Rational Polarization from Joint Inference: Theory and Experiment

Working Draft

Jacob VanNattan

April 2026

Abstract

Two fully Bayesian agents with different priors over a state and identical, correct priors over the complete signal structure can update in opposite directions from the same evidence. The mechanism, which we call joint inference, requires heterogeneity only in priors over the state. We give a complete characterization of when joint inference alone is sufficient to generate rational polarization. Experimental results from an unincentivized pilot follow the theory. Subjects correctly use their prior beliefs about the state to infer the informativeness of each signal, weighting it more heavily the closer it is to their prior. These asymmetric weights on ex ante identical sources then propagate into the subjects' posterior beliefs about the state.

1 Introduction

Two people exposed to the same evidence can update their initial views in opposite directions, each becoming more extreme. Lord et al. (1979) documented this experimentally in the context of capital punishment research; subjects with opposing prior views, presented with the same mixed evidence on deterrence, became more polarized rather than less. The pattern has been replicated across domains from evidence about the JFK assassination (McHoskey, 1995) to biological explanations of homosexuality (Munro and Ditto, 1997) and has generated a substantial theoretical literature on when and why agents holding different beliefs can rationally update in opposite directions from identical information.

Three classes of explanations dominate the literature, each departing from the rational Bayesian benchmark with standard preferences in a different way. The first appeals to biases, misspecification, or bounded rationality: agents process evidence with systematic distortions (Rabin and Schrag, 1999), aggregate signals incorrectly (Fryer et al., 2019), or misperceive selection in the information shared through social networks (Bowen et al., 2023). The second appeals to non-standard preferences: agents derive utility directly from their beliefs and update accordingly (Bénabou and Tirole, 2002; Bénabou, 2015). The third retains both rational updating and standard preferences but introduces heterogeneity in priors beyond the state itself — either on auxiliary dimensions that affect signal interpretation (Jern et al., 2014; Benoît and Dubra, 2019; Danenberg, 2026) or on the informativeness of the signals themselves (Acemoglu et al., 2016).

We present a fully rational alternative that produces polarization without any of these deviations. Two Bayesian agents update rationally, have standard preferences, and share an identical, correct prior over the complete information structure. The only heterogeneity is over the state itself. The mechanism, which we call *joint inference*, is full Bayesian inference over the state and signal informativeness jointly: each agent places more posterior weight on interpretations of the evidence consistent with her own prior on the state, and these heterogeneous posterior beliefs about signal informativeness propagate back into her posterior over the state. Heterogeneous beliefs about signal informativeness arise as an output of joint inference, not an input.

Consider a political analyst who thinks the Democrat is leading by a few points. Two new polls land: Poll A shows the Democrat ahead, Poll B shows the Republican ahead. Suppose that she doesn't know the specifics of how either poll was conducted and has no reason to trust one over the other. Her prior treats each poll as ex ante equally informative.

In order to use this information to update her belief about the race, Bayes' rule requires that she also assess the informativeness of each poll. Poll A is consistent with her prior: if Poll A is the more informative one, the data are unsurprising. If Poll B is the more informative one, her own read has to be substantially wrong. Her posterior places more weight on Poll A being the more informative poll, and she updates toward a stronger Democratic lead.

A second analyst who started the day believing the Republican was leading runs the symmetric but opposing calculation: Poll B is now the one consistent with her prior, and she updates toward a stronger Republican lead. Both analysts updated rationally and shared the same prior over informativeness; neither exhibited bias nor motivated reasoning. Each analyst's prior on the state shifted her posterior on signal informativeness, which in turn shifted her posterior on the state. This fully rational behavior is precisely the joint inference mechanism.

Two recent papers share the joint inference machinery of agents simultaneously updating about a state and the informativeness of their sources. Gentzkow et al. (2025) use a similar mechanism within a dynamic model of misspecified learning over time, where small biases in agents' reasoning amplify into divergent trust in the limit. Pilgrim et al. (2024) use a similar mechanism within a model of bounded rationality, where an independence approximation over a binary hypothesis generates confirmation bias and polarization. Neither model produces polarization in its rational benchmark; our central theorem accounts for both null results.

This main theoretical result is a complete characterization: polarization is possible under a given signal structure if and only if the marginal likelihood of the state given the data is multimodal at some realization in the support. Three cases rule it out: a single signal, signal informativeness known ex ante, or a binary state space. Each identifies one of the mechanism's three necessary ingredients: at least two signals, at least three states, and uncertainty about signal informativeness. These conditions are present in most realistic information environments, and we provide Normal and Binomial examples that exhibit polarization by direct construction.

We test the mechanism experimentally using a design we call the *Imposter Task*, built on a classical balls-and-urns setup. A subject draws balls from an urn and forms a prior on the urn's

composition. She then observes two external reports: a "partner" report drawn from the same urn, and an "imposter" report drawn from an independent urn. Labels are hidden; the subject jointly infers the composition of the urn and the identity of the partner. A three-stage elicitation separates the inference into prior formation, identity inference, and composition, allowing us to locate the behavioral wedge between subjects' beliefs and fully Bayesian benchmarks at each stage.

In an unincentivized 30-subject pilot, a mixture decomposition between the joint inference Bayesian benchmark and the no joint inference but otherwise Bayesian alternative places subjects 87% of the way toward the joint inference benchmark; we cannot reject full joint inference. A full-scale implementation is planned for Fall 2026.

Section 2 presents the general framework. Section 3 develops the main theorem and surrounding results. Section 4 describes the experimental design. Section 5 reports results. Section 6 concludes.

2 Setup

Two agents $i \in \{a, b\}$ learn about a realized state θ belonging to an ordered set of possible states Θ from a finite set of signals. Agents hold possibly different priors over θ but the same correct joint prior over signal informativeness. Each agent i has prior π_i over θ with mean $\mu_i = \mathbb{E}_{\pi_i}[\theta]$.

There are $n \geq 1$ signals indexed $j = 1, \ldots, n$. Signal j is drawn from a likelihood $L(\cdot \mid \theta, \lambda_j)$ indexed by an *informativeness* parameter $\lambda_j \in \Lambda$. Conditional on θ and $(\lambda_1, \ldots, \lambda_n)$, signals are independent. The informativeness vector $(\lambda_1, \ldots, \lambda_n)$ is drawn from a joint distribution F on Λ^n. Each agent's prior on $(\lambda_1, \ldots, \lambda_n)$ is F, the true data-generating distribution.

Each agent knows their own prior π_i, the distribution F, and the likelihood structure, and observes $(x_1, \ldots, x_n)$ but not $(\lambda_1, \ldots, \lambda_n)$. Agents do not interact; the two-agent framing is a device for comparing the posteriors that two different state priors would produce given the same evidence. Each agent forms a posterior over the state by full Bayesian inference:

$$p_i(\theta, \lambda_1, \ldots, \lambda_n \mid x_1, \ldots, x_n) \propto \pi_i(\theta)\, f(\lambda_1, \ldots, \lambda_n) \prod_{j=1}^{n} L(x_j \mid \theta, \lambda_j).$$

Agent i's posterior over θ, marginalizing out λ, is $\hat{\pi}_i(\cdot \mid x)$, with mean $\hat{\mu}_i \equiv \mathbb{E}_{\hat{\pi}_i}[\theta \mid x_1, \ldots, x_n]$.

Definition 1 (Polarization)**.** *Polarization* occurs *at signal realization* $x = (x_1, \ldots, x_n)$ *given priors* $\pi_a \preceq_{FOSD} \pi_b$ *if*

$$\hat{\pi}_a(\cdot \mid x) \prec_{FOSD} \pi_a \preceq_{FOSD} \pi_b \prec_{FOSD} \hat{\pi}_b(\cdot \mid x),$$

with both posterior shifts strict.

Polarization is possible *under signal structure* (Θ, Λ, L, F) *if there are priors* $\pi_a \preceq_{FOSD} \pi_b$ *and a signal realization* x *in the support of the marginal data distribution at which polarization occurs.*

The first part of Definition 1 is the ex post event: a particular x and a particular pair of priors yield divergent updates. The second part is the ex ante structural property: the signal structure admits some realization x and prior pair (π_a, π_b) at which polarization occurs, so polarization

happens with positive probability under the data-generating distribution. The mean inequality $\hat{\mu}_a < \mu_a \leq \mu_b < \hat{\mu}_b$ follows immediately from strict FOSD shifts.

3 Theoretical results

Let $\tilde{L}(x \mid \theta) = \int \prod_j L(x_j \mid \theta, \lambda_j)\, dF(\lambda)$ denote the marginal likelihood of θ given x under the agent's joint prior over informativeness, where $\lambda = (\lambda_1, \ldots, \lambda_n)$.

3.1 Main result

Definition 2 (Multimodality)**.** *A real-valued function g on Θ is* multimodal *if there exist $\theta_1 < \theta_2 < \theta_3$ in Θ with $g(\theta_1) > g(\theta_2)$ and $g(\theta_3) > g(\theta_2)$. Otherwise g is* unimodal*.*

Theorem 1. *Polarization is possible under signal structure (Θ, Λ, L, F) if and only if there is some x in the support of the marginal data distribution at which $\tilde{L}(x \mid \cdot)$ is multimodal in θ.*

Proof sketch. Fix x and write $L(\theta) := \tilde{L}(x \mid \theta)$.

($\Leftarrow$). If L is multimodal, it has two peaks separated by a valley. Place π_a on a two-point support consisting of the left peak and the valley, and π_b on a two-point support consisting of the valley and the right peak. Each peak has strictly higher likelihood than the valley, so Bayes shifts π_a's mass onto the left peak (a strict downward FOSD shift) and π_b's mass onto the right peak (a strict upward FOSD shift). The two priors share the valley point and are FOSD-ordered to begin with; the posteriors are pulled apart at x.

($\Rightarrow$). Suppose L is unimodal at peak θ^*. Bayes amplifies prior mass in the "above-average" region $\{\theta : L(\theta) \geq Z_\pi\}$ (where Z_π is the prior-weighted average of L) and erodes mass outside it. For π_a to shift downward in FOSD, she must have prior mass in the right tail beyond her above-average region, where Bayes can trim it; symmetrically π_b must have mass in the left tail beyond hers. Combined with $\pi_a \preceq_{\text{FOSD}} \pi_b$, these requirements force agent b's above-average region to lie shifted right of agent a's at both endpoints. But unimodality makes $\{L \geq Z\}$ a contiguous block that widens as Z shrinks, so any two such regions are nested by their normalizers, not shifted past each other. The geometry is inconsistent. The full proof appears in Appendices A.1 and A.2.

□

Empirical work typically operationalizes polarization through mean shifts rather than the full CDF inequalities of Definition 1. We refer to the strictly weaker condition $\hat{\mu}_a < \mu_a \leq \mu_b < \hat{\mu}_b$, with $\mu_i = \mathbb{E}_{\pi_i}[\theta]$ and $\hat{\mu}_i = \mathbb{E}_{\hat{\pi}_i}[\theta \mid x]$, as *mean polarization*. The sufficiency direction of Theorem 1 carries over as an immediate corollary, since strict FOSD shifts imply the corresponding ordering of means. The necessity direction does not: unimodal $\tilde{L}(x \mid \cdot)$ can admit priors with $\hat{\mu}_a < \mu_a \leq \mu_b < \hat{\mu}_b$. We retain the FOSD definition as primary because it is the natural notion for comparing distributions and yields the clean characterization of Theorem 1.

3.2 Impossibility cases

We impose standard regularity on the component likelihood: for each (x, λ), the map $\theta \mapsto L(x \mid \theta, \lambda)$ is unimodal with mode at a λ-invariant summary statistic $s(x)$, and log-concave in θ. Standard parametric families satisfy both.

Proposition 1. *In any of the following cases, $\tilde{L}(x \mid \cdot)$ is unimodal in θ at every x, so polarization is not possible under the signal structure by Theorem 1:*

(i) $n = 1$.

(ii) Signal informativeness is known ex ante: both agents form their posterior using a common fixed value $\hat{\lambda} \in \Lambda^n$.

(iii) $|\Theta| = 2$.

Proof. (i). $\tilde{L}(x \mid \theta) = \int L(x \mid \theta, \lambda)\, dF(\lambda)$ is a non-negative mixture of functions each unimodal in θ with common mode at $s(x)$. The mixture is non-decreasing on $(-\infty, s(x)]$ and non-increasing on $[s(x), \infty)$, hence unimodal at $s(x)$. (Component likelihoods constant in θ trivially satisfy this with any mode.)

(ii). The marginal likelihood collapses to $\prod_j L(x_j \mid \theta, \hat{\lambda}_j)$, a product of log-concave functions, hence log-concave and therefore unimodal in θ.

(iii). A function on a two-point set is unimodal by definition; multimodality requires at least three values. □

Each case removes an individually necessary ingredient: (i) the second signal that lets the marginal likelihood mix two informativeness configurations, (ii) the inference over λ that connects the prior on θ to weights on those configurations, (iii) the third state that allows the marginal likelihood to admit multiple modes. This is why Pilgrim et al. (2024), who study a binary hypothesis, find that the rational Bayesian benchmark in their setting is incompatible with polarization. The signal structure in Gentzkow et al. (2025) does not fit cleanly into any of these three cases, but renders the marginal likelihood unimodal by construction, which is why their unbiased benchmark also produces no polarization.

3.3 Support restrictions

So far we have characterized when polarization is possible in terms of the information environment. We now turn to what this requires of the priors. Proposition 2 states the support restrictions implicit in Theorem 1 and shows that they are quantitatively negligible; full-support priors can achieve mean polarization with arbitrarily small violation of FOSD polarization.

Proposition 2. *Let multimodal $\tilde{L}(x \mid \cdot)$ admit polarization at x. For any prior π, let t_1^π and t_2^π denote the leftmost and rightmost points of $\{\theta : \tilde{L}(x \mid \theta) \geq Z_\pi\}$ with $Z_\pi := \mathbb{E}_\pi[\tilde{L}]$.*

(i) Any priors $\pi_a \preceq_{FOSD} \pi_b$ that polarize at x satisfy

$$supp(\pi_a) \subseteq [t_1^a, \infty), \quad \pi_a((t_2^a, \infty)) > 0, \quad supp(\pi_b) \subseteq (-\infty, t_2^b], \quad \pi_b((-\infty, t_1^b)) > 0.$$

(ii) For any $\varepsilon > 0$, there exist full-support priors $\pi_a^\varepsilon \preceq_{FOSD} \pi_b^\varepsilon$ with $\hat{\mu}_a^\varepsilon < \mu_a^\varepsilon \leq \mu_b^\varepsilon < \hat{\mu}_b^\varepsilon$ and

$$\sup_t [F_{\pi_a^\varepsilon}(t) - F_{\hat{\pi}_a^\varepsilon}(t)]_+ < \varepsilon, \qquad \sup_t [F_{\hat{\pi}_b^\varepsilon}(t) - F_{\pi_b^\varepsilon}(t)]_+ < \varepsilon.$$

Proof sketch. (i) The support inclusion $\text{supp}(\pi_a) \subseteq [t_1^a, \infty)$ rules out mass in agent a's wrong tail, where $\tilde{L} < Z_a$ and Bayes erodes mass in the direction opposite the desired downward FOSD shift. The positive-mass condition $\pi_a((t_2^a, \infty)) > 0$ rules out priors entirely contained in their above-average region, in which case $Z_a = \mathbb{E}_{\pi_a}[\tilde{L}]$ forces $\tilde{L} = Z_a$ a.s. and the posterior equals the prior. The constraints on π_b are symmetric. (ii) Mix any polarizing pair from Theorem 1 with ε weight on a full-support distribution; FOSD violations are $O(\varepsilon)$. The full proof appears in Appendix A.3. □

Together the two cases in (i) rule out unbounded full-support priors (mass arbitrarily far in the wrong tail) and bounded full-support priors when $\tilde{L}$ peaks reach the boundaries (no room beyond t_2^a). The Normal and Binomial examples in Section 3.4 hit each case respectively. In both cases the restrictions are quantitatively negligible. The examples use natural full-support priors, exhibit clear mean polarization, and have FOSD violations on the order of 10^{-10}.

3.4 Possibility examples

Two parametric examples illustrate that the joint inference mechanism is not specific to any particular signal structure. Both use full-support priors and exhibit mean polarization with negligible FOSD violations, as in Proposition 2(ii).

Normal example. Take $\Theta = \mathbb{R}$, $\Lambda = (0, \infty)$ (precision), likelihood $x_j \mid \theta, \lambda_j \sim \mathcal{N}(\theta, \lambda_j^{-1})$, $F = \nu \otimes \nu$ with $\nu = \frac{1}{2}\delta_{0.1} + \frac{1}{2}\delta_{10}$, and signals $(x_1, x_2) = (-2, +2)$. The marginal likelihood is multimodal in θ, with peaks near ± 1.96 corresponding to the configurations in which exactly one signal is informative. Take $\pi_a \sim \mathcal{N}(-1.1, 0.16)$ and $\pi_b \sim \mathcal{N}(1.1, 0.16)$, both with full support on $\mathbb{R}$. Direct computation yields posterior means $\hat{\mu}_a = -1.37$ and $\hat{\mu}_b = +1.37$ against prior means $\mu_a = -1.10$ and $\mu_b = +1.10$; the FOSD violation is on the order of 10^{-10} (Figure 1(a)).

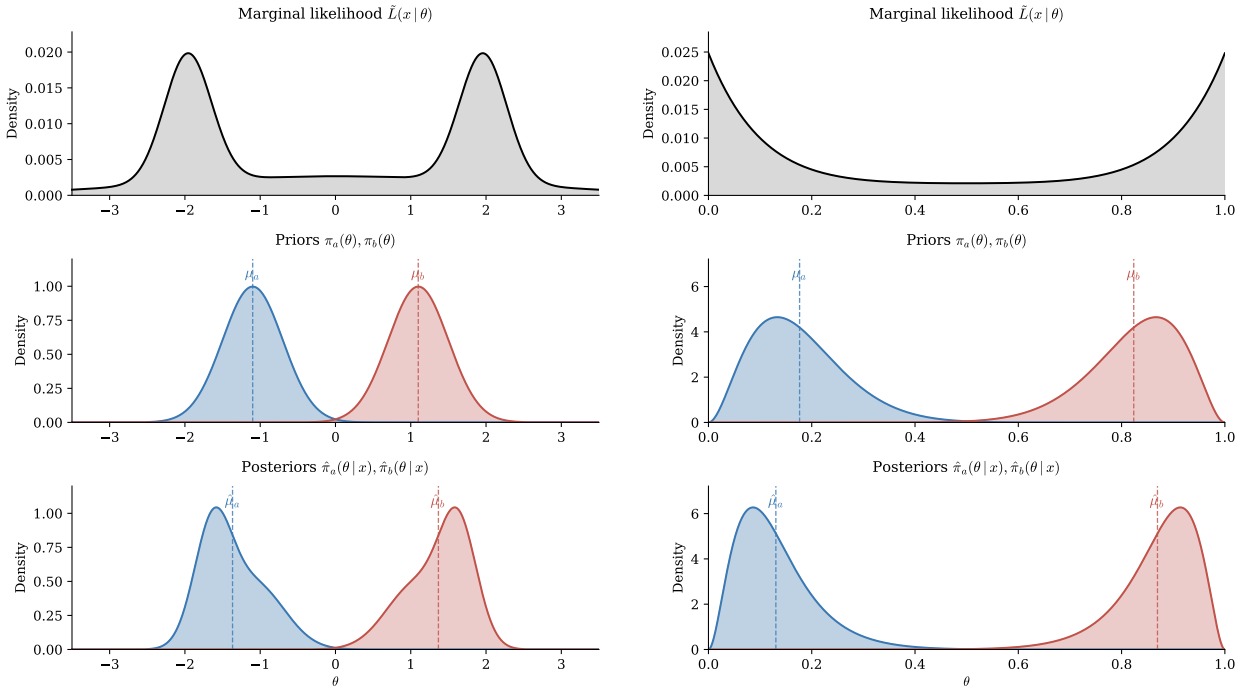
Binomial example. Take $\Theta = [0, 1]$, $\Lambda = [0, 1]$, likelihood the mixture

$$x_j \mid \theta, \lambda_j \sim \lambda_j \cdot \text{Binomial}(m, \theta) + (1 - \lambda_j) \cdot \text{BetaBinomial}(m, 1, 1),$$

$m = 10$, $F = \nu \otimes \nu$ with $\nu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$, and signals $(x_1, x_2) = (0, 10)$. Each λ_j is binary, corresponding to signal j being fully informative ($\lambda_j = 1$) or pure noise ($\lambda_j = 0$). The posterior over (λ_1, λ_2) for each agent is:

(λ_1, λ_2)	$\Pr_a(\cdot \mid x_1, x_2)$	$\Pr_b(\cdot \mid x_1, x_2)$
$(1, 1)$	0.000	0.000
$(1, 0)$	0.703	0.000
$(0, 1)$	0.000	0.703
$(0, 0)$	0.297	0.297

Both agents rule out $(1, 1)$ as incompatible with the asymmetric data and put almost no weight on the wrong-side interpretation. With priors $\pi_a \sim \text{Beta}(3, 14)$ and $\pi_b \sim \text{Beta}(14, 3)$, both with full support on $(0, 1)$, posterior means are $\hat{\mu}_a \approx 0.131$ and $\hat{\mu}_b \approx 0.869$, against prior means $\mu_a \approx 0.176$ and $\mu_b \approx 0.824$. The mean inequality $\hat{\mu}_a < \mu_a \leq \mu_b < \hat{\mu}_b$ holds; the FOSD violation is on the order of 10^{-10}. See Figure 1(b).



(a) Normal: full-support Normal priors.

(b) Binomial: full-support Beta priors.

Figure 1: Two parametric examples. Each panel shows the marginal likelihood (top), priors (middle), and posteriors (bottom). Both panels use full-support priors and exhibit negligible FOSD violations illustrating Proposition 2(ii).

3.5 Asymptotic learning

Proposition 3. *Suppose each signal is iid from $\tilde{L}(\cdot \mid \theta_0)$, θ_0 lies in the Kullback–Leibler support of π_a and π_b, and the family $\{\tilde{L}(\cdot \mid \theta) : \theta \in \Theta\}$ admits uniformly consistent tests of θ_0 against $\{\theta : |\theta - \theta_0| > \varepsilon\}$ for every $\varepsilon > 0$. Then*

$$\hat{\mu}_a^{(n)}, \hat{\mu}_b^{(n)} \to \theta_0 \quad \textit{a.s. under } \theta_0.$$

Proof. By the posterior consistency theorem of Schwartz (1965), each agent's posterior on θ converges weakly to δ_{θ_0} a.s. under θ_0. For bounded Θ, weak convergence gives $\hat{\mu}_i^{(n)} \to \theta_0$ a.s. directly; for unbounded Θ, the same conclusion holds whenever the posterior sequence is uniformly integrable, which is satisfied in the parametric examples of Section 3.4. □

Polarization under joint inference is a finite-signal phenomenon. This separates the mechanism from existing accounts of polarization. Models with biases, bounded rationality, or belief-based utility sustain polarization by blocking convergence to the truth. Models with auxiliary heterogeneity admit polarization only when convergence on the auxiliary dimensions is itself blocked. Joint inference produces polarization without blocking convergence: with shared correct priors over informativeness, posteriors over λ collapse onto the truth as signals accumulate, and disagreement on θ vanishes with them. Disagreement remains a finite-signal feature of correct Bayesian updating rather than a permanent deviation from it.

4 Experimental design

We implement the Binomial setup in an experimental design we call the Imposter Task, built on a classical balls-and-urns framework. Figure 2 summarizes the data-generating process.

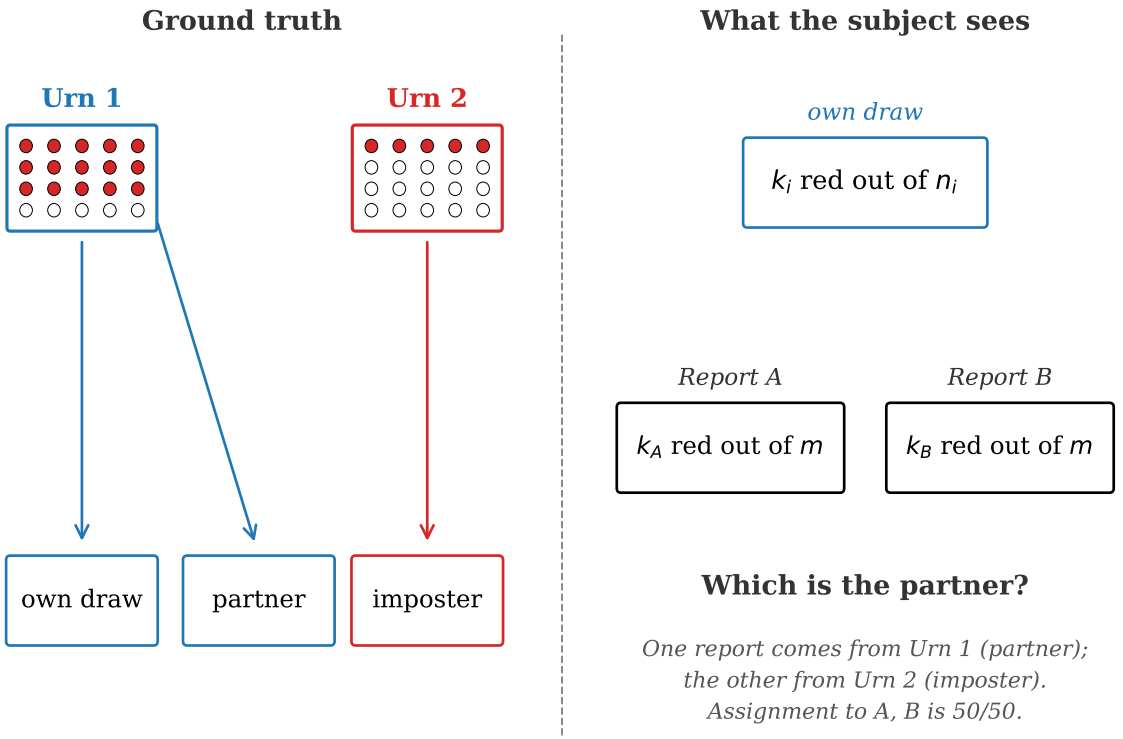


Figure 2: Imposter Task: data-generating process. The subject's own draw and one of the two external reports come from Urn 1 (the partner); the other report comes from Urn 2 (the imposter). Reports are displayed as A and B in random order with labels hidden.

Each trial begins with two urns containing r_1, r_2 red balls out of 100, drawn independently from a uniform prior on $\{1, \ldots, 99\}$. The state of interest is r_1 (equivalently, the urn-1 red proportion $\theta = r_1/100$). The subject's own draw of $n_i \in \{10, 50\}$ balls from Urn 1, observing k_i red, endows her with a prior π_i over the state, with Bayesian benchmark Beta$(1 + k_i, 1 + n_i - k_i)$ under an initial Beta$(1, 1)$. Subjects with different realized own draws form different priors, which is exactly the heterogeneity the model in Section 2 requires.

The subject then observes two external reports, each a count of red balls from $m = 10$ draws with replacement: one from Urn 1 (the partner) and one from Urn 2 (the imposter). These two reports are the signals x_1, x_2 in the framework of Section 2, with $\lambda_j = 1$ for the partner report (informative) and $\lambda_j = 0$ for the imposter report (noise). Reports are labeled A and B in random order; the subject does not observe which is which, so F is supported on $\{(1, 0), (0, 1)\}$ with equal probability: exactly one report is informative about the state.

The non-iid F used here differs from the iid version in Section 3.4, yet remains non-degenerate. It simplifies the subject's task: she infers a single binary parameter ("which report is the partner?") rather than four (λ_1, λ_2) configurations, and her belief over report informativeness admits a one-dimensional elicitation. The marginal likelihood remains multimodal in θ, so polarization is achievable by Theorem 1. Direct computation under the parameters of the Binomial example gives posterior means of 0.111 and 0.889, slightly sharper than the iid case because "both signals uninformative" is ruled out.

The subject's inference is elicited in three stages within each trial:

Q1. *Prior formation.* After the own draw, the subject estimates r_1.

Q2. *Identity inference.* After seeing both reports, the subject indicates the probability that report A came from Urn 1 (the partner report), via a 21-row price-list-style choice table (switching point yields an interval for the indicated probability).

Q3. *Composition.* After seeing both reports, the subject re-estimates r_1.

One of the three elicitations is selected at random for payment on each trial. Q1 and Q3 are incentivized by quadratic loss against the realized r_1; Q2 is incentivized by the choice table's realized payoff.

We compute stage-wise Bayesian benchmarks using the Beta-Binomial approximation to the discrete-uniform prior on r_1. Under Beta$(1, 1)$ (uniform on $[0, 1]$), the Bayesian posterior on $r_1/100$ given (k_i, n_i) is Beta$(1 + k_i, 1 + n_i - k_i)$, with mean $Q1^*(k_i, n_i) = (1 + k_i)/(2 + n_i)$. The Bayesian identity posterior $Q2^*(Q1, k_a, k_b)$ given the subject's stated Q1 is computed from Beta-Binomial marginal likelihoods; the Bayesian composition $Q3^*(Q1, Q2, k_a, k_b)$ is the Q2-weighted average of posterior means under each identity assignment.[1]

[1]The approximation to the exact discrete prior is tight: maximum error 0.005 for $Q1^*$ and 0.014 for $Q2^*$, with median error essentially zero across the design grid; the errors concentrate at extreme own draws ($k_i \in \{0, n_i\}$) and are negligible relative to the reported regression estimates.

The central test is a mixture decomposition at Stage 3 that directly asks whether subjects engage in joint inference. Let $Q3^{\text{noJI}}$ denote the Bayesian composition that does not update beliefs about identity from the realizations and treats both reports as equally likely to be the partner ($Q2 = 1/2$); let $Q3^{\text{JI}}$ denote the full joint inference Bayesian composition under the realization-contingent $Q2^*$. Both are computed using the subject's stated Q1. The mixture regression

$$Q3_t = \alpha \cdot Q3_t^{\text{JI}} + (1 - \alpha) \cdot Q3_t^{\text{noJI}} + \varepsilon_t$$

measures the relative weighting between the two Bayesian benchmarks: $\alpha = 1$ corresponds to the full joint inference Bayesian response, while $\alpha = 0$ corresponds to an otherwise fully Bayesian agent who either does not update her belief over the informativeness of the reports from their realizations or does not allow this updated belief to propagate into her updated (Q3) belief about the state. Both of these $\alpha = 0$ interpretations result in the same no joint inference benchmark.

Three further regressions isolate the wedge between subject behavior and the Bayesian benchmark at each stage. Each regresses the subject's update at that stage on the Bayesian update, both centered at the relevant prior; the coefficient measures the share of the Bayesian update the subject implements, with 0 corresponding to no update and 1 to the full Bayesian update.

At Stage 1, variables are centered at 1/2 (the uniform prior mean):

$$Q1_t - \tfrac{1}{2} = \lambda_0 + \lambda_1 \left(Q1_t^* - \tfrac{1}{2}\right) + \varepsilon_t.$$

At Stage 2, variables are centered at 1/2 (the prior identity belief under uniform random labeling):

$$Q2_t - \tfrac{1}{2} = \gamma_0 + \gamma_1 \left(Q2_t^* - \tfrac{1}{2}\right) + \varepsilon_t.$$

$Q2^*$ is computed using the subject's stated Q1, so γ_1 isolates identity-inference error from prior-formation error. At Stage 3, the outcome is centered at the subject's stated Q1:

$$Q3_t - Q1_t = \beta_0 + \beta_1 \left(Q3_t^* - Q1_t\right) + \eta_t,$$

with $Q3^*$ computed using the subject's stated Q1 and Q2. The coefficient β_1 isolates composition error from identity and prior errors.

The pilot reported in Section 5 was conducted in a single classroom session with 30 subjects and 8 trials each (240 trials total). The pilot was unincentivized, though the scoring rule above was applied for diagnostic purposes; under that scoring subjects would have earned an average of \$5.88 each for approximately 20 minutes of work (median \$6.22, range \$2.63 to \$7.81). The pilot identifies the mechanism components at the individual trial level: stage-wise updating and the joint inference mixture. A direct test of paired divergence would utilize matched (k_a, k_b) across subjects with dispersed Q1, so that the lower-Q1 subject should update down past her Q1 while the higher-Q1 subject updates up past hers. This is a consideration for the full experiment.

5 Empirical results

5.1 Behavioral decomposition

The central empirical question is whether subjects engage in joint inference of the type we describe. The mixture regression defined in Section 4 contrasts subject behavior against the joint inference Bayesian benchmark ($\alpha = 1$) and the no joint inference alternative ($\alpha = 0$).

The pooled estimate is $\hat{\alpha} = 0.87$ (SE 0.14, $p < 10^{-9}$ against zero; we cannot reject $\alpha = 1$); see Table 1. The estimate is stable across the two prior-strength conditions: $\hat{\alpha} = 0.87$ (SE 0.08) at $n_i = 10$ and $\hat{\alpha} = 0.82$ (SE 0.18) at $n_i = 50$.

Sample	$\hat{\alpha}$	SE	p_0	R^2
Pooled	0.87	0.14	$< 10^{-9}$	0.39
$n_i = 10$	0.87	0.08	$< 10^{-7}$	0.50
$n_i = 50$	0.82	0.18	$< 10^{-5}$	0.14

Table 1: Joint inference mixture at Stage 3. $\hat{\alpha}$ measures the relative weighting between the joint inference Bayesian benchmark ($\alpha = 1$) and the no updating on signal informativeness alternative ($\alpha = 0$). p_0 is the p-value for $\alpha = 0$. Standard errors clustered at the subject level.

Figure 3 reports the per-subject distribution of $\hat{\alpha}$ across the full 30-subject sample. The distribution concentrates near one with a long left tail; a small number of non-responsive subjects sit near zero. The median $\hat{\alpha}$ is 1.04, and 26 of 30 subjects (87%) have $\hat{\alpha} > 0.5$.

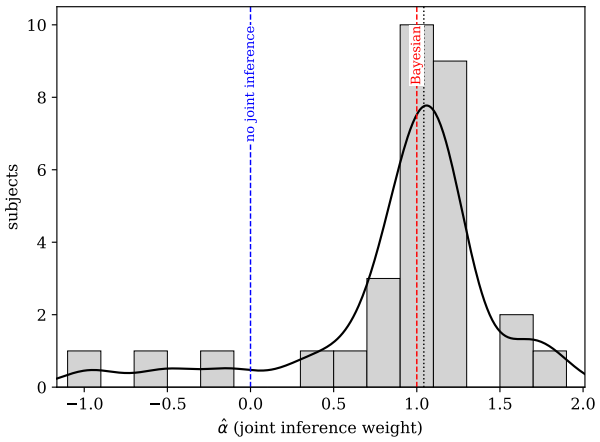


Figure 3: Per-subject distribution of $\hat{\alpha}$. $\hat{\alpha} = 1$ matches the full joint inference Bayesian; $\hat{\alpha} = 0$ matches the no-updating alternative. Bin size = .20. The black dotted line marks the median.

5.2 Stage-wise decomposition

The mixture result establishes that subjects engage in behavior consistent with joint inference overall. The stage-wise regressions defined in Section 4, with results in Table 2, localize where this inference is fully Bayesian and where it is attenuated.

Stage 1 fits essentially as fully Bayesian: $\hat{\lambda}_1 = 1.00$ (SE 0.03, $p < 10^{-100}$ against zero; we cannot reject the Bayesian null $\lambda_1 = 1$, $p = 0.91$) with $R^2 = 0.90$. Subjects form their prior over the state from their own draw in line with the Bayesian prediction.

Stage 2 shows joint inference operating but attenuated: $\hat{\gamma}_1 = 0.66$ (SE 0.06, $p < 10^{-28}$ against zero; we reject the Bayesian null $\gamma_1 = 1$ at $p < 10^{-7}$) with $R^2 = 0.56$. Subjects update identity beliefs from the realizations in the Bayesian direction, but the magnitude of the update is attenuated relative to the Bayesian benchmark.

Stage 3 is noisier but consistent with the same pattern: $\hat{\beta}_1 = 0.79$ (SE 0.14, $p < 10^{-7}$ against zero; we cannot reject $\beta_1 = 1$, $p = 0.13$), $R^2 = 0.40$. The point estimate suggests attenuation in the same direction as Stage 2, though the data are not sufficient to reject the Bayesian null.

The pattern across stages is that the Stage 1 prior formation is Bayesian while Stages 2 and 3 show joint inference operating in the predicted direction with attenuation toward the prior. This is consistent with the conservatism in belief updating documented in the existing literature.

Stage	Coefficient	SE	p_0	R^2
1. Prior formation ($\hat{\lambda}_1$)	1.00	0.03	$< 10^{-100}$	0.90
2. Identity inference ($\hat{\gamma}_1$)	0.66	0.06	$< 10^{-28}$	0.56
3. Composition ($\hat{\beta}_1$)	0.79	0.14	$< 10^{-7}$	0.40

Table 2: Stage-wise decomposition. Each row reports one regression from Section 4. p_0 is the p-value against the null that the coefficient is zero. Standard errors clustered at the subject level.

5.3 Discussion

The pilot's central finding is that subjects engage in joint inference. The pooled mixture estimate $\hat{\alpha} = 0.87$ is statistically indistinguishable from full joint inference and clearly distinguishable from the no joint inference alternative. Without instruction in Bayesian inference, subjects use their prior beliefs about the state to infer which signal is more likely to be informative and feed that inference back into their state estimate. The mechanism may sound elaborate to describe, but subjects appear to utilize it without difficulty.

Free-text responses to an optional post-task prompt — "What strategy did you use to solve these problems?" — reinforce this picture. Of the 21 subjects who responded, 13 describe a procedure consistent with joint inference. One subject wrote: "I always assumed the number of red balls in the sample was proportional to the random sample I had drawn. I then assumed that my partner's report that was more similar to my initial drawn sample was the correct report. I then basically assumed the true number of red balls was somewhere between what I drew in my original sample and what I assumed to be my partner's correct report."

Another, more compactly: "I tended to trust my first sample to predict the quantity (by multiplying by 10 or 2). Then, I would choose the partner who's sample was closest to mine, and change my final guess (slightly) either higher or lower depending on the closest partners sample." These responses are describing the joint inference procedure in plain English. The remaining 8

responses don't describe alternative procedures; they are simply too brief or vague to classify (one subject wrote "Math.").

The stage-wise pattern is consistent with the same picture, with one wrinkle: subjects perform the joint inference in the predicted direction at every stage, but the magnitudes at Stages 2 and 3 are attenuated relative to the Bayesian benchmark. A subject who is Bayesian at Stage 3 given her stated Q2 ($\beta_1 = 1$) would, by linearity of $Q3^*$ in $Q2$, produce an implied $\hat{\alpha}$ equal to $\hat{\gamma}_1 = 0.66$. The direct estimate $\hat{\alpha} = 0.87$ is higher, so subjects' Stage 3 responses load more on joint inference than their stated Q2 alone predicts. The gap localizes the attenuation to the Stage 2 elicitation rather than to subjects' internal inference.

The Stage 2 elicitation grid has 5% resolution, truncating stated probabilities into $[0.025, 0.975]$ as we take the mean of the identified interval as the subject's point estimate. At $n_i = 50$, the Bayesian $Q2^*$ lies outside this range on 34% of trials (versus 20% at $n_i = 10$); on those trials the subject cannot express the implied Bayesian belief even if she holds it, and the attenuation gap is correspondingly wider in the cell where grid saturation bites harder ($\hat{\gamma}_1 = 0.61$ at $n_i = 50$ versus 0.72 at $n_i = 10$). Finer Q2 resolution is a design focus for the full-scale experiment.

The central empirical claim that subjects engage in joint inference holds independently of the stage-wise measurement concerns. Both $Q3^{\text{JI}}$ and $Q3^{\text{noJI}}$ are Bayesian benchmarks computed without conditioning on the subject's stated Q2, so the direct mixture sidesteps the grid issue entirely. Q1 measurement error does enter the mixture decomposition, but is negligible given that Stage 1 is essentially Bayesian. It also enters both benchmarks identically, so any residual noise shifts the two reference points together rather than driving the relative weighting between them.

6 Conclusion

The polarization literature has interpreted divergent updating from common evidence as a signature of irrationality. We have shown that the same pattern arises from full Bayesian inference whenever the marginal likelihood is multimodal at some signal realization, a condition requiring multiple signals of uncertain informativeness and more than two possible states. These key ingredients are native to most realistic information environments. States of interest are rarely binary, evidence usually arrives piecemeal, and the informativeness of each piece is almost never known with certainty ex ante. In a setting meeting these requirements but stripped of emotional content, we find that subjects engage in joint inference.

Revisited in our terms, the canonical study of Lord et al. (1979) has all three ingredients. The state (the magnitude of the deterrent effect) is elicited on a scale from -8 to $+8$. There are 2 signals (the two studies), and the informativeness of each unlabeled study is uncertain and ex ante identical. Subjects performing rational joint inference over the deterrent effect and the informativeness of the studies should rate the study consistent with their prior as more informative, with this asymmetric weighting propagating back into their beliefs about the state. We should not be surprised that these subjects came away from the experiment with more extreme versions of their initial views.

A Proof of Theorem 1

Theorem 1. *Polarization is possible under signal structure* (Θ, Λ, L, F) *if and only if there is some* x *in the support of the marginal data distribution at which* $\tilde{L}(x \mid \cdot)$ *is multimodal in* θ*.*

We prove the pointwise version at a fixed x in each direction; the existential statement follows by quantifying over x in the support of the marginal data distribution.

A.1 Sufficiency

We construct polarizing priors at any x where $\tilde{L}(x \mid \cdot)$ is multimodal. Fix such an x and write $L(\theta) := \tilde{L}(x \mid \theta)$. The construction proceeds in four steps.

Step 1: Pick three points and define priors

By Definition 2, L admits points $\theta_1 < \theta_2 < \theta_3$ with

$$L(\theta_1) > L(\theta_2) \qquad \text{and} \qquad L(\theta_3) > L(\theta_2).$$

Define two-point priors that exploit the valley at θ_2:

$$\pi_a = w\,\delta_{\theta_1} + (1-w)\,\delta_{\theta_2}, \qquad \pi_b = s\,\delta_{\theta_2} + (1-s)\,\delta_{\theta_3},$$

for any $w, s \in (0,1)$. Each prior splits mass between a high-likelihood point and the valley θ_2, leaving room for Bayes to redistribute (Figure 4).

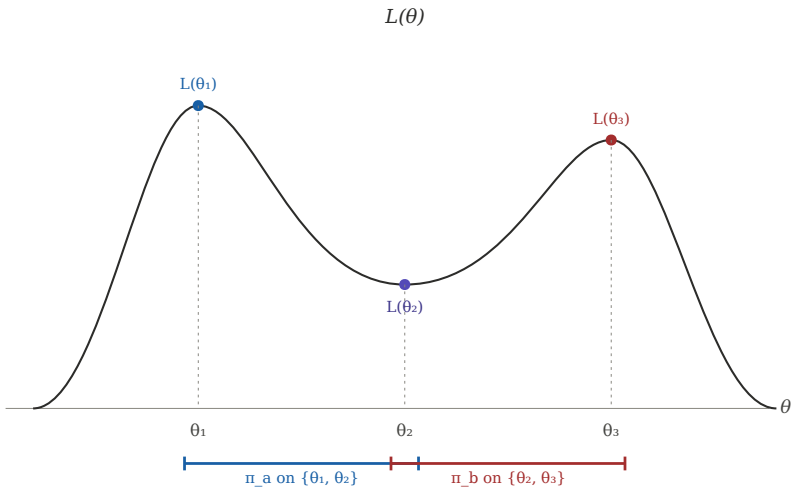


Figure 4: Sufficiency setup. Three points $\theta_1 < \theta_2 < \theta_3$ with $L(\theta_2)$ strictly below both $L(\theta_1)$ and $L(\theta_3)$. Prior π_a is supported on $\{\theta_1, \theta_2\}$, prior π_b on $\{\theta_2, \theta_3\}$.

Step 2: $\pi_a \preceq_{\mathbf{FOSD}} \pi_b$

Compare CDFs at each of the three support points. At θ_1: $F_{\pi_a}(\theta_1) = w \geq 0 = F_{\pi_b}(\theta_1)$. At θ_2: $F_{\pi_a}(\theta_2) = 1 \geq s = F_{\pi_b}(\theta_2)$. At θ_3: $F_{\pi_a}(\theta_3) = 1 = F_{\pi_b}(\theta_3)$. Between the support points the CDFs are constant, and outside the joint support they agree at 0 or 1. So $F_{\pi_a}(t) \geq F_{\pi_b}(t)$ for all t, hence $\pi_a \preceq_{\text{FOSD}} \pi_b$.

Step 3: Bayes pulls the posteriors apart

Apply Bayes to each prior. For agent a, the posterior mass on θ_1 is

$$\hat{\pi}_a(\theta_1) = \frac{w\, L(\theta_1)}{w\, L(\theta_1) + (1-w)\, L(\theta_2)} = \frac{w}{w + (1-w)\, L(\theta_2)/L(\theta_1)} > w,$$

where the second equality divides by $L(\theta_1)$ and the inequality uses $L(\theta_2)/L(\theta_1) < 1$. So $F_{\hat{\pi}_a}(\theta_1) > F_{\pi_a}(\theta_1)$. Since π_a and $\hat{\pi}_a$ share the support $\{\theta_1, \theta_2\}$, both CDFs equal 0 for $t < \theta_1$ and 1 for $t \geq \theta_2$, so the strict inequality at θ_1 extends to $F_{\hat{\pi}_a} \geq F_{\pi_a}$ everywhere with strict inequality on $[\theta_1, \theta_2)$, hence $\hat{\pi}_a \prec_{\text{FOSD}} \pi_a$.

For agent b, by the symmetric argument,

$$\hat{\pi}_b(\theta_3) = \frac{(1-s)\, L(\theta_3)}{s\, L(\theta_2) + (1-s)\, L(\theta_3)} > 1 - s,$$

so $F_{\hat{\pi}_b}(\theta_2) < F_{\pi_b}(\theta_2)$, and the same support argument gives $\pi_b \prec_{\text{FOSD}} \hat{\pi}_b$.

Step 4: Polarization

Stacking the four FOSD relations:

$$\hat{\pi}_a \prec_{\text{FOSD}} \pi_a \preceq_{\text{FOSD}} \pi_b \prec_{\text{FOSD}} \hat{\pi}_b.$$

This is precisely Definition 1: the constructed priors polarize at x (Figure 5). □

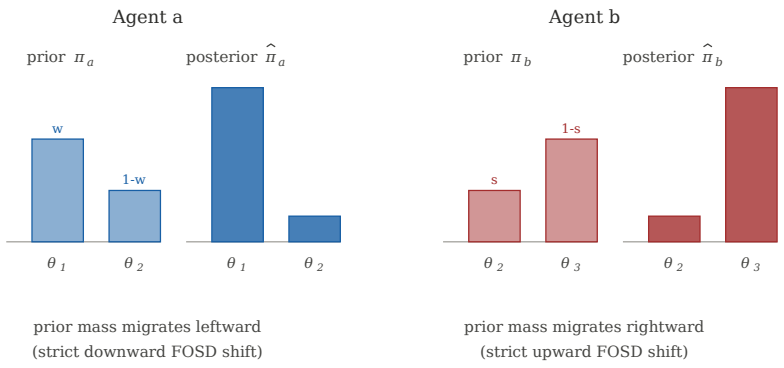


Figure 5: Sufficiency: prior-to-posterior shifts. Bayes pulls π_a's mass onto θ_1 (strict downward FOSD shift) and π_b's mass onto θ_3 (strict upward FOSD shift).

A.2 Necessity

Conversely, fix x and suppose $L(\theta) := \tilde{L}(x \mid \theta)$ is unimodal with peak θ^*. We show no priors π_a, π_b polarize at x.

Under unimodality, the level set $\{L \geq Z\}$ is a contiguous block of Θ (an interval when Θ is continuous), so the argument below covers both continuous and discrete Θ.

Geometric setup

For each prior π, the normalizer $Z_\pi = \mathbb{E}_\pi[L]$ is the prior's average likelihood. Define the *above-average region*

$$[t_1^\pi, t_2^\pi] = \{\theta : L(\theta) \geq Z_\pi\}.$$

Under unimodality, this is a single interval containing the peak θ^*. Outside this interval, $L < Z_\pi$, so the Bayes ratio $L(\theta)/Z_\pi < 1$ and prior mass there gets eroded. Inside, $L \geq Z_\pi$, and prior mass gets amplified.

Two agents, two different normalizers Z_a and Z_b, two different above-average regions $[t_1^a, t_2^a]$ and $[t_1^b, t_2^b]$. The picture is one bell with two horizontal slices, defining four endpoints on the θ axis (Figure 6).

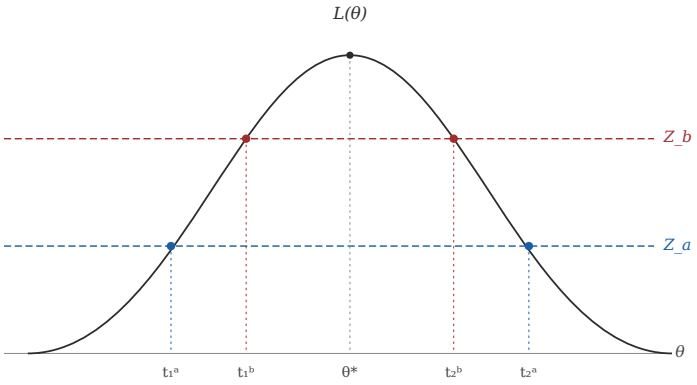


Figure 6: Setup: a unimodal L peaked at θ^*. Each prior π defines its own above-average region $[t_1^\pi, t_2^\pi]$ via the horizontal slice at height Z_π. Under unimodality, smaller Z gives a wider interval containing the larger one.

Suppose for contradiction that priors $\pi_a \preceq_{\text{FOSD}} \pi_b$ polarize at x — that is, $\hat{\pi}_a \prec_{\text{FOSD}} \pi_a$ and $\pi_b \prec_{\text{FOSD}} \hat{\pi}_b$ are both strict. We constrain where each prior's mass can sit relative to the four endpoints, then show the constraints are inconsistent with unimodality. The argument proceeds in six steps.

Step 1: π_a has no mass below t_1^a

A downward FOSD shift $\hat{\pi}_a \prec_{\text{FOSD}} \pi_a$ means the posterior CDF lies above the prior CDF everywhere: $F_{\hat{\pi}_a}(t) \geq F_{\pi_a}(t)$ for all t, with strict inequality somewhere.

Evaluate this difference at $t = t_1^a$:

$$F_{\hat{\pi}_a}(t_1^a) - F_{\pi_a}(t_1^a) = \int_{-\infty}^{t_1^a} \pi_a(d\theta) \left[\frac{L(\theta)}{Z_a} - 1 \right].$$

Below t_1^a, by definition of the above-average region, $L(\theta) < Z_a$, so the integrand is negative everywhere π_a has mass. If π_a has any mass below t_1^a, this integral is strictly negative, contradicting $F_{\hat{\pi}_a}(t_1^a) - F_{\pi_a}(t_1^a) \geq 0$. So π_a has no mass below t_1^a.

The intuition: a prior shifting down cannot have mass to the left of its above-average region, because that mass sits in a low-likelihood zone where Bayes erodes it, pushing the posterior CDF down — the wrong direction for a downward FOSD shift (Figure 7).

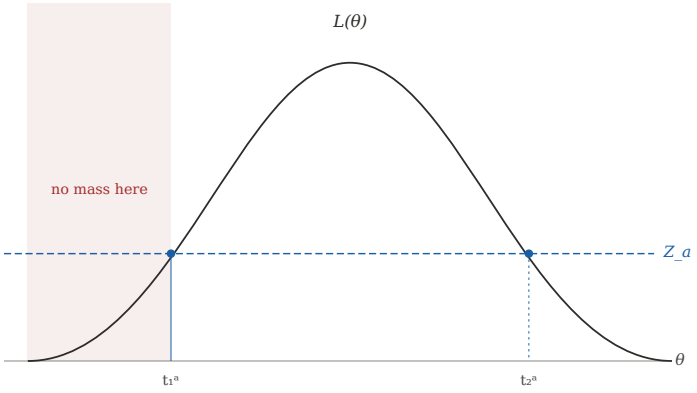


Figure 7: Step 1: π_a has no mass below t_1^a.

Step 2: π_a has positive mass strictly above t_2^a

Step 1 established $\text{supp}(\pi_a) \subseteq [t_1^a, \infty)$. Suppose for contradiction that all of π_a's mass sits inside $[t_1^a, t_2^a]$. On this interval, $L(\theta) \geq Z_a$ everywhere π_a has mass.

But $Z_a = \mathbb{E}_{\pi_a}[L]$ is the π_a-mean of L, and $L \geq Z_a$ on $\text{supp}(\pi_a)$ forces $L = Z_a$ π_a-a.s.

Then the Bayes ratio $L/Z_a = 1$ everywhere on the support, the posterior equals the prior, and there is no shift at all. This contradicts the strict downward shift.

So π_a must have positive mass outside $[t_1^a, t_2^a]$. Combined with Step 1, the only direction available is rightward, strictly above t_2^a (Figure 8).

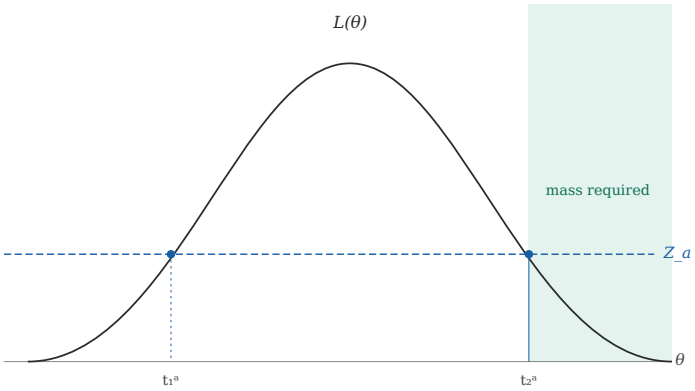


Figure 8: Step 2: π_a has positive mass strictly above t_2^a.

A prior shifting *down* in FOSD must have its support extending *up* past the high-likelihood region, with mass there that Bayes erodes back toward the peak. The downward FOSD shift comes from trimming the right tail.

Step 3: Symmetric statements for π_b

Mirror Steps 1 and 2 with the signs flipped. Agent b's posterior shifts strictly up: $\pi_b \prec_{\text{FOSD}} \hat{\pi}_b$.

By the mirror of Step 1, π_b has no mass above t_2^b: high-side mass would get eroded by Bayes, pushing the posterior CDF *up* relative to the prior CDF, the wrong direction for an upward FOSD shift. Hence $\text{supp}(\pi_b) \subseteq (-\infty, t_2^b]$.

By the mirror of Step 2, π_b has positive mass strictly below t_1^b. Otherwise all mass would sit in $[t_1^b, t_2^b]$, forcing $L = Z_b$ on the support and yielding no shift.

The picture is the symmetric one (Figure 9): π_b's support sits in $(-\infty, t_2^b]$, with mass spilling out below t_1^b. The upward FOSD shift comes from Bayes trimming the left tail.

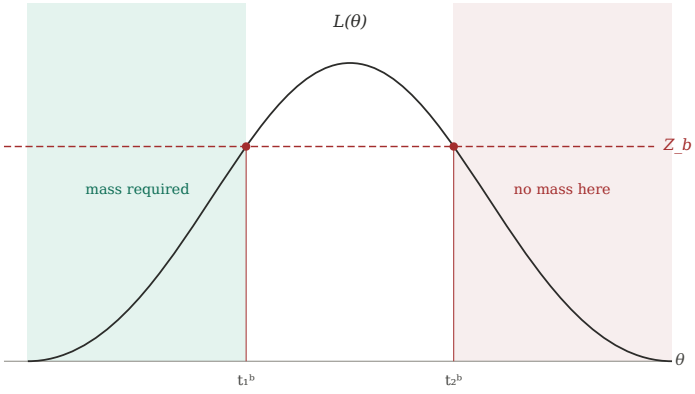


Figure 9: Step 3: symmetric constraints on π_b.

Step 4: FOSD couples the two supports

Now we bring in the FOSD relationship $\pi_a \preceq_{\text{FOSD}} \pi_b$, which says $F_{\pi_a}(t) \geq F_{\pi_b}(t)$ for all t.

Apply this just below t_1^a. By Step 1, $F_{\pi_a}(t_1^a - \varepsilon) = 0$ for any $\varepsilon > 0$. FOSD gives $F_{\pi_b}(t_1^a - \varepsilon) \leq F_{\pi_a}(t_1^a - \varepsilon) = 0$. Since CDFs are nonnegative, $F_{\pi_b}(t_1^a - \varepsilon) = 0$. So π_b has no mass below $t_1^a - \varepsilon$. Letting $\varepsilon \to 0$: $\text{supp}(\pi_b) \subseteq [t_1^a, \infty)$.

The symmetric argument applied at t_2^b gives $\text{supp}(\pi_a) \subseteq (-\infty, t_2^b]$.

The intuition: FOSD says π_a is stochastically below π_b. So π_b's leftmost mass cannot lie left of π_a's leftmost mass, and π_a's rightmost mass cannot lie right of π_b's rightmost mass. Both supports get squeezed into $[t_1^a, t_2^b]$ (Figure 10).

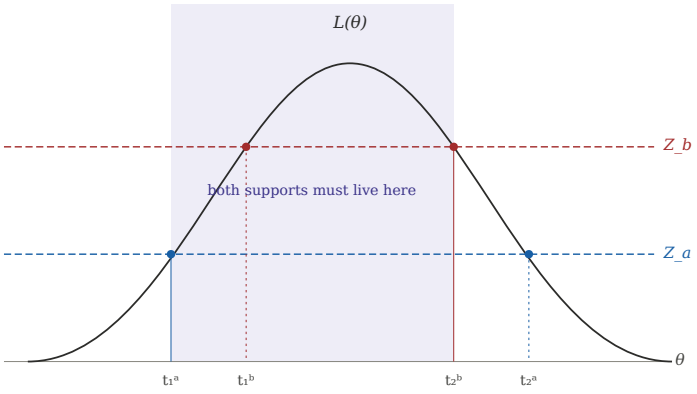


Figure 10: Step 4: FOSD couples the two supports. Both must lie within $[t_1^a, t_2^b]$.

Step 5: Endpoint orderings

Combine the support constraints to extract orderings on the four endpoints.

First. By Step 3, π_b has positive mass below t_1^b. By Step 4, $\text{supp}(\pi_b) \subseteq [t_1^a, \infty)$. So π_b has positive mass on $[t_1^a, t_1^b)$, which requires

$$t_1^a < t_1^b.$$

Second. By Step 2, π_a has positive mass above t_2^a. By Step 4, $\text{supp}(\pi_a) \subseteq (-\infty, t_2^b]$. So π_a has positive mass on $(t_2^a, t_2^b]$, which requires

$$t_2^a < t_2^b.$$

Both endpoints of b's above-average interval lie strictly to the right of the corresponding endpoints of a's above-average interval. The two intervals are not nested but *shifted past each other*.

Step 6: Unimodality forces nesting; contradiction

Under unimodality, the two above-average regions $[t_1^a, t_2^a]$ and $[t_1^b, t_2^b]$ are nested: the one with smaller Z contains the one with larger Z. Step 5 says b's region is shifted right of a's at both

endpoints, which is incompatible with nesting. We make this precise by reading off L at the level-set boundaries.

From the left endpoints. $t_1^a \in \{L \geq Z_a\}$ gives $L(t_1^a) \geq Z_a$. And $t_1^a < t_1^b = \inf\{L \geq Z_b\}$ gives $t_1^a \notin \{L \geq Z_b\}$, hence $L(t_1^a) < Z_b$. Combining:

$$Z_a \leq L(t_1^a) < Z_b, \qquad \text{so} \qquad Z_a < Z_b.$$

From the right endpoints. The mirror argument: $t_2^b \in \{L \geq Z_b\}$ gives $L(t_2^b) \geq Z_b$, and $t_2^b > t_2^a = \sup\{L \geq Z_a\}$ gives $L(t_2^b) < Z_a$. So

$$Z_b \leq L(t_2^b) < Z_a, \qquad \text{so} \qquad Z_b < Z_a.$$

But $Z_a < Z_b$ and $Z_b < Z_a$ cannot both hold. Contradiction (Figure 11). □

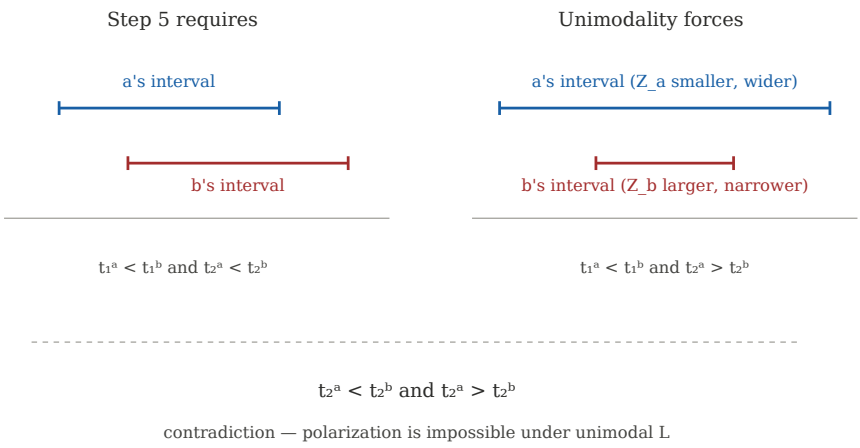


Figure 11: Step 6: Step 5's required ordering ($t_2^a < t_2^b$) contradicts what unimodality forces ($t_2^a > t_2^b$). Polarization is impossible under unimodal L.

A.3 Proof of Proposition 2

Proposition 2. *Let multimodal $\tilde{L}(x \mid \cdot)$ admit polarization at x. For any prior π, let t_1^π and t_2^π denote the leftmost and rightmost points of $\{\theta : \tilde{L}(x \mid \theta) \geq Z_\pi\}$ with $Z_\pi := \mathbb{E}_\pi[\tilde{L}]$.*

(i) *Any priors $\pi_a \preceq_{FOSD} \pi_b$ that polarize at x satisfy*

$$supp(\pi_a) \subseteq [t_1^a, \infty), \quad \pi_a((t_2^a, \infty)) > 0, \quad supp(\pi_b) \subseteq (-\infty, t_2^b], \quad \pi_b((-\infty, t_1^b)) > 0.$$

(ii) *For any $\varepsilon > 0$, there exist full-support priors $\pi_a^\varepsilon \preceq_{FOSD} \pi_b^\varepsilon$ with $\hat{\mu}_a^\varepsilon < \mu_a^\varepsilon \leq \mu_b^\varepsilon < \hat{\mu}_b^\varepsilon$ and*

$$\sup_t [F_{\pi_a^\varepsilon}(t) - F_{\hat{\pi}_a^\varepsilon}(t)]_+ < \varepsilon, \qquad \sup_t [F_{\hat{\pi}_b^\varepsilon}(t) - F_{\pi_b^\varepsilon}(t)]_+ < \varepsilon.$$

Part (i) follows directly from Steps 1, 2, and 3 of Appendix A.2: those arguments use only the strict FOSD shifts and the definitions of t_1^π, t_2^π, not unimodality of $\tilde{L}$. Step 1 gives $\text{supp}(\pi_a) \subseteq [t_1^a, \infty)$; Step 2 gives $\pi_a((t_2^a, \infty)) > 0$; Step 3 gives the mirror statements for π_b.

Part (ii) is a perturbation argument. We construct full-support priors with arbitrarily small FOSD violations by mixing a polarizing pair with a full-support distribution. The argument proceeds in five steps.

Step 1: Define mixture priors

By Theorem 1, multimodality of $\tilde{L}(x \mid \cdot)$ implies the existence of a polarizing pair $\pi_a^\star \preceq_{\text{FOSD}} \pi_b^\star$; the construction in Appendix A.1 produces such a pair with bounded support. Let U be any full-support distribution on Θ with finite mean. For $\varepsilon \in (0,1)$ define

$$\pi_i^\varepsilon = (1-\varepsilon)\pi_i^\star + \varepsilon U, \qquad i \in \{a, b\}.$$

Each π_i^ε has full support since U does.

Step 2: FOSD ordering is preserved

Mixture preserves the FOSD ordering between the priors:

$$F_{\pi_a^\varepsilon} = (1-\varepsilon)F_{\pi_a^\star} + \varepsilon F_U \geq (1-\varepsilon)F_{\pi_b^\star} + \varepsilon F_U = F_{\pi_b^\varepsilon}$$

pointwise, where the inequality uses $\pi_a^\star \preceq_{\text{FOSD}} \pi_b^\star$. Hence $\pi_a^\varepsilon \preceq_{\text{FOSD}} \pi_b^\varepsilon$.

Step 3: Posteriors and means are continuous in ε

The marginal likelihood normalizes the mixture as

$$\hat{\pi}_i^\varepsilon = \frac{(1-\varepsilon)Z_i^\star}{(1-\varepsilon)Z_i^\star + \varepsilon Z_U}\,\hat{\pi}_i^\star + \frac{\varepsilon Z_U}{(1-\varepsilon)Z_i^\star + \varepsilon Z_U}\,\hat{U},$$

where $Z_\pi := \mathbb{E}_\pi[\tilde{L}]$ and $\hat{U}$ is the posterior under U. The mixture weights are continuous in ε at $\varepsilon = 0$, with limits 1 and 0 respectively, so $\hat{\pi}_i^\varepsilon \to \hat{\pi}_i^\star$ weakly and $\hat{\mu}_i^\varepsilon \to \hat{\mu}_i^\star$ as $\varepsilon \to 0$. Prior means are linear in ε: $\mu_i^\varepsilon = (1-\varepsilon)\mu_i^\star + \varepsilon \mathbb{E}_U[\theta]$, so $\mu_i^\varepsilon \to \mu_i^\star$ as well.

Step 4: Mean inequality extends to a neighborhood of $\varepsilon = 0$

The polarizing pair satisfies the strict mean inequality $\hat{\mu}_a^\star < \mu_a^\star \leq \mu_b^\star < \hat{\mu}_b^\star$ at $\varepsilon = 0$. By continuity from Step 3, the inequality $\hat{\mu}_a^\varepsilon < \mu_a^\varepsilon \leq \mu_b^\varepsilon < \hat{\mu}_b^\varepsilon$ extends to a neighborhood of $\varepsilon = 0$.

Step 5: FOSD violations are $O(\varepsilon)$

At $\varepsilon = 0$, $\hat{\pi}_i^\star \prec_{\text{FOSD}} \pi_i^\star$ holds strictly, so

$$\sup_t [F_{\pi_a^\star}(t) - F_{\hat{\pi}_a^\star}(t)]_+ = 0, \qquad \sup_t [F_{\hat{\pi}_b^\star}(t) - F_{\pi_b^\star}(t)]_+ = 0.$$

The εU contributions to π_i^ε and $\hat{\pi}_i^\varepsilon$ each scale linearly in ε, so the FOSD violation suprema are $O(\varepsilon)$. Choose ε small enough that both the strict mean inequalities of Step 4 and the FOSD violation bounds in the proposition statement hold. $\square$

References

Daron Acemoglu, Victor Chernozhukov, and Muhamet Yildiz. Fragility of asymptotic agreement under Bayesian learning. *Theoretical Economics*, 11(1):187–225, 2016.

Roland Bénabou. The economics of motivated beliefs. *Revue d'économie politique*, 125(5):665–685, 2015.

Roland Bénabou and Jean Tirole. Self-confidence and personal motivation. *Quarterly Journal of Economics*, 117(3):871–915, 2002.

Jean-Pierre Benoît and Juan Dubra. Apparent bias: What does attitude polarization show? *International Economic Review*, 60(4):1675–1703, 2019.

T Renee Bowen, Danil Dmitriev, and Simone Galperti. Learning from shared news: When abundant information leads to belief polarization. *Quarterly Journal of Economics*, 138(2):955–1000, 2023.

Tuval Danenberg. Bayesian polarization. *Working paper*, 2026.

Roland G. Fryer, Philipp Harms, and Matthew O. Jackson. Updating beliefs when evidence is open to interpretation: Implications for bias and polarization. *Journal of the European Economic Association*, 17(5):1470–1501, 2019.

Matthew Gentzkow, Michael B. Wong, and Allen T. Zhang. Ideological bias and trust in information sources. *American Economic Journal: Microeconomics*, 17(2):162–213, 2025.

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

Charles G Lord, Lee Ross, and Mark R Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11):2098–2109, 1979.

John W. McHoskey. Case closed? on the John F. Kennedy assassination: Biased assimilation of evidence and attitude polarization. *Basic and Applied Social Psychology*, 17(3):395–409, 1995.

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

Charlie Pilgrim, Adam Sanborn, Eugene Malthouse, and Thomas T. Hills. Confirmation bias emerges from an approximation to Bayesian reasoning. *Cognition*, 245:105693, 2024.

Matthew Rabin and Joel L Schrag. First impressions matter: A model of confirmatory bias. *Quarterly Journal of Economics*, 114(1):37–82, 1999.

Lorraine Schwartz. On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26, 1965.