Reasonably Polarized (Technical Appendix)

Kevin Dorst
kevindorst@pitt.edu
Comments welcome!

Version: 31st October 2020

This is the technical appendix to my blog series *Reasonably Polarized: Why politics is more rational than you think*. It develops the technical details underlying the various arguments, and also addresses further questions and concerns that might come up.

It's a work in progress, and is constantly being revised. Feedback is most appreciated! You can find the most up-to-date version of the appendix here.

Contents

1 Introduction

Here's a link to the full post; it was published on September 5, 2020.

> **Post Synopsis:** I introduced the example of me and Becca, and how we predictably polarized in our political leanings. I argued, using this case, that the standard irrationalist explanation of polarization doesn't work, and sketched how a *rational* explanation would go, previewing the argument to come.
>
> **Appendix Summary:** I don't want to bog down this high-level post with too many details, as those are on their way in the coming weeks. Instead, I'll just provide some relevant links and citations.

Here are some **links to other versions of the core argument this post sets out**:

- 'Why Rational People Polarize' in *Phenomenal World*. This was an early version of the idea that ambiguous evidence can lead to rational polarization.
- 'A Plea for Political Empathy'. This was the opening piece to my *Stranger Apologies* blog (of which this RP-series is a part), on why irrationalist explanations of polarization lead to demonization.
- 'Why the the other side is more rational than you think'. This piece should come out any day now in *Arc Digital*; it develops in more detail the argument that (1) we can't sensibly blame polarization on irrationality, and (2) that we *can* understand it as rationally-caused.

Here are some citations to **articles that endorse the "standard (irrationalist) story" I criticized in the post:**

- Klein (2014), 'How politics makes us stupid'.
- Klein (2020), *Why We're Polarized*.

  > Two comments. (1) This is a great book, and I highly recommend it. As we'll see, I buy a good amount of what Klein says when it comes to the empirical story—where I'm skeptical is the way he thinks irrational 'identity-protective cognition' explains polarization.
  >
  > (2) Klein *says* he's giving us a rational story: "The American political system... is full of rational actors making rational decisions given the incentives they face. We are a collection of functional parts whose efforts combine into a dysfunctional whole" (xvii). Though I like the sound of that, it turns out that what he means is that individuals are ***pragmatically* rational**—for example, if you really care about your Republican identity, it makes sense to ignore facts that are inconvenient for that identity, because maintaining your identity is more important than getting to the truth. As will become

clearer in Post 4, what I'm interested—and what matters for how we think about the other side—in is whether polarization can be ***epistemically* rational**, i.e. whether people *who care about the truth* can nevertheless wind up predictably polarized.

- Achen and Bartels (2017), *Democracy for Realists*.
- Taber and Lodge (2006), 'Motivated Skepticism in the Evaluation of Political Beliefs'.
- Kahan et al. (2017), 'Motivated Numeracy and Enlightened Self-Government'.
- Nguyen (2018), 'Escape the echo chamber'.

  This is a great piece: it makes some good distinctions between echo chambers and filter bubbles, and it gives a nuanced picture of the (ir)rationality of falling into an echo chamber (or cult). Still, I think it's fair to consider it an irrationalist narrative.
- Lazer et al. (2018), 'The Science of Fake News'.
- Pennycook and Rand (2019), 'Why Do People Fall for Fake News?'
- Van Heuvelen (2007), 'The Internet is making us stupid'.
- Robson (2018), 'The myth of the online echo chamber'.
- Koerth (2019), 'Why Partisans Look At The Same Evidence On Ukraine And See Wildly Different Things'.
- Carmichael (2017), 'Political Polarization Is A Psychology Problem'.

Here are citations to **some other work that's critical of irrationalist explanations of political disagreement:**

- Jern et al. (2014), 'Belief polarization is not always irrational'.
- Benoît and Dubra (2019), 'Apparent Bias: What does attitude polarization show?'
- Singer et al. (2019), 'Rational social and political polarization'.
- Whittlestone (2017), *The importance of making assumptions: why confirmation is not necessarily a bias.*

  See this blog post for an accessible summary of her thesis.
- O'Connor and Weatherall (2018), 'Scientific Polarization'.
- Engber (2018), 'LOL Something Matters'.
- Lepoutre (2020), 'Democratic Group Cognition'.
- Landemore (2017), *Democratic reason: Politics, collective intelligence, and the rule of the many.*

2 How to Polarize Rational People

Here's a link to the full post, published on September 12, 2020.

> **Post Synopsis:** I described an experiment designed to show how it's possible to polarize people using ambiguous evidence. To do this, I introduced *word-completion tasks* in which you're asked to determine whether a given letter-string is completable by an English word. The key point about this task is that it provides *asymmetrically ambiguous evidence*: it's easier to know what to think if there *is* a completion than if there's *not* a completion. As a result, we can split people into groups—the Headsers and the Tailsers—such that Headsers are better at recognizing cases where a coin lands heads, Tailsers are better at recognizing cases where it lands tails, and as a result they predictably polarize.
>
> **Appendix Summary:** Here I'll report in more detail the results of the study I ran confirming this prediction. A pre-registration form for the study is available here.

250 participants were recruited through Prolific (107 F/139 M/4 Other; mean age = 27.06). Subjects were randomly divided into an Ambiguous (A) and Unambiguous (U) condition. Within each condition, they were further (randomly) 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.

The A group was informed about how word-completion 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. Each participant was presented with four independent word-completion tasks. In each, they were first told that a coin was flipped to determine (as per the rule above) whether the letter-string they next saw would be completable, and asked how confident they were that it was completable. They used a $0-100\%$ slider to rate this confidence, which they were given standard instructions about how to use. This first ("prior" confidence) question for each toss was an attention check, and participants were instructed to answer "50%" at this stage, since they had not received any evidence. It was pre-registered that I would exclude data from participants who failed two or more of these attention checks. (All in all, data from 25 subjects out of 250 were excluded for these reasons.)

After each check, the participant was presented with some evidence about the coin toss. Of the four tosses each participant saw, two landed heads and two landed tail, so each saw two completable strings and two uncompletable strings, in random orders. 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}.[1] After seeing their string for 7 seconds, the participants were asked how confident they were that it was completable, and presented with a slider between 0–100%.[2]

The U group, in contrast, was told that each toss of the coin would be used to determine the contents of the urn. 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 make clear they were different urns. Again, each toss started by telling them a new coin had been tossed, and asking how confident they were that it landed heads (U-Tsers: how confident they were that it landed tails). This was an attention check as those above; they were instructed to answer "50%", and data from subjects who failed two or more was omitted. Subjects were then told what color marble came from a single random draw of the urn, and asked how confident they were that the coin landed heads (U-Tsers: tails). Each subject saw four separate coin-toss/urn pairs; three of the four revealed a non-black marble, while the fourth revealed a black one—simulating the expected rate of black marbles if the coin landed heads/tails 50% of the time, and the marbles were drawn at their expected rate.

The U group was so designed in order to test the hypothesis that it is *ambiguity* of evidence that drives the polarization effect. As can be seen, there is a structural similarity but also a structural dissimilarity between the A-group and the U-group. The similarity is that both groups have some chance of getting decisive evidence in favor of a hypothesis (finding a completion for the A group; seeing a black marble for the U-group), and some chance of getting weak evidence against that hypothesis (failing to find a completion for the A group; seeing a non-black marble for the U-group).

The dissimilarity is that subjects in the U-group are, in principle, able to know what they should do with this evidence—a straightforward Bayesian calculation says that if you're a U-Hser and see a black marble, you should assign probability 1 to the coin landing heads, and if you see a non-black marble, you should assign probability $\frac{1}{3}$ to it having landed heads.[3] In contrast, with the word completion task when you

[1] No doubt it would be useful to run a study using a bigger sample of letter-strings; obviously, they must be chosen with some care, as completely random strings like 'X_TNO_' will standardly be too obvious.

[2] Pilot studies indicated that when people were asked how confident they were that the coin landed *heads* or *tails*, subjects were substantially confused, often—it seems, reversing the scale or not moving it from 50%. This makes sense, as there is a pragmatic oddness and an extra cognitive load in asking about whether the coin landed heads (tails), when that is known to be equivalent to whether the string was completable. For this reason, I elicited their opinion about whether the string was completable, and used that reported number to infer their confidence in heads/tails on the given flip, based on which group they were in.

[3] $P(H|\textit{non-black}) = \frac{P(\textit{non-black}|H)\cdot P(H)}{P(\textit{non-black}|H)\cdot P(H)+P(\textit{non-black}|\neg H)\cdot P(\neg H)} = \frac{0.5\cdot 0.5}{0.5\cdot 0.5+0.5\cdot 1} = \frac{1}{3}$.

do find a word, you should of course be certain that it's completable; but when you *don't* find a word, it is much harder to know what opinion you should have. This is a theoretical prediction we'll go more into after we get a theory of ambiguous evidence on the table—but the basic idea is that you should be unsure whether you *should* think of a completion, and therefore unsure what evidence you actually received—unsure whether you should be sure the string is completable. (Staring at '_EAR_T', you think, "I don't see one; but should I?" When 'learnt' pops into you're head, you may well think, "Ah, I should've seen that!") Because of this uncertainty, subjects will (rationally) be unsure how much to lower their confidence that the string is completable.

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). It was 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 (by-item), along with 95% confidence intervals, displayed in Figure 1:

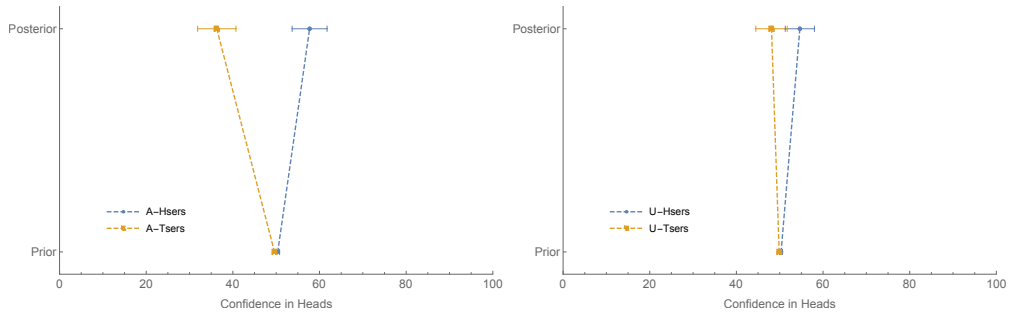


Figure 1: Mean prior and posterior confidence in heads in ambiguous- (left) and unambiguous-evidence (right) conditions. Bars represent 95% confidence intervals.

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, 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 U-Tser posteriors (M =48.10, $SD = 28.47$) were above A-Tser posteriors (M = 36.29, SD = 31.04), 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 expected: 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]. The former dominates the latter, indicating 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 ambiguous (= 1) or unambiguous (= 0) group. Analyzing the results using a 2 (valence: Headser vs. Tailser) by 2 (ambiguity: ambiguous vs. unambiguous) ANOVA indicated that there was a main effect of valence ($F(1, 899) = 46.47$, $p < 0.001$), a main effect of ambiguity ($F(1, 899) = 4.31$, $p = 0.038$), and (as should've been predicted) an interaction effect between valence and ambiguity ($F(1, 899) = 14.57$, $p < 0.001$), indicating that the divergence between Headsers and Tailsers was exacerbated by having ambiguous evidence.

(b) Meanwhile, the empirically bootstrapped 95% confidence interval for the difference 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. As mentioned, this condition had a Cohen's d effect size of 0.699. Notably, 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) $p = 0.009$, but the effect size was smaller ($d = 0.236$).

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 participant's 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.[4] If they are Bayesian in their confidence, this average confidence equals their estimate for the proportion of times the coin landed heads.[5] In fact, using these numbers (and assuming they treated each coin flip independently, as instructed) we could calculate how their opinions would've evolved in any proposition about the coin-flips if they were Bayesian.

Thus I was able to track how their estimate of the proportion of heads (as well as other measures of their beliefs about heads) evolved over time. In other words, we can re-run the above data by pooling responses within subjects and calculating them 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, since the variance of the data has dropped since we've pooled data within subjects; Prediction 5 is still the only non-significant effect).

Using this, we can calculate the trajectories of their the mean estimate of the proportion of heads (i.e. calculate the mean of the subjects' average confidence in heads at each stage in the experiment), as reported in the blog post; see Figure 2 for this evolution in both the A and U groups.

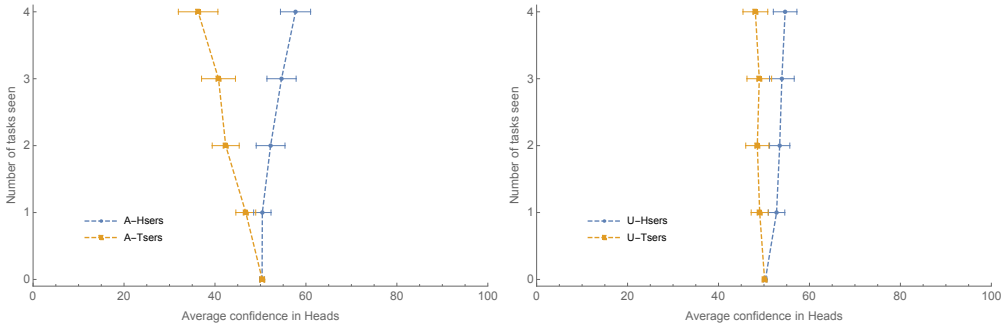


Figure 2: Mean confidence in heads trajectories as subjects view more tasks. Left is ambiguous condition, right is unambiguous condition. Bars represent 95% confidence intervals.

As can be seen, the ambiguous group continues to diverge, while the unambiguous group does not. Using this pooled-within-subject data, a one-sided independent samples t-test indicated that there is an even larger difference between A-Hser posteriors ($M = 57.71, SD = 12.26$) and A-Tser posteriors ($M = 36.29, SD = 14.95$), with $t(101) = 7.98$,

[4] I.e. at stage 0 average their priors for all tosses; at stage 1, average their posterior for the first toss they saw with their priors from the 3 remaining; at stage 2, average their posteriors for the first two tosses they saw with their priors from the remaining 2, etc.

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

$p < 0.001$, and $d = 1.577$. The 95% confidence interval for the difference is now [16.02, 26.82]. (Meanwhile, a two-sided t-test for the difference between U-Hser posteriors ($M = 54.64, SD = 10.19$) and U-Tser posteriors ($M = 48.10, SD = 10.53$) was again significant, but again with a smaller effect size: $t(120) = 3.49$, $p = 0.0034$, $d = 0.631$. The 95% confidence interval for the difference between U-Hsers and U-Tsers was [2.82, 10.26]—again dominated by the A-group's difference confidence interval; likewise, a 2x2 ANOVA again indicated significant main and interaction effects; etc.)

This time-series data allows us to see the divergence in other ways. For instance, consider claims of the form,"there were at least $\boldsymbol{x}$ heads", for x ranging from 1 to 4, and we see the diverging trajectories in the ambiguous condition on the left of Figure 3 (page 10), along with the smaller divergences of the unambiguous condition on the right.

The crucial question: what drives the polarization? The full theory of this will have to wait till we get the theory of ambiguous evidence on the table. But we can start by getting a few things on the table.

First, **the effect is *not* being driven by an asymmetry between strong and weak evidence**. The U group was set up to mirror this asymmetry, but they did not display nearly as strong a polarization effect. Moreover, there is reason to expect that the effect they *did* experience was a "response bias" effect.[6]

Rather, what drives the effect has to do with the *ambiguity* of the evidence—the fact that it's hard to know what to think when you don't find a completion than when you do find one. There is evidence for this in the data. For example, we can divide the A-group cases in which they found a word (operationalized as: they had credence 100 that there was a completion; more on this in a few weeks) and didn't, and then calculate the expected variance in opinion (by weighting the two by what proportion of trials found word) if there *is* a word compared to when there isn't. Likewise, we can divide the U-group into cases in which they saw a black marble vs. didn't, and then calculate the expected variance in opinion if there *is* a black marble vs. if there isn't. The fact that there is an "ambiguity asymmetry" should mean that the expected variance in opinion is asymmetric around the completable/not-completable distinction, but that it is not asymmetric around the black-marble/no-black-marble distinction.

This is what we find. For the A-group: the expected variance in opinion when a word-completion task is completable was 383.9, whereas when it *wasn't* completable it was 746.7. In contrast, for the U-group: the expected variance in a opinion when there was a black marble was 239.0, whereas when there *wasn't* was 266.1.

[6] U-Hsers were asked "How confident are you that the coin landed heads?", whereas U-Tsers were asked "How confident are you that the coin landed tails?". (This mirrors the fact that A-Hsers and A-Tsers were both asked how confident they were that there was a completion—for the former, that question is equivalent to asking for their confidence in heads; for the latter, it's equivalent to asking for their confidence in tails.) It seems probable that this difference in questions drove what effect there was.

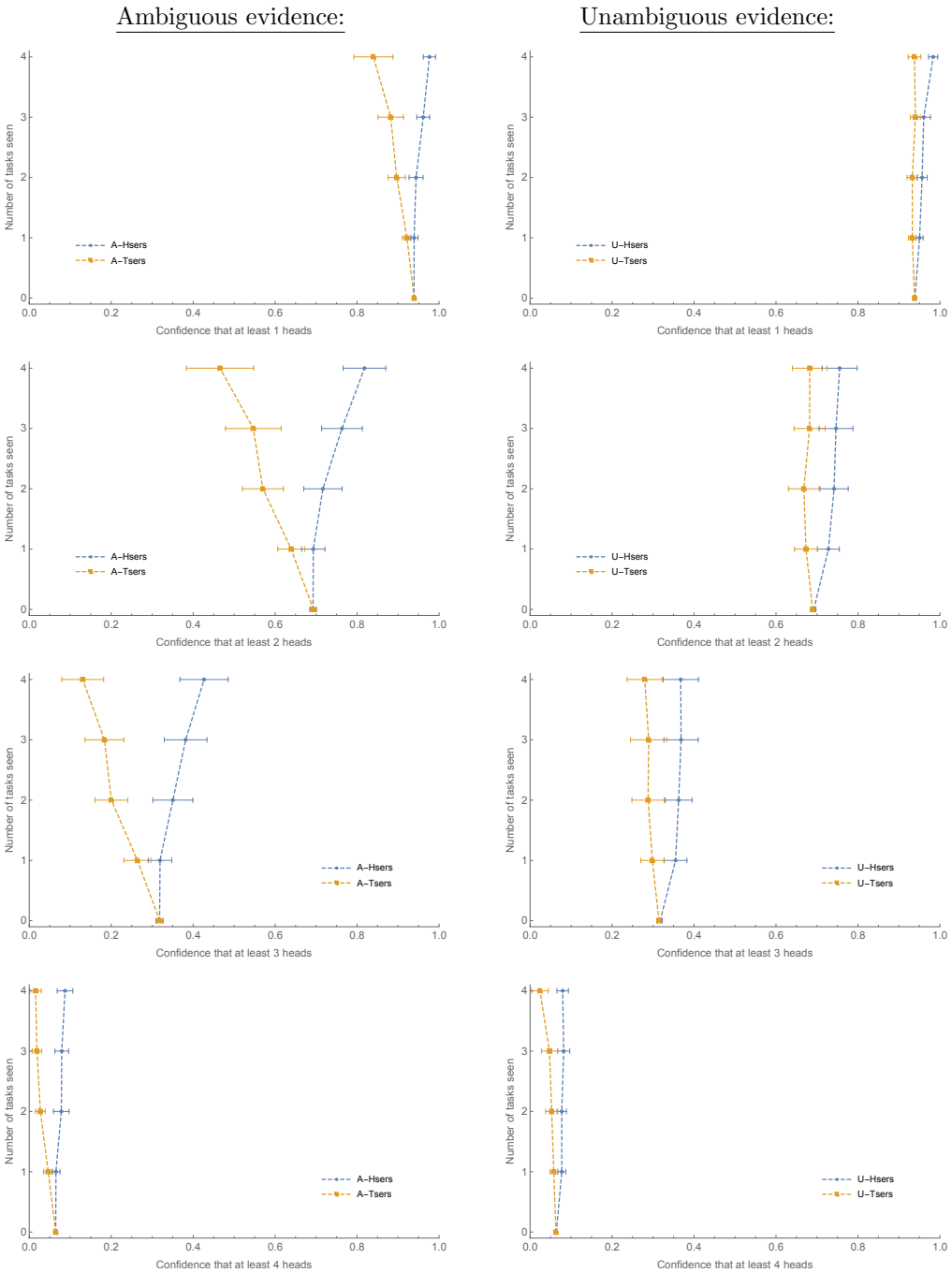


Figure 3: Mean confidence in "at least x heads" as view more tasks, for x ranging from 1–4. Left side is ambiguous condition; right side is unambiguous condition.

I'll argue, in future posts, that this asymmetry in ambiguity in the A-group case is what drives the polarizing effect, for it makes Headsers better at recognizing heads-cases, and Tailsers better at recognizing tails-cases. For now, we can simply report that the data fits with this explanation. When we pool across subjects and divide trials into heads-cases and tails-cases, here are the average posteriors in the groups:

Average confidence it landed heads across cases:

	A-Hsers	A-Tsers	U-Hsers	U-Tsers
Overall:	57.7	36.29	54.64	48.10
Heads cases:	67.42	47.73	66.89	59.95
Tails cases:	48.00	24.84	42.39	36.25

As can be seen, Headsers are better at recognizing heads-cases, Tailsers are better at recognizing tails-cases—and these differences are especially stark amongst the A group. After the fact, I decided to look at these differences statistically (so these tests were not pre-registered). A one-sided t-test found that amongst heads cases, A-Hsers ($M = 67.42, SD = 30.38$) had a mean significantly above A-Tsers ($M = 47.74, SD = 27.54$), with $t(204) = 4.84$, $p < 0.001$, $d = 0.677$. (A-Tsers mean confidence of 47.74 is not significantly different from 50: two-sided t-test yielded $t(95) = -0.80$, $p = 0.423$.) Meanwhile, amongst all tails-cases, A-Hsers ($M = 48.00, SD = 27.10$) were again significantly higher than A-Tsers ($M = 24.84, SD = 30.22$), with $t(204) = 5.80$, $p < 0.001$, $d = 0.810$. (A-Hsers mean confidence of 48.00 is not significantly different from 50: two-sided t-test yielded $t(109) = -0.77$, $p = 0.440$.)

In contrast, these asymmetries were smaller for the unambiguous condition. The difference between U-Hsers ($M = 66.89, SD = 30.27$) and U-Tsers ($M = 59.95, SD = 17.16$) in heads-cases was significant but small: two-sided t-test revealed $t(195) = 2.21$, $p = 0.028$, $d = 0.281$. Meanwhile the difference in tails-cases between U-Hsers ($M = 42.39, SD = 15.43$) and U-Tsers ($M = 36.25, SD = 32.42$) was not significant: $t(169) = 1.88$, $p = 0.062$.

3 How We Polarized

Here's a link to the full post, published on September 19, 2020.

> **Post Synopsis:** Polarization has always been with us. A set of basic psychological and sociological mechanisms explain why human societies have always been characterized by both *local conformity* and *global diversity*: there tends to be agreement within small social circles, but disagreement between them. As a result, when people go off on different life trajectories, it's always been normal for their attitudes to drift apart. What's *changed* in recent decades is that a series of factors have come together to align these various mechanisms and kick them into overdrive. As a result, now when people like me and Becca go off on different life trajectories, their opinions diverge in predictable and consistent directions—and do so faster and farther than before. **In a slogan:** normally, polarization is a random walk; recently, it's been transformed into a feedback loop.
>
> **Appendix Summary:** There is a *ton* more that could be said. But to avoid getting bogged down in too many details, I'll just address the one part of my empirical story that I expect to be somewhat controversial—namely, whether American's *opinions* have polarized in recent decades.

There is, of course, a large debate in political science (and elsewhere) about what the proper characterization of recent polarization polarization is, and in what sense things have changed. Two of the three core mechanisms I described in the post are uncontroversial: Americans have definitely become increasingly ideologically sorted (Fiorina 2016)[7], and they have definitely had increasingly negative feelings toward those on the other side (Iyengar et al. 2019).

The claim I made that might be controversial is that Americans have also experienced **attitude polarization**, where that means they have become more extreme in their political opinions. Some authors contest this. They (rightfully) resist overblown claims that the "political center" has disappeared—pointing out that although opinions are now increasingly sorted by party, the distribution of opinions on particular policies (things like: "The maximum tax rate should be __%"; "The state should supply healthcare"; etc.) have not changed dramatically (Fiorina 2016). It is sometimes inferred that this means that American's political opinions have *not* polarized—they've just "sorted."

[7] In my post I appealed to the claim that parties have also sorted *geographically*, with an increasing urban-rural divide (Bishop 2009). This is somewhat controversial (Abrams and Fiorina 2012), but I take the evidence to show that some substantial increase in urban/rural sorting has occurred (Brownstein 2016). For instance, in 1916, there was no correlation between population per square mile and the Democratic share of the two-party vote; in 2016, there was a robust and steep positive correlation (Klein 2020, 40).

This, I think, is a mistake—and it is one due to an overly narrow view of what counts as a "political opinion." Yes, it's true that if you ask people "What do you think of policy X?", and you do not tell them whether a Democrat or a Republican is proposing the policy, the distribution of opinions has not shifted dramatically over time. But if you instead ask them "What do you think of the Democratic Party's proposal of policy X?", then their opinions *have* shifted over time. For example, the split between Republican and Democratic presidential approval ratings has become much more extreme over the past half-century (Gao and Smith 2016).

The point is a conceptual one: although it is useful to see what people's opinions about policies in the abstract are, in any concrete disagreement over political policies, it will be common knowledge which party is proposing them. Moreover, the following two claims are *different* claims:

1) Policy X will be good for the country if implemented.
2) The Democrat's proposal of policy X will be good for the country is implemented.

Since they are different claims, it can be perfectly sensible to have very different attitudes toward them. (If you really trust Democrats, it makes sense to be more confident of (2) than (1); if you really distrust them, it makes sense to be less confident of (2) than (1).) Thus opinions about policies-proposed-by-party-Y *are* political opinions; they are political opinions that matter; and they are ones that *have* gotten increasingly polarized over the decades.

There is another way to make this point, this time focusing on affective polarization. It is uncontroversial that (for instance) Democrats have increasingly negative feelings toward Republicans. But it should also be uncontroversial that such feelings will be highly correlated with various beliefs about members of these parties (Haidt 2012; Ryan 2014). For example, if for the past few decades we had asked Democrats how confident they were in claims like "Republicans tend to be selfish" or "Republicans tend not to listen to information with an open mind," we know that *these* ratings would also be going up over time. Since Republican's don't have increasingly negative views of *themselves*, Republican's opinions on these points will not be going up—at least not at the same rate. That means we expect attitude polarization on certain political opinions—for instance, those about the qualities (selfishness, open-mindedness, etc.) that partisans possess. Those are political opinions *par excellence*: if you were to drop me off in a foreign country I knew nothing about, and I had to figure out which political party to support, the first thing I would want to know is what the members of that political party are like!

Upshot: I think it should be uncontroversial that whenever we have predictable ideological sorting and predictable affective polarization, we *thereby* will get certain forms of predictable attitude polarization as well, wherein people's political opinions become increasingly divergent.

4 What is *Rational* Polarization?

Here's the link to the full post, published on September 26, 2020.

> **Post Synopsis:** I explained how my project focuses on whether polarization can be *epistemically* rational—meaning that it is to be expected from people who are doing the best they can to get to the truth—and offered a theory of epistemic rationality within which to situate my argument. In particular, I (1) introduced **unambiguous Bayesianism** as the standard theory of epistemic rationality and explained why it forbids predictable polarization (Fact 4.4), (2) showed that any theory that allows ambiguous evidence will *permit* predictable polarization (Fact 4.7), (3) introduced my favored theory (**Rationality as Value**), which says that epistemically rational transitions are those that satisfy the *value of evidence*, and showed that it generalizes unambiguous Bayesianism (Fact 4.10) and also permits ambiguous (and, therefore, predictably polarizing) evidence (Theorem 4.13).
>
> **Appendix Summary:** In this appendix entry, I'll formalize and prove the facts stated in the blog post. §4.1 focuses on modeling ambiguous evidence and predictable polarization, and proving a tight connection between the two. §4.2 focuses on formalizing the Rationality as Value constraint, and showing how this implies unambiguous Bayesianism when evidence is assumed to be unambiguous, but allows predictable polarization once we give up that assumption.

4.1 Ambiguous evidence and predictable polarization

The class of models we'll be working with are all Bayesian, in the sense that they all represent rational beliefs with probability functions. The models I will employ are probabilistic generalizations of the possible-worlds models used in modal and epistemic logic (Hintikka 1962; Kripke 1963). The key idea is that the rational probabilities vary across worlds, and when evidence is ambiguous, the rational probability function can be unsure what the rational probability function is. For an introduction to these models see (Dorst 2019, 2020b) and the citations therein.

The models are sometimes called **(dynamic) probabilities frames** $\langle W, \mathcal{P}^1, \mathcal{P}^2 \rangle$.[8] They are used for modeling the epistemic state of a given person (say, you) at two times, 1 and 2. W is a finite set of possibilities ("worlds"), thought of as a partition of logical space that is fine-grained enough for the modeling purposes at hand. Claims/propositions/events are represented as ways the world could be, i.e. subsets of

[8] "Frames" following standard modal-logic terminology, since these structures do not contain an interpretation function for a formal language. This is because, as we'll see, the formal language would get cumbersome and is not necessary.

W. I'll use $p, q, r...$ as variables for these propositions. Logical operations are handled via set theory: $p \wedge q = p \cap q$, $\neg p = W \setminus p$, $p \rightarrow q = (W \setminus p) \cup q$, etc. p is true at world w iff $w \in p$; or, more generally, p is true at $q \subseteq W$ (i.e. q entails p) iff $q \subseteq p$.

Each $\mathcal{P}^i : W \rightarrow \Delta(W)$ is a function from worlds $w \in W$ to probability function $\mathcal{P}^i_w$ defined over the subsets of W. The interpretation is that $\mathcal{P}^i_w$ captures the rational opinions for you to have at time i, given the evidence you have then. This is to be understood as a definite description: you can be unsure what the rational degrees of confidence are, given your evidence, which is just to say you can be unsure whether you're in a world a world where they are one thing, or a different world in which they're different.

Importantly, $\mathcal{P}^i$ is *not* intended to represent the opinions you actually have; it is intended to represent the opinions you *should* have, given your evidence. Thus even if you are in fact rational and know what opinions you in fact have, you can be unsure what $\mathcal{P}^i$ is because you can be unsure whether you are rational (see Dorst 2019).

Given a frame $\langle W, \mathcal{P}^1, \mathcal{P}^2 \rangle$, we use $\mathcal{P}^i$ to define propositions about what opinions are rational for you, thus allowing us to represent higher-order opinions (opinions about what opinions are rational) as well as first-order opinions (opinions about the world) seamlessly. For instance, $[P^i(q) \geq t] := \{w : \mathcal{P}^i_w(q) \geq t\}$ is the set of worlds where the rational credence function for you at time i assigns credence at least t to the proposition q. Thus, for instance, $[P^1(P^2(q) \geq 0.5) = 0.3]$ is true at a world w iff $\mathcal{P}^1_w(P^2(q) \geq 0.5) = 0.3$, iff $\mathcal{P}^1_w(\{x : \mathcal{P}^2_x(q) \geq 0.5) = 0.3$. Since the inner probability claim is always a set of worlds, we can always "unpack" it so that probabilities of probabilities are always probabilities of propositions, and thus are always well-defined.

For simplicity of definitions and results, I will also assume that our probabilities are always **reflexive**, meaning that for all w, i: $\mathcal{P}^i_w(w) > 0$. This is equivalent to the claim that your evidence never warrants being certain ins a falsehood: $[P^i(q) = 1] \rightarrow q$ is true at every world in for every i, q in every frame.[9]

Given these models, we can formally define what it is for your evidence to be ambiguous:

Definition 4.1. $\mathcal{P}^i$ is **ambiguous** at world w iff there is a proposition q such that $\mathcal{P}^i_w(P^i(q) = t) < 1$ for all t. $\mathcal{P}^i$ is **potentially ambiguous** iff it is ambiguous at some world.

Informally, evidence i is ambiguous iff, given that evidence you should be unsure what opinions it warrants having—iff it warrants higher-order uncertainty. In any such case, there will, of course, be some t such that $\mathcal{P}^i_w(q) = t$. Thus at world w, it's true that $P^i(q) = t$, but also true that $P^i(P^i(q) = t) < 1$: you should be t-confident of q, but you should be less than certain that you should.

[9]Everything I say and prove could be generalized beyond this case, but many of the results and definitions get more tedious to state.

Colloquially: your evidence is ambiguous when, given that evidence, you should be unsure what you should think. Even more colloquially: evidence is ambiguous when it doesn't wear its verdicts on its sleeve, i.e. it's hard (even for rational people) to know what to do with it.

Next, we need to define what I mean to say that evidence is *predictably polarizing*. Since our frames encode higher-order opinions, we can ask what how confident you expect the future rational opinions to be in some claim, and compare that to how confident you should be now.

For any random variable (function from worlds to numbers) X, let $\mathbb{E}^i_w[X] := \sum_{t\in\mathbb{R}} \mathcal{P}^i_w(X = t) \cdot t$ be $\mathcal{P}^i_w$'s expectation of X, i.e. a weighted average of the various possible values of X, with weights determined by how likely they are to obtain. (We can use this function to define propositions about rational expectations as above, e.g. $[E^2[X] = t] := \{w : \mathbb{E}^2_w[X] = t\}$.) Note that for any proposition q, $P^i(q)$ is such a random variable—give it a world w, and it'll output how likely q is according to evidence i at that world, i.e. $\mathcal{P}^i_w(q)$. Thus we can compare how confident you should be at (at world i) at time 1—namely, $\mathcal{P}^1_w(q)$—with how confident you should (at time 1) expect that you should be at time 2—namely, $\mathbb{E}^1_w[P^2(q)]$. The evidence you'll receive at time 2 is predictably polarizing if you expect it to push the rational confidence in a particular direction:

Definition 4.2. $\mathcal{P}^2$ is **predictably polarizing** relative to $\mathcal{P}^1$ iff, at some world w, for some q, $\mathcal{P}^1_w(q) \neq \mathbb{E}^1_w[P^2(q)]$ (so either $\mathcal{P}^1_w(q) > \mathbb{E}^1_w[P^2(q)]$, or $\mathcal{P}^1_w(q) < \mathbb{E}^1_w[P^2(q)]$).

Why *this* definition of predictable polarization? This is a good question, on which much could be said, but here I'll just make a few remarks.[10]

Let $\mathcal{P}^i$ be the rational opinions for me at time i; let $\mathcal{Q}^i$ be the rational opinions for you at time i (modeled in the same way), let $\mathbb{F}^i_w$ be *your* corresponding expectation function at world w and time i, and let Q^i be a definite description for your credence function at time i. Suppose we know that we agree at time 1, i.e. that $\mathcal{P}^1 = \mathcal{Q}^1$. Then the divergence between our opinion in q at time 1 is zero: for all q, $P^1(q) - Q^1(q) = 0$. Now suppose that neither of us will get predictably polarizing evidence. Then whatever we might learn, our prior best estimate will *still* be that the divergence between our opinion on every q will be zero: at every w, $\mathbb{E}^1_w[P^2(q)] = P^1_w(q) = Q^1_w(q) = \mathbb{E}^1_w[Q^2(q)]$, and therefore by linearity of expectations $\mathbb{E}^1_w[P^2(q) - Q^2(q)] = 0$ (and likewise for $\mathbb{F}^1_w$). In short: if we agree and we don't get predictably polarizing evidence, we can't predict that our opinions will diverge in a particular direction. In contrast, if we *do* get predictably polarizing evidence, then we can—for instance, if $\mathbb{E}^1_w[P^2(q)] > P^1_w(q) = Q^1_w(q) > \mathbb{F}^1_w[Q^2(q)]$, then if we share opinions at time 1, it follows that $\mathbb{E}^1_w[P^2(q) - Q^2(q)] > 0 = P^1_w(q) - Q^1(q)$. Thus we can both agree on some political claim q at time 1, but we can both expect that by time 2, I'll be more confident of q than you will be.

[10] Salow (2018) makes a convincing case for it, albeit in a slightly different context.

Moreover, even if we have *different* initial opinions, if I won't receive predictably polarizing evidence, then insofar as I expect our disagreements to grow larger, this must be because I expect *your* opinion to move away from mine, and expect my opinion to stay fixed: $\mathbb{E}^1_w[(P^2(q) - Q^2(q)) - (P^1(q) - Q^1(q))] = \mathbb{E}^1_w[P^2(q) - P^1(q)] + \mathbb{E}^1_w[Q^1(q) - Q^2(q)] = 0 + \mathbb{E}^1_w[Q^1(q) - Q^2(q)]$. This is unlike the polarization we see in real life, in which I can expect that my *own* opinion will get more extreme in one direction, and that your opinion will get more extreme in another.

A different question you may have is why care about the *expectational* sense of "predictable polarization", rather than (say) a version which says that divergence in a particular direction is "predictable" if it's highly likely you'll diverge in that direction. In general, I say let a thousand flowers bloom, so this may be a useful definition of "predictable polarization" for some purposes. But notice that there is a close connection between the two, through the law of large numbers.

Suppose neither of us will get predictably polarizing evidence in my sense. Moreover, suppose that there is class of people ("Democrats") who are going to get the same type of evidence as I will, and there is a different class of people ("Republicans") who will get the same type of evidence as you will. Let $P^{2,1}, P^{2,2}, ..., P^{2,n}$ be the future credence functions of the former group, and $Q^{2,1}, Q^{2,2}, ..., Q^{2,m}$ be the future credence functions of the latter. Suppose further that we all have the same opinions at time 1. Then it'll often be appropriate to treat the $P^{2,i}$ as i.i.d. with respect to my initial credences $\mathcal{P}^1_w$, and likewise for $Q^{2,i}$. By the law of large numbers, it follows that, as n gets large, I should (at time 1, i.e. according to $\mathcal{P}^1_w$) become arbitrarily confident that the average of the $P^{2,i}(q)$ will be very close to $\mathbb{E}^1_w[P^{2,i}(q)]$, which by the lack of predictable polarization equals $P^1_w(q)$. Likewise, as m gets large, for the average of the $Q^{2,i}$. In other words, without predictable polarization in the sense defined above, I must be quite confident that the *population* of people-like-me will, on average, be as confident of q are they are initially—and that this average will not diverge from the average confidence of people-like-you.

Much more could be said, but for now I'll leave it there.

The next step is to prove a close connection between potentially ambiguous evidence and predictably polarizing evidence in the sense defined.

First, let's formulate **unambiguous Bayesianism**—what I called the "standard theory" of epistemic rationality—and show that it rules out predictably polarizing evidence. The standard theory is that we can represent someone's beliefs with a probability function π that is fixed and known, and when someone learns something that evidence comes in the form of a partition Π which the agent will condition on (e.g. Weisberg 2017).[11] Note two facts. First, if π is fixed, that means it doesn't vary across worlds,

[11] *Notation:* I'll use 'P^1' etc. as definite descriptions for probability functions (i.e. as probability functions which vary across worlds), and use lowercase greek letters (π, δ,...) for rigid designators for probability functions, i.e. for probability functions which don't vary across worlds.

so $\{w : \pi(q) = t\}$ will either be W (if $\pi(q) = t$) or $\emptyset$ (if $\pi(q) \neq t$). Thus there is no uncertainty about what the rational credence function, π is. Second, if for every world $w \in \Pi(w)$, the posterior rational credence is gotten by updating π on $\Pi(w)$, then all the worlds in $\Pi(w)$ will agree on the rational posteriors, and thus every world in this set will be certain of what the rational posteriors are.

Because of this, the following definition is faithful to the standard Bayesian line (compare van Benthem 2011; van Ditmarsch et al. 2015, Ch. 4):

Definition 4.3. A frame $\langle W, \mathcal{P}^1, \mathcal{P}^2\rangle$ is **unambiguous Bayesian** iff $\mathcal{P}^1$ is unambiguous, and there is some partition Π of W such that for all w: $\mathcal{P}^2_w(\cdot) = \mathcal{P}^1_w(\cdot|\Pi(w))$.

Here's the first result:

Fact 4.4 ("Fact 1", in the blog post)**.** If $\langle W, \mathcal{P}^1, \mathcal{P}^2\rangle$ is unambiguous Bayesian, then $\mathcal{P}^2$ is not predictably polarizing (Weisberg 2007; Briggs 2009; Salow 2018).

Proof Sketch. Take an arbitrary world w. Since $\mathcal{P}^1$ is unambiguous (and reflexive), it follows that $\mathcal{P}^1_w(P^1 = \mathcal{P}^1_w) = 1$, so all worlds assigned positive probability by $\mathcal{P}^1_w$ agree on the rational credence function at time 1. Let $E^1_w := \{x : \mathcal{P}^1_w(x) > 0\}$. Since the frame is unambiguous Bayesian, there is a partition Π such that for all $x \in E^1_w$, $\mathcal{P}^2_x(\cdot) = \mathcal{P}^1_w(\cdot|\Pi(x))$, and hence that if $x \in \Pi_i \in \Pi$, then $\mathcal{P}^2_x(q) = \mathcal{P}^1_w(q|\Pi_i)$, and therefore $\mathbb{E}^1_w[P^2(q)|\Pi_i) = P^1_w(q|\Pi_i)$. Since Π partitions E^1_w, that means, for any $q \subseteq W$:

$$\begin{aligned}
\mathbb{E}^1_w[P^2(q)] &= \sum_{\Pi_i \in \Pi} \mathcal{P}^1_w(\Pi_i) \cdot \mathbb{E}^1_w[P^2(q)|\Pi_i] && \text{(total expectation)}\\
&= \sum_{\Pi_i \in \Pi} \mathcal{P}^1_w(\Pi_i) \cdot \mathcal{P}^1_w(q|\Pi_i)\\
&= \mathcal{P}^1_w(q). && \text{(total probability)}
\end{aligned}$$

□

An immediate consequence are the points mentioned above: if both you and I will have opinions that evolve in an unambiguous Bayesian way, then (1) if we share opinions initially then we can't expect to diverge in a given direction, and (2) insofar as I expect us to diverge, it's completely because I expect *your* confidence to shift. Precisely: let $\langle W, \mathcal{P}^1, \mathcal{P}^2, \mathcal{Q}^1, \mathcal{Q}^2\rangle$ be a two-person frame. Then:

Corollary 4.5. If both $\mathcal{P}^i$ and $\mathcal{Q}^i$ are unambiguous Bayesian, then:

(1) If $\mathcal{P}^1_w = \mathcal{Q}^1_w$ at all w, then $\mathbb{E}^1_w[P^2(q) - Q^2(q)] = 0$ at all w; and
(2) Regardless of what $\mathcal{Q}^1$ is, $\mathbb{E}^1_w[(P^2(q) - Q^2(q)) - (P^1(q) - Q^1(q))] = \mathbb{E}^1_w[Q^1(q) - Q^2(q)]$.

In short: unambiguous Bayesianism makes predictable polarization impossible.

It turns out that the constraint that evidence is unambiguous is crucial in this result. In fact, *whenever* evidence is ambiguous, it can be predictably polarizing.

The first step in showing this is to note that if $\mathcal{P}^2$ is potentially ambiguous, then this implies that there must be some world at which $\mathbb{E}^2[P^2(q)] \neq P^2(q)$, i.e. the time-2 rational confidence does not equal the time-2 expectation of the time-2 rational confidence:

Theorem 4.6 (Samet 2000)**.** If $\mathcal{P}^2$ is potentially ambiguous, then there is some w and q such that $\mathbb{E}^2_w[P^2(q)] \neq \mathcal{P}^2_w(q)$.

Proof Sketch. Supposing that for all w, q, $\mathbb{E}^2_w[P^2(q)] = \mathcal{P}^2_w(q)$, we show that $\mathcal{P}^2$ must be unambiguous. Note that $\mathcal{P}^2$ can be viewed as a (finite) Markov chain with W the states and $\mathcal{P}^2_w(w')$ the transition probabilities. As such, we can partition W into its communicating classes, $E_1, ..., E_n$, plus perhaps a set of transient states E_0. The claim that for all w, q, $\mathbb{E}^2_w[P^2(q)] = \mathcal{P}^2_w(q)$ is equivalent to the claim that for all w, $\mathcal{P}^2_w$ is a stationary distribution with respect to the Markov chain; i.e. where M is its transition matrix and v_w is the vector corresponding to $\mathcal{P}^2_w$, $v_w M = v_w$. Every E_i has a unique stationary distribution π_i, therefore if $w \in E_i$, $\mathcal{P}^i_w = \pi_i$, and therefore (since $\pi_i(E_i) = 1$) $\mathcal{P}^2_w(P^2 = \pi_i) = 1$. Moreover, E_0 must be empty, for any stationary of M assigns 0 probability to all transient states, and by definition $\mathcal{P}^2_w(w) > 0$, so since $\mathcal{P}^i_w$ must be stationary, $w \notin E_0$. Thus at *all* w, $\mathcal{P}^2_w(P^2 = \pi_i) = 1$, which implies that $\mathcal{P}^2$ is *not* potentially ambiguous. □

For an elementary proof of this result (without appeal to Markov chain convergence, etc.), see Dorst (2019).

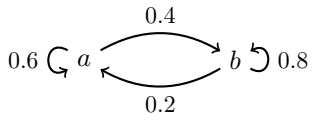
Using this result, it's easy to find frames that are predictably polarizing:

Fact 4.7 (**"Fact 2"**, in the blog post)**.** If $\mathcal{P}^2$ is potentially ambiguous, then there are frames $\langle W, \mathcal{P}^1, \mathcal{P}^2 \rangle$ on which $\mathcal{P}^2$ is predictably polarizing for $\mathcal{P}^1$.

Proof Sketch. By Theorem 4.6, there is a w and q such that $\mathbb{E}^2_w[P^2(q)] \neq \mathcal{P}^2_w(q)$; therefore any frame on which $\mathcal{P}^1_w = \mathcal{P}^2_w$ will be one on which $\mathcal{P}^2$ is predictably polarizing for $\mathcal{P}^1$. In fact, *any* $\mathcal{P}^1_w$ that is not stationary with respect to the transition-matrix corresponding to $\mathcal{P}^2$ will do. □

Example 4.8. Let our dynamic frame have $W = \{a, b\}$; for all w, $\mathcal{P}^1_w(a) = 0.5$, and $\mathcal{P}^2$ as described in the below Markov diagram or (equivalently) transition matrix:

Note that $\mathcal{P}^1_w = (0.5, 0.5)$, but $(0.5, 0.5)\begin{pmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{pmatrix} = (0.4, 0.6)$, which is just to say that $\mathbb{E}^1_w[P^2(b)] = 0.6 > 0.5 = P^1_w(b)$, so this frame is predictably polarizing on $\{b\}$. You start out (certain that you should be) 50% confident of b, but your best estimate is that



$$\begin{pmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{pmatrix}$$

Figure 4: A probability frame.

after getting the incoming evidence $\mathcal{P}^2$, your posterior confidence should be higher in $\{b\}$.

Note two things. First, this predictable shift is driven by an asymmetry in ambiguity: $\mathcal{P}^2_a$ is more ambiguous than $\mathcal{P}^2_b$, since it is less certain of what P^2 is: $\mathcal{P}^2_a$ is 60-40 split between $[P^2 = \mathcal{P}^2_a]$ and $[P^2 = \mathcal{P}^2_b]$, while $\mathcal{P}^2_b$ is 20-80 split between these two.

Second, despite this predictable shift in opinion, $\mathcal{P}^2$ *accuracy-dominates* $\mathcal{P}^1$: at every world w, $\mathcal{P}^2_w$ is uniformly more confident of all truths and less confident of all falsehoods than $\mathcal{P}^1_w$. (At a, $\mathcal{P}^2_a(a) = 0.6 > 0.5 = \mathcal{P}^1_a(a)$, and at b, $\mathcal{P}^2_b(b) = 0.8 > 0.5 = \mathcal{P}^1_b(b)$.) As such, the transition from $\mathcal{P}^1$ to $\mathcal{P}^2$ validates the **value of evidence**.

I turn now to formalizing and generalizing this notion, so that we can state **Rationality as Value** more precisely, and explain how it subsumes unambiguous Bayesianism and permits predictable polarization.

4.2 Rationality as Value

The basic idea behind the value of evidence is this. Given some options, the rational option is that which maximizes expected value relative to the rational credence function. If you have are unsure what the rational credence function is (or will be), then we can consider the expected value of "doing the rational thing—whatever it turns out to be." Evidence is valuable iff, no matter what options you have, the expected value of doing the rational thing given that evidence is never lower than simply ignoring the evidence and picking an option. The basic idea was made famous by Good (1967); see Salow (2020) for a modern exposition, as well as Huttegger (2014); Ahmed and Salow (2018); Das (2020). Here I'll follow most closely the formalism and results of Dorst (2020a).

Here's how to formalize the value of evidence. Note that W is a partition of logical space, and thus can be thought of as a question (Hamblin 1976; Roberts 2012). Thus when we use such a model, we are implicitly relativizing both the agent's beliefs and the options they face to a question—namely, "Which $w \in W$ is actual?" In the coming weeks, this relativization to questions will be important, so I'll make it explicit in what follows.

Given a question W, and a decision of which of a set of options to perform, standard decision theory treats your options $O_1, ..., O^n$ as simply functions from worlds (answers to the question) w to numbers $O_i(w)$ representing the amount of value (or "utility") that option O_i yields at w. In other words, your options are simply a (I'll assume finite)

set of random variables whose values are determined by the cells of the partition W. In such a case, I'll say that the decision between the O_i is a **decision based on** W.

The rational thing to do, given question W, a rational credence function π, and these options, is to choose an option O_i that maximizes expected value relative to π: an option O_i such that, for all O_j, $\mathbb{E}_\pi[O_i] \geq \mathbb{E}_\pi[O_j]$ (where $\mathbb{E}_\pi[X] := \sum_t \pi(X = t) \cdot t$).

In our probability frames, the rational credence function for you to have varies across worlds. Thus, given options $O_1, ..., O_n$, we can consider the option "doing the rational thing, whatever it is". Let r be a function from probability functions π to options O_i that maximize expected utility relative to π—it is a policy for choosing expectedly-best options based on your probabilities. Since "P^i" is a definite description for "the rational credence function at time i", that means that "$r(P^i)$" is a definite description for "the option chosen by policy r given the rational credence function P^i, whatever it is". Thus we can treat $r(P^i)$ as a random variable as well, such that $r(P^i)(w) = r(\mathcal{P}^i_w)(w)$, i.e. the actual value, at w, of picking an option $r(\mathcal{P}^i_w)$ that maximizes expected value according to $\mathcal{P}^i_w$.

Say that a probability function π *values* transitioning to credence function P^i, relative to question W, iff, for *any* decision based on W, the expected value of (1) first transitioning and then rationally choosing an option is always higher than (2) simply choosing an option without first transitioning. Formally:

Definition 4.9. A probability function π **values** P^i relative to question W iff, for any finite set of options $O_1, ..., O_n$ based on W and policy r of choosing maximal-expected-utility options, $E_\pi[r(P^i)] \geq E_\pi[O_j]$, for all O_j. A probability frame $\langle W, \mathcal{P}^1, \mathcal{P}^2 \rangle$ validates the **value of evidence** iff, if $i \leq k$, then for all w, P^i_w values P^k.[12]

The theory I stated in the blog post, now stated precisely, is:

> **Rationality as Value:** A transition from beliefs π to beliefs P^i is rational iff, π values P^i relative to the live question W.

The remaining unanalyzed notion here is what counts as a "live" question W. Usually in decision theory this question is not directly addressed, but "modeled around" by using a partition/question that is fine enough to represent what's of interest for modeling purposes. We could do the same, and simply always make sure that we use a W fine enough to capture all relevant distinctions. However, I'll argue next week that there is an interesting and important sense in which it makes sense to say that a given transition is (at least boundedly) rational so long as it it is valuable relative to a relevant but relatively coarse-grained question.

The first result we can show is that Rationality as Value subsumes unambiguous Bayesianism if we assume that evidence is never ambiguous:

[12] The constraint $i \leq k$ makes it so that at time 1 you value both the time-1 rational credences and time-2 ones, while at time 2 you value only the time-2 ones.

Fact 4.10 ("**Fact 3**" in the blog post)**.** If $\mathcal{P}^1$ and $\mathcal{P}^2$ are not potentially ambiguous, then if $\langle W, \mathcal{P}^1, \mathcal{P}^2\rangle$ validates the value of evidence, then it is an unambiguous Bayesian model.

Proof Sketch. We need to show that there is some partition Π such that, for all w, $\mathcal{P}^2_w(\cdot) = \mathcal{P}^1_w(\cdot|\Pi(w))$. Let our partition $\Pi = \{[P^2 = \pi_1], ..., [P^2 = \pi_n]\}$, i.e. be the possible future rational credence functions. By reflexivity, every w in $[P^2 = \pi_1]$ has $\mathcal{P}^2_w(P^2 = \mathcal{P}^2_w) > 0$. Since $\mathcal{P}^2$ is unambiguous, it follows that $\mathcal{P}^2_w(P^2 = \mathcal{P}^2_w) = 1$, so that for all x: if $\mathcal{P}^2_w(x) > 0$, then $\mathcal{P}^2_x = \mathcal{P}^2_w$. Suppose, for reductio, there is an x such that $\mathcal{P}^2_x(\cdot) \neq \mathcal{P}^1_x(\cdot|\Pi(x)) = \mathcal{P}^1_x(\cdot|P^2 = \mathcal{P}^2_x)$. WLOG, we can assume that, for some q, $\mathcal{P}^2_x(q) = t > t - \epsilon = \mathcal{P}^1_x(q|P^2 = \mathcal{P}^2_x)$, for some $\epsilon > 0$. Since every world in this partition cell has the same credence function, that means $[P^2 = \mathcal{P}^2_x] \subseteq [P^2(q \wedge [P^2 = \mathcal{P}^2_x]) \geq t]$.

We construct a value-of-evidence failure as follows. Let $O_1 = 0$ everywhere, and for $a > 0$, let

$$O_2 = \begin{cases} 1 - t + a & \text{if } q \wedge [P^2 = \mathcal{P}^2_x] \\ -t & \text{if } \neg q \wedge [P^2 = \mathcal{P}^2_x] \\ 0 & \text{otherwise} \end{cases}$$

Let r be any rational-option function, i.e. function from π to an option in $\{O_1, O_2\}$ that maximize expected utility relative to π. Note that $r(\mathcal{P}^2_x) = O_2$, since $\mathbb{E}^2_x[O_2] = t \cdot (1 - t + a) + (1 - t)(-t) = ta > 0$, while $\mathbb{E}^2_x[O_1] = 0$. This means $\mathbb{E}^1_x[r(P^2)|P^2 = \mathcal{P}^2_x] = \mathbb{E}^1_x[O_2|P^2 = \mathcal{P}^2_x]$. Outside of $[P^2 = \mathcal{P}^2_x]$, both options yield 0 value everywhere, thus $\mathbb{E}^1_x[r(P^2)] \geq 0$ iff $\mathbb{E}^1_x[r(P^2)|P^2 = \mathcal{P}^2_x] \geq 0$, iff $\mathbb{E}^1_x[O_2|P^2 = \mathcal{P}^2_x] \geq 0$. We show that this can be made false by choosing a small enough.

We know $\mathcal{P}^1_x(q \wedge [P^2 = \mathcal{P}^2_x]|P^2 = \mathcal{P}^2_x) = \mathcal{P}^1_x(q|P^2 = \mathcal{P}^2_x) = t - \epsilon$. Thus

$$\begin{aligned} \mathbb{E}^1_x[O_2|P^2 = \mathcal{P}^2_x] &= (t - \epsilon)(1 - t + a) + (1 - t + \epsilon)(-t) \\ &= t - t^2 + ta - \epsilon + t\epsilon - \epsilon a - t + t^2 - t\epsilon \\ &= ta - \epsilon - \epsilon a = a(t - \epsilon) - \epsilon. \end{aligned}$$

Sending $a \to 0$, this will go negative. Thus $\mathbb{E}^1_x[O_2|P^2 = \mathcal{P}^2_x] < 0$, and hence, by the above, $\mathbb{E}^1_x[r(P^2)] < 0 = \mathbb{E}^1_x[O_2]$, meaning that $\mathcal{P}^1_x$ does not value P^2, so $\langle W, \mathcal{P}^1, \mathcal{P}^2\rangle$ does not validate the value of evidence. □

An immediate corollary of Fact 4.4 and Fact 4.10 is that if evidence is unambiguous and satisfies the value of evidence, it is not predictably polarizing.

The final result we need to establish ("**Fact 4**" in the blog post) is that Rationality as Value permits ambiguous evidence—and, therefore, predictably polarizing evidence. We've already seen an example of this in Example 4.8, but for the arguments to come we will need a much more general characterization of transitions that satisfy Rationality as Value.

A fully general characterization of the value of evidence in the context of ambiguous evidence has been an open question for years. In a new paper some co-authors and I have finally have gotten such a result (Dorst et al. 2020), but its contours remain to be fully explored; moreover, we do not yet have tractable algorithms for generating and testing such general models.

Instead we can use a result proved by Geanakoplos (1989) (and generalized in Dorst 2020a) to characterize a tractable sub-class of the frames the validate the value of evidence.

Let the i-**neighborhood** of w be the set of worlds assigned positive probability by evidence i at w: $E^i_w := \{x : \mathcal{P}^i_w(x) > 0\}$. Instead of focusing on general belief-transitions between a probability function π and P^i, we will focus on those that can be obtained by conditioning π on the various i-neighborhoods—this will be our tractable subclass of probability frames:

Definition 4.11. $\langle W, \pi, \mathcal{P}^i\rangle$ is a **conditioning update** iff, for all $w \in W$, $\mathcal{P}^i_w(\cdot) = \pi(\cdot|E^i_w)$. $\langle W, \mathcal{P}^1, \mathcal{P}^2\rangle$ represents a (dynamic) **conditioning frame**[13] iff there is a π such that both $\langle W, \pi, \mathcal{P}^1\rangle$ and $\langle W, \pi, \mathcal{P}^2\rangle$ are conditioning updates.

Note that if $\langle W, \pi, \mathcal{P}^i\rangle$ is a conditioning update, then $\mathcal{P}^i$ is not potentially ambiguous iff, the i-neighborhoods form a partition of $E := \{w : \pi(w) > 0\}$. Such an update is unambiguous Bayesian. However, conditioning updates allow for ambiguity whenever the i-neighborhoods are not partitional (cf. Williamson 2000, 2008, 2014, 2019; Cresto 2012; Lasonen-Aarnio 2013; Ahmed and Salow 2018; Salow 2018, 2019; Das 2020).

Definition 4.12. $\langle W, \pi, \mathcal{P}^i\rangle$ is a **forest update**[14] iff it is a conditioning update and within $E := \{w : \pi(w) > 0\}$, we have:

(1) E^i is *reflexive*: for all $w \in E$, $w \in E^i_w$;
(2) E^i is *transitive*: for all $w, x \in E$, if $x \in E^i_w$, then $E^i_x \subseteq E^i_w$; and
(3) E^i is *nested*: for all x, y in E, either $E^i_x \subseteq E^i_y$ or $E^i_x \supseteq E^i_y$ or $E^i_x \cap E^i_y = \emptyset$.

Given these definitions, we can state a characterization of the value of evidence within conditioning updates:

Theorem 4.13 (Geanakoplos 1989; this is "**Fact 4**" in the blog post)**.** Assume $\langle W, \pi, \mathcal{P}^i\rangle$ is a conditioning update. Then π values $\mathcal{P}^i$ iff $\langle W, \pi, \mathcal{P}^i\rangle$ is a forest update.

(The proof is complicated; see (Geanakoplos 1989, Theorem 1) for one direction of it; (Dorst 2020a, Theorem 7.4) generalizes the condition to entire frames, and proves both directions.)

[13] A.k.a. a "prior frame" in the terminology of Dorst (2020a,b).
[14] So called because the the binary relation xRy iff $\mathcal{P}^i_x(y) > 0$ has the structure of a forest once we mod out on equivalent worlds; see (Dorst 2020a, Appendix A).

Since any forest update which is not a partition is one in which $\mathcal{P}^i$ is ambiguous, this generates a large class of ambiguous-but-valuable updates. Moreover, many (in fact *all*, I believe—but won't prove it here) such updates in involve predictable polarization, as can be seen in the following simple example:

Example 4.14. $W = \{a, b\}$; for all w, $\mathcal{P}^1_w(a) = 0.5$, and $E^2_a = \{a, b\}$ while $E^2_b = \{b\}$. The frame thus looks like this, with arrows from w representing the worlds visible under E^2_w:

Thus $\mathcal{P}^2_a = \mathcal{P}^1_w$, while $\mathcal{P}^2_b = \mathcal{P}^1_w(\cdot|\{b\})$. Note that $\mathcal{P}^1_w(b) = 0.5$, but $\mathbb{E}^1_w[P^2(b)] = 0.5 \cdot 0.5 + 0.5 \cdot 1 = 0.75$, so this update is predictably polarizing on $\{b\}$.

In some of the arguments to come, will use Theorem 4.13's characterization of ambiguous-but-valuable evidence, along with various tractable characterizations of forest updates (e.g. Dorst 2020a, Theorem 5.8) to generate large, random updates that are ambiguous but satisfy the value of evidence, in order to track various statistical properties that they have.

5 Profound, Persistent and Predictable Polarization

Here's the link to the full post, published on October 3, 2020.

> **Post Synopsis:** I explained how, in principle, the mechanism of ambiguous evidence can lead to predictable polarization that is both **profound** (both sides disagree massively) and **persistent** (neither side loses confidence upon discovering this disagreement). This forms the theoretical foundation on which I will build the argument that the empirical mechanisms that drive real-world polarization are rational.
>
> **Appendix Summary:** In this appendix entry, I'll (1) offer several variant models of the word-completion task, explaining why they all validate the value of evidence, and (2) argue that by iterating such tasks, we can arrive at profound, persistent polarization through rational mechanisms.

5.1 Models of the word-completion task

In the blog post, the model I used of the word-completion task was this one:

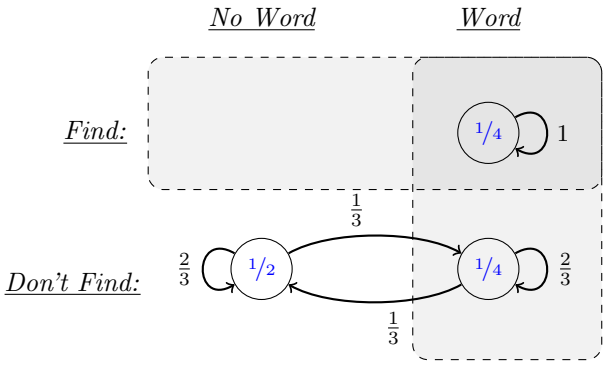


Figure 5: A graded asymmetry in the word-completion task.

There are three possibilities: wf, where there's a word and you find one (top right); $w\overline{f}$, where there's a word and you don't find one (bottom right); and $\overline{w}\overline{f}$, where there's no word and you don't find one (bottom left). Before getting the evidence, you should be $1/2$ confident you'll end up in $\overline{w}\overline{f}$, and $1/4$ confident you'll end up in each of $w\overline{f}$ and wf—as indicated by the blue numbers. Thus in our probability frame, for all possibilities w $\mathcal{P}^1_w(wf) = \mathcal{P}^1_w(w\overline{f}) = 1/4$, while $\mathcal{P}^1_w(\overline{w}\overline{f}) = 1/2$.

Meanwhile, after getting the evidence, if there's a word and you find one, you should be certain of this: $\mathcal{P}^2_{wf}(wf) = 1$. If there's no word and you don't find one, you should be $\frac{2}{3}$ confident of that, and $\frac{1}{3}$ confident that there is a word and you didn't find one: $\mathcal{P}^2_{\overline{w}\overline{f}}(\overline{w}\overline{f}) = \frac{2}{3}$ and $\mathcal{P}^2_{\overline{w}\overline{f}}(w\overline{f}) = \frac{1}{3}$. If there's a word but you don't find one, you should

be $\frac{2}{3}$ confident of this and $\frac{1}{3}$ confident that there's no word and you didn't find one: $\mathcal{P}^2_{w\overline{f}}(w\overline{f}) = \frac{2}{3}$, while $\mathcal{P}^2_{w\overline{f}}(\overline{w}\overline{f}) = \frac{1}{3}$.

Note that evidence is ambiguous when you don't find a word, because at both possibilities you should leave open that the rational credence that there's a world might be $\frac{1}{3}$ (if you're at $\overline{w}\overline{f}$) or $\frac{2}{3}$ (if you're at $w\overline{f}$).

Intuitively, we can see that this frame will satisfy the value of evidence because in transitioning from $\mathcal{P}^1$ to $\mathcal{P}^2$, each world gets uniformly more confident in truths and less confident in falsehoods, since the probabilities become more centered on the actual world.

Formally, the easiest way to prove this is to invoke a theorem from Dorst et al. (2020). For any candidate for the rational credence function $\mathcal{P}^i_w$, let $\widehat{\mathcal{P}}^i_w := \mathcal{P}(\cdot|P^i = \mathcal{P}_w)$ be the credence function that would be rational were $\mathcal{P}^i_w$ to be **informed** that it was the rational credence function—i.e. were its higher-order doubts about its ambiguous evidence to be removed (cf. Elga 2013; Dorst 2019; Stalnaker 2019). Say that $\mathcal{P}^i$ is **class-convex** iff for every w, $\mathcal{P}^i_w$ is in the convex hull of $\{\mathcal{P}^i_x : x \in W \text{ and } \mathcal{P}^i_x \neq \mathcal{P}^i_w\} \cup \{\widehat{\mathcal{P}}^i_w\}$. In other words, $\mathcal{P}^i_w$ can be obtained by a mixture of these other probability functions: there are some $\lambda_{wx} \geq 0$ which sum to 1 such that $\mathcal{P}^i_w = \lambda_{ww}\widehat{P}^i_w + \sum_{\mathcal{P}^i_x : \mathcal{P}^i_x \neq \mathcal{P}^i_w} \lambda_{wx}\mathcal{P}^i_x$. The characterization of the value of evidence in general, which generalizes Theorem 4.13, is:

Theorem 5.1 (Dorst et al. 2020)**.** *π values $\mathcal{P}^i$ relative to W iff $\langle W, \mathcal{P}^i\rangle$ is class-convex and π is in the convex hull of $\{\mathcal{P}^i_w : w \in W\}$.*[15]

Given this, it's not hard to see that, in our word-completion frame, $\mathcal{P}^1_w$ values $\mathcal{P}^2$. $\mathcal{P}^2$ is class-convex because for each w, $\widehat{\mathcal{P}}^2_w(w) = 1$ (removing all evidential ambiguities makes you certain of which possibility you're in), and:

- $\mathcal{P}^2_{wf} = \widehat{\mathcal{P}}^2_{wf}$
- $\mathcal{P}^2_{w\overline{f}} = \frac{1}{2}\widehat{\mathcal{P}}^2_{w\overline{f}} + \frac{1}{2}\mathcal{P}^2_{\overline{w}\overline{f}}$, since $\frac{1}{2}(0,1) + \frac{1}{2}(\frac{2}{3}, \frac{1}{3}) = (\frac{1}{3}, \frac{2}{3})$.
- $\mathcal{P}^2_{\overline{w}\overline{f}} = \frac{1}{2}\widehat{\mathcal{P}}^2_{\overline{w}\overline{f}} + \frac{1}{2}\mathcal{P}^2_{w\overline{f}}$, since $\frac{1}{2}(1,0) + \frac{1}{2}(\frac{1}{3}, \frac{2}{3}) = (\frac{2}{3}, \frac{1}{3})$.

Meanwhile, $\mathcal{P}^1_w$ is in the convex hull of the $\{\mathcal{P}^2_x\}$ since $(1/2, 1/4, 1/4) = \frac{3}{4}(\frac{2}{3}, \frac{1}{3}, 0) + 0(\frac{1}{3}, \frac{2}{3}, 0) + \frac{1}{4}(0, 0, 1)$.

Thus this model validates the value of evidence.

As we've seen, it's predictably polarizing because, letting $word = \{wf, w\overline{f}\}$, we have that $\mathcal{P}^1_w(word) = \frac{1}{2}$ but $\mathbb{E}^1_w[P^2(word)] = \frac{1}{2}\cdot\frac{1}{3} + \frac{1}{4}\cdot\frac{2}{3} + \frac{1}{4}\cdot 1 = \frac{7}{12} > \frac{1}{2}$.

There are many, many other models we could use to illustrate the same point. Figure 6 gives a simple one that uses a conditioning update (Definition 4.11), i.e. where $\mathcal{P}^2_w$ is always recoverably from $\mathcal{P}^1_w$ by conditioning on some proposition.

In this model, whenever there's a word, you should be sure that there is—even if you don't in fact find it, you *should* find it. $\mathcal{P}^1_w(word) = \frac{1}{2}$ but $\mathbb{E}^1_w[P^2(word)] = \frac{1}{2}\cdot\frac{1}{3} + \frac{1}{2}\cdot 1 =$

[15] Note: this assumes that $\mathcal{P}^i$ is reflexive; if it is not, then we simply restrict W to the set of worlds assigned positive probability by π; all of these must be reflexive for π to value them.

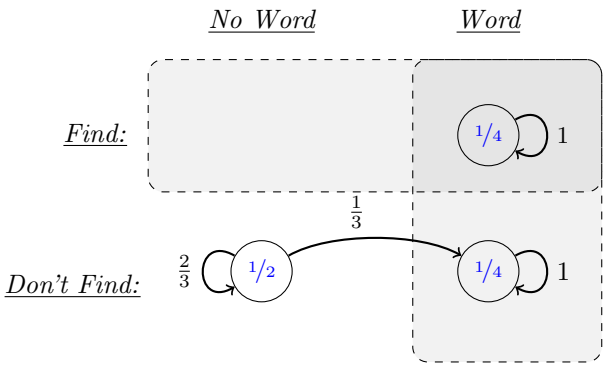


Figure 6: A conditioning asymmetry in the word-completion task.

$\frac{2}{3}$. It validates the value of evidence by Theorem 4.13 because it is a forest update (Definition 4.12).

Figure 7 gives a more realistic conditioning update, in which there is a range of levels confidence it might be rational to have if you don't find a word.

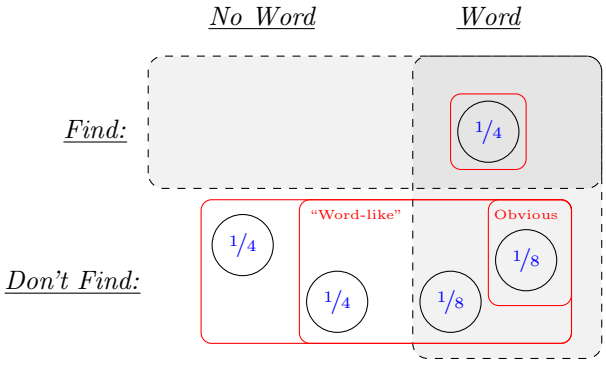


Figure 7: A complex conditioning asymmetry in the word-completion task; here posterior probabilities at w are obtained by conditioning the prior (blue fractions) on the smallest red rectangle that contains w.

In this model, if there's no word and you don't find one there's some chance you should be $\frac{2}{3}$ confident of this, and also some chance you should only be $\frac{1}{2}$ confident of this; meanwhile, if there is a word and you don't find it, there's some chance you should be $\frac{1}{2}$ confident of this, and some chance you should be certain of this (you should find a word). Labeling the 4 possibilities in the *Don't Find* section a, b, c, d from left to right, we have $\mathcal{P}^2_a(word) = \frac{1}{3}$, $\mathcal{P}^2_b(word) = \mathcal{P}^2_c(word) = \frac{1}{2}$, and $\mathcal{P}^2_d(word) = 1$. Interpretively, we might think of $\{b, c, d\}$ as possibilities where the letter-string looks "word-like", and $\{d\}$ as one where there is *obviously* a completion which you should find (even though, in fact, you don't).

In this model, $\mathbb{E}^1_w[P^2(word)] = \frac{1}{4} \cdot \frac{1}{3} + \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{8} \cdot \frac{1}{2} + \frac{1}{8} \cdot 1 + \frac{1}{4} \cdot 1 = {}^{31}/_{48} \approx 0.65$, while

of course $\mathcal{P}^1_w(word) = \frac{1}{2}$. Again, this model validates the value of evidence because it is a forest update (Definition 4.12).

In all these cases, we can use the resulting evidence to polarize people by dividing them up into Headsers and Tailsers. Using the simple graded asymmetric model from Figure 5, this yields the following different models for Headsers and Tailsers:

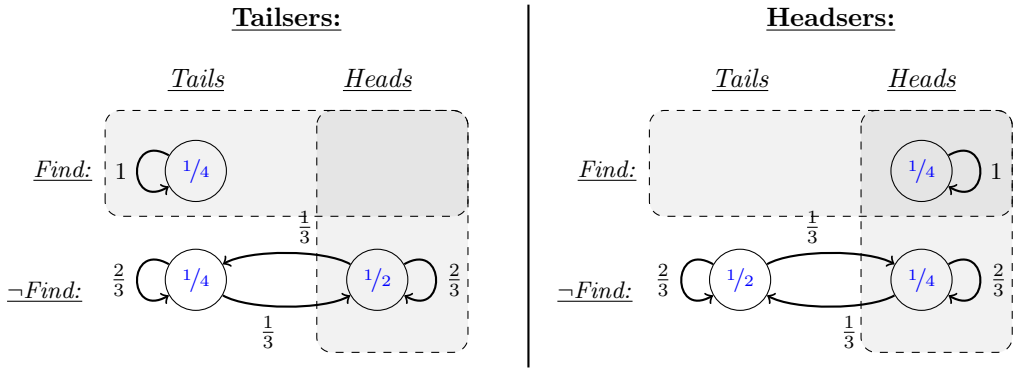


Figure 8: Headsers vs. Tailsers in simple graded-asymmetry model of word-completion task

In this model, both Headsers and Tailsers start out 50% confident the the coin will land heads, but if it does then Tailser should be $\frac{2}{3}$ confident it did while Headsers should be either $\frac{2}{3}$ or 1, while if it doesn't land heads then Tailsers should be either 0 or $\frac{1}{3}$ confident that it did, while Headsers should be $\frac{1}{3}$. In expectation, Headsers should end up $\frac{7}{12}$ confident of heads, while Tailsers should end up $\frac{5}{12}$ confident of it.

Similar lessons could be derived using any of our variant models of the word-completion task above.

5.2 Profound and Persistent Polarization

How can we use these models to generate profound and persistent polarization? The basic idea is to iterate this process n times with n independent coin flips and word-completion tasks for both Headsers and Tailsers. As n grows large, if the evidence in each task works as above, then we can all expect Headsers to become confident that the proportion of heads is roughly $\frac{7}{12}$, while Tailsers will become confident it's roughly $\frac{5}{12}$.

Focus on a given Headser, though the same reasoning will apply to Tailsers. Let Q_i be the question of what happens on the ith toss: will the coin land heads, and will our Headser find a completion or not? Thus $Q_i = \{H_iF_i, H_i\overline{F}_i, \overline{H}_i\overline{F}_i\}$. Let $\sqcap_i Q_i$ be the partition obtained by crossing all the Q_i. The key formal result is that even if the Q_i are logically and evidentially independent at all times (so that the rational credence in the answers to some of the Q_i have no bearing on the rational credence in the others), if

evidence is ambiguous then it can nevertheless be the case that the evidential transition is valuable with respect to *each* Q_i, but predictably and profoundly polarizing with respect to their join $\sqcap_i Q_i$. (And, as a result, is *not* valuable with respect to $\sqcap_i Q_i$.) This is impossible when evidence is unambiguous.[16]

More precisely, let the series of rational credence functions $\mathcal{P}^1, ..., \mathcal{P}^n$ be defined as follows, offering a simple iteration of our above model for our Headser:

Simple Headser iteration:

- For $i > k$, for all w: $\mathcal{P}^k_w(H_iF_i) = \mathcal{P}^k_w(H_i\overline{F}_i) = \frac{1}{4}$, while $\mathcal{P}^k_w(\overline{H}_i\overline{F}_i) = \frac{1}{2}$.
- For $i \leq k$:
  - If $w \in H_iF_i$, then $\mathcal{P}^k_w(H_iF_i) = 1$;
  - If $w \in H_i\overline{F}_i$, then $\mathcal{P}^k_w(H_i\overline{F}_i) = \frac{2}{3}$ and $\mathcal{P}^k_w(\overline{H}_i\overline{F}_i) = \frac{1}{3}$;
  - If $w \in \overline{H}_i\overline{F}_i$, then $\mathcal{P}^k_w(\overline{H}_i\overline{F}_i) = \frac{2}{3}$ and $\mathcal{P}^k_w(H_i\overline{F}_i) = \frac{1}{3}$.

As a result, for each k, $\langle Q_k, \mathcal{P}^{k-1}, \mathcal{P}^k\rangle$ is equivalent to the simple graded-asymmetry model for our Headser represented on the right of Figure 8. Since these models all validate the value of evidence, that means that each time the Headser is presented with a word-completion task, the evidence they are presented is valuable with respect to the question they are then presented with—namely, "Will this coin land heads, and will I find a heads?"

Moreover, the answers to the Q_i (as well as facts about the rational credence in those answers) are mutually independent at all times k.

It follows, I claim, that at each time k, it is (at least boundedly) rational to obtain evidence $\mathcal{P}^k$ at time k. There are a variety of ways to support this conclusion, but the basic idea is simple enough. People cannot consider all the distinctions generated by fine-grained question like $\sqcap_i Q_i$, for such distinctions lead to a combinatorial explosion that makes probabilistic inference (or, really, any inference) over them intractable (Dagum and Luby 1993). With only 10 coin flips, that is already $3^{10} = 59046$ possibilities to track. Any sense of "rational" in which its an open question whether humans are reasoning rationally clearly is not one on which they must be expected to track all these distinctions; instead, some more minimal or "bounded" notion of rationality is the operative one (Simon 1956, 1976; Cherniak 1986). A plausible such notion is this: when faced with some evidence about a question that is independent of all other relevant questions, it is rational to gather that evidence if *with respect to that question*, the evidence is valuable. If so, then each step in the transition from $\mathcal{P}^1$

[16] *Proof sketch:* If evidence is unambiguous and valuable with respect to each Q_i, then for each $q_i \in Q_i$, $\mathcal{P}^0_w(q_i|P^n = \pi) = \pi(q_i)$ (Elga 2013; Gallow 2017; Dorst 2020a). To show that it's valuable with respect to $\sqcap_i Q_i$, we need to show that for any $q_{k_i} \in Q_i$: $\mathcal{P}^0_w(q_{k_1} \cap ... \cap q_{k_n}|P^n = \pi) = \pi(q_{k_1} \cap ... \cap q_{k_n})$. An induction on the size of n, plus independence gives us that for any such π, $\pi(q_{k_1} \cap ... \cap q_{k_n}) = \pi(q_{k_1}) \cdots \pi(q_{k_n})$, and that $\mathcal{P}^0_w(q_{k_1} \cap ... \cap q_{k_n}|P^n = \pi) = \mathcal{P}^0_w(q_{k_1} \cap ... \cap q_{k_{n-1}}|P^n = \pi) \cdot \mathcal{P}^0_w(q_{k_n}|P^n = \pi, q_{k_1} \cap ... \cap q_{k_n}) = \pi(q_{k_1}) \cdots \pi(q_{k_{n-1}}) \cdot \mathcal{P}^0_w(q_{k_n}|P^n = \pi) = \pi(q_{k_1}) \cdots \pi(q_{k_n}) = \pi(q_{k_1} \cap ... \cap q_{k_n})$.

to $\mathcal{P}^2$ to... to $\mathcal{P}^n$ is rational. In effect, the claim is that people *can't* assess whether the evidence is valuable with respect to all possible questions (or even the fine-grained question $\sqcap_i Q_i$), so instead they must simply check whether the evidence is "locally" valuable with respect to the relevant questions at issue.

UPDATE: I've recently discovered that stronger result is possible, though I don't yet have the time to write out all the details. What's true is this. Let Q be the question of how all the coins land in all the tosses, i.e. the partition $\{H_1H_2....H_n, H_1...H_{n-1}T_n, T_1...T_n\}$. Then a slight variant on the above series of transitions has the following result. At each time i, P^i values the transition to P^{i+1} with respect to question Q—so at each stage you expect that you're getting more accurate in your beliefs about Q. Yet P^0 expects with high confidence that P^n will be strongly polarized, and thus P^0 does not value P^n. In effect, we have a diachronic tragedy where at each stage you're doing what makes sense in gathering the evidence, but by the end you've foreseeably gotten much less accurate about the overall distribution of heads. The trick, which involves technicalities I can't add just yet but will add soon, is to "zero out" your higher-order uncertainty about Q in between each coin that's presented. More details forthcoming.

On this way of setting things up, I think it should be relatively uncontroversial that there's a substantive sense in which each transition is epistemically rational, despite the fact that collectively they are predictably, profoundly polarizing.

So let's suppose that gathering the evidence at each stage is rational. What happens our Headser goes through all these transitions? Consider what can be expected at time 0, before she's looked at any of the tasks. Since each of these tasks are independent, for each $1 \leq i \leq n$, $P^n(H_i)$ is an i.i.d. random variable with respect to $\mathcal{P}^0_w$, such that $\mathbb{E}^0_w[P^n(H_i)] = \frac{7}{12}$. By the weak law of large numbers, it follows that at $n \to \infty$, $\mathcal{P}^0$ become arbitrarily confident that the mean of the $P^n(H_i)$ is arbitrarily close to $\frac{7}{12}$. Thus at the starting stage, both Headsers and Tailsers will be arbitrarily confident that Headsers will end up with an average confidence in heads (across coin flips) close to $\frac{7}{12}$. By exactly parallel reasoning, they both can be arbitrarily confident that Tailsers will end up with an average credence in heads close to $\frac{5}{12}$.

Using this, we can establish that they will predictably (with arbitrarily high probability) be such that Headsers will end up arbitrarily confident of some proposition q, while Tailsers will end up arbitrarily confident in its negation.

Precisely, let ***Mostly-Heads*** = *more than half the tosses landed heads.* We know (with high probability) that Headsers will wind up with an average confidence in heads around $\frac{7}{12}$. If follows that they will be arbitrarily confident that roughly $\frac{7}{12}$ of the coins landed heads. To see this, let A be the set of H_i which they should assign credence 1 to, B be the set they should assign credence $\frac{2}{3}$ to, and C be the set they should assign credence $\frac{1}{3}$ to. Since the coins are independent, the H_i in each of these sets are, relative to P^n_H, i.i.d., and therefore P^n_H will be arbitrarily confident that roughly $\frac{2}{3}$ of the tosses

in set B landed heads and that roughly $\frac{1}{3}$ of the tosses in C landed heads (and they will be certain that all those in A landed heads). At the outset we are arbitrarily confident that $|A| \approx \frac{1}{4}n$, $|B| \approx \frac{1}{4}n$, and $|C| \approx \frac{1}{2}n$. Thus we are arbitrarily confident that Headsers will end up arbitrarily confident that roughly $\frac{1}{4}n + \frac{2}{3} \cdot \frac{1}{4}n + \frac{1}{3} \cdot \frac{1}{2}n$ of the coins landed heads, i.e. that roughly $\frac{7}{12}n$ of the coins landed heads.

It follows that we can be arbitrarily confident at the outset that Headers should end up arbitrarily confident of *Mostly-Heads*: $\mathcal{P}^0_w(P^n_H(\textit{Mostly-Heads}) \approx 1) \approx 1$. By parallel reasoning, we can also be arbitrarily confident that Tailsers will end up arbitrarily confident that *Mostly-Heads* is *false*. Letting P^n_T be the posterior rational Tailser confidence, we have $\mathcal{P}^0_w(P^n_T(\textit{Mostly-Heads}) \approx 0) \approx 1$.

That is predictable, **profound** polarization: there are propositions such that Headsers will, predictably, become arbitrarily more confident of them than Tailsers are.

A similar line of reasoning shows that this polarization will be *persistent*, in the sense that learning of these disagreements will not dislodge them. In particular, a variant on the above argument shows that even after Headsers have gone though the process, and now should be confident that roughly $\frac{7}{12}$ of the coins landed heads, they should still be arbitrarily confident that Tailsers should be arbitrarily confident that less than $\frac{1}{2}$ of them landed heads. In particular, Headsers should think that on roughly $\frac{2.5}{12}$ of the tosses, Tailsers should have credence 0, on roughly $\frac{2.5}{12}$ they should have credence $\frac{1}{3}$, and on the remaining $\frac{7}{12}$ they should have credence $\frac{2}{3}$. By the same law-of-large-numbers argument, they then should be arbitrarily confident that Tailsers should be arbitrarily confident that roughly $\frac{2.5}{12} \cdot \frac{1}{3} + \frac{7}{12}\frac{2}{3} = \frac{11}{24} \approx 0.458$ of the coins landed heads. Thus Headsers should be arbitrarily confident that Tailsers should be arbitrarily confident that *Mostly-Heads* is false: $P^n_H(P^n_T(\textit{Mostly-Heads}) \approx 0) \approx 1$.

It follows that Headsers should be unmoved when they find out that Tailsers disagree with them: $P^n_H(\textit{Mostly-Heads} \mid P^n_T(\textit{Mostly-Heads}) \approx 0) \approx P^n_H(\textit{Mostly-Heads}) \approx 1$. A parallel argument establishes that, likewise, Tailsers should be unmoved when they discover that Headsers should be confident of Mostly-Heads: $P^n_T(\textit{Mostly-Heads} \mid P^n_H(\textit{Mostly-Heads}) \approx 1) \approx P^n_T(\textit{Mostly-Heads}) \approx 0$.

Thus, through rational processing of ambiguous evidence, it's possible to for people who agree on everything at time 0 to predict with arbitrary confidence that one of them should wind up very confident of q, the other should wind up very confident of $\neg q$, and that neither of them should be surprised or moved when they discover that this disagreement has come to pass. That is how, I claim, predictable, profound, persistent polarization can be rational.

In the rest of this series, I'll argue that this theoretical possibility is not an idle one: rational processing of ambiguous evidence can help explain many of the real-world processes that drive polarization.

6 Confirmation Bias as Avoiding Ambiguity

Here's the link to the full post, published on October 17, 2020.

> **Post Synopsis:** I explained how a form of confirmation bias known as **biased assimilation**—the tendency to interpret evidence in a way that favors your prior beliefs—is often the rational response to conflicting evidence. This is because (1) assessing evidence requires doing a form of **cognitive search** which, like our word-completion tasks, engenders ambiguity, (2) often the way to make your beliefs most accurate is to avoid such ambiguity, and (3) this corresponds to doing cognitive searches that are expected to confirm your prior beliefs.
>
> **Appendix Summary:** In this entry I'll give a formal definition of rational confirmation bias (§6.1) and describe the details of the simulations mentioned in the blog post (§6.2).

6.1 Rational Confirmation Bias

For summaries of the literature on confirmation bias, see Nickerson (1998); Whittlestone (2017). For results on biased assimilation in particular, see Lord et al. (1979); Plous (1991); Baron (1995); Kuhn and Lao (1996); Munro and Ditto (1997); Taber and Lodge (2006).

On a natural definition, confirmation bias can be rational if and only if evidence can be ambiguous. After all, confirmation bias is meant to be a tendency to gather and interpret evidence in a way that *can be expected to confirm your prior beliefs*, i.e. can be expected to move your rational probabilities in a particular direction. That is just what I proved, in §4, was possible if and only if evidence is ambiguous. In particular, recall that unambiguous Bayesian models of rational polarization entail that *no* way of getting evidence can be predictably polarizing, since your opinions should always equal your expectation of the opinions you should have after getting and interpreting the evidence (Fact 4.4). Moreover, once evidence is ambiguous, this is always possible (Fact 4.7).

More precisely, say that your strategy of gathering and interpreting evidence is **confirmatory for** q between time 1 and time 2 iff at time 1 you should expect the rational credence in q to go up between times 1 and 2: $\mathbb{E}^1[P^2(q) - P^1(q)] > 0$. Then it follows immediately from Fact 4.4 that if evidence is always unambiguous-Bayesian, then no strategy is ever confirmatory, for any q (cf. White 2006; Titelbaum 2010; Salow 2018; Gallow 2019; Das 2020).

Moreover, it follows from Fact 4.7, many such strategies will be confirmatory. For instance, a choice to look at a word-completion task, and thereby receive evidence that

should induce the belief-transition in any of the models offered in §5.1, would be a confirmatory strategy for the claim that the task is completable.

Confirmation bias can then be defined as a tendency to prefer ways of gathering and interpreting evidence that are confirmatory for your prior beliefs. Following the theory of rationality given in §4, such strategies are *epistemically rational* if they satisfy the value of evidence (Definition 4.9) with respect to the live question you care about.

This distinguishes my claim about rational confirmation bias from many others in the literature. First, any models which use unambiguous-Bayesian models—those where credences are represented with precise probability and updated by conditioning on a partition—cannot allow for confirmation bias in this sense (Kelly 2008; Hahn and Harris 2014; Jern et al. 2014; Pallavicini et al. 2018; Benoît and Dubra 2019).[17] Second, many models of confirmation bias (or polarization) involve deviations from unambiguous Bayesianism in ways that lead to clear violations of the the *value of evidence*, for example when agents "double-update" on signals (Rabin and Schrag 1999; Fryer et al. 2019), forget question-relevant evidence (Singer et al. 2019), fail to track correlations between different pieces of evidence (Loh and Phelan 2018), ignore certain types of evidence (Hegselmann and Krause 2002), or fail to reason through the logical consequences of their evidence (Stone 2020). For other models it is not straightforwardly obvious: it is not straightforward whether the updating mechanisms used in O'Connor and Weatherall (2018); Weatherall and O'Connor (2020) can be given a value-validating interpretation; I believe the model of ambiguity given in Baliga et al. (2013) violates the sure-thing principle and therefore the value of evidence, but I am not yet certain about that.

(Dear Reader: Do you know of other rational-seeming models of confirmation bias I haven't discussed? Please send me references! Email me at kevindorst@pitt.edu.)

6.2 Cognitive Search Models and Simulations

Cognitive search models are generalizations of the word-completion task model from Figure 5. Such a model has the structure given in Figure 9:

Blue numbers in cells are prior probabilities of ending up in them; labeled arrows from a cell are posterior probabilities if you end up in that cell. $\pi(C)$ is the prior probability of the search being completable; $\pi(F|C)$ is the probability of finding a completion given that it's completable; and a is a constant between 0 and $1-\pi(C)$. As can be seen, when the the search is not completable, you should update by conditioning your credence on not finding one. When it is completable and you find one, you should update by conditioning your credence on finding one. And when it is completable but you don't find one, you should update it by conditioning your credence on not finding one, but

[17] No such model can offer confirmatory evidence in my sense. However, there are other senses in which they can allow for predictable polarization, e.g. allowing that you and I can predict that one of us will become more confident in q, while the other will become less confident—but we cannot predict which will move in which direction.

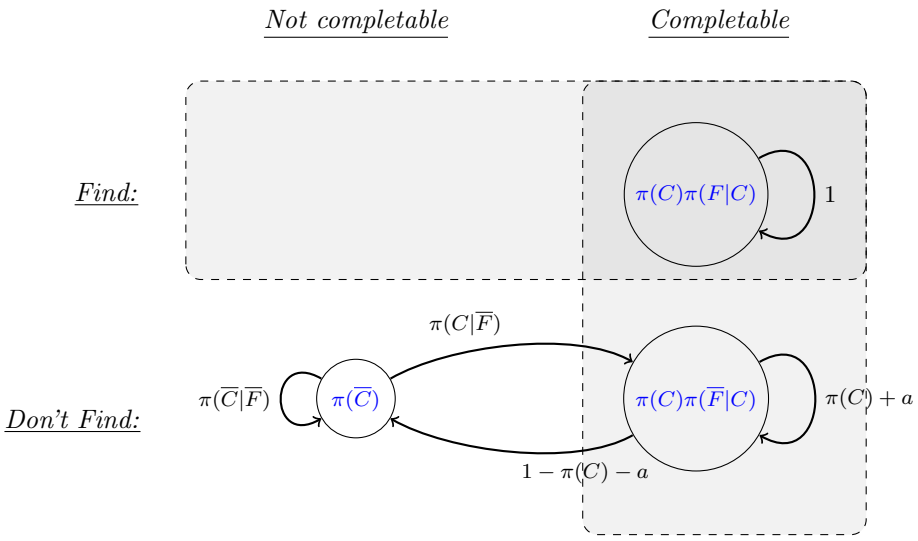


Figure 9: An arbitrary cognitive-search model

then shifting your credence somewhat toward it being completable—enough so that your credence in C ends up higher than it was previously.[18]

Any such cognitive search models is confirmatory for the claim that the search is completable, in the sense that $\pi(C) < \mathbb{E}_\pi[P(C)]$, where P is the posterior probability. Moreover, every such transition validate the value of evidence, since the model is class convex and π is in the convex hull of the posteriors (Theorem 5.1). More intuitively, note that the posterior is always more accurate about every cell of the partition than the prior.

The final ingredient is expected accuracy. Here I used the well-known Brier score, which uses the squared distance between probability and the truth-value of a proposition to measure that function's inaccuracy with respect to that proposition (de Finetti 1974; Joyce 1998; Pettigrew 2016); see Pettigrew (2019) for an accessible overview.

For computational tractability, I used the partition-based version of this. Given a probability function δ and a question Q (i.e. a partition of the state space) with n members, and a member of that partition $q_i \in Q$, we calculate δ's **Brier inaccuracy** as its average squared distance from the truth (q_i) across cells of the partition:

$$\mathcal{B}_Q(\delta, q_i) = \sum_{j \neq i} \frac{1}{n}(\delta(q_j) - 0)^2 + \frac{1}{n}(\delta(q_i) - 1)^2$$

A probability function's **Brier accuracy** with respect to q_i is simply 1 minus it's inaccuracy: $\mathcal{A}_Q(\delta, q_i) := 1 - \mathcal{B}_Q(\delta, q_i)$.

[18] NOTE TO SELF: similar results would hold up if we made this $\pi(C|\overline{F}) + a$; investigate details.

The **expected accuracy** of a cognitive search (relative to π) is simply π's probability that it will wind up in any of the possibilities, times the posterior accuracy if so. Let w be the worlds of our frame, and let $Q(w) \in Q$ be the partition-cell that w is a part of. Then

$$\mathbb{E}_\pi[\mathcal{A}] = \sum_{w \in W} \pi(w) \cdot \mathcal{A}_Q(\mathcal{P}_w, Q(w))$$

, where, as in §4 $\mathcal{P}_w$ is the posterior probability function at world w, as indicated by the arrows from possibilities in Figure 9.

To run my simulations, I first randomly generated 10,000 cognitive search models by randomly choosing values for $\pi(C)$, $\pi(F|C)$, and a, and recording both the probability of finding a word if there is one, as well as the expected accuracy of doing the cognitive search. The result is given in Figure 10. The line is the minimal least-squares line, with

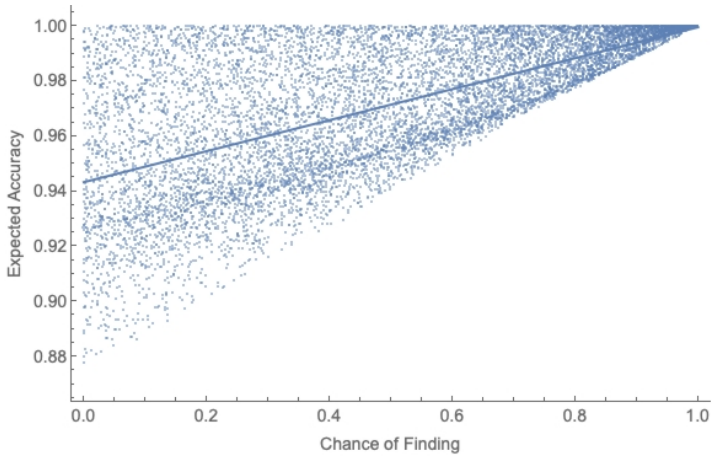


Figure 10: Correlation in random cognitive search models between the chance of finding a completion if there is one with expected accuracy of doing the search.

an $R^2 = 0.43$, though of course there is substantial heteroscedasticity.

This is the argument that how likely a cognitive search is to lead to ambiguous evidence (i.e. to not find a completion) is negatively correlated with accuracy, and therefore that a concern for accuracy can help guide your choice of which cognitive search to do.

In particular, I proceeded to model agents who are given a choice between a series of cognitive searches as follows. Each agent started out with an estimate of the proportion of pieces of evidence that favored a given proposition q. At each stage, they were presented with a pair of randomly-generated cognitive search models. The *confirmatory search* was one in which a "completable" outcome was one in which the piece of evidence favored q; the *disconfirmatory* search was one in which a "completable" outcome was

one in which the piece of evidence favored $\neg q$. The agent chose which search to do by calculating which one had higher expected accuracy. There then was an objective chance of 50% that the search was completable (irrespective of the agent's belief about which search was completable), but the agent's beliefs about how likely they were to find a completion if there was one were accurate (so $\pi(F|C)$ equals the objective chance of finding given that it's completable).

If the search was completable and they found a completion, they updated to certainty that this piece of evidence told in favor of or against q—depending on which search they were doing. If the search was completable but they didn't find one, they updated to the posterior probability of a completion, i.e. that the evidence was in favor (against) q, of $\pi(C) + a$. If the search was uncompletable, they updated to a posterior confidence that the search was completable, i.e. that the evidence was in favor (against) q, of $\pi(C|\overline{F})$.

Regardless of how they updated about this particular piece of evidence, they then updated their overall estimate of the proportion of evidence favoring q as Bayesians do: their new estimate of the proportion of the arguments they've seen that favor q is a weighted average of their prior estimate (weighted by how many bits of evidence it contained) and their probability (estimate) for how likely this new piece of evidence favors q (weighted by 1).[19] The process repeats 2000 times.

There are two groups of agents: a "pro" group (like Becca), and a "con" group (like me). Although probabilities of finding are generated randomly, the "pro" group is 20% more likely to find a flaw in the evidence that tells against q than that which tells in favor; vice versa for the "con" group. (The qualitative results are robust to variations in this parameter.) As a result, via the correlations shown above, the expected-accuracy-based choice of cognitive search is more likely to lead the "pro" group to scrutinize the evidence disfavoring q, and the "con" group to scrutinize the evidence that favors q.

I simulated 20 "pro" agents and 20 "con" agents, and plotted the trajectories of their estimates of the proportion of favorable evidence they've seen below in Figure 11. Although the absolute gap between the two agents is not large, note that as more and more arguments come in, they will become arbitrarily confident of these estimates. Thus "pro" agents will be nearly sure that *more than 50% of the pieces of evidence tell in favor of* q, while "con" agents will be nearly sure that that claim is false.

[19] To avoid excessive movements at the outset, I start all agents as having seen 200 pieces of evidence initially and having a 50%-favoring estimate.

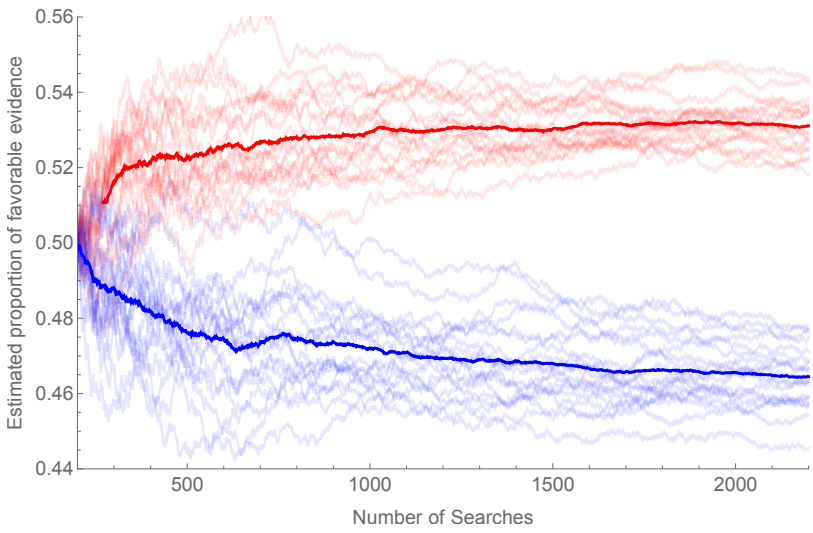


Figure 11: Simulation 2000 cognitive searches for 20 "pro" agents (red lines) who are more likely to find flaws in detracting arguments, and for 20 "con" agents (blue lines) who are more likely to find flaws in supporting arguments. All agents choose to do the search at each stage that maximizes expected accuracy. Thick lines are averages for each group.

7 Why Arguments Polarize Us

Here's the link to the full post, published on October 31, 2020.

> **Post Synopsis:** A prime driver of divergence—the **group polarization effect**—is due to the fact that different groups are presented with different arguments. Though this can seem obviously rational, I explained (appealing to the results of §4 and Salow 2018) that such polarization can't be *predictable* unless arguments present ambiguous evidence. I sketched a simple model of how they can do so, and showed simulations of how this model leads to predictable polarization.
>
> **Appendix Summary:** Here I'll describe the formal model of asymmetrically-ambiguous arguments, and explain how the simulations use this model.

7.1 Formal model of arguments

Imagine you are presented with an argument in favor of a claim q. That argument is either *good*—in the sense that, all things considered, it should raise your credence in q—-or *bad*—in the sense that, all things considered, it should *lower* your credence in q. In the simplest version of such a situation you have a current credence, $\pi(q) = t$, and this credence should either go up to t^+ or down to t^-.

If the argument is unambiguous, this cannot have an expected shifting effect on your opinion in q, as we've seen in §4. But suppose instead the argument is asymmetrically ambiguous—if the argument is good, you should be relatively confident it's good; but if the argument's bad, you should be relatively unsure whether it's good or bad.

Letting P be the future rational credence function, recall (§5.1) that the *informed* rational credence function $\widehat{P}$ is the one it'd be rational to have is all your higher-order uncertainty were removed: $\widehat{\mathcal{P}}_w := \mathcal{P}_w(\cdot|P = \mathcal{P}_w)$. Supposing your prior is $\pi(q) = t$, there are $t^+ > t$ and $t^- < t$ such that the argument is ***good*** iff $\widehat{P}(q) = t^+$, and ***bad*** iff $\widehat{P}(q) = t^-$.

The value of evidence entails that $\pi(q) = \mathbb{E}_\pi[\widehat{P}(q)]$, i.e. your current credence equals your expectation of the future informed credence (Dorst 2019; Stalnaker 2019). But we can still have an expected shift in rational credence because the evidence is ambiguous, and so $P(q) \neq \widehat{P}(q)$.

In particular, a valuable *favoring* argument model is one in which $t = \pi(q) = \pi(\mathit{good}) \cdot t^+ + \pi(\mathit{bad}) \cdot t^-$, and in which at any $b \in \mathit{bad}$, $\mathcal{P}_b(\mathit{bad}) \geq \pi(\mathit{bad})$, and at any $g \in \mathit{good}$, $\mathcal{P}_g(\mathit{good}) \geq \mathcal{P}_b(\mathit{bad})$ and $\mathcal{P}_g(\mathit{good}) \geq \pi(\mathit{good})$. The latter constraint ensures that although the evidence is valuable (so $\mathcal{P}_b(\mathit{bad})$ must go up from $\pi(\mathit{bad})$, and likewise for $\mathcal{P}_g(\mathit{good})$), in the good case you should be more confident of the good case than in the bad case you should be of the bad case.

In a diagram, these models look like Figure 12:

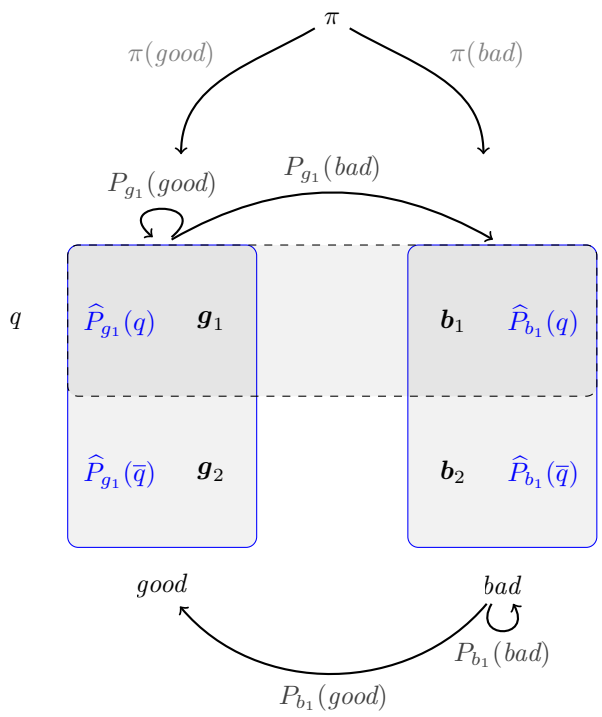


Figure 12: An argument model. All worlds agree on probabilities within blue cells, hence they are labeled directly. $P_{g_1} = P_{g_2}$ and $P_{b_1} = P_{b_2}$.

A *disfavoring* argument model is formally the same, except that when the argument is good it *lowers* the rational credence, so $\widehat{P}_g(q) < \pi(q)$

All such models validate the value of evidence, so π expects to become more accurate (about q, as well as everything else) by transitioning to them. Nevertheless, most favorable models are confirmatory for q—simulations of randomly generated models (pulled uniformly at random from those that meet the above constraints) suggest around 80% of them have $\mathbb{E}_\pi[P(q)] > \pi(q)$. Vice versa for disfavoring models.

7.2 Simulations of arguments

For my simulations, I made two groups of agents: 20 who repeatedly received favorable arguments for q, 20 who received disfavorable ones. All agents began 50% confident of q. In each iteration, they were presented with a an argument model. In addition to the above constraints, to prevent wild oscillations of opinion and to simulate hardening of opinion as more evidence comes in, I made it so that the maximal difference between $\widehat{P}_g(q)$ and $\pi(q)$ was always $\frac{0.1}{n}$, where n is the total number of arguments seen by

this agent (including the current one). Within all these constraints, the models were generated uniformly at random.

Once the argument was generated, I made it so that the agent had a 50% chance of their posterior credence in q shifting to $\mathcal{P}_g(q)$, and 50% shifting to $\mathcal{P}_b(g)$. This is to simulate the assumption that, as a matter of fact, exactly 50% of the arguments point in each direction.

This was a single iteration; on the next iteration, the agent's prior credence was set to its posterior from the previous iteration, and the process repeated. Each agent witnessed 1000 arguments.

The results from a typical situation appear in Figure 13.

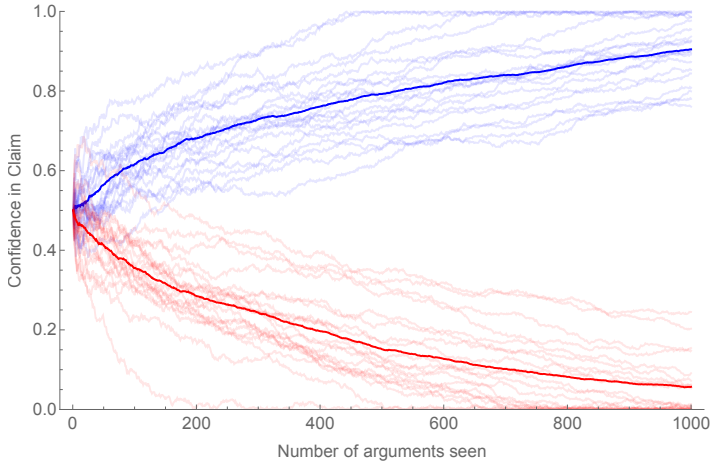


Figure 13: Simulation of 20 (blue) agents receiving pro arguments and 20 (red) agents receiving con arguments for q, when in fact 50% of the arguments point in each direction.

Notably, although ambiguity asymmetries in arguments are a force for polarization, they are not an insurmountable force. As the proportion of arguments that are actually evidence for q (i.e. are good if they are in favor of q, and are bad if they are against it) moves away from 50%, polarization can eventually be overcome. Figure 14 displays the results for rates of arguments providing evidence for q varying from 0.5 to 0.8.

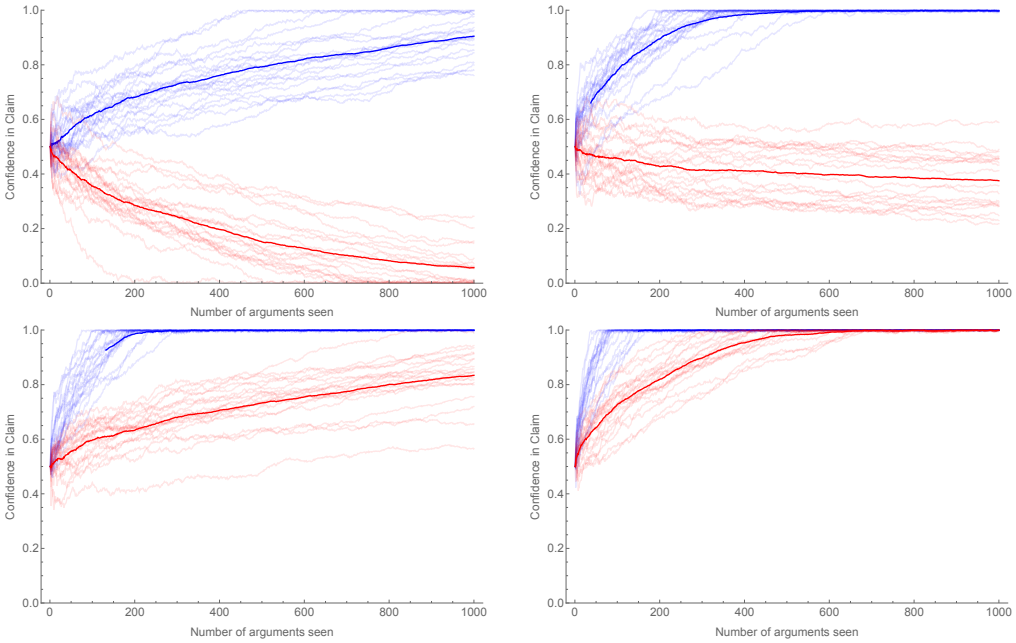


Figure 14: Simulations varying the actual proportion of arguments that support q: top left is 50%; top right is 60%, bottom left is 70%, and bottom right is 80%.

References

Abrams, Samuel J. and Fiorina, Morris P., 2012. 'The big sort that wasn't: A skeptical reexamination'. *PS - Political Science and Politics*, 45(2):203–210.

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

Ahmed, Arif and Salow, Bernhard, 2018. 'Don't Look Now'. *British Journal for the Philosophy of Science*, To appear.

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

Baron, Jonathan, 1995. 'Myside Bias in Thinking About Abortion'. *Thinking & Reasoning*, 1(3):221–235.

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

Bishop, Bill, 2009. *The big sort: Why the clustering of like-minded America is tearing us apart*. Houghton Mifflin Harcourt.

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

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

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

Cherniak, Christopher, 1986. *Minimal rationality*. Mit Press.

Cresto, Eleonora, 2012. 'A Defense of Temperate Epistemic Transparency'. *Journal of Philosophical Logic*, 41(6):923–955.

Dagum, Paul and Luby, Michael, 1993. 'Approximating probabilistic inference in Bayesian belief networks is NP-hard'. *Artificial intelligence*, 60(1):141–153.

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

de Finetti, Bruno, 1974. *Theory of Probability*. John Wiley and Sons.

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

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

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

Dorst, Kevin, Levinstein, Benjamin, and Salow, Bernhard, 2020. 'Of Modest Value'. *Philosophical Perspectives*, To appear.

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

Engber, Daniel, 2018. 'LOL something matters'. *Slate*, 8:1–18.

Fiorina, Morris P, 2016. 'The Political Parties Have Sorted'. *Hoover Institute*, 2(2):1–20.

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

Gallow, J. Dmitri, 2017. 'Local & Global Experts'.

———, 2019. 'Updating for externalists'. *Noûs*, (November 2018):1–30.

Gao, George and Smith, Samantha, 2016. 'Presidential job approval ratings from Ike to Obama'. Technical report.

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

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

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

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

Hamblin, Charles L, 1976. 'Questions in montague english'. In *Montague grammar*,

247–259. Elsevier.

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

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

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

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

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

Joyce, James M, 1998. 'A Nonpragmatic Vindication of Probabilism'. *Philosophy of Science*, 65(4):575–603.

Kahan, Dan M., Peters, Ellen, Dawson, Erica Cantrell, and Slovic, Paul, 2017. 'Motivated numeracy and enlightened self-government'. *Behavioural Public Policy*, 1:54–86.

Kelly, Thomas, 2008. 'Disagreement, Dogmatism, and Belief Polarization'. *The Journal of Philosophy*, 105(10):611–633.

Klein, Ezra, 2014. 'How politics makes us stupid'. *Vox*, 1–14.

———, 2020. *Why We're Polarized*. Profile Books.

Koerth, Maggie, 2019. 'Why Partisans Look At The Same Evidence On Ukraine And See Wildly Different Things'. *FiveThirtyEight*.

Kripke, Saul A, 1963. 'Semantical analysis of modal logic i normal modal propositional calculi'. *Mathematical Logic Quarterly*, 9(56):67–96.

Kuhn, Deanna and Lao, Joseph, 1996. 'Effects of Evidence on Atittudes: is Polarization the Norm?' *Psycholohical Science*, 7(2):115–120.

Landemore, Hélène, 2017. *Democratic reason: Politics, collective intelligence, and the rule of the many*. Princeton University Press.

Lasonen-Aarnio, Maria, 2013. 'Disagreement and evidential attenuation'. *Nous*, 47(4):767–794.

Lazer, David, Baum, Matthew, Benkler, Jochai, Berinsky, Adam, Greenhill, Kelly, Metzger, Miriam, Nyhan, Brendan, Pennycook, G., Rothschild, David, Sunstein, Cass, Thorson, Emily, Watts, Duncan, and Zittrain, Jonathan, 2018. 'The science of fake news'. *Science*, 359(6380):1094–1096.

Lepoutre, Maxime, 2020. 'Democratic Group Cognition'. *Philosophy & Public Affairs*, 48(1):40–78.

Loh, Isaac and Phelan, Gregory, 2018. 'D IMENSIONALITY AND D ISAGREEMENT : Asymptotic belief divergence in response to common information'. 1–52.

Lord, Charles G., Ross, Lee, and Lepper, Mark R., 1979. 'Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence'.

Journal of Personality and Social Psychology, 37(11):2098–2109.

Munro, Geoffrey D and Ditto, Peter H, 1997. 'Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information'. *Personality and Social Psychology Bulletin*, 23(6):636–653.

Nguyen, C. Thi, 2018. 'Escape the echo chamber'. *Aeon*.

Nickerson, Raymond S., 1998. 'Confirmation bias: A ubiquitous phenomenon in many guises.' *Review of General Psychology*, 2(2):175–220.

O'Connor, Cailin and Weatherall, James Owen, 2018. 'Scientific Polarization'. *European Journal for Philosophy of Science*, 8(3):855–875.

Pallavicini, Josefine, Hallsson, Bjørn, and Kappel, Klemens, 2018. *Polarization in groups of Bayesian agents*. Springer Netherlands.

Pennycook, By Gordon and Rand, David, 2019. 'Why Do People Fall for Fake News? Are they blinded by their political passions? Or are they just intellectually lazy?'

Pettigrew, Richard, 2016. 'Jamesian Epistemology Formalized: An Explication of 'The Will to Believe''. *Episteme*, 13(3):253–268.

———, 2019. 'Epistemic Utility Arguments for Probabilism'.

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

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

Roberts, Craige, 2012. 'Information structure in discourse: Towards an integrated formal theory of pragmatics'. *Semantics and Pragmatics*, 5(6):1–69.

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

Ryan, Timothy J, 2014. 'Reconsidering moral issues in politics'. *The Journal of Politics*, 76(2):380–397.

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

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

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

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

Simon, Herbert A., 1956. 'Rational Choice and the Structure of the Environment'. *Psychological Review*, 63(2):129–138.

Simon, Herbert A, 1976. 'From substantive to procedural rationality'. In *25 years of economic theory*, 65–86. Springer.

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

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

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

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

Titelbaum, Michael G., 2010. 'Tell me you love me: Bootstrapping, externalism, and no-lose epistemology'. *Philosophical Studies*, 149(1):119–134.

van Benthem, Johan, 2011. *Logical Dynamics of Information and Interaction*. Cambridge University Press.

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

Van Heuvelen, Ben, 2007. 'The Internet is Making us Stupid'. *Salon*.

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

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

———, 2017. 'Formal Epistemology'.

White, Roger, 2006. 'Problems for Dogmatism'. *Philosophical Studies*, 131:525–557.

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

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

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

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

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