cognitive search model is any in which:

- $\pi(q|F) = \pi(q|C)$.
  (The existence of a flaw is what affects the probability of $q$, not whether you find it.)
- For any $w \in N$: $P_w = \pi(\cdot|\neg F)$.
  (If there’s no flaw, all you learn is that you didn’t find one.)
- For any $w \in F$: $P_w = \pi(\cdot|F)$
  (If you find a flaw, you learn exactly that.)
- For any $x,y \in C$: $P_x = P_y$, $P_x(F) = 0$, and $P_x(C) \geq \pi(C|\neg F)$.
  (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 are a simple generalization of the course-grained model of the word-search task from Figure 2—they can likewise be seen as the result of coarse-graining a more detailed model, e.g. one that distinguishes whether the flaw is obvious or not. For $x \in C$ and $y \in N$, we must have $P_x(C) \geq P_y(C)$ to satisfy the value of evidence; when $P_x = P_y$, the model is unambiguous and just consists in conditioning on whether or not you found a completion; but when $P_x(C) > P_y(C)$, the evidence is ambiguous (since $P_y(x) > 0$ and $P_x \neq P_y$)—and this is the ambiguity that 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 function `csQModel` takes one such set of parameters and generates the resulting cognitive-search model. The function `getBBCondQCSModel` 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 random 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.
C. COMPUTATIONAL DETAILS

Given a cognitive search model and some posterior probability function $P_w$, we can get the (Brier) inaccuracy of that function at $w$ by taking the mean squared distance between it’s probability of each world $x$ and the truth-value of $\{x\}$ at $w$. (We use this form of the Brier score, rather than summing across all propositions, for computational tractability.) Thus getGlobPartitionInAcc takes a probability frame (specified using a stochastic matrix, where row $i$ column $j$ equals $P_i(j)$) and a world $w$, and outputs the Brier inaccuracy of $P_w$ at $w$. By subtracting this number from 1 we get a Brier measure of the accuracy of $P_w$. And by taking the expectation of this value, according to our prior $\pi$, we get $\pi$’s expected accuracy of the posterior rational credence function for the update encoded in this model.

We can then test the correlation between the probability of finding a flaw if there is one, 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, 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 of which would raise your credence in $q$ if you found a flaw, the other of which would lower it), such that the 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 in means grew, the proportion of pairs in which expected accuracy favors scrutinizing the incongruent study grew as well—i.e. selective scrutiny became more and more common. This led to the plot on the right of Figure 5 (page 25).

This finally puts us in a position to run a full 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.

There are a variety of choice-points in how to run such simulations. Although variations on the theme will lead to the same result, here are the ones I made. First, 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; red agents are pulling (uniformly) from $[0.1 + \text{findGap}, 0.9]$ and blue agents are pulling (uniformly) from $[0.1, 0.9 - \text{findGap}]$. This parameter findGap can range from 0 (where there’s no different between the groups) to 0.8 (where the difference is maximal). The simulation I displayed is with a gap of 0.5, but generally the rate of polarization grows as the gap increases.

Second, the amount agent’s 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 confidence-trajectories.

The result of running this simulation with these parameters and 300 pairs of studies are displayed in Figure 6 (page 26).

Robustness. We can check for robustness in two ways. First, by simulating 100 red (“pro”) agents and 100 blue (“con”) agents with these parameters, we get estimates for their posterior average credences at $0.600$ (95% confidence interval = [0.575, 0.624]) and $0.390$ (95% confidence interval = [0.370, 0.411]), respectively.

Obviously these exact numbers will vary as we vary the parameters in the simulation. Thus we can also check for robustness by varying these parameters. At the end of section (1) on Cognitive Search in the Mathematica notebook, cross-variations on findGap and hardenSpeed are run, showing that as findGap grows (up to a point) and hardenSpeed shrinks, polarization becomes more extreme.

C.2 Argument Models (§5)

This subsection explains the simple models of arguments used in §5, before introducing scrutiny of arguments. We are to imagine you know that you are about to be presented with an argument in favor of a given claim ($q$). The general form of such models divides the worlds into two classes, depending on whether the argument is good ($G$) or bad ($B$). If the argument is good, it’s rational to increase your confidence in $q$; if not, you’re rational to decrease it. For simplicity, we assume there are only two posteriors you could end up with; moreover, we assume the argument will be more ambiguous if it’s bad.

Thus where $\pi$ is the prior and $P$ is the posterior (with $P_w$ its realization at world $w$), an argument-for-$q$ model is any in which $\{G, B\}$ is a partition and in which:

- $\pi(q|G) > \pi(q) > \pi(q|B)$  
  (If the argument is good, $q$ is more likely to be true; if not, it’s not.)

- For any $x, y$: if $x, y \in G$, then $P_x = P_y$; and if $x, y \in B$, then $P_x = P_y$  
  (Whether the argument is good or bad fully determines the rational posterior.)

- $\exists a, b > 0, a \geq b$: if $x \in G$ and $y \in B$, $P_x(G) = \pi(G) + a$ and $P_y(B) = \pi(B) + b$.  
  (Whether the argument is good or bad, your confidence should shift toward the truth; but—since good arguments are easier to recognize—it should shift more if the argument is good than if it’s bad.)

In such models in which there are two posteriors—$P_g$ for worlds in $G$ and $P_b$ for worlds in $B$—$\pi$ values the update iff it is less extreme than (in the convex hull of) $P_g$ and $P_b$, and $P_g(G) \geq P_b(G)$ (and, by symmetry, $P_b(B) \geq P_g(B)$). This follows from the above specification; the only additional constraint we add is that $P_g(G)$ shifts more from $\pi(G)$ than $P_b(B)$ does from $\pi(B)$.

The simplest argument 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 $\pi(q)$
C. COMPUTATIONAL DETAILS

(prior\(Q\), \(\pi(q|G)\) (g\(\text{Inf}\)), \(P_g(G)\) for \(g \in G\) (g\(\text{Conf}\)), \(\pi(q|B)\) (g\(\text{Inf}\)), and \(P_b(B)\) for \(b \in B\) (b\(\text{Conf}\)).

An argument favors \(q\) if it’s an argument-for-\(q\) model, so \(\pi(q|G) > \pi(q)\), and the shift in confidence about the argument is higher if \(G\) than if \(B\); an argument disfavors \(q\) if it’s an argument-for-\(\neg q\) model, so \(\pi(q|G) < \pi(q)\), and the shift in confidence about the argument is higher if \(G\) than if \(B\). To generate random instances of such arguments, use getRandFavShiftArgModel and getRandDisShiftArgModel respectively. Both take a prior probability of \(q\), a constraint on how far this probability might shift as a result of seeing the argument, and a prior probability that the argument is good.

Given such functions, we can simulate presenting a group of (red) agents with random arguments that favor \(q\), and a separate group of (blue) agents with random arguments that disfavor \(q\). Again, there are a variety of choice-points in how to run such simulations. First, I assume agents always have accurate beliefs about how likely the arguments they’re presented with are to be good or bad. Second, I assumed all arguments are equally likely to be good—\(\pi(G)\) was drawn uniformly from \([0, 1]\). Finally, in addition to questions about the number of agents and arguments to simulate, we can modify how much in principle 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 (\(\text{baseShift}\)) of 0.2; the result is the plot displayed in Figure 8.

The code also allows for simulations to vary the rate at which each group of agents is presented with good arguments, in particular using fav\(GBound\) to lower-bound the probability that a red group-member’s argument is good (\(\pi(G)\) drawn from \([\text{favGBound}, 1]\)) and upper-bound the probability that a blue-member’s is (\(\pi(G)\) drawn from \([0, 1 - \text{favGBound}]\)). The code runs simulations with 30 agents and 50 arguments, with the above parameters for possible shifts and hardening speed, with fav\(GBound\) 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 intervening levels it seems to move both groups move towards belief in \(q\) (still with a substantial gap), and at high levels it seems to reduce both the degree of belief-change and of polarization.

**Robustness.** To check for robustness, I first ran the above simulation with the same parameters, first for 100 red (favorable-argument) agents and then for 100 blue (disfavorable-argument) agents to get estimates for their mean posteriors. This resulted in an estimate for the posterior average confidence of favorable-argument agents at 0.663 (with 95% confidence interval = [0.642, 0.684]), and an estimate for the posterior average confidence of disfavorable-argument agents at 0.341 (with 95% confidence interval = [0.321, 0.363]).

Obviously the rate and reliability of polarization will vary with key parameters. Thus the second way we can check for robustness if by varying base\(\text{Shift}\) and harden\(\text{Speed}\) systematically. The end of the Robustness subsection of section (2) on Argument models in the Mathematica notebook does this, finding that as base\(\text{Shift}\) grows and harden\(\text{Shift}\) shrinks, the amount and rate of polarization grows. All simulations resulted in the average red agent having posterior confidence above the average blue agent.

56
C.3 Argument-Scrutiny Models (§5)

This subsection explains how to combine the simple argument-models of §5 with the cognitive-search models of scrutiny given in §4 to yield argument-scrutiny models. As discussed in the main text, we begin with an argument-model favoring some claim, and then give the agent the choice to either scrutinize that argument or not. If she does not, the model remains the same and she updates as in §C.2; if she does scrutinize it—searching for a flaw in the argument—the scenarios where the argument is bad (B) split into two. In one set of possibilities (F) she finds a flaw with the argument; in another (C) there is a flaw but she doesn’t find it (the search is Completable), and another—namely, the set of worlds where the argument is good—there is not flaw (N = G).

More precisely, given an argument model as described in §C.2, with a prior $\pi$ and posterior $P$—realized as $P_g$ if the argument is good and $P_b$ if it’s bad—scrutinizing it generates a cognitive search model with the partition $\{F, C, N\}$ fixing the posterior $P$ as specified in §C.1, and the following parameters:

- $\pi(q|F) = \pi(q|C) = \pi(q|B)$.
  (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.)

- $\pi(q|N) = \pi(q|G)$
  (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 $P_x(C) \geq P_b(C|\neg F)$
  (If there’s a flaw that you didn’t find, your confidence 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 if you had.)

The only subtle constraint is the third one. This ensures that, compared with the original argument model, not finding a flaw that is indeed there provides no more evidence against there being a flaw than simply conditioning on not finding one would, in keeping with our treatment of what happens in $N$-possibilities in cognitive-search models. When $P_x(C) = P_b(C|\neg F)$, scrutiny adds no additional ambiguity over-and-above that already present in the argument model; when $P_x(C) > P_b(C|\neg F)$, the divergence between $P_x$ (for $x \in C$) and $P_y$ for $y \in N$ grows, increasing the ambiguity. (This corresponds, intuitively, to how likely you think it is that you should’ve found a flaw that was there, even if you didn’t.)

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 $P_x(C)$ approaches 1, over and above $P_b(C|\neg F)$ (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...
C. COMPUTATIONAL DETAILS

and constrain this. I used the same parameters for generating arguments that I used in §C.2 (except this time all agents favored $q$), and ran four versions of the scrutiny simulation. Since scrutiny introduces more noise into the simulations, all used 50 agents (to better see the average trend) and 100 arguments.

In version (1), scrutinizing agents never find a flaw even if there is one ($p_{\text{Find}} = 0$), and the scrutiny adds no ambiguity ($j_{\text{Shift}} = 0$). Such scrutiny does not change the original argument-model at all, 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 30.

In version (2), scrutinizing agents always find a flaw if there is one ($p_{\text{Find}} = 1$), meaning that scrutiny removes all ambiguity from the argument. (The $j_{\text{Shift}}$ parameter has no effect in this case; I set it to pull uniformly at random from $[0, 1]$.) Since such scrutiny changes the model to an unambiguous one, by Theorem 2.1, scrutinizing agents do not expectedly polarize from their priors of 0.5—as seen in the top right of Figure 9 on page 30.

In version (3), scrutinizing agents sometimes find a flaw if there is one ($p_{\text{Find}}$ pulled randomly from $[0, 1]$), and scrutiny introduces a small degree of ambiguity ($j_{\text{Shift}}$ pulled randomly from $[0, 0.5]$). The result is that agents who scrutinize predictably polarize is the same direction as those that don’t, but less so—as seen in the bottom left of Figure 9 on page 30.

In version (4), scrutinizing agents again sometimes find a flaw if there is one ($p_{\text{Find}}$ pulled randomly from $[0, 1]$), and scrutiny introduces a substantial ambiguity ($j_{\text{Shift}}$ pulled randomly from $[0, 1]$). The result is that agents who scrutinize predictably polarize in the opposite direction of those that don’t—as seen in the bottom right of Figure 9 on page 30.

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 0.663, with 95% confidence interval of $[0.642, 0.684]$. To check that the results in the above simulations (1)–(4) were robust, I ran the same parameters but 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 0.649, with a 95% confidence interval of $[0.636, 0.663]$, indicating that scrutinizing agents shift to a comparable degree to those who don’t scrutinize, as expected. In version (2), the mean posterior was 0.506, with a 95% confidence interval of $[0.473, 0.539]$, indicating that agents do not predictably shift from their priors of 0.5, as expected. In version (3), the mean posterior was 0.529, with a 95% confidence interval of $[0.506, 0.553]$. As this is well below the confidence interval of $[0.642, 0.684]$ for those who don’t scrutinize, this confirms our expectation that such scrutiny dampens polarization. In version (4), the mean posterior was 0.461, with a 95% confidence interval of $[0.434, 0.488]$. As this is below the initial credences of 0.5, this confirms our expectation that such scrutiny reverses the direction of polarization.
References


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


Carmichael, Chloe, 2017. ‘Political Polarization Is A Psychology Problem’.


REFERENCES
Rizzo, Mario J and Whitman, Glen, 2019. 'Informal Logical Fallacies.' In R.J. Sternberg, H.L. Roediger, and D.F. Ramstead, F. P., 1990. 'Weight or the value of knowledge'.

Ross, Lee, 2012. 'Reflections on Biased Assimilation and Belief Polarization'.

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

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

———, 2020. 'The Value of Evidence'. In Maria Lasonen-Aarnio and Clayton Littlejohn, eds.,


Schkade, David, Sunstein, Cass R., and Hastie, Reid, 2010. 'WHEN DELIBERATION PRODUCES EXTREMISM'.

Schueler, John, 1967. 'Selective exposure to information: A critical review'.

Schultheis, Ginger, 2018. 'Living on the Edge: Against Epistemic Permissivism'.


———, 2020. ‘Just a Big Misunderstanding? Bias and Bayesian Affective Polarization’. International Economic