Rational Polarization: Appendices

Kevin Dorst
kevindorst@pitt.edu
September 2021

Word count: 12,967 words

A Analytical Details

This appendix contains the model theory for the higher-order probabilities and ambiguous evidence used throughout, as well as proofs of all theorems.

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

Following the practice of standard epistemic logic (Hintikka 1962; van Ditmarsch et al. 2015), we give a semantics for higher-order probability by using a (finite) structure that can identify higher-order claims (of any order) with events, i.e. sets of worlds, i.e. propositions. A **probability frame** $\langle W, \{P^i\}_{i \in N} \rangle$ consists of a (finite) set of worlds W and a set of functions P^i from worlds $w \in W$ to probability functions P^i_w defined over all subsets of W, so that $P^i : W \to \Delta(W)$. Thus 'P^i' can be thought of as a *description* of a probability function—it picks out different such functions in different worlds. In our case, it'll always be interpreted as 'the rational credence function (for some particular agent) at some particular time i'. Note that 'P^i_w' is a rigid designator that picks out the (unique) probability function that P^i associates with a given world w. In cases where we're only concerned about one description of a probability function, I'll drop the index and use P, P_w, etc. In addition, I'll often enrich the structure of a probability frame with one or more (rigidly designated) probability functions, denoted $\pi, \delta, \rho, \eta, ...$[1]

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

We use P to identify facts about probabilities as sets of worlds in the frame, thus allowing us to 'unravel' higher-order probability claims to be sets of worlds. Thus for any $q \subseteq W$

[1] For more uses of (structures like) probability frames, see Gaifman 1988; Samet 2000; Williamson 2000, 2014, 2019; Schervish et al. 2004; Lasonen-Aarnio 2013, 2015; Campbell-Moore 2016; Salow 2018, 2019; Das 2020a,b; Dorst 2020a; Dorst et al. 2021. See Williamson 2008 and Dorst 2019, 2020b for summaries.

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

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

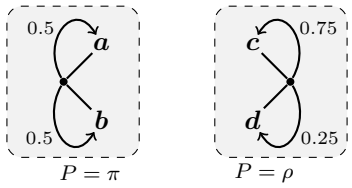


Figure 1: An unambiguous frame. π assigns 0.5 to a and 0.5 to b; ρ assigns 0.75 to c and 0.25 to d.

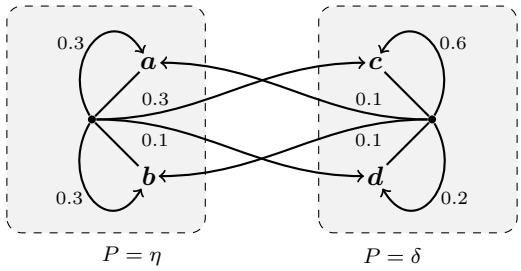


Figure 2: An ambiguous frame. η assigns 0.3 to a, to b, and to c, and 0.1 to d; δ assigns 0.1 to a and to b, 0.6 to c, and 0.2 to d. Thus $\eta(P = \eta) = 0.6$ and $\eta(P = \delta) = 0.4$, while $\delta(P = \eta) = 0.2$ and $\delta(P = \delta) = 0.8$.

When is a transition from an initial probability function π, to a posterior P—updating to different probabilities P_w in different worlds w—a rational update? If you should expect it to lead you to make better decisions, no matter what decision you face, then I claimed it is. We can formalize this as follows (following Dorst et al. 2021). Given a probability frame $\langle W, P\rangle$, let an **option** O be a function from worlds w to numbers $O(w)$ (so options are random variables), representing the utility that would be achieved by taking option O at w. Let a **decision problem** be a finite set of options $\mathcal{O}$. Let a **strategy** S be a way of choosing options based on P's probabilities, i.e. a function from w to $S_w \in \mathcal{O}$ such that $S_w = S_x$ whenever $P_w = P_x$. P **recommends** a strategy S for $\mathcal{O}$ iff S always selects an option that maximizes expected value according to P. Recalling that $\mathbb{E}_\pi(O) = \sum_{t\in\mathbb{R}} \pi(O = t) \cdot t$ is the

expected value of option O according to π, a strategy is recommended by P iff for all w and $O \in \mathcal{O}$: $\mathbb{E}_{P_w}(S_w) \geq \mathbb{E}_{P_w}(O)$. Then π **values** P iff, for every decision problem, the expected value of following a strategy recommended by P is at least as high as taking the option that maximizes expected value according to π. Abusing notation slightly so that $\mathbb{E}_\pi(S)$ is the expected value of following a strategy S, i.e. $\mathbb{E}_\pi(S) := \sum_w \pi(w)S_w(w)$, that is:

> **Value:** π values P iff $\forall\mathcal{O}$: if P recommends S for $\mathcal{O}$, then $\forall O \in \mathcal{O}$: $\mathbb{E}_\pi(S) \geq \mathbb{E}_\pi(O)$. π values P iff, for any decision problem, π prefers to let P decide on π's behalf, rather than make the decision itself.

Dorst et al. 2021 argue that Value is what it takes for π to defer to P, and give a characterization of what this amounts to (Theorem 5.2). For our purposes, the relevant points are that π values P iff no (fixed-option[2]) Dutch book can be constructed for updating from π to P (i.e. the transition cannot be made to lead to a sure loss), iff on every generally strictly proper[3] way of measuring the accuracy of estimates (Schervish 1989; Levinstein 2020; Campbell-Moore 2020), π expects P to be more accurate than itself.

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

Now think of a probability function π over a set W of size $|W| = n$ as a point in Euclidean n-space, i.e. a vector in which entry i is $\pi(w_i)$. The **convex hull** of a set of such probability functions $\pi_1, .., \pi_n$ is the set of points obtainable by averaging them; formally, $CH\{\pi_1, ..., \pi_n\} = \{\delta : \exists\lambda_i \geq 0 \text{ and } \sum\lambda_i = 1 \text{ such that } \delta = \sum\lambda_i\pi_i\}$. Let $C_w = \{\rho : P_w(P = \rho) > 0\}$ be the set of candidate probability functions that P_w thinks might be P, and $C_w^- = C_w - \{P_w\}$ be the set of such functions other than P_w. Then say that P_w is **modestly informed** iff it is an average of its informed self $\widehat{P}_w$ along with the other candidates C_w^-, i.e. iff P_w is in the convex hull of $\{\widehat{P}_w\} \cup C_w^-$. Finally let $C_\pi = \{P_w : \pi(P = P_w) > 0\}$ be the set of candidates that π leaves open might be P. Then we have:

Theorem A.1. π values P iff each P_w in C_π is modestly informed, and π is in the convex hull of C_π.

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

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

Proof. If π values P, then each $P_w \in C_\pi$ is modestly informed and π is in their convex hull. This implies that for any $P_w \in C_\pi$, $P_w(P = P_w) > 0$ (Dorst et al. 2021, Lemma 7.2.5).

[2] A *fixed-option* Dutch book is one in which options are random variables as above, and thus a situation in which you have the same options at every world. See Dorst et al. 2021, §1.

[3] A proper scoring rule is one such that every (rigidly designated) probability function expects itself to be at least as accurate as any other. They are the standard ways of measuring (in)accuracy—see e.g. Schervish 1989; Pettigrew 2016; Campbell-Moore 2020; Campbell-Moore and Levinstein 2020; Levinstein 2020.

If P is unambiguous, then each P_w is certain of what P is, and since $P_w(P = P_w) > 0$, it follows that $P_w(P = P_w) = 1$. Since π is in the convex hull of the P_w, it follows that for each P_w, $\pi(\cdot|P = P_w) = P_w$. (Taking any q, we know $\pi(q|P = P_w) = \frac{\pi(q \wedge [P=P_w])}{\pi(P=P_w)} = \frac{\sum_i \lambda_i P_i(q \wedge [P=P_w])}{\sum_i \lambda_i P_i(P=P_w)} = \frac{\lambda_w P_w(q \wedge [P=P_w])}{\lambda_w P_w(P=P_w)} = \frac{P_w(q \wedge [P=P_w])}{P_w(P=P_w)} = P_w(q|P = P_w) = P_w(q)$, where the third and final equalities comes from the fact that for all i, $P_i(P = P_i) = 1$.) It follows immediately that, for any q:

$$\begin{aligned}
\pi(q) &= \sum_{P_w} \pi(P = P_w)\pi(q \,|\, P = P_w) && \text{(Total probability)} \\
&= \sum_{P_w} \pi(P = P_w)P_w(q) \\
&= \sum_{P_w} \pi(P = P_w)\mathbb{E}_\pi(P(q) \,|\, P = P_w) \;=\; \mathbb{E}_\pi(P(q)) && \text{(Total expectation)}
\end{aligned}$$

□

Recall that we say that P is **valuable** iff there is some ρ that values it while also leaving open all its potential realizations, i.e. for all $w \in W$: $\rho(P = P_w) > 0$. Then:

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

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

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

Thus we know that there is a ρ_i and q such that $\rho_i(q) \neq \mathbb{E}_{\rho_i}(P(q))$. Since $\rho_i(q) < \mathbb{E}_{\rho_i}(P(q))$ iff $\pi(\neg q) > \mathbb{E}_{\rho_i}(P(\neg q))$, WLOG assume $\rho_i(q) < \mathbb{E}_{\rho_i}(P(q))$. Equivalently, where $\mathbb{1}_q$ is the indicator function of q (1 at $w \in q$, 0 elsewhere), $\mathbb{E}_{\rho_i}(P(q) - \mathbb{1}_q) > 0$. We want to show that there are uncountably many δ such that δ values P and yet $\mathbb{E}_\delta(P(q) - \mathbb{1}_q) > 0$. Pick some ρ_i that maximizes $\mathbb{E}_{\rho_i}(P(q) - \mathbb{1}_q)$ within the frame (the frame is finite, so there is one), and any other $\rho_j \neq \rho_i$ (there must be at least one other, since P is ambiguous).

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

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

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

We first show that the (coarse-grained) model of word-search tasks in Figure 2 (page 19) is such that the prior η values the posterior H. Let $W = \{n, c, f\}$, where $f =$ *Find*, $c =$ *¬Find&Completable*, and $n =$ *¬Completable*. Recall that $\eta(f) = \eta(c) = \frac{1}{4}$ and $\eta(n) = \frac{1}{2}$, so that (thinking of probability functions as vectors over W), $\eta = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$, while $H_f = (0,0,1)$, $H_c = (\frac{1}{3}, \frac{2}{3}, 0)$, and $H_n = (\frac{2}{3}, \frac{1}{3}, 0)$. By Theorem A.1, we must show that each H_i is modestly informed and that η is in their convex hull.

Note that $H_f(H = H_f) = 1$, so $H_f = \widehat{H}_f$ and hence is modestly informed. Meanwhile, $\widehat{H}_c = (0,1,0)$ and $\widehat{H}_n = (1,0,0)$. Thus $\frac{1}{2}\widehat{H}_c + \frac{1}{2}H_n = \frac{1}{2}(0,1,0) + \frac{1}{2}(\frac{2}{3}, \frac{1}{3}, 0) = (0, \frac{3}{6}, 0) + (\frac{1}{3}, \frac{1}{6}, 0) = (\frac{1}{3}, \frac{2}{3}, 0) = H_c$, so H_c is modestly informed. Similarly, $\frac{1}{2}\widehat{H}_n + \frac{1}{2}H_c = H_n$, so H_n is also modestly informed. Finally note that $\frac{1}{4}(0,0,1) + 0(\frac{1}{3}, \frac{2}{3}, 0) + \frac{3}{4}(\frac{2}{3}, \frac{1}{3}, 0) = (0,0,\frac{1}{4}) + (\frac{2}{4}, \frac{1}{4}, 0) = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4}) = \eta$, so η is in the convex hull of $\{H_f, H_c, H_n\}$. By Theorem A.1, this shows that η values H.

Next turn to the model of the word-search task in Figure 1 (page 18). Recall that there are four possibilities, label them $W = \{n, c, o, f\}$, for *not Completable*, *Completable&¬Find*, *Obvious&¬Find*, and *Find*, respectively. Recall that $\eta = (\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{4})$, while:

$H_f = (0,0,0,1)$;
$H_o = (0,0,1,0)$; and
$H_c = H_n = (\frac{2}{3}, \frac{1}{6}, \frac{1}{6}, 0)$.

Clearly H_f and H_o are modestly informed. Note that $\widehat{H}_c = \widehat{H}_n = (\frac{4}{5}, \frac{1}{5}, 0, 0)$, and that $\frac{5}{6}\widehat{H}_c + \frac{1}{6}H_o = \frac{5}{6}(\frac{4}{5}, \frac{1}{5}, 0, 0) + \frac{1}{6}(0,0,1,0) = (\frac{4}{6}, \frac{1}{6}, 0, 0) + (0,0,\frac{1}{6}, 0) = (\frac{2}{3}, \frac{1}{6}, \frac{1}{6}, 0) = H_c$, so H_c (and likewise H_n) is modestly informed. Finally, notice that $\frac{1}{4}H_f + \frac{3}{4}H_n = \frac{1}{4}(0,0,0,1) + \frac{3}{4}(\frac{2}{3}, \frac{1}{6}, \frac{1}{6}, 0) = (\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{4}) = \eta$, so η is in the convex hull of $\{H_n, H_c, H_o, H_f\}$. By Theorem A.1, this shows that η values H.

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

Let a question Q be a partition of W, and let $Q(w)$ be the partition-cell of $w \in W$. A proposition $p \subseteq W$ is *about* Q iff $p = \bigcup_i q_i$ for $q_i \in Q$, i.e. iff p is a partial answer to the question Q (Hamblin 1976; Roberts 2012). Recall that a decision problem $\mathcal{O}$ is any set

of options (i.e. random variables—functions from worlds to numbers) on W. Say that an option O is Q-*measurable* iff Q settles the value of O, i.e. for all w, w', if $Q(w) = Q(w')$, then $O(w) = O(w')$. Say that $\mathcal{O}_Q$ is a *decision about* Q iff each of its options is Q-measurable. Then: π *values* P *with respect to* Q iff it prefers to let P decide any decision about Q:

> **Value wrt Q:** π values P with respect to Q iff: for every decision problem $\mathcal{O}_Q$ about Q, if P recommends S for $\mathcal{O}_Q$, then $\forall O \in \mathcal{O}_Q : \mathbb{E}_\pi(S) \geq \mathbb{E}_\pi(O)$.
> π values P with respect to Q iff, for any decision about Q, it prefers to let P decide on π's behalf, rather than make the decision itself.

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

Given this, I'll now turn to proving that updates that are valuable with respect to Q can nevertheless lead to predictable, profound polarization about Q:

Theorem 3.1. There is a sequence of probability functions H^0, $\overline{H^0}$, H^1, $\overline{H^1}$....,H^n, $\overline{H^n}$, a partition Q, and a proposition $h = \bigcup_i q_i$ (for some $q_i \in Q$) such that, as $n \to \infty$:

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

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

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

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

Over W we lay some partitions. Let

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

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

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

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

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

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

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

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

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

Machinery in place, we can define the series of probability functions $H^0, \overline{H^0}, H^1, \overline{H^1}, ...,$ $H^n, \overline{H^n}$ that represent Haleys rational opinions over time. (H^i is that right after completing the ith word-search task, while $\overline{H^i}$ is some time after that, when she's forgotten the string

and so consolidated her higher-order uncertainty.) Recall that H^i is a description (so it picks out different probability functions at different worlds), whereas H^i_w is a rigid designator (that always picks out the function that H^i associates with w).

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

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

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

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

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

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

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

If $f_{i+1} \in w$, then $H^{i+1}_w = \overline{H^i_w}[0, 0, 1]_{i+1}$;
If $c_{i+1} \in w$ **or** $n'_{i+1} \in w$, then $H^{i+1}_w = \overline{H^i_w}[\frac{1}{3}, \frac{2}{3}, 0]_{i+1}$;
If $n_{i+1} \in w$ **or** $c'_{i+1} \in w$, then $H^{i+1}_w = \overline{H^i_w}[\frac{2}{3}, \frac{1}{3}, 0]_{i+1}$;

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

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

Proof. (1) Suppose $\overline{H^i_w}(x) > 0$. By definition, $\overline{H^i_w}(x) = H^i_w(x||H^i = H^i_w) > 0$. By the definition of imaging, $x \in [H^i = H^i_w]$, i.e. $H^i_x = H^i_w$. Thus $\overline{H^i_x} = H^i_x(\cdot||H^i = H^i_x) = H^i_w(\cdot||H^i = H^i_w) = \overline{H^i_w}$. Since x was arbitrary, $\overline{H^i_w}(\overline{H^i} = \overline{H^i_w}) = 1$.

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

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

Lemma 3.1.2. For any $w, x, i < j$, if $n'_j \in x$ or $c'_j \in x$, then $H^i_w(x) = 0$ and $\overline{H^i_w}(x) = 0$.

Proof. By induction. *Base case:* By construction, $H^0_w(U) = 0$, so $H^0_w(x) = 0$. Since $\overline{H^0_x} = H^0_x$, likewise for $\overline{H^0_x}$. *Induction:* Supposing it holds for all w with $k < i$, we show it holds for i. Since $H^i_w = \overline{H^{i-1}_w}[a_1, a_2, a_3]_i$, and this doesn't raise any probabilities from 0, since

(by induction) $\overline{H_w^{i-1}}(x) = 0$, likewise $H_w^i(x) = 0$. Now suppose, for reductio, $\overline{H_w^i}(x) > 0$. Thus there must be a y such that $H_w^i(y) > 0$ and $g_y(H^i = H_w^i) = x$. But since H_w^i didn't assign positive probability to any world with n_j' or c_j' in it, those are not in y and yet they are in x. If $H_y^i = H_w^i$, then (by Strong Centering) $g_y(H^i = H_w^i) = y$, so this is impossible; hence $H_y^i \neq H_w^i$. Since $H_w^i(y) > 0$, and if $w \in f_i$ then H_w^i would be higher-order certain, it must be that either (i) $w \in U_c^i$ and $y \in U_n^i$, or (ii) $w \in U_n^i$ and $y \in U_c^i$. Since we must've had $\overline{H_w^{i-1}}(y) > 0$, by the inductive hypothesis we know either $c_i \in y$ or $n_i \in y$ (not $c_i' \in y$ nor $n_i' \in y$). So if (i), then $y' = \langle y_1, ..., n_i', ..., y_n\rangle$—which swaps out n_i' for n_i in y and is a world that is in the same Q-cell as y—updates the same as w so $H_{y'}^i = H_w^i$. Since y' agrees with the end-sequence of y more than x does (since $n_j' \in x$ or $c_j' \in x$), by Sequence-Respecting, $g_y(H^i = H_w^i) \neq x$—contradiction. If (ii), parallel reasoning works substituting c_i' into y, completing the proof. □

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

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

Proof. By Lemma 3.1.1 (2), $\overline{H_x^{i-1}} = \overline{H_y^{i-1}}$. By Lemma 3.1.2 and the fact that H_x^i preserves $\overline{H_x^{i-1}}$'s certainties, neither $c_i' \in y$ nor $n_i' \in y$; hence either $f_i \in y$ or $c_i \in y$ or $n_i \in y$.

If $f_i \in x$, then of course $f_i \in y$ and so $H_y^i = H_x^i$, meaning that by Strong Centering $g_y(H^i = H_x^i) = y$, establishing the result.

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

$$\begin{aligned} H_{y'}^i &= \overline{H_{y'}^{i-1}}[\tfrac{1}{3}, \tfrac{2}{3}, 0]_i \\ &= \overline{H_x^{i-1}}[\tfrac{1}{3}, \tfrac{2}{3}, 0]_i \;=\; H_x^i. \end{aligned}$$

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

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

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

Proof. By construction:

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

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

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

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

Lemma 3.1.6. For all w and i, $\overline{H^i_w}$ values H^{i+1}_w.

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

(1) Taking an arbitrary $x \in S^i_w$, we show that H^{i+1}_x is modestly informed. By Lemma 3.1.1 (1), note that since $\overline{H^i_w}(x)$ is higher-order certain, $\overline{H^i_x} = \overline{H^i_w}$. Now either (i) $f_{i+1} \in x$, or (ii) $c_{i+1} \in x$ or $n'_{i+1} \in x$, or (iii) $n_{i+1} \in x$ or $c'_{i+1} \in x$. Supposing (i), then $H^{i+1}_x = \overline{H^i_x}[0, 0, 1]_{i+1}$, meaning $H^{i+1}_x(F_i) = 1$ so that if $H^{i+1}_x(y) > 0$, then $f_{i+1} \in y$, and $H^{i+1}_y = H^{i+1}_x$. Hence $H^{i+1}_x(H^{i+1} = H^{i+1}_x) = 1$, so trivially H^{i+1} is modestly informed. On the other hand, if (ii) holds then $H^{i+1}_x = \overline{H^i_x}[\frac{1}{3}, \frac{2}{3}, 0]_{i+1} = \overline{H^i_w}[\frac{1}{3}, \frac{2}{3}, 0]_{i+1}$—label this function π_c. If (iii) holds, then $H^{i+1}_x = \overline{H^i_w}[\frac{2}{3}, \frac{1}{3}, 0]_{i+1}$—label this function π_n. Note that π_c and π_n both assign 1 to S^i_w, and also assign 1 to $[H^{i+1} = \pi_c] \vee [H^{i+1} = \pi_n]$. Now, since by Lemma 3.1.2 we have that $\overline{H^i_w}$ assigns 0 to any world with n'_{i+1} or c'_{i+1} in it, it follows that π_c and π_n do too, and hence that:

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

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

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

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

(2) We now show that $\overline{H^i_w}$ is in their convex hull. Note that by Lemma 3.1.5 and total probability,

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

Now notice that:

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

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

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

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

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

Proof. By Lemma 3.1.4, for any $q \in Q$: $H^i_x(\overline{H^i}(q) = H^i(q)) = 1$. It follows that for any decision-problem $\mathcal{O}_Q$ based on Q, H^i recommends strategy S for $\mathcal{O}_Q$ iff $\overline{H^i}$ recommends S for $\mathcal{O}_Q$. Since, by Corollary 3.1.7, H^i_x values H^i, it follows immediately that H^i_x values $\overline{H^i}$ with respect to Q. □

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

Recall that h = *more than half the coins land heads* is a proposition about Q, and that for each $\mathit{Heads}_i = F_i \cup C_i$, $H^i(\mathit{Heads}_i) = \frac{1}{2}$, mutually independently. Thus letting $\#h$ be a random variable for the number of coins that land heads, $H^0(\#h = k)$ is a binomial distribution with parameters $\frac{1}{2}$ and n. Since each sequence of heads and tails is equally likely, and as $n \to \infty$ the proportion of sequences with more than half heads tends to 1/2, the third bullet-point follows: $H^0(h) \approx \frac{1}{2}$, and since H^0 is higher-order certain, $H^0(H^0(h) \approx \frac{1}{2}) = 1$.

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

Lemma 3.1.9. With $\mathit{Heads}_i = F_i \cup C_i$, we have, for all w, i, H^0_w assigns probability 1 to:

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

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

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

Since $H^0(U) = 0$, so H^0_w assigns 0 to any world with n'_i or c'_i in it, it suffices to show that $\overline{H^n}$ follow the same pattern. By Lemma 3.1.5, each $\overline{H^j}$ treats the Q_k as mutually independent, so by definition none of the later Jeffrey-shifts—for $j \geq i$, the update from $\overline{H^j}$ to H^{j+1}—change the probabilities in Q_k. And by Lemma 3.1.4, none of the consolidations (from H^j to $\overline{H^j}$) do so either. Thus $\overline{H^n}$ follows the above pattern as well, establishing the result. □

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

Lemma 3.1.10. For any $\epsilon > 0$, as $n \to \infty$, $H^0(H^n(h) \geq 1 - \epsilon) \to 1$.

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

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

Thus by the weak law of large numbers, as $n \to \infty$ we have:

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

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

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

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

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

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

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

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

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

Lemma 3.1.10 establishes the fourth.

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

Corollary 3.2. There are two sequences of probability functions H^0, $\overline{H^0}$,..., $\overline{H^n}$ and T^0, $\overline{T^0}$,...,$\overline{T^n}$, a partition Q and a proposition $h = \bigcup_i q_i$ (for some $q_i \in Q$) such that, as $n \to \infty$:

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

Proof. All but the final bullet-point are straightforward generalizations of the proofs of Theorem 3.1, gotten by dividing possibilities further to track which updates T^i goes through, consolidating throughout the process in a way that maintains opinions about Q, and adding the partitions $Q_i^t = \{F_i^i, C_i^t, N_i^t\}$, where $F_i^t \cup C_i^t = N_i$ and $N_i^t = F_i \cup C_i$. By doing so, we create a model in which both H^0 and T^0 are (correctly) certain that:

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

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

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

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

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

and that

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

To establish the final bullet-point, of persistent polarization, notice that by the weak law of large numbers, both H^0 and T^0 are arbitrarily confident that (where $I_{F^t} = \{i : Q_i^t = F_i^t\}$, etc.) $|I_F| \approx \frac{n}{4}$ & $|I_C| \approx \frac{n}{4}$ & $|I_{F^t}| \approx \frac{n}{4}$ & $|I_{C^t}| \approx \frac{n}{4}$. Supposing this obtains, we show

that the resulting polarization on H^n (and, by parallel reasoning, T^n) is persistent—which suffices to show that it is predictable, profound, and persistent.[4]

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

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

ii) For all $i \in I_C$, since $H^n(C_i) = \frac{2}{3}$ and $H^n(N_i) = \frac{1}{3}$, so $H^n(N_i \& F_t) = H^n(N_i \& C_t) = \frac{1}{6}$, we have: $H^n(T^n(\mathit{Heads}_i) = \frac{2}{3}) = \frac{2}{3}$, $H^n(T^n(\mathit{Heads}_i) = 0) = \frac{1}{6}$, and $H^n(T^n(\mathit{Heads}_i) = \frac{1}{3}) = \frac{1}{6}$.
Therefore, if $\pi = H^n$, for all $i \in I_C$, $\mathbb{E}_\pi(T^n(\mathit{Heads}_i)) = \frac{2}{3}(\frac{2}{3}) + \frac{1}{6}(\frac{1}{3}) = \frac{1}{2}$. Since H^n treats the $T^n(\mathit{Heads}_i)$ as independent, by the weak law of large numbers, as $n \to \infty$, $H^n(\sum_{i \in I_C} \frac{T^n(\mathit{Heads}_i)}{|I_C|} \approx \frac{1}{2}) \to 1$.

iii) For all $i \in I_N$, since $H^n(C_i) = \frac{1}{3}$ and $H^n(N_i) = \frac{2}{3}$, so $H^n(N_i \& F_t) = H^n(N_i \& C_t) = \frac{1}{3}$, we have: $H^n(T^n(\mathit{Heads}_i) = \frac{2}{3}) = \frac{1}{3}$, $H^n(T^n(\mathit{Heads}_i) = 0) = \frac{1}{3}$, and $H^n(T^n(\mathit{Heads}_i) = \frac{1}{3}) = \frac{1}{3}$.
Therefore, if $\pi = H^n$, for all $i \in I_N$, $\mathbb{E}_\pi(T^n(\mathit{Heads}_i)) = \frac{1}{3}(\frac{2}{3}) + \frac{1}{3}(\frac{1}{3}) = \frac{1}{3}$. Since H^n treats the $T^n(\mathit{Heads}_i)$ as independent, by the weak law of large numbers, as $n \to \infty$, $H^n(\sum_{i \in I_N} \frac{T^n(\mathit{Heads}_i)}{|I_N|} \approx \frac{1}{3}) \to 1$.

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

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

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

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

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

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

B Experimental Details

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

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

250 participants were recruited through Prolific (107 F/139 M/4 Other; mean age = 27.06; pre-registration here: https://aspredicted.org/8jg3e.pdf).[5] Subjects were (pseudo)randomly divided into Ambiguous (A) and Unambiguous (U) conditions. Within each condition, they were further (pseudorandomly) divided into "Headsers" and "Tailsers". I will abbreviate the groups "**A-Hsers**"; "**A-Tsers**"; "**U-Hsers**", and "**U-Tsers**". Each group was told they'd be given evidence about a series of independent, fair coin tosses. Both groups were given standard instructions about how to use a 0–100% scale to rate their confidence in the answer to a yes/no question, and then given specialized instructions.

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

The U group was given instructions about how the urn task worked. For U-Hsers, if the coin landed heads then the urn contained 1 black marble and 1 non-black marble; if it landed tails, it contained two non-black marbles. (For U-Tsers, 'heads' and 'tails' were reversed.) The colors of the non-black marbles changed across trials to emphasize that they were different urns.

Both groups saw four independent tasks, and were asked before and afterwards how confident they were in the outcome.[6] The pre-task question was an attention-check, wherein they were instructed to move the slider to 50%; it was pre-registered that I would exclude participants who failed two or more of these attention-checks. In total, 25 of 250 participants were excluded in this way.

The order of the tasks was randomized. Each subject in the A-group saw two completable and two uncompletable strings. (The completable strings were randomly drawn from the list, {FO_E_T, ST_ _N, FR_ _L} (forest/foment; stain/stern; frail/frill); the uncompletable strings were drawn from the list, {TR_P_R, ST_ _RE, P_G_ER}.) Each subject in the U-group saw 3 tasks in which a non-black marble was drawn, and 1 in which a black marble was (simulating the expected rate of drawing black marbles from a fair coin and urn).

From the responses of each group to each question, I calculated their prior and posterior confidence that the coin landed heads in each toss (for Hsers, this was the number they reported as their confidence; for Tsers, it was obtained by subtracting this number from 100). I pooled such responses across all participants and items to calculate the following statistics. (*Note:* As discussed below, we obtain more statistical power if we group *by participant* and calculate their mean confidence as they view more tasks; those stronger statistics were what was reported in the main text in §3.2, page 21.)

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

[5] I made several mistakes at the pre-registration phase: (1) failing to realize I had collected time-series data for individual participant's average confidence—which allowed me to increase statistical power over merely pooling all participants' judgments—and (2) failing to plan both the ANOVA and difference-of-difference confidence intervals that could further confirm my second prediction.The main text reported the results after correcting this; here I report both pre-registered and post-hoc tests (the upshots are the same).

[6] The A-group was asked how confident they were that "that the string is completable"—equivalent to "coin landed heads" for A-Hsers, and "coin landed tails" for A-Tsers. The U-group was asked how confident they were that the coin landed heads (U-Hsers) or tails (U-Tsers).

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

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

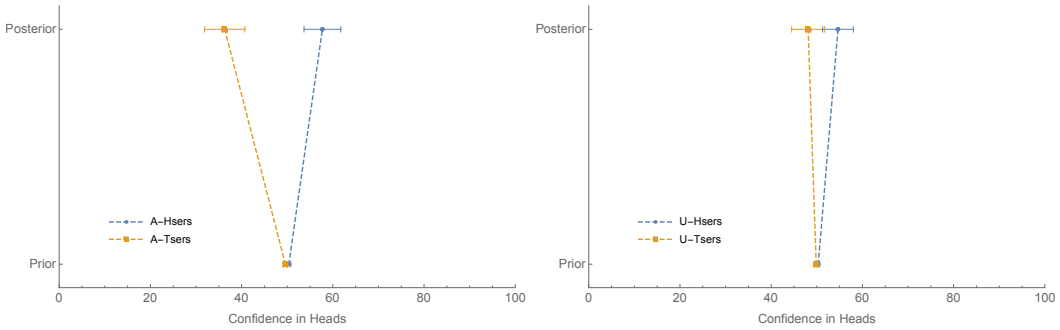


Figure 3: Mean prior and posterior confidence in heads in A (left) and U (right) conditions. Bars represent 95% confidence intervals.

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 A-Tser posteriors (M = 36.29, SD = 31.04) were below U-Tser posteriors (M =48.10, $SD = 28.47$), with $t(393) = 4.07$, $p < 0.001$, $d = 0.398$.

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

What *should've* been planned, I later realized, was to do (a) a 2×2 ANOVA, and (b) an empirically bootstrapped 95% confidence interval for the *difference* between the differences between A-Hsers/A-Tsers and U-Hsers/U-Tsers.

(a) Let **valence** be the variable for whether the subject was a Headser (= 1) or Tailser (= 0), and **ambiguity** be the variable for whether the subject was in the A (= 1) or U (= 0) group. Analyzing the results using a 2 (valence: Hser vs. Tser) by 2 (ambiguity: A vs. U) ANOVA indicated that there was a main effect of valence ($F(1, 899) = 46.47$, $p < 0.001$, $\eta^2 = 0.048$), a marginally significant main effect of ambiguity ($F(1, 899) = 4.31$, $p = 0.038$, $\eta^2 = 0.005$), and (as should've been predicted) an interaction effect between valence and ambiguity ($F(1, 899) = 14.57$, $p < 0.001$, $\eta^2 = 0.015$), indicating that the divergence between Headsers and 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. And while there *was* a significant difference between U-Hser posteriors (M = 54.64, SD = 26.93) and U-Tser posteriors (M = 48.10, SD = 28.47), with $t(486) = 2.61$ and (two-sided) $p = 0.009$, the effect size was smaller ($d = 0.236$) than for the difference between A-Hser and A-Tser posteriors (as mentioned, $d = 0.699$).

A further oversight on my part at the pre-registration phase was that I only realized after the fact that I actually had access to *time-series data* about how the participants' confidence evolves over time. In particular, using their priors and posteriors for each of the four coin tosses, I could calculate their average confidence in heads after seeing n bits of evidence, for n ranging from 0 to 4.[7] (For Bayesians, this average confidence equals their estimate for the proportion of times the coin landed heads.[8])

In other words, we can re-run the above statistics by pooling responses within subjects at each stage in their progression through the experiment. All the predicted results above hold true with this way of carving up the data (with universally lower p-values and higher effect sizes, since the variance of the data has dropped; Prediction 5 is still the only non-significant effect). These are the statistics I reported in the main text (§3.2, page 21).

A supplemental prediction was intended to probe the hypothesis that (something like) the model in Figure 2 is driving the effect. Within the ambiguous condition, I predicted that amongst those who *didn't* find a completion, the average confidence that their string was completable would be higher if it *was* completable (bottom right possibility of Figure 2) than if it wasn't (bottom left). This would indicate sensitivity to whether or not there was a word, over and above whether or not they found one.

To test this, in addition to recording their confidence, the experiment explicitly asked subjects in the ambiguous condition whether they found a completion. This data failed to confirm the supplemental prediction ($t(243) = 1.11$, $p = 0.13$, one-sided). However, I noticed

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

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

that a substantial proportion of people who *claimed* to have found a word did not have the extreme confidence that they should've if so (39% of them were less than 95% confident there was a completion, and 25% of them were less than 80%), indicating that self-reports of 'finding' might've been unreliable. If we rerun the data by operationalizing 'finding' as 'reporting 100% confidence there's a completion', the prediction is confirmed.[9]

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

First, since it's natural to expect ambiguity—uncertainty about how to react to evidence—to cause *variance* in people's opinions, we should expect the word-search condition to have more variance than the urn condition. This is what we find. Restricting attention to those with weak (so, potentially ambiguous) evidence—those who did not find a completion, or who did not see a black marble—the variance in opinions was much higher in the ambiguous condition than in the unambiguous one. This can be seen in the plots in Figure 15, and is confirmed by tests for equality of variance.[10] (Notice that the there remains a nontrivial amount of variance even in the unambiguous condition. Thus it may be that low levels of ambiguity—people being unsure how confident to be in response to a red marble—could be driving the small degree of polarization found in the unambiguous condition.)

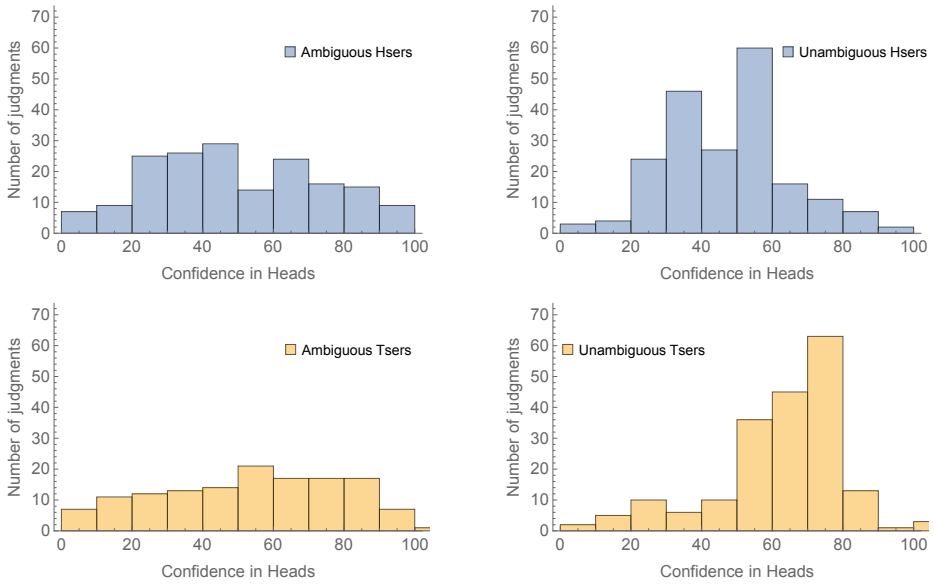


Figure 4: Histograms of judgments given weak evidence. Left: word-search. Right: urn.

Second, recall that the above (§3) model predicts that the mechanism that will drive polarization is that the accuracy of the rational opinions will increase *asymmetrically*: Headsers will be better at recognizing heads-cases; Tailsers will be better at recognizing tails-cases. As

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

[10] Word-search Headsers' variance was 563.33, while urn Headsers' was 285.28, Conover $= 5.40$, $p < 0.001$; and word-search Tailsers' was 606.78, while urn Tailsers' was 321.88, Conover $= 5.44$, $p < 0.001$.

can be seen in Figure 16, this is what we find. When presented with uncompletable strings (*Tails* cases for Headsers; *Heads* cases for Tailsers), neither group's average posterior moved significantly from their priors of 50%; but when they saw a completable string, it moved significantly in the direction of the truth. These data confirm the hypothesis that asymmetric accuracy-increases in rational credence can lead to polarization.[11] (The mean squared errors of their average prior vs. posterior is: for Headsers, $0.5(1-0.5037)^2+0.5(0-0.5034)^2 = 0.250$ vs. $0.5(1-0.6742)^2+0.5(0-0.4773)^2 = 0.167$; for Tailsers, 0.253 vs. 0.166.)

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

Figure 5: Ambiguous condition, mean prior and posterior confidence in *Heads*, by cases.
* = not significantly different from 50%.

C Computational Details

This appendix contains the details of the simulations used in §§4–5. It can be read on its own, or in tandem with the Mathematica notebook (https://github.com/kevindorst/RP_notebook) which contains a working version of all code.

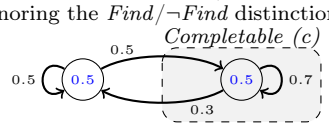
C.1 Cognitive Search Models (§4)

This subsection explains the generalizations of the (course-grained) word-search-task models that I call *cognitive search models*. We are to imagine an agent doing a cognitive search for flaws in a piece of evidence that purports to support or tell against a given proposition q. The general form of such a model divides the worlds into 3 classes, depending on whether the agent finds a flaw (F), there is a flaw that they don't find (C; the search is "Completable"), or there is no flaw (N). Within each class are (at least) two worlds that have the same posteriors, but which differ on whether the target proposition q is true.

Letting π be the prior and P the posterior (with P_w its realization at world w), a cognitive search model is any in which:

- $\pi(q|F) = \pi(q|C)$.
  (The existence of a flaw is what affects the probability of q, not whether you find it.)
- For any $w \in N$: $P_w = \pi(\cdot|\neg F)$.
  (If there's no flaw, all you learn is that you didn't find one.)

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



- For any $w \in F$: $P_w = \pi(\cdot|F)$
  (If you find a flaw, you learn exactly that.)
- For any $x, y \in C$: $P_x = P_y$, $P_x(F) = 0$, and $P_x(C) \geq \pi(C|\neg F)$.
  (If there is a flaw that you don't find, that determines the rational credence; you learn that you didn't find one; and you assign at least as much credence to this possibility as you would if that were all you learned.)

Such models are a simple generalization of the course-grained model of the word-search task from Figure 2—they can likewise be seen as the result of coarse-graining a more detailed model, e.g. one that distinguishes whether the flaw is obvious or not. For $x \in C$ and $y \in N$, we must have $P_x(C) \geq P_y(C)$ to satisfy the value of evidence; when $P_x = P_y$, the model is unambiguous and just consists in conditioning on whether or not you found a completion; but when $P_x(C) > P_y(C)$, the evidence is ambiguous (since $P_y(x) > 0$ and $P_x \neq P_y$)—and this is the ambiguity that leads to expectable polarization.

The simplest cognitive search models consist of 6 worlds (two in each of F, C, and N) plus a prior over them. (In Mathematica, we represent this with a 7-world frame in which the first world encodes the prior and is assigned probability 0 by all worlds, including itself.) Such models can be parameterized in a variety of ways; the funtion `csQModel` takes one such set of parameters and generates the resulting cognitive-search model. The function `getBBCondQCSModel` takes a prior in q, the degree to which finding a flaw would move it, and a probability of finding a flaw, and outputs a random cognitive search model by generating a random probability of there being a flaw (uniform from [0,1]), and then using that and the above to fix all the other parameters in a cognitive-search model.

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

We can then test the correlation between the probability of finding a flaw if there is one, and the expected accuracy of the update. There are a variety of ways to run such simulations. One issue is that when the `gBump` is large, i.e. the searches might shift your credence quite a bit, that introduces noise in the correlation. Thus I constrained such bumps to be small (as they will be in ensuing simulations), between 0 and 0.2. To minimize noise, I also fixed the prior in q at 0.5—but similar results are obtained by setting it to any other number. This simulation led to the plot on the left of Figure 5 (page 25).

Given this, we can test what proportion of the time expected accuracy favors scrutinizing incongruent studies rather than congruent ones, as a function of how much more likely you are (on average) to find extant flaws in the former than the latter. The simulations I ran fixed a given prior in q, and then generated pairs of cognitive-search models (one of which would raise your credence in q if you found a flaw, the other of which would lower it), such

that the the probability of finding a flaw was pulled from distributions with steadily higher means for the incongruent study and steadily lower means for the congruent one. As the gap in means grew, the proportion of pairs in which expected accuracy favors scrutinizing the incongruent study grew as well—i.e. selective scrutiny became more and more common. This led to the plot on the right of Figure 5 (page 25).

This finally puts us in a position to run a full simulation of two groups of agents, presented with pairs of studies, but one group (red) is better at finding flaws with studies that tell against q, while the other group (blue) is better at finding flaws with those that tell in favor of q. At each stage, each agent chooses which study to scrutinize based on which one they expect to make them most accurate, and then updates their credences with probability matching the various outcomes of that update-model.

There are a variety of choice-points in how to run such simulations. Although variations on the theme will lead to the same result, here are the ones I made. First, agents always have accurate beliefs about how likely they are to find a flaw in each study; this probability varies from a minimum of 0.1 to a maximum of 0.9; red agents are pulling (uniformly) from $[0.1 + \texttt{findGap}, 0.9]$ and blue agents are pulling (uniformly) from $[0.1, 0.9 - \texttt{findGap}]$. This parameter `findGap` can range from 0 (where there's no different between the groups) to 0.8 (where the difference is maximal). The simulation I displayed is with a gap of 0.5, but generally the rate of polarization grows as the gap increases.

Second, the amount agent's credences would move if they found a flaw in the study was limited to an initial upper bound (of 0.125), which was steadily lowered as agents saw more studies and the "weight" behind their credence in q was correspondingly increased. `hardenSpeed` is a parameter that controls how quickly agents harden in opinions; the smaller it is, the more polarization generally results but also the more chaotic their confidence-trajectories.

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

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

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

C.2 Argument Models (§5)

This subsection explains the simple models of arguments used in §5, before introducing scrutiny of arguments. We are to imagine you know that you are about to be presented with an argument in favor of a given claim (q). The general form of such models divides the worlds into two classes, depending on whether the argument is good (G) or bad (B). If the

argument is good, it's rational to increase your confidence in q; if it's bad, you're rational to decrease it. For simplicity, we assume there are only two posteriors you could end up with; moreover, we assume the argument will be more ambiguous if it's bad.

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

- $\pi(q|G) > \pi(q) > \pi(q|B)$
  (If the argument is good, q is more likely to be true; if not, it's not.)
- For any x, y: if $x, y \in G$, then $P_x = P_y$; and if $x, y \in B$, then $P_x = P_y$
  (Whether the argument is good or bad fully determines the rational posterior.)
- $\exists a, b > 0$, $a \geq b$: if $x \in G$ and $y \in B$, $P_x(G) = \pi(G) + a$ and $P_y(B) = \pi(B) + b$.
  (Whether the argument is good or bad, your confidence should shift toward the truth; but—since good arguments are easier to recognize—it should shift *more* if the argument is good than if it's bad.)

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

The simplest argument models consist of 4 worlds (two in each of G and B) plus a prior over them. (In Mathematica, we represent this with a 5-world frame in which the first world encodes the prior and is assigned probability 0 by all worlds, including itself). Such models can be parameterized in a variety of ways; the function `getArgModel` does so using $\pi(q)$ (`priorQ`), $\pi(q|G)$ (`gInf`), $P_g(G)$ for $g \in G$ (`gConf`), $\pi(q|B)$ (`gInf`), and $P_b(B)$ for $b \in B$ (`bConf`).

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

Given such functions, we can simulate presenting a group of (red) agents with random arguments that favor q, and a separate group of (blue) agents with random arguments that disfavor q. Again, there are a variety of choice-points in how to run such simulations. First, I assume agents always have accurate beliefs about how likely the arguments they're presented with are to be good or bad. Second, I assumed all arguments are equally likely to be good—$\pi(G)$ was drawn uniformly from $[0, 1]$. Finally, in addition to questions about the number of agents and arguments to simulate, we can modify how much in principle arguments could initially shift opinions, and how quickly agent's opinions "harden" (become less susceptible to change with new arguments).

I simulated the result of 20 agents in each group, each witnessing 100 (different) random arguments, with an initial maximum potential shift (`baseShift`) of 0.2; the result is the plot displayed in Figure 8.

The code also allows for simulations to vary the rate at which each group of agents is presented with good arguments, in particular using `favGBound` to lower-bound the probability that a red group-member's argument is good ($\pi(G)$ drawn from $[\texttt{favGBound}, 1]$) and upper-bound the probability that a blue-member's is ($\pi(G)$ drawn from $[0, 1 - \texttt{favGBound}]$). The code runs simulations with 30 agents and 50 arguments, with the above parameters for possible shifts and hardening speed, with `favGBound` at 0, 0.25, 0.5, 0.75, and 0.95. The effects of varying this parameter are not straightforward—at low levels it does little; at intervening levels it seems to move both groups move towards belief in q (still with a substantial gap), and at high levels it seems to reduce both the degree of belief-change and of polarization.

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

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

C.3 Argument-Scrutiny Models (§5)

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

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

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

would be if you learned the argument was good.)

- If $x \in C$, then $P_x(C) \geq P_b(C|\neg F)$
  (If there's a flaw that you don't find, your confidence that there is should be at least as great as it should be if you didn't scrutinize and updated your beliefs accordingly, and then conditioned on the claim that you wouldn't have found a flaw if you had.)

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

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

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

In version (1), scrutinizing agents never find a flaw even if there is one (`pFind` = 0), and the scrutiny adds no ambiguity (`jShift` = 0). Such scrutiny does not change the original argument-model at all, and so agents who scrutinize polarize as much and in the same direction as those who don't—as seen in the top left of Figure 9 on page 30.

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

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

In version (4), scrutinizing agents again sometimes find a flaw if there is one (`pFind` pulled randomly from $[0, 1]$), and scrutiny introduces a *substantial* ambiguity (`jShift` pulled

randomly from $[0, 1]$). The result is that agents who scrutinize predictably polarize in the *opposite* direction of those that don't—as seen in the bottom right of Figure 9 on page 30.

Robustness. Recall that pro agents in this simulation are identical to those from the main simulation of §C.2, meaning we have estimates for their mean posteriors with these parameters at 0.663, with 95% confidence interval of $[0.642, 0.684]$. To check that the results in the above simulations (1)–(4) were robust, I ran the same parameters but with 200 con agents and calculated estimates and confidence intervals for their posteriors. The results are as expected.

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

References

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

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

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

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

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

Anglin, Stephanie M., 2019. 'Do beliefs yield to evidence? Examining belief perseverance vs. change in response to congruent empirical findings'. *Journal of Experimental Social Psychology*, 82(February):176–199.

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

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

Asch, Solomon E., 1955. 'Opinions and Social Pressure'. *Scientific American*, 193(5):31–35.

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

Aumann, Robert J, 1999. 'Interactive epistemology II: probability'. *International Journal of Game Theory*, 28(3):301–314.

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

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

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

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

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

Becher, Tony and Trowler, Paul, 2001. *Academic tribes and territories*. McGraw-Hill Education (UK).

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

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

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

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

Brauer, Markus, Judd, Charles M., and Gliner, Melissa D., 1995. 'The effects of repeated expressions on attitude polarization during group discussions.' *Journal of Personality and Social Psychology*, 68(6):1014–1029.

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

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

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

Briggs, Ray, 2009. 'The Anatomy of the Big Bad Bug'. *Nous*, 43(3):428–449.

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

Brown, Roger, 1986. *Social Psychology*. Free Press.

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

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

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

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

———, 2020. 'Accuracy, Estimates, and Representation Results'.

Campbell-Moore, Catrin and Levinstein, Benjamin A, 2020. 'Strict Propriety is Weak'. *Analysis*, To Appear.

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

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

Carr, Jennifer Rose, 2019. 'A modesty proposal'. *Synthese*, 1–21.

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

Chater, Nick and Oaksford, Mike, 1999. 'Ten years of the rational analysis of cognition'. *Trends in Cognitive Sciences*, 3(2):57–65.

Christensen, David, 2007. 'Epistemology of Disagreement: The Good News'. *Philosophical Review*, 116(2):187–217.

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

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

Cohen, Geoffrey L, 2003. 'Party over policy: The dominating impact of group influence on political beliefs.' *Journal of personality and social psychology*, 85(5):808.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Feldman, Richard, 2007. 'Reasonable religious disagreements'. In Louise Antony, ed., *Philosophers Without Gods: Meditations on Atheism and the Secular*, 194–214. Oxford University Press.

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

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

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

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

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

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

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

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

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

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

Geanakoplos, John, 1989. 'Game Theory Without Partitions, and Applications to Speculation and Consensus'. *Research in Economics*, Cowles Fou(914).

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

———, 2018. 'The Bias Bias in Behavioral Economics'. *Review of Behavioral Economics*, 5(3-4):303–336.
Gilovich, Thomas, 1983. 'Biased evaluation and persistence in gambling.' *Journal of personality and social psychology*, 44(6):1110—-1126.
———, 1991. *How We Know What Isn't So*. Free Press.
Glaser, Markus and Weber, Martin, 2010. 'Overconfidence'. *Behavioral finance: Investors, corporations, and markets*, 241–258.
Good, I J, 1967. 'On the Principle of Total Evidence'. *The British Journal for the Philosophy of Science*, 17(4):319–321.
———, 1974. 'A Little Learning Can be Dangerous'. *The British Journal for the Philosophy of Science*, 25(4):340–342.
Gopnik, Alison, 1996. 'The Scientist as Child'. *Philosophy of Science*, 63(December):485–514.
Griffiths, Thomas L., Chater, Nick, Norris, Dennis, and Pouget, Alexandre, 2012. 'How the bayesians got their beliefs (and what those beliefs actually are): Comment on bowers and davis (2012)'. *Psychological Bulletin*, 138(3):415–422.
Griffiths, Thomas L., Lieder, Falk, and Goodman, Noah D., 2015. 'Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic'. *Topics in Cognitive Science*, 7(2):217–229.
Grönlund, Kimmo, Herne, Kaisa, and Setälä, Maija, 2015. 'Does enclave deliberation polarize opinions?' *Political Behavior*, 37(4):995–1020.
Hahn, Ulrike and Harris, Adam J.L., 2014. 'What Does It Mean to be Biased. Motivated Reasoning and Rationality.' In *Psychology of Learning and Motivation - Advances in Research and Theory*, volume 61, 41–102.
Haidt, Jonathan, 2012. *The righteous mind: Why good people are divided by politics and religion*. Vintage.
Halpern, Joseph Y, 2010. 'I don't want to think about it now: Decision theory with costly computation'. In *Twelfth international conference on the principles of knowledge representation and reasoning*.
Hamblin, Charles L, 1976. 'Questions in montague english'. In *Montague grammar*, 247–259. Elsevier.
Hart, William, Albarracín, Dolores, Eagly, Alice H, Brechan, Inge, Lindberg, Matthew J, and Merrill, Lisa, 2009. 'Feeling validated versus being correct: a meta-analysis of selective exposure to information.' *Psychological bulletin*, 135(4):555.
Harvey, Nigel, 1997. 'Confidence in judgment'. *Trends in cognitive sciences*, 1(2):78–82.
Hastie, Reid and Dawes, Robyn M, 2009. *Rational choice in an uncertain world: The psychology of judgment and decision making*. Sage Publications.
Hedden, Brian, 2015. 'Options and Diachronic Tragedy'. *Philosophy and Phenomenological Research*, 90(2):423–451.
Hegselmann, Rainer and Krause, Ulrich, 2002. 'Opinion dynamics and bounded confidence: Models, analysis and simulation'. *Jasss*, 5(3).
Henderson, Leah, 2021. 'Higher-Order Evidence and Losing One's Conviction'. *Noûs*, To appear.
Hintikka, Jaako, 1962. *Knowledge and Belief*. Cornell University Press.
Hoffrage, Ulrich, 2004. 'Overconfidence'. In *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*, 235–254.
Horowitz, Sophie, 2013. 'Immoderately rational'. *Philosophical Studies*, 167(1):41–56.
———, 2014. 'Epistemic Akrasia'. *Noûs*, 48(4):718–744.
———, 2019. 'The Truth Problem for Permissivism'. *The Journal of Philosophy*, cxvi(5):237—-262.
Isenberg, Daniel J., 1986. 'Group Polarization. A Critical Review and Meta-Analysis'. *Journal of Personality and Social Psychology*, 50(6):1141–1151.
Iyengar, Shanto, Lelkes, Yphtach, Levendusky, Matthew, Malhotra, Neil, and Westwood, Sean J., 2019. 'The origins and consequences of affective polarization in the United States'. *Annual Review of Political Science*, 22:129–146.
Iyengar, Shanto, Sood, Gaurav, and Lelkes, Yphtach, 2012. 'Affect, not ideology: A social identity perspective on polarization'. *Public Opinion Quarterly*, 76(3):405–431.
Jackson, Elizabeth, 2021. 'A Defense of Intrapersonal Belief Permissivism'. *Episteme*, 18(2):313–327.
Jamieson, Kathleen Hall and Cappella, Joseph N, 2008. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press.
Jeffrey, Richard C, 1990. *The logic of decision*. University of Chicago press.
Jern, Alan, Chang, Kai Min K., and Kemp, Charles, 2014. 'Belief polarization is not always irrational'. *Psychological Review*, 121(2):206–224.
Johnson, Dominic D P, 2009. *Overconfidence and war*. Harvard University Press.
Jost, John T., Glaser, Jack, Kruglanski, Arie W., and Sulloway, Frank J., 2003. 'Political Conservatism as Motivated Social Cognition'. *Psychological Bulletin*, 129(3):339–375.
Joyce, James M, 2009. 'Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief'. In Franz Huber and Christoph Schmidt-Petri, eds., *Degrees of Belief*, 263–297. Springer.
———, 2010. 'A Defense of Imprecise Credences in Inference and Decision Making'. *Philosophical Perspectives*, 24(1):282–323.
Kadane, Joseph B, Schervish, Mark, and Seidenfeld, Teddy, 2008. 'Is ignorance bliss?' *The Journal of Philosophy*, 105(1):5–36.
Kadane, Joseph B., Schervish, Mark J., and Seidenfeld, Teddy, 1996. 'Reasoning to a foregone conclusion'. *Journal of the American Statistical Association*, 91(435):1228–1235.
Kahan, D M, 2018. 'Why smart people are vulnerable to putting tribe before truth'. *Scientific American*.
Kahan, Dan M., 2013. 'Ideology, motivated reasoning, and cognitive reflection'. *Judgment and Decision Making*, 8(4):407–424.
Kahan, Dan M., Peters, Ellen, Dawson, Erica Cantrell, and Slovic, Paul, 2017. 'Motivated numeracy and enlightened self-government'. *Behavioural Public Policy*, 1:54–86.
Kahan, Dan M., Peters, Ellen, Wittlin, Maggie, Slovic, Paul, Ouellette, Lisa Larrimore, Braman, Donald, and Mandel, Gregory, 2012. 'The polarizing impact of science literacy and numeracy on perceived climate change

risks'. *Nature Climate Change*, 2(10):732–735.
Kahneman, Daniel, 2011. *Thinking Fast and Slow*. Farrar, Straus, and Giroux.
Kahneman, Daniel, Slovic, Paul, and Tversky, Amos, eds., 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press.
Kahneman, Daniel and Tversky, Amos, 1996. 'On the reality of cognitive illusions.'
Kelly, Thomas, 2005. 'The epistemic significance of disagreement'. In John Hawthorne and Tamar Szabó Gendler, eds., *Oxford Studies in Epistemology*, volume 1, 167–196. Oxford University Press.
———, 2008. 'Disagreement, Dogmatism, and Belief Polarization'. *The Journal of Philosophy*, 105(10):611–633.
———, 2010. 'Peer disagreement and higher order evidence'. In Alvin I Goldman and Dennis Whitcomb, eds., *Social Epistemology: Essential Readings*, 183–217. Oxford University Press.
Kinney, David and Bright, Liam, 2021. 'Elite Group Ignorance'. *Philosophy and Phenomenological Research*, To appear.
Klaczynski, Paul A and Narasimham, Gayathri, 1998. 'Development of scientific reasoning biases: Cognitive versus ego-protective explanations.' *Developmental Psychology*, 34(1):175.
Klein, Ezra, 2014. 'How politics makes us stupid'. *Vox*, 1–14.
———, 2020. *Why We're Polarized*. Profile Books.
Koerth, Maggie, 2019. 'Why Partisans Look At The Same Evidence On Ukraine And See Wildly Different Things'. *FiveThirtyEight*.
Koriat, Asher, Lichtenstein, Sarah, and Fischhoff, Baruch, 1980. 'Journal of Experimental Psychology : Human Learning and Memory Reasons for Confidence'. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2):107–118.
Kossinets, Gueorgi and Watts, Duncan J., 2009. 'Origins of Homophily in an Evolving Social Network'. *American Journal of Sociology*, 115(2):405–450.
Krueger, Joachim I and Massey, Adam L, 2009. 'A rational reconstruction of misbehavior'. *Social Cognition*, 27(5):786–812.
Kuhn, Deanna and Lao, Joseph, 1996. 'Effects of Evidence on Atittudes: is Polarization the Norm?' *Psycholohical Science*, 7(2):115–120.
Kuhn, Thomas, 1962. *The Structure of Scientific Revolutions*. The University of Chicago Press.
Kunda, Ziva, 1990. 'The case for motivated reasoning'. *Psychological Bulletin*, 108(3):480–498.
Lakoff, George, 1997. *Moral politics: What conservatives know that liberals don't*. University of Chicago Press.
Lasonen-Aarnio, Maria, 2013. 'Disagreement and Evidential Attenuation'. *Nous*, 47(4):767–794.
———, 2014. 'Higher-order evidence and the limits of defeat'. *Philosophy and Phenomenological Research*, 8(2):314–345.
———, 2015. 'New Rational Reflection and Internalism about Rationality'. In Tamar Szabó Gendler and John Hawthorne, eds., *Oxford Studies in Epistemology*, volume 5, 145–171. Oxford University Press.
Lazer, David, Baum, Matthew, Benkler, Jochai, Berinsky, Adam, Greenhill, Kelly, Metzger, Miriam, Nyhan, Brendan, Pennycook, G., Rothschild, David, Sunstein, Cass, Thorson, Emily, Watts, Duncan, and Zittrain, Jonathan, 2018. 'The science of fake news'. *Science*, 359(6380):1094–1096.
Le Mens, Gaël and Denrell, Jerker, 2011. 'Rational Learning and Information Sampling: On the " Naivety" Assumption in Sampling Explanations of Judgment Biases'. *Psychological Review*, 118(2):379–392.
Lederman, Harey, 2015. 'People with Common Priors Can Agree to Disagree'. *The Review of Symbolic Logic*, 8(1):11–45.
Levi, Isaac, 1974. 'On indeterminate probabilities'. *The Journal of Philosophy*, 71(13):391–418.
Levinstein, Benjamin A., 2020. 'Accuracy, Deference, and Chance'.
Lewis, David, 1976. 'Probabilities of Conditionals and Conditional Probabilities'. *The Philosophical Review*, 85(3):297–315.
Liberman, Akiva and Chaiken, Shelly, 1992. 'Defensive processing of personally relevant health messages'. *Personality and Social Psychology Bulletin*, 18(6):669–679.
Lichtenstein, Sarah, Fischhoff, Baruch, and Phillips, Lawrence D., 1982. 'Calibration of probabilities: The state of the art to 1980'. In Daniel Kahneman, Paul Slovic, and Amos Tversky, eds., *Judgment under Uncertainty*, 306–334. Cambridge University Press.
Lieder, Falk and Griffiths, Thomas L., 2019. 'Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources'. *Behavioral and Brain Sciences*.
Lilienfeld, Scott O, Ammirati, Rachel, and Landfield, Kristin, 2009. 'Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare?' *Perspectives on psychological science*, 4(4):390–398.
Liu, ChengâĂŘHong, 2017. 'Evaluating arguments during instigations of defence motivation and accuracy motivation'. *British Journal of Psychology*, 108(2):296–317.
Loh, Isaac and Phelan, Gregory, 2019. 'Dimensionality and disagreement: Asymptotic belief divergence in response to common information'. *International Economic Review*, 60(4):1861–1876.
Lord, Charles G, Lepper, Mark R, and Preston, Elizabeth, 1984. 'Considering the opposite: a corrective strategy for social judgment.' *Journal of personality and social psychology*, 47(6):1231.
Lord, Charles G., Ross, Lee, and Lepper, Mark R., 1979. 'Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence'. *Journal of Personality and Social Psychology*, 37(11):2098–2109.
Lord, Charles G and Taylor, Cheryl A, 2009. 'Biased assimilation: Effects of assumptions and expectations on the interpretation of new evidence'. *Social and Personality Psychology Compass*, 3(5):827–841.
Lundgren, Sharon R and Prislin, Radmila, 1998. 'Motivated cognitive processing and attitude change'. *Personality and Social Psychology Bulletin*, 24(7):715–726.
Mandelbaum, Eric, 2018. 'Troubles with Bayesianism: An introduction to the psychological immune system'. *Mind & Language*, 1–17.

Mäs, Michael and Flache, Andreas, 2013. 'Differentiation without distancing. explaining bi-polarization of opinions without negative influence'. *PLoS ONE*, 8(11).

Mason, Lilliana, 2016. 'A cross-cutting calm: How social sorting drives affective polarization'. *Public Opinion Quarterly*, 80(Specialissue1):351–377.

———, 2018. *Uncivil agreement: How politics became our identity*. University of Chicago Press.

Matheson, Jonathan, 2015. *The epistemic significance of disagreement*. Springer.

May, Regine S., 1991. 'Overconfidence in Overconfidence'. In Attila Chik/'an, ed., *Progress in Decision, Utility and Risk Theory*, 67–74. Springer Science + Business Media.

McGrath, Sarah, 2008. 'Moral Disagreement and Moral Expertise'. *Oxford Studies in Metaethics: Volume III*, 87–108.

McHoskey, John W., 1995. 'Case Closed? On the John F. Kennedy Assassination: Biased Assimilation of Evidence and Attitude Polarization'. *Basic and Applied Social Psychology*, 17(3):395–409.

McKenzie, Craig R M, 2004. 'Framing effects in inference tasksâĂŤand why they are normatively defensible'. *Memory & cognition*, 32(6):874–885.

Mcpherson, Miller, Smith-lovin, Lynn, and Cook, James M, 2001. 'Birds of a Feather: Homophily in Social Networks'. *Annual Review of Sociology*, 27:415–444.

McWilliams, Emily C., 2019. *Evidentialism and belief polarization*. 8. Springer Netherlands.

Mercier, Hugo, 2017. 'Confirmation bias—Myside bias.' *Cognitive illusions: Intriguing phenomena in thinking, judgment and memory, 2nd ed.*, 99–114.

———, 2020. *Not born yesterday*. Princeton University Press.

Mercier, Hugo and Sperber, Dan, 2011. 'Why do humans reason ? Arguments for an argumentative theory'. 57–111.

———, 2017. *The enigma of reason*. Harvard University Press.

Miller, Arthur G., McHoskey, John W., Bane, Cynthia M., and Dowd, Timothy G., 1993. 'The attitude polarization phenomenon: Role of response measure, attitude extremity, and behavioral consequences of reported attitude change.' *Journal of Personality and Social Psychology*, 64(4):561–574.

Mills, Charles W., 2007. 'White ignorance'. *Race and Epistemologies of Ignorance*, 13–38.

Moore, Don A, Tenney, Elizabeth R, and Haran, Uriel, 2015. 'Overprecision in judgment'. *The Wiley Blackwell handbook of judgment and decision making*, 2:182–209.

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

Murray, Mark, 2018. 'Poll: 58 percent say gun ownership increases safety'.

Myers, David G., 1975. 'Discussion-Induced Attitude Polarization'. *Human Relations*, 28(8):699–714.

———, 2012. *Social Psychology*. McGraw-Hill Education.

Myers, David G and Bishop, George D, 1970. 'Discussion effects on racial attitudes'. *Science*, 169(3947):778–779.

Myers, David G. and Lamm, Helmut, 1976. 'The group polarization phenomenon'. *Psychological Bulletin*, 83(4):602–627.

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

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

Nielsen, Michael and Stewart, Rush T, 2021. 'Persistent Disagreement and Polarization in a Bayesian Setting'. *British Journal for the Philosophy of Science*, 72(1):51–78.

Nimark, Kristoffer P. and Sundaresan, Savitar, 2019. 'Inattention and belief polarization'. *Journal of Economic Theory*, 180:203–228.

Nyhan, Brendan and Reifler, Jason, 2010. 'When corrections fail: The persistence of political misperceptions'. *Political Behavior*, 32(2):303–330.

Oaksford, Mike and Chater, Nick, 1994. 'A Rational Analysis of the Selection Task as Optimal Data Selection'. *Psychological Review*, 101(4):608–631.

———, 1998. *Rational models of cognition*. Oxford University Press Oxford.

———, 2007. *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

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

———, 2019. *The Misinformation Age: How False Beliefs Spread Cailin O'Connor James Owen Weatherall*. Yale Press.

Olsson, Erik J, 2013. 'A Bayesian simulation model of group deliberation and polarization'. In *Bayesian argumentation*, 113–133. Springer.

Ortoleva, Pietro and Snowberg, Erik, 2015. 'Overconfidence in political behavior'. *American Economic Review*, 105(2):504–535.

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

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

Pennycook, Gordon and Rand, David G., 2019. 'Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning'. *Cognition*, 188(September 2017):39–50.

Peterson, CAMERON R. and Beach, LEE R., 1967. 'Man As an Intuitive Statistician'. *Psychological Bulletin*, 68(1):29–46.

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

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

Petty, Richard E. and Wegener, Duane T., 1998. 'Attitude change: Multiple roles for persuasion variables'. *The*

handbook of social psychology, 323–390.
Plous, Scott, 1991. 'Biases in the assimilation of technological breakdowns: Do accidents make us safer?' *Journal of Applied Social Psychology*, 21(13):1058–1082.
Podgorski, Abelard, 2016. 'Dynamic permissivism'. *Philosophical Studies*, 173(7):1923–1939.
Proietti, Carlo, 2017. 'The dynamics of group polarization'. In A. Baltag, J. Seligman, and T. Yamada, eds., *International Workshop on Logic, Rationality and Interaction*, volume 10455, 195–208.
Pronin, Emily, 2008. 'How We See Ourselves and How We See Others'. *Science*, 320(16):1177–1180.
Rabin, Matthew and Schrag, Joel, 1999. 'First impressions matter: a model of confirmatory bias'. *Quarterly Journal of Economics*, (February):37–82.
Ramsey, F. P., 1990. 'Weight or the value of knowledge'. *British Journal for the Philosophy of Science*, 41(1):1–4.
Risen, Jane and Gilovich, Thomas, 2007. 'Informal Logical Fallacies.' In R.J. Sternberg, H.L. Roediger, and D.F. Halpern, eds., *Critical thinking in psychology*, 110–130. Cambridge University Press.
Rizzo, Mario J and Whitman, Glen, 2019. *Escaping paternalism: Rationality, behavioral economics, and public policy*. Cambridge University Press.
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'.
Rogers, Kayleigh, 2020. 'Americans Were Primed To Believe The Current Onslaught Of Disinformation'.
Ross, Lee, 2012. 'Reflections on Biased Assimilation and Belief Polarization'. *Critical Review*, 24(2):233–245.
Salow, Bernhard, 2018. 'The Externalist's Guide to Fishing for Compliments'. *Mind*, 127(507):691–728.
———, 2019. 'Elusive Externalism'. *Mind*, 128(510):397–427.
———, 2020. 'The Value of Evidence'. In Maria Lasonen-Aarnio and Clayton Littlejohn, eds., *The Routledge Handbook for the Philosophy of Evidence*. Routledge.
Samet, Dov, 1999. 'Bayesianism without learning'. *Research in Economics*, 53:227–242.
———, 2000. 'Quantified Beliefs and Believed Quantities'. *Journal of Economic Theory*, 95(2):169–185.
Savage, Leonard J, 1954. *The Foundations of Statistics*. Wiley Publications in Statistics.
Schervish, M. J., Seidenfeld, T., and Kadane, J.B., 2004. 'Stopping to Reflect'. *The Journal of Philosophy*, 101(6):315–322.
Schervish, Mark J, 1989. 'A general method for comparing probability assessors'. *The annals of statistics*, 17(4):1856–1879.
Schkade, David, Sunstein, Cass R., and Hastie, Reid, 2010. 'WHEN DELIBERATION PRODUCES EXTREMISM'. *Critical Review*, 22(2-3):227–252.
Schoenfield, Miriam, 2012. 'Chilling out on epistemic rationality'. *Philosophical Studies*, 158(2):197–219.
———, 2014. 'Permission to Believe: Why Permissivism is True and What it Tells Us About Irrelevant Influences on Belief'. *Nous*, 48(2):193–218.
———, 2015. 'A Dilemma for Calibrationism'. *Philosophy and Phenomenological Research*, 91(2):425–455.
———, 2017. 'Conditionalization Does Not (In General) Maximize Expected Accuracy'. *Mind*, 126(504):1155–1187.
———, 2018. 'An Accuracy Based Approach to Higher Order Evidence'. *Philosophy and Phenomenological Research*, 96(3):690–715.
Schuette, Robert A and Fazio, Russell H, 1995. 'Attitude accessibility and motivation as determinants of biased processing: A test of the MODE model'. *Personality and Social Psychology Bulletin*, 21(7):704–710.
Schultheis, Ginger, 2018. 'Living on the Edge: Against Epistemic Permissivism'. *Mind*, 127(507):863–879.
Sears, David O and Freedman, Jonathan L, 1967. 'Selective exposure to information: A critical review'. *Public Opinion Quarterly*, 31(2):194–213.
Seidenfeld, Teddy, 1981. 'Levi on the dogma of randomization in experiments'. In *Henry E. Kyburg, Jr. & Isaac Levi*, 263–291. Springer.
Seidenfeld, Teddy and Wasserman, Larry, 1993. 'Dilation for Sets of Probabilities'. *The Annals of Statistics*, 21(3):1139–1154.
Setiya, Kieran, 2012. *Knowing right from wrong*. OUP Oxford.
Siegel, Susanna, 2021. 'The Problem of Culturally Normal Beliefs'. In Robin Celikates, Sally Haslanger, and Jason Stanley, eds., *Ideology: New Essays*, To Appear. Oxford University Press.
Simpson, Robert Mark, 2017. 'Permissivism and the arbitrariness objection'. *Episteme*, 14(4):519–538.
Singer, Daniel J, Bramson, Aaron, Grim, Patrick, Holman, Bennett, Jung, Jiin, Kovaka, Karen, Ranginani, Anika, and Berger, William J, 2019. 'Rational social and political polarization'. *Philosophical Studies*, 176(9):2243–2267.
Sliwa, Paulina and Horowitz, Sophie, 2015. 'Respecting *All* the Evidence'. *Philosophical Studies*, 172(11):2835–2858.
Smith, Laura G. E. and Postmes, Tom, 2011. 'The power of talk: Developing discriminatory group norms through discussion'. *British Journal of Social Psychology*, 50(2):193–215.
Solomon, Miriam, 1992. 'Scientific Rationality and Human Reasoning'. *Philosophy of Science*, 59(3):439–455.
Stafford, Tom, 2015. *For argument's sake: evidence that reason can change minds*. Smashwords Edition.
———, 2020. 'Evidence for the rationalisation phenomenon is exaggerated'. *Behavioral and Brain Sciences*, 43.
Stalnaker, Robert, 1968. 'A Theory of Conditionals'. In Nicholas Rescher, ed., *Studies in Logical Theory*, 98–112. Oxford University Press.
———, 2019. 'Rational Reflection, and the Notorious Unmarked Clock'. In *Knowledge and Conditionals: Essays on the Structure of Inquiry*, 99–112. Oxford University Press.
Stangor, Charles and Walinga, Jennifer, 2014. *Introduction to psychology*. BCcampus, BC Open Textbook Project.
Stanovich, Keith E., 2020. *The Bias That Divide Us: The Science and Politics of Myside Thinking*. MIT Press.
Stone, Daniel F., 2019. 'âĂIJUnmotivated biasâĂİ and partisan hostility: Empirical evidence'. *Journal of Behavioral and Experimental Economics*, 79(August):12–26.
———, 2020. 'Just a Big Misunderstanding? Bias and Bayesian Affective Polarization'. *International Economic*

Review, 61(1):189–217.
Sunstein, C, 2009. *Going to Extremes: How Like Minds Unite and Divide*. Oxford University Press.
Sunstein, Cass R, 2000. 'Deliberative trouble? Why groups go to extremes'. *The Yale Law Journal*, 110(1).
———, 2007. 'Group polarization and 12 angry men'. *Negotiation Journal*, 23(4):443–447.
———, 2017. *#Republic: Divided democracy in the age of social media*. Princeton University Press.
Sutherland, Stuart, 1992. *Irrationality: The enemy within*. Constable and Company.
Taber, Charles S, Cann, Damon, and Kucsova, Simona, 2009. 'The motivated processing of political arguments'. *Political Behavior*, 31(2):137–155.
Taber, Charles S and Lodge, Milton, 2006. 'Motivated Skepticism in the Evaluation of Political Beliefs'. *American Journal of Political Science*, 50(3):755–769.
Talisse, Robert B, 2019. *Overdoing democracy: Why we must put politics in its place*. Oxford University Press.
Tenenbaum, Joshua B and Griffiths, Thomas L, 2006. 'Optimal Predictions in Everyday Cognition'. *Psychological Science*, 17(9):767–773.
Tenenbaum, Joshua B, Kemp, Charles, Griffiths, Thomas L, and Goodman, Noah D, 2011. 'How to grow a mind: Statistics, structure, and abstraction'. *science*, 331(6022):1279–1285.
Tesser, Abraham, Martin, Leonard, and Mendolia, Marilyn, 1995. 'The impact of thought on attitude extremity and attitude-behavior consistency.'
Thaler, Richard H., 2015. *Misbehaving: The Making of Behavioural Economics*. Penguin.
Titelbaum, Michael, 2015. 'Rationality's Fixed Point (or: In Defense of Right Reason)'. In Tamar Szabó Gendler and John Hawthorne, eds., *Oxford Studies in Epistemology*, volume 5, 253–292. Oxford University Press.
Todd, Peter M, Hills, Thomas T, Robbins, Trevor W, and Lupp, Julia, 2012. *Cognitive search: Evolution, algorithms, and the brain*, volume 9. MIT press.
Toplak, Maggie E and Stanovich, Keith E, 2003. 'Associations between myside bias on an informal reasoning task and amount of postâĂŘsecondary education'. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 17(7):851–860.
Tversky, Amos and Kahneman, Daniel, 1974. 'Judgment under uncertainty: Heuristics and biases'. *Science*, 185(4157):1124–1131.
van der Maas, Han L J, Dalege, Jonas, and Waldorp, Lourens, 2020. 'The polarization within and across individuals: the hierarchical Ising opinion model'. *Journal of Complex Networks*, 8(2).
van Ditmarsch, Hans, Halpern, Joseph Y, van der Hoek, Wiebe, and Kooi, Barteld, 2015. *Handbook of Epistemic Logic*. College Publications.
van Fraassen, Bas, 1984. 'Belief and the Will'. *The Journal of Philosophy*, 81(5):235–256.
Van Heuvelen, Ben, 2007. 'The Internet is Making us Stupid'. *Salon*.
van Prooijen, Jan-Willem and Krouwel, André P M, 2019. 'Psychological Features of Extreme Political Ideologies'. *Current Directions in Psychological Science*, 28(2):159–163.
Van Swol, Lyn M., 2007. 'Perceived importance of information: The effects of mentioning information, shared information bias, ownership bias, reiteration, and confirmation bias'. *Group Processes and Intergroup Relations*, 10(2):239–256.
Vavova, Katia, 2014. 'Confidence, Evidence, and Disagreement'. *Erkenntnis*, 79(S1):173–183.
———, 2016. 'Irrelevant Influences'. *Philosophy and Phenomenological Research*, To appear.
———, 2018. 'Irrelevant Influences'. *Philosophy and Phenomenological Research*, 96(1):134–152.
Vinokur, Amiram and Burstein, Eugene, 1974. 'Effects of partially shared persuasive arguments on group-induced shifts: A group-problem-solving approach.' *Journal of Personality and Social Psychology*, 29(3):305–315.
Vosoughi, Soroush, Roy, Deb, and Aral, Sinan, 2018. 'The spread of true and false news online'. *Science*, 359(6380):1146–1151.
Weatherall, James Owen and O'Connor, Cailin, 2020. 'Endogenous epistemic factionalization'. *Synthese*, 1–23.
Weisberg, Jonathan, 2007. 'Conditionalization, reflection, and self-knowledge'. *Philosophical Studies*, 135(2):179–197.
White, Roger, 2005. 'Epistemic Permissiveness'. *Philosophical Perspectives*, 19(1):445–459.
———, 2009. 'Evidential Symmetry and Mushy Credence'. In Tamar Szabó Gendler and John Hawthorne, eds., *Oxford Studies in Epistemology*, volume 3, 161–186. Oxford University Press.
———, 2010. 'You Just Believe that Because...' *Philosophical Perspectives*, 24:573–615.
Whittlestone, Jess, 2017. 'The importance of making assumptions : why confirmation is not necessarily a bias'. (July).
Wilkinson, Will, 2018. 'The Density Divide: Urbanization, Polarization, and Populist Backlash'. Technical report, The Niskanen Center.
Williamson, Timothy, 2000. *Knowledge and its Limits*. Oxford University Press.
———, 2008. 'Why Epistemology Cannot be Operationalized'. In Quentin Smith, ed., *Epistemology: New Essays*, 277–300. Oxford University Press.
———, 2014. 'Very Improbable Knowing'. *Erkenntnis*, 79(5):971–999.
———, 2019. 'Evidence of Evidence in Epistemic Logic'. In Mattias Skipper and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidence: New Essays*, 265–297. Oxford University Press.
Wilson, Andrea, 2014. 'Bounded Memory and Biases in Information Processing'. *Econometrica*, 82(6):2257–2294.
Wolfe, Christopher R and Britt, M Anne, 2008. 'The locus of the myside bias in written argumentation'. *Thinking & reasoning*, 14(1):1–27.
Wolfers, Justin, 2014. 'How Confirmation Bias Can Lead to a Spinning of Wheels'.
Ye, Ru, 2019. 'The Arbitrariness Objection against Permissivism'. *Episteme*, (2019):1–20.
Zollman, Kevin, 2021. *Network Epistemology: How our Social Connections Shape Knolwedge*.