Rational Polarization from Joint Inference: Theory and Experiment

Working Draft

Jacob VanNattan
Department of Economics, Washington University in St. Louis
j.d.vannattan@wustl.edu

August 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, and, for the case of Normal priors with common variance, of exactly which configurations of prior beliefs polarize. Experimental results from an unincentivized pilot follow the theory. Subjects use their prior beliefs about the state to infer the informativeness of each report, weighting the report closer to their prior more heavily. These asymmetric weights on ex ante symmetric reports 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 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 in their initial beliefs 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 does not 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 must 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 characterization accounts for both null results.

Our most general result is that rational polarization is possible under a given information environment if and only if the marginal likelihood of the state is multimodal at some attainable data realization. Three concrete cases rule polarization out for any pair of priors: 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, uncertainty about signal informativeness, and at least three states. These conditions are present in most realistic information environments, and we

construct Normal and Binomial examples that exhibit such polarization.

Because it constrains only the environment, this characterization is existential in the priors: it says when some pair of beliefs polarizes, not which pairs do. Pinning down the polarizing beliefs themselves calls for structure on the priors. For the natural case of Normal priors with shared variance, the polarizing configurations form an explicit, full-dimensional region in the three-dimensional space of the two prior means and their common variance. We find that polarization is no longer possible once prior variance is sufficiently large, while widening the spread of the prior means has a non-monotonic impact.

These results also provide an audit of what existing polarization experiments identify. Because the three necessary ingredients can often be read directly off an experimental design, we provide a simple screen for when the joint-inference mechanism studied here is ruled out. Applying that screen to canonical studies, we find that nearly every design contains all three necessary ingredients, so their design features do not exclude rational joint inference as an explanation for divergence. This does not establish that each experiment's marginal likelihood is multimodal, nor does it negate other evidence of motivated processing collected in some studies. Our conclusion is rather that polarization itself generally does not identify biased assimilation in this literature, complementing the Bayesian identification critique of Benoît and Dubra (2019).

We test the mechanism experimentally using a design we call the *Imposter Task*, a classical balls-and-urns setup that removes the emotionally charged content present in canonical polarization experiments. A subject draws red and blue balls from an urn and forms a prior on the urn's color 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 so that 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 a 30-subject pilot study, 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 theoretical results. Section 4 reviews the existing experimental evidence through the lens of joint inference. Section 5 describes the experimental design motivated by this framework. Section 6 reports pilot results and discusses findings. Section 7 concludes. Technical proofs are relegated to the appendix.

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 $\lambda = (\lambda_1, \ldots, \lambda_n)$ 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 her 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:

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

where f denotes the density or probability mass function of F. Agent i's posterior over θ, marginalizing out λ, is $\hat{\pi}_i(\cdot \mid x)$, with mean $\hat{\mu}_i \equiv \mathbb{E}_{\hat{\pi}_i}[\theta \mid x_1, \ldots, x_n]$. Given the realized state θ, the data $x = (x_1, \ldots, x_n)$ are drawn from $\tilde{L}(x \mid \theta) := \int \prod_{j=1}^{n} L(x_j \mid \theta, \lambda_j)\, dF(\lambda)$. As a function of x this is the *marginal data distribution*; as a function of θ it is the *marginal likelihood* through which all of our results run, since marginalizing λ out of the joint posterior gives $\hat{\pi}_i(\theta \mid x) \propto \pi_i(\theta)\, \tilde{L}(x \mid \theta)$. A realization x is in the support of the marginal data distribution if it lies in this distribution's support for some $\theta \in \Theta$.

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

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

FOSD 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 it occurs.*

Definition 2 (Mean polarization)**.** *Mean polarization* occurs *at signal realization* $x = (x_1, \ldots, x_n)$ *given priors* π_a, π_b *with* $\mu_a \leq \mu_b$ *if*

$$\hat{\mu}_a < \mu_a \leq \mu_b < \hat{\mu}_b.$$

Mean polarization is possible *under signal structure* (Θ, Λ, L, F) *if there are priors* π_a, π_b *with* $\mu_a \leq \mu_b$ *and a signal realization* x *in the support of the marginal data distribution at which it occurs.*

The two definitions track the same divergence at different resolutions. FOSD polarization pulls the full posterior CDFs apart in stochastic order; mean polarization asks only that the posterior means diverge which is the coarser notion that empirical work typically operationalizes. FOSD polarization is the strictly stronger of the two, since strict FOSD shifts imply the corresponding ordering of means.

We state the general characterization in FOSD terms because it is the sharper distribution-level notion and yields a clean equivalence; the mean-polarization counterpart, which requires

added structure on the priors, follows in Section 3.3. In each definition the first clause is the ex post event, a particular x and prior pair yielding divergent updates, and the second is the ex ante structural property, the signal structure admitting some such x and (π_a, π_b), so that polarization has the potential to occur under the information environment.

Throughout the theoretical analysis we maintain the following regularity: Θ is a closed subset of $\mathbb{R}$, and at every x in the support of the marginal data distribution, the map $\theta \mapsto \tilde{L}(x \mid \theta)$ is upper semicontinuous and, when Θ is unbounded, vanishes as $|\theta| \to \infty$. These conditions make every superlevel set $\{\theta : \tilde{L}(x \mid \theta) \geq z\}$ with $z > 0$ compact, so that when nonempty it contains its leftmost and rightmost points, which the level-set arguments to come require; they also make $\tilde{L}(x \mid \cdot)$ bounded. When we specialize to Normal priors on $\Theta = \mathbb{R}$ in Section 3.3, we additionally maintain that $\tilde{L}(x \mid \cdot)$ is integrable and absolutely continuous with an integrable derivative $\tilde{L}'$; we call these the *smoothness conditions*. The maintained regularity holds trivially for finite Θ, and both tiers are immediate in the examples of Section 3.5.

3 Theoretical results

All results in this section concern a fixed signal structure (Θ, Λ, L, F) and its marginal likelihood $\tilde{L}$, under the maintained regularity described in Section 2.

3.1 The general characterization

Definition 3 (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. *FOSD 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 and π_b's mass onto the right peak. 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, not shifted past each other. The geometry is inconsistent. The full proof appears in Appendix A.1. □

Theorem 1 asks nothing of the priors. The pair $\pi_a \preceq_{\text{FOSD}} \pi_b$ may be any two distributions on Θ, discrete or continuous, parametric or not; the FOSD ordering serves only to label the agents, and whether polarization is possible turns entirely on the environment through $\tilde{L}$. The price of this generality is that the characterization is purely existential in the priors: it guarantees a polarizing pair at every multimodal $\tilde{L}$ but is silent on which pairs polarize.

The theorem also has an immediate one-directional corollary for mean polarization. Since strict FOSD shifts imply the corresponding ordering of means, the sufficiency direction carries over: a multimodal $\tilde{L}(x \mid \cdot)$ admits mean polarization as well. The necessity direction does not: with unrestricted priors, unimodality no longer rules out mean polarization. Restoring necessity, as well as describing the set of polarizing pairs, requires committing to a class of priors, as we do in Section 3.3 for Normal priors with shared variance.

3.2 Support restrictions

So far we have characterized when FOSD polarization is possible in terms of the information environment. We now turn to what this requires of the priors. Proposition 1 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 1. *Let $\tilde{L}(x \mid \cdot)$ be multimodal. For any prior π with $Z_\pi := \mathbb{E}_\pi[\tilde{L}] > 0$, let t_1^π and t_2^π denote the leftmost and rightmost points of $\{\theta : \tilde{L}(x \mid \theta) \geq Z_\pi\}$.*

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

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

(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 $\text{supp}(\pi_a) \subseteq [t_1^a, \infty)$ is Step 1 of the necessity argument for Theorem 1: below t_1^a we have $\tilde{L} < Z_a$, so Bayes erodes any mass there, pushing the posterior CDF the wrong way for a downward FOSD shift. The constraint on π_b is the mirror image. (ii) Mix any FOSD-polarizing pair from Theorem 1 with sufficiently small weight on a suitably chosen full-support finite-mean distribution; the resulting FOSD violations vanish with the mixing weight. The full proof appears in Appendix A.2. □

Taken together, the support conditions rule out a full-support FOSD-polarizing pair on any unbounded state space: if Θ is unbounded below, a full-support π_a necessarily places mass below the finite cutoff t_1^a, while if Θ is unbounded above, a full-support π_b necessarily places mass above the finite cutoff t_2^b. The Normal example of Section 3.5 is of this kind. On a bounded state space the support conditions can be vacuous: if agent a's above-average region extends to the lower endpoint

of Θ, then t_1^a is that endpoint and there is no wrong tail below it to clear, and symmetrically for b at the upper endpoint. Where they do bind, part (ii) shows that full-support priors come arbitrarily close to FOSD polarization. The Normal and Binomial examples to follow use natural full-support priors, exhibit clear mean polarization, and have FOSD violations on the order of 10^{-5} and 10^{-11} respectively, illustrating part (ii).

These support restrictions also delineate our result from the impossibility of Baliga et al. (2013), who show that Bayesian agents with common-support priors over a one-dimensional state, updating from a shared signal likelihood, cannot polarize in first-order stochastic dominance. The FOSD-polarizing priors of Theorem 1 have non-common support, consistent with the support restrictions of part (i); under common support only the mean polarization of part (ii) survives.

3.3 Mean polarization under Normal priors

Describing which priors polarize, rather than merely whether some pair does, calls for structure on the priors. For Normal priors with shared variance that structure restores the necessity that fails for general priors: multimodality of $\tilde{L}$ becomes necessary as well as sufficient for mean polarization, and the polarizing configurations form an explicit, full-dimensional region in the space of prior means of the agents and their common variance. This subsection establishes the equivalence and characterizes that region.

Throughout this subsection $\Theta = \mathbb{R}$, priors are Normal, and the smoothness conditions of Section 2 are in force; they justify the convolution and integration-by-parts identities below.

Theorem 2 (Mean polarization under Normal priors)**.** *Mean polarization is possible under signal structure (Θ, Λ, L, F) with Normal priors of shared variance, $\pi_a = \mathcal{N}(\mu_a, \sigma^2)$ and $\pi_b = \mathcal{N}(\mu_b, \sigma^2)$, 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. We prove the pointwise version at a fixed x; the existential statement follows by quantifying over x in the support of the marginal data distribution. For a Normal prior $\mathcal{N}(\mu, \sigma^2)$, a Gaussian identity and integration by parts express the posterior mean shift as $\hat{\mu} - \mu = \sigma^2(f_{\sigma^2} * \tilde{L}')(\mu)/Z$, where f_{σ^2} is the prior density centered at the origin and $Z > 0$ is the posterior normalizer. Its sign is that of the smoothed derivative $(f_{\sigma^2} * \tilde{L}')(\mu)$, so mean polarization with $\mu_a < \mu_b$ requires a $(-, +)$ sign change in $f_{\sigma^2} * \tilde{L}'$.

The Gaussian is a Pólya frequency kernel, so by the variation-diminishing theorem (Karlin, 1968) smoothing cannot create additional sign changes. If $\tilde{L}$ is multimodal, choose $\theta_1 < \theta_2 < \theta_3$ as in Definition 3. Absolute continuity gives a Lebesgue point $\mu_a \in (\theta_1, \theta_2)$ with $\tilde{L}'(\mu_a) < 0$ and a Lebesgue point $\mu_b \in (\theta_2, \theta_3)$ with $\tilde{L}'(\mu_b) > 0$; Gaussian smoothing preserves these signs for sufficiently small σ, giving a mean-polarizing Normal pair. If $\tilde{L}$ is unimodal, $\tilde{L}'$ has at most one sign change, from $+$ to $-$, so its Gaussian smoothing has at most one sign change and cannot contain the $(-, +)$ crossing required for mean polarization. The full proof is in Appendix A.3. □

Theorem 2 is the mean-polarization counterpart of Theorem 1, restricted to Normal priors with shared variance. Multimodality of $\tilde{L}$ is the common structural condition: necessary and sufficient for FOSD polarization under general priors, and necessary and sufficient for mean polarization within the Normal shared-variance class. The two results trade structure on the priors against structure on the notion of polarization, and multimodality is the common condition that survives the trade. It is always sufficient, delivering a polarizing pair under either notion with general priors; what varies is necessity. The strong FOSD notion secures necessity for arbitrary priors; weakening to mean polarization forfeits it in general, and restricting the priors to the Normal shared-variance class restores it.

The sufficiency argument for Theorem 2 delivers a mean-polarizing pair at any x where $\tilde{L}(x \mid \cdot)$ is multimodal; the next result identifies exactly which configurations mean-polarize. For multimodal $\tilde{L}(x \mid \cdot)$, define the mean polarization region

$$P(\tilde{L}) := \{(\mu_a, \mu_b, \sigma^2) : \mu_a < \mu_b,\ \sigma^2 > 0,\ (f_{\sigma^2} * \tilde{L}')(\mu_a) < 0 < (f_{\sigma^2} * \tilde{L}')(\mu_b)\}$$

and the critical variance

$$\sigma^2_{\text{crit}}(\tilde{L}) := \sup\{\sigma^2 > 0 : f_{\sigma^2} * \tilde{L}' \text{ has a } (-,+) \text{ sign change}\}.$$

In words, $f_{\sigma^2} * \tilde{L}'$ is the derivative of the likelihood as smoothed at the prior's scale, so membership in $P(\tilde{L})$ places μ_a on a falling stretch of the smoothed evidence and μ_b on a rising one, and σ^2_{crit} is the supremum of the prior variances for which the smoothed evidence retains an interior valley i.e. its multimodality.

Theorem 3 (Mean polarization region)**.** *Fix x with $\tilde{L}(x \mid \cdot)$ multimodal, and Normal priors $\pi_a = \mathcal{N}(\mu_a, \sigma^2)$, $\pi_b = \mathcal{N}(\mu_b, \sigma^2)$ with $\mu_a < \mu_b$.*

(i) Mean polarization occurs at x given π_a, π_b if and only if $(\mu_a, \mu_b, \sigma^2) \in P(\tilde{L})$.

(ii) $P(\tilde{L})$ is open and non-empty.

(iii) $P(\tilde{L}) \subseteq \{\sigma^2 < \sigma^2_{\text{crit}}(\tilde{L})\}$, and if $\tilde{L}(x \mid \cdot)$ has compact support or is a finite mixture of Gaussian densities, then $\sigma^2_{\text{crit}}(\tilde{L}) < \infty$.

Proof sketch. The same convolution identity gives the exact criterion in the statement; openness is continuity in (μ, σ^2) and non-emptiness is the sufficiency argument for Theorem 2. Any polarizing variance carries a $(-,+)$ crossing, and continuity in (μ, σ^2) pushes that crossing to slightly larger variances, so the supremum defining σ^2_{crit} strictly exceeds it: mean polarization requires $\sigma^2 < \sigma^2_{\text{crit}}$. For the finiteness of the ceiling, the Gaussian semigroup makes the number of sign changes of $f_{\sigma^2} * \tilde{L}'$ non-increasing in σ^2: additional smoothing can only erase crossings. Under either added hypothesis, Lemma 1 shows that large σ leaves only the global $(+,-)$ shape, so the threshold variance σ^2_{crit} beyond which no $(-,+)$ crossing remains is finite. The full proof is in Appendix A.4. □

The added hypothesis in part (iii) is not redundant: eventual unimodality of the smoothed likelihood can fail under the maintained regularity alone. The bump train $\tilde{L}(\theta) = \sum_{k\geq 1} 2^{-k} f_1(\theta - 2^k)$ is integrable, absolutely continuous with integrable derivative, and vanishes at infinity, yet at every σ the bump at 2^k, for 2^k large relative to σ, produces a local mode of $f_{\sigma^2} * \tilde{L}$, since the remaining mass lies at distance at least 2^{k-1} and contributes exponentially less there than the bump's own peak; so $f_{\sigma^2} * \tilde{L}'$ retains a $(-,+)$ crossing at every σ^2 and $\sigma^2_{\text{crit}} = \infty$. Lemma 1 in Appendix A.4 shows that compact support or a finite Gaussian mixture structure rules this out; the latter covers the Normal example of Section 3.5, whose marginal likelihood is a four-component Gaussian mixture in θ.

The mean polarization region $P(\tilde{L})$ is a genuinely three-dimensional object, open by Theorem 3 and hence of positive Lebesgue measure in (μ_a, μ_b, σ^2)-space, with the three coordinates coupled rather than separable. The polarization-prone configurations of prior means depend on the prior variance, and vice versa: the admissible region of prior means deforms as σ grows and, when $\sigma^2_{\text{crit}} < \infty$, disappears entirely at and above σ^2_{crit}.

The bimodal case gives the simplest geometric picture, which the next result makes exact at every variance: each slice of $P(\tilde{L})$ is a box cut out by the peaks and valley of the likelihood smoothed at the prior's own scale, and the box exists precisely below the critical variance. For $\tilde{L}(x \mid \cdot)$ with three or more modes each slice admits a similar description, as a union of such products over admissible descending-ascending interval pairs of the smoothed likelihood.

Proposition 2 (Comparative statics)**.** *Fix x with $\tilde{L}(x \mid \cdot)$ bimodal, with peaks $\theta_1 < \theta_3$ and valley θ_2: strictly increasing on $(-\infty, \theta_1]$, strictly decreasing on $[\theta_1, \theta_2]$, strictly increasing on $[\theta_2, \theta_3]$, and strictly decreasing on $[\theta_3, \infty)$. Consider Normal priors with shared variance σ^2 and means $\mu_a < \mu_b$, and write $g_\sigma := f_{\sigma^2} * \tilde{L}(x \mid \cdot)$ for the smoothed likelihood.*

(i) Some pair mean-polarizes at variance σ^2 if and only if $\sigma^2 < \sigma^2_{\text{crit}}(\tilde{L})$.

(ii) For every $\sigma^2 < \sigma^2_{\text{crit}}(\tilde{L})$, the smoothed likelihood g_σ is bimodal, with peaks $a_\sigma < c_\sigma$ and valley b_σ, and, off the isolated zeros of g'_σ, the pair mean-polarizes at variance σ^2 if and only if $a_\sigma < \mu_a < b_\sigma < \mu_b < c_\sigma$. For $\sigma^2 \geq \sigma^2_{\text{crit}}(\tilde{L})$, g_σ is unimodal.

(iii) If $\theta_1 < \mu_a < \theta_2 < \mu_b < \theta_3$, the pair mean-polarizes at all sufficiently small σ^2; if $\mu_a \notin [\theta_1, \theta_2]$ or $\mu_b \notin [\theta_2, \theta_3]$, it does not mean-polarize at all sufficiently small σ^2.

Proof sketch. Since g_σ is nonnegative and vanishes at $\pm\infty$, the sign sequence of $g'_\sigma = f_{\sigma^2} * \tilde{L}'$ must begin with $+$ and end with $-$, and by variation diminishing it has at most the three sign changes of $\tilde{L}'$: either the pattern $(+,-)$, making g_σ unimodal and polarization impossible, or $(+,-,+,-)$, making g_σ bimodal. The semigroup makes the bimodal case a down-set in σ^2 whose supremum is σ^2_{crit}, giving (i) and the dichotomy in (ii). In the bimodal case the sign of g'_σ is determined by position relative to $(a_\sigma, b_\sigma, c_\sigma)$, and the criterion of Theorem 3(i) converts those signs into the box. Part (iii) is the $\sigma \to 0$ limit: at each mean strictly inside a monotone segment of $\tilde{L}$, the concentrating smoothing takes the sign of that segment, which pins the criterion down on the open box and off its closure. The full proof is in Appendix A.5. □

A prior of variance σ^2 reads the evidence at resolution σ, and the pair mean-polarizes exactly when the two means fall on opposite sides of the valley of the smoothed likelihood, each between the valley and its nearer peak. Position in this landscape, not the distance between the priors, is what matters. A pair straddling the valley polarizes however small the gap, a pair on the same side cannot mean-polarize however far apart, and a mean beyond either peak is pulled back toward it since from there the perceived evidence points only back toward the peak. Raising σ^2 coarsens the resolution until, at σ^2_{crit}, the bimodality of the evidence washes out and polarization becomes impossible. Rational polarization thus concentrates where the evidence is divided at the resolution of the agents' own confidence and their priors take opposite sides.

The shared variance assumption in Theorem 2 is essential. For concreteness, take the asymmetric unimodal $\tilde{L}(x \mid \theta) = e^{-\theta^2/2}$ for $\theta \leq 0$ and $e^{-\theta^2/18}$ for $\theta > 0$, with a single mode at 0 and a heavier right tail. A confident agent, $\mathcal{N}(1.1, 1)$, and a diffuse agent, $\mathcal{N}(1.4, 25)$, have posterior means 1.03 and 1.47 against prior means 1.1 and 1.4, so they mean-polarize under a unimodal $\tilde{L}$: the confident prior is pulled toward the mode, the diffuse prior outward into the heavy tail.

With equal variances, the smoothed derivative has at most one $(+,-)$ crossing, so the two agents cannot exhibit the opposing outward shifts required for mean polarization; heterogeneous precision is required for the split in this example. This example also relies on an asymmetric $\tilde{L}$: for symmetric unimodal $\tilde{L}$, $f_{\sigma^2} * \tilde{L}'$ is odd and single-crossing at the mode for every σ^2, and unequal variances cannot separate the shifts. The identity $sf(s) = -\sigma^2 f'(s)$ holds only for the Gaussian, so the argument used in this section does not extend to non-Normal priors.

3.4 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. Many standard location and count families, including the examples used below, satisfy both.

Proposition 3. *In any of the following cases, $\tilde{L}(x \mid \cdot)$ is unimodal in θ at every x, so FOSD polarization is not possible under the signal structure (Theorem 1). In the continuous-state cases (i) and (ii), mean polarization is also impossible with Normal priors of shared variance under the smoothness conditions of Theorem 2; in the binary-state case (iii), mean polarization is impossible for any priors:*

(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. Moreover, on a two-point ordered state space a distribution is indexed by its probability on the high state and its mean is an affine increasing function of that probability. Mean polarization therefore coincides with FOSD polarization and is impossible as well.

In the continuous-state cases (i) and (ii), the conclusion of Theorem 2 applies whenever $\tilde{L}(x \mid \cdot)$ also satisfies the smoothness conditions described in Section 2. In case (ii), a finite product of absolutely continuous log-concave components vanishing in the tails inherits both directly. In case (i), it suffices in addition that each component likelihood be absolutely continuous in θ and that $\int |\partial_\theta L(x \mid \theta, \lambda)| \, dF(\lambda)$ be integrable in θ: Fubini's theorem then gives that the mixture is absolutely continuous with $\tilde{L}' = \int \partial_\theta L(x \mid \cdot, \lambda) \, dF(\lambda)$. □

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 rational benchmark of Gentzkow et al. (2025) is accounted for on two margins. Within a period, their agents share a common prior over the state, so the heterogeneous state priors our characterization requires are absent by construction. Across periods, their agents learn source accuracies, so informativeness uncertainty vanishes and case (ii) closes the channel in the limit. As shown in Proposition 4, joint inference polarization is a finite-signal phenomenon that correct learning extinguishes. Their benchmark therefore cannot polarize within periods or in the long run.

3.5 Possibility examples

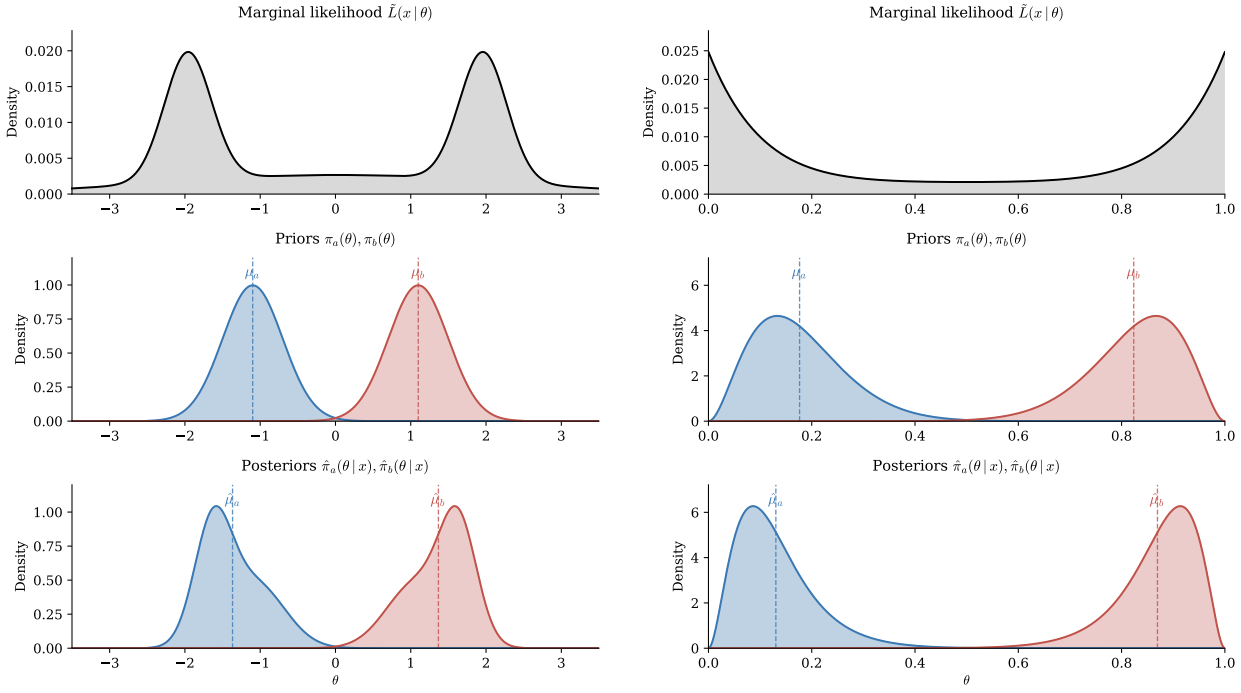
Two parametric examples illustrate that the joint inference mechanism is not specific to any particular signal structure. Both use full-support priors and exhibit mean polarization with negligible FOSD violations, as in Proposition 1(ii).

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 $\nu = \frac{1}{2}\delta_{0.1} + \frac{1}{2}\delta_{10}$ (where δ_z denotes a unit point mass at z), 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 $\pi_a = \mathcal{N}(-1.1, 0.16)$ and $\pi_b = \mathcal{N}(1.1, 0.16)$, both with full support on $\mathbb{R}$. The prior pair $(\mu_a, \mu_b, \sigma^2) = (-1.1, 1.1, 0.16)$ lies in the mean polarization region $P(\tilde{L})$ of Theorem 3. 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^{-5}. See Figure 1(a).

Binomial example. Take $\Theta = [0,1]$, $\Lambda = [0,1]$, likelihood $x_j \mid \theta, \lambda_j \sim \lambda_j \cdot \text{Binomial}(m,\theta) + (1-\lambda_j) \cdot \text{BetaBinomial}(m,1,1), m = 10$, $F = \nu \otimes \nu$ 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 essentially rule out $(1,1)$ as incompatible with the asymmetric data and put almost no weight on the wrong-side interpretation. With priors $\pi_a = \text{Beta}(3,14)$ and $\pi_b = \text{Beta}(14,3)$, 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^{-11}. 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 1(ii).

3.6 Asymptotic learning

Suppose signals accumulate under an iid informativeness structure: $F = \nu^{\otimes n}$, so that given θ the signals are iid with per-signal marginal $\tilde{\ell}(\cdot \mid \theta) = \int L(\cdot \mid \theta, \lambda)\, d\nu(\lambda)$, and the data are generated at θ_0. Assume the standard regularity for posterior consistency: θ_0 lies in the Kullback–Leibler support of π_a and π_b, the family $\{\tilde{\ell}(\cdot \mid \theta) : \theta \in \Theta\}$ admits uniformly consistent tests of θ_0 against $\{\theta : |\theta - \theta_0| > \varepsilon\}$ for every $\varepsilon > 0$, and, if Θ is unbounded, the posterior sequences are uniformly integrable. Both parametric examples of Section 3.5 satisfy these conditions.

Proposition 4. *Under these conditions,*

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

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; boundedness of Θ or uniform integrability upgrades weak convergence to convergence of the posterior means. □

The point of Proposition 4 is that 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 a shared correct prior over informativeness, posteriors over θ collapse onto the truth as signals accumulate, and the heterogeneous beliefs about signal informativeness that drive polarization vanish with them. Disagreement in the joint inference framework remains a finite-signal feature of correct Bayesian updating rather than a permanent deviation from it.

4 Existing experimental evidence

Much of the experimental literature documenting polarization from shared evidence follows a common protocol: subjects with different prior positions on an issue are shown a common body of evidence and their positions or evaluations are then re-measured. Many of these studies find that subjects with different priors update their positions in opposite directions, and the prominent interpretations involve departures from Bayesian rationality while our work thus far establishes a rational alternative. Furthermore, our theoretical results identify coarse features of an information environment that rule out the joint-inference mechanism studied here, and therefore clarify what can and cannot be identified from subject behavior in these experiments.

The screen turns on three features of the information environment: whether subjects see more than one piece of evidence, whether the quality of that evidence is uncertain, and whether the quantity under dispute can take more than two values. Under the regularity conditions of Proposition 3, each is necessary for the marginal likelihood to be multimodal. Their absence therefore closes our

rational channel. Their presence, however, is only a necessary-condition screen. We claim only that these studies leave joint inference open as a possibility and, therefore, do not cleanly identify the presence of bias.

Table 1 applies this screen to canonical studies. The three ingredient columns are read from each design, granting the regularity conditions of Section 3.4. "Open" means no case of Proposition 3 applies, so the design features alone do not rule out rational joint inference. "Closed" means our mechanism is ruled out under the stated coding. The final column records the study's polarization outcome and how it was measured, since elicitation has been shown to matter for the phenomenon (Miller et al., 1993; Kuhn and Lao, 1996).

Study	Evidence shown	$n \geq 2$	λ unknown	$\lvert\Theta\rvert \geq 3$	Channel	Divergence
Lord et al. (1979)	two opposing studies	Yes	Yes	Yes	Open	Yes (reported change)
Plous (1991), Study 1	one breakdown dossier	No*	Yes	Yes	Closed (i)*	Yes (reported change)
McHoskey (1995)	mixed argument set	Yes	Yes	Yes	Open	Yes (reported change)
Munro and Ditto (1997)	opposing fabricated studies	Yes	Yes	Yes	Open	Yes (perceived change only)
Taber and Lodge (2006)	pro and con argument pool	Yes	Yes	Yes	Open	Yes (attitude re-rating)
Corner et al. (2012)	two opposing editorials	Yes	Yes	Yes	Open	No (co-movement)
Fryer et al. (2019)	mixed research summaries	Yes	Yes	Yes	Open†	Yes (interpretation and belief re-rating)

Table 1: Open means no case of Proposition 3 applies; it is a necessary-condition screen, not a claim that the realized marginal likelihood is multimodal. Closed (i) names the case of Proposition 3 that rules out joint inference under the stated coding. The asterisk and dagger mark coding judgments discussed in the text.

The archetype is Lord et al. (1979). Proponents and opponents of capital punishment each read two purported studies, one finding a deterrent effect and one finding none, with different empirical designs whose probative value subjects were asked to assess. Subjects rated the study congenial to their position as more convincing and reported attitudes that had moved further apart. McHoskey (1995) found a related pattern after giving conspiracy and lone-gunman believers a common set of arguments on the Kennedy assassination. Munro and Ditto (1997) found prior-congruent evaluations of two fabricated studies bearing on a sexual-orientation stereotype; their evidence of attitude polarization was restricted to perceived attitude change. None of these designs is ruled out by Proposition 3. That fact does not prove that the exact stimuli generated a multimodal marginal

likelihood, but it means that divergence and prior-congruent evaluations are not, by themselves, sufficient to distinguish biased assimilation from Bayesian inference in which priors affect posterior assessments of informativeness.

Some studies provide process evidence beyond the polarization outcome, and our audit should not be read as dismissing it. In Taber and Lodge (2006), subjects evaluated pro and con arguments about affirmative action and gun control under forced exposure and, in a separate task, chose which arguments to inspect. The paper documents prior-congruent evaluations and counterarguing of contrary evidence, as well as selective exposure when choice was available. Joint inference can speak directly to prior-dependent evaluations and the resulting belief movement, but it does not explain counterargument production or information-seeking behavior. We therefore treat the polarization outcome as non-diagnostic, not the paper as a whole. Similarly, Munro and Ditto (1997) report affective responses to attitude-inconsistent evidence that lie outside our mechanism. Corner et al. (2012) find prior-dependent evaluations of opposing climate editorials but no attitude polarization. Both more- and less-skeptical subjects became more skeptical after exposure, an outcome compatible with the co-movement allowed by Proposition 2.

Plous (1991) requires study-level coding. In Study 1, subjects read a dossier about one technological breakdown—either the Three Mile Island accident or the 1980 false missile alerts—assembled from several excerpts and factual components. If that dossier is treated as one signal, Proposition 3(i) closes our channel. If its distinct components are treated as separate pieces of evidence whose diagnostic value must be assessed, the single-signal restriction no longer clearly applies. Moreover, the paper contains additional studies; Study 2 supplements the false-alert material with four additional breakdowns. The asterisk in Table 1 records this coding issue rather than resolving it in our favor.

The theory of Fryer et al. (2019) deliberately departs from full Bayesian updating in a binary-state model: ambiguous signals are interpreted and then stored in their interpreted form without the possibility for re-interpretation. Their experiment is richer than that binary theoretical benchmark. Subjects state a prior belief, interpret a sequence of research summaries on climate change or the death penalty, and then answer the initial belief question again. The experiment therefore contains both prior-dependent signal interpretation and re-elicited belief movement. Fryer et al. explicitly note that an experiment cannot rule out Bayesian updating under an arbitrarily rich worldview; what their design rejects is a more restricted one-dimensional Bayesian benchmark.

In summary, the polarization outcome in these studies generally does not, on its own, identify biased assimilation. Most of the canonical designs pass the necessary-condition screen for joint inference, while the cleanest apparent exception depends on how the evidence in one study is partitioned into signals. At the same time, several papers collect additional evidence—affect, counterarguing, selective exposure, and other process measures—that our theorem does not rationalize and that may bear independently on motivated processing.

The identification problem for the polarization outcome is therefore that rational and non-rational channels can be open at the same time. Our design changes that comparison. The Imposter

Task preserves the joint-inference channel while removing the identity-laden content and belief-based utility motives emphasized in much of the prior literature. It does not rule out generic mistakes, heuristics, or misunderstanding. Instead, its staged elicitation is designed to identify whether and to what extent subjects perform the specific Bayesian joint inference over the state and signal quality that our theory describes without ruling out every possible alternative to explain polarization.

5 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.

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 with replacement from Urn 1, observing k_i red, endows her with a prior π_i over the state, with Bayesian benchmark $\text{Beta}(1+k_i, 1+n_i-k_i)$ under an initial $\text{Beta}(1,1)$. Subjects with different realized own draws form different priors, which is exactly the heterogeneity the model in Section 2 requires.

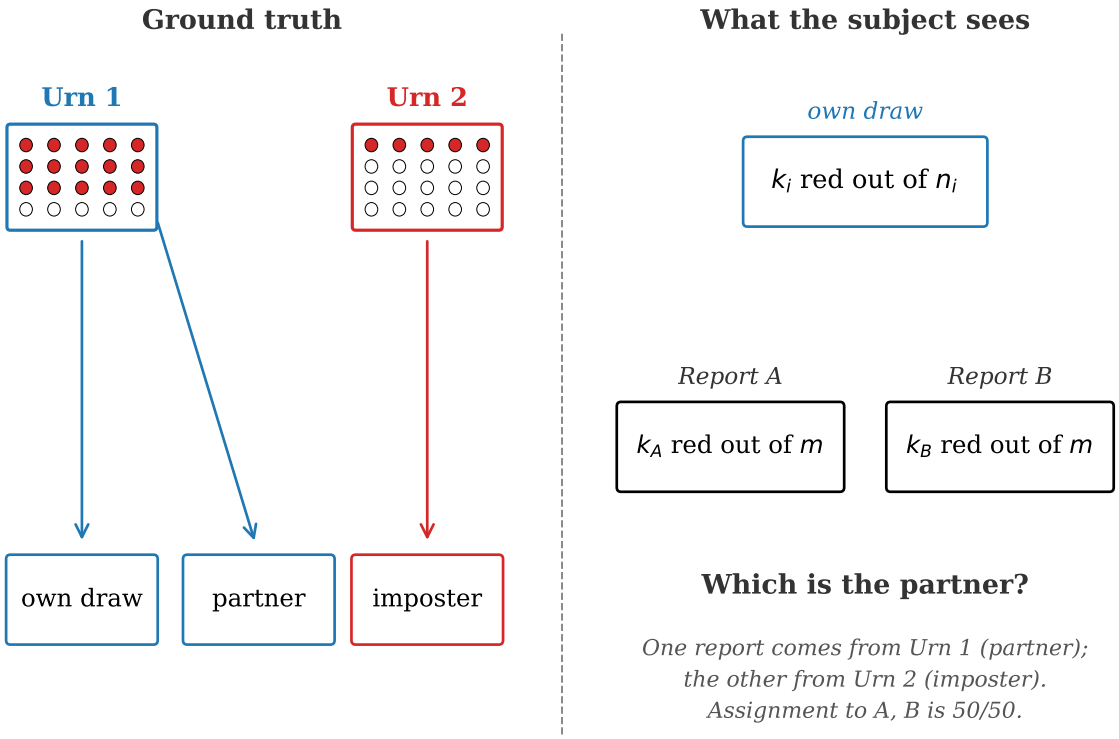


Figure 2: Imposter Task: data-generating process. The subject's own draw and one of the two external reports come from Urn 1 (the partner); the other report comes from Urn 2 (the imposter). Reports are displayed as A and B in random order with labels hidden.

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 3.5, 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 FOSD 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. Stated estimates of r_1 are rescaled to proportions, so $Q1$ and $Q3$ lie in $[0,1]$ throughout. 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]

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

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

[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.

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):

$$Q1_t - \tfrac{1}{2} = \kappa_0 + \kappa_1 \left(Q1_t^* - \tfrac{1}{2}\right) + \varepsilon_t.$$

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 $Q1$:

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

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 6 was conducted in a single classroom session with 30 undergraduate student subjects and 8 trials each (240 trials total). The pilot was unincentivized in that subjects were given arbitrary points which did not correspond to actual payment. The scoring rule above was still 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 full-scale and properly incentivized experiment is planned for Fall 2026.

6 Empirical results

6.1 Behavioral decomposition

The central empirical question is whether subjects engage in joint inference of the type we describe. The mixture regression defined in Section 5 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, $p < 0.001$ against zero; we cannot reject $\alpha = 1$); see Table 2. 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	< 0.001	0.39
$n_i = 10$	0.87	0.08	< 0.001	0.50
$n_i = 50$	0.82	0.18	< 0.001	0.14

Table 2: Joint inference mixture at Stage 3. $\hat{\alpha}$ measures the relative weighting between the joint inference Bayesian benchmark ($\alpha = 1$) and the no joint inference 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 $\hat{\alpha} > 0.5$. We do not drop any outliers though three subjects appear to be in the far left tail of the distribution.

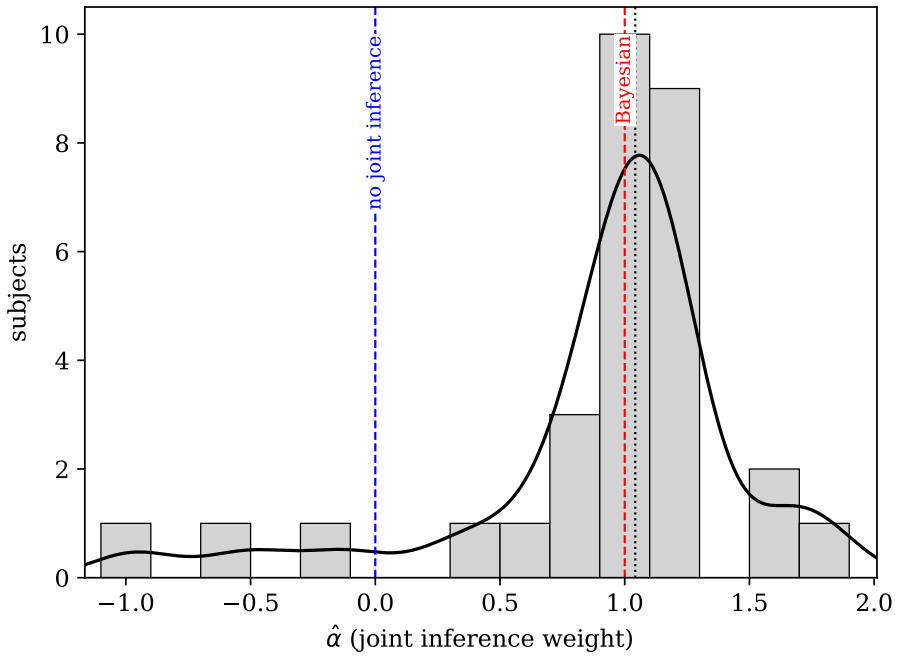


Figure 3: Per-subject distribution of $\hat{\alpha}$. $\hat{\alpha} = 1$ matches the full joint inference Bayesian; $\hat{\alpha} = 0$ matches the no joint inference alternative. Bin size = 0.20. The black dotted line marks the median.

6.2 Stage-wise decomposition

The mixture result establishes that subjects engage in behavior consistent with joint inference overall. The stage-wise regressions defined in Section 5, with results in Table 3, localize where this inference is fully Bayesian and where it is attenuated.

Stage 1 is essentially fully Bayesian: $\hat{\kappa}_1 = 1.00$ (SE 0.03, $p < 0.001$ against zero; we cannot reject the Bayesian null $\kappa_1 = 1$, $p = 0.91$) 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, $p < 0.001$ against zero; we reject the Bayesian null $\gamma_1 = 1$ at $p < 0.001$) 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, $p < 0.001$ 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 (Edwards, 1968).

Stage	Coefficient	SE	p_0	R^2
1. Prior formation ($\hat{\kappa}_1$)	1.00	0.03	< 0.001	0.90
2. Identity inference ($\hat{\gamma}_1$)	0.66	0.06	< 0.001	0.56
3. Composition ($\hat{\beta}_1$)	0.79	0.14	< 0.001	0.40

Table 3: Stage-wise decomposition. Each row reports one regression from Section 5. p_0 is the p-value against the null that the coefficient is zero. Standard errors clustered at the subject level.

6.3 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 use 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 do not describe alternative procedures; they are simply too brief or vague to classify (one subject wrote "Math.").

The stage-wise pattern tells the same story. 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. If the Stage 2 attenuation were a common proportional scaling, then 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 $[0.025, 0.975]$ since 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, yet this inability to pin down precisely where subjects' beliefs sit in the extreme $[0.00, 0.05]$ and $[0.95, 1.00]$ intervals does not prevent us from identifying that their beliefs are statistically different from 0.50.

Hence 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 Stage 1 is essentially Bayesian with $R^2 = 0.90$, suggesting that this concern is not of first-order importance in the current results.

7 Conclusion

A canonical interpretation of divergent updating from common evidence is biased or otherwise non-Bayesian processing. We have shown that the same pattern can arise from full Bayesian inference whenever the marginal likelihood is multimodal at some signal realization. This single structural condition is necessary and sufficient for FOSD polarization under general priors and for mean polarization under Normal priors with shared variance. Imposing the structure of the Normal shared-variance class lets us further characterize the set of polarizing beliefs as an explicit, full-dimensional region in the space of prior means and common variance.

Multimodality of the marginal likelihood is then clearly the key structural condition behind these results. Under the regularity conditions of Proposition 3, it requires multiple signals of uncertain informativeness and more than two possible states. Together, these are the ingredients through which an information environment can generate multimodality, and they are native to many realistic settings. States of interest are rarely binary, evidence usually arrives piecemeal, and the informativeness of each piece is almost never known with certainty ex ante. Our experimental setting satisfies these requirements and has a multimodal marginal likelihood; stripped of emotional content, it provides preliminary evidence that subjects do engage in joint inference.

Through this lens, the polarization outcome in the existing experimental literature provides weaker identification of biased assimilation than is often assumed. Nearly every canonical design passes the three-ingredient necessary-condition screen for joint inference. This does not deny the independent process evidence for motivated reasoning reported in some of these papers. It isolates a narrower point: divergence of beliefs from common evidence, by itself, does not identify its behavioral mechanism nor any departure from Bayesian rationality.

The canonical study of Lord et al. (1979) illustrates the distinction. Subjects report beliefs about the deterrent efficacy of capital punishment on a scale from -8 to $+8$ and evaluate two purported studies with opposing conclusions and different empirical designs. The design therefore contains multiple pieces of evidence whose probative value subjects must assess. A Bayesian who jointly infers the deterrent effect and the informativeness of the evidence can, in principle, evaluate prior-congruent evidence as more likely to be accurate and propagate that assessment back into beliefs about the state. We should not be surprised that these subjects came away from the experiment with more extreme views.

A Complete Proofs

A.1 Proof of Theorem 1

Theorem 1. *FOSD 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. See Figure 4.

Sufficiency

We construct FOSD-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 3, 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(a)).

Step 2: $\pi_a \preceq_{FOSD} \pi_b$. 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 $\pi_b \prec_{\text{FOSD}} \hat{\pi}_b$.

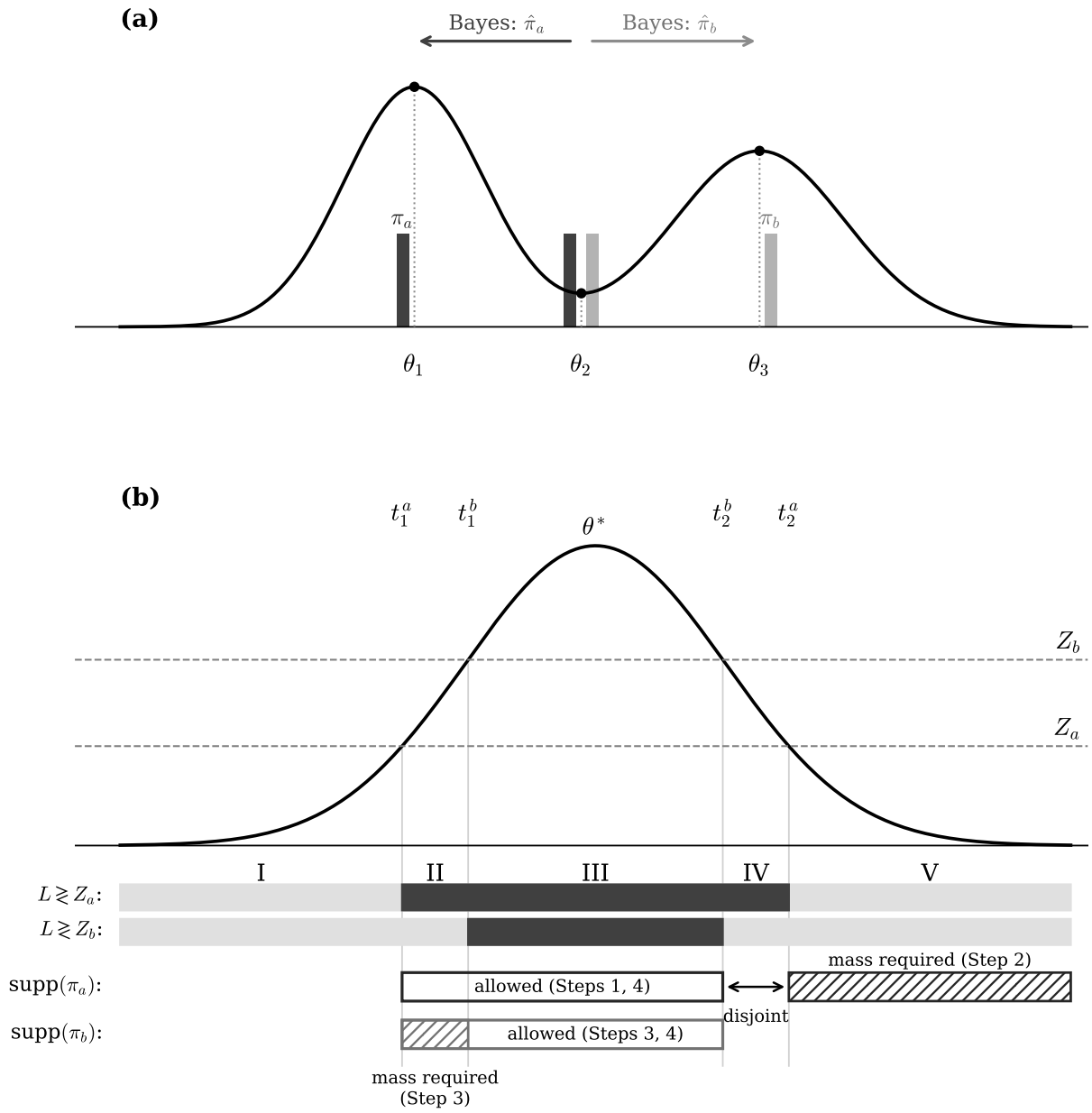


Figure 4: Proof of Theorem 1. Panel (a), sufficiency: a multimodal L with $\theta_1 < \theta_2 < \theta_3$, the two-point priors π_a (on $\{\theta_1, \theta_2\}$) and π_b (on $\{\theta_2, \theta_3\}$), and the direction in which Bayes moves each prior's valley mass. Panel (b), necessity: a unimodal L with the two above-average regions drawn in the nested configuration that unimodality forces, shown for $Z_a < Z_b$; the mirror case is symmetric. The four endpoints partition Θ into regions I through V: I $= (-\infty, t_1^a)$, II $= [t_1^a, t_1^b)$, III $= [t_1^b, t_2^b]$, IV $= (t_2^b, t_2^a]$, V $= (t_2^a, \infty)$. The two strips mark where Bayes amplifies (dark, $L \geq Z$) or erodes (light, $L < Z$) each agent's prior mass. The outlined bars are the supports permitted by Steps 1, 3, and 4; the hatched zones are the masses required by Steps 2 and 3. Agent b's required zone lies inside its permitted support; agent a's lies strictly outside it, which is the contradiction of Step 6.

Step 4: FOSD 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 FOSD-polarize at x (Figure 4(a)). □

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 FOSD-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*

$$A_\pi := \{\theta \in \Theta : L(\theta) \geq Z_\pi\} = \Theta \cap [t_1^\pi, t_2^\pi],$$

where t_1^π and t_2^π are its leftmost and rightmost points. Under unimodality, A_π is a single order interval in Θ containing the peak θ^*. (By the maintained regularity of Section 2, this superlevel set is compact and contains its endpoints.) For brevity below, we refer to $[t_1^\pi, t_2^\pi]$ as the above-average interval, with intersection with Θ understood. Outside this region, $L < Z_\pi$, 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, which partition Θ into the regions I through V of Figure 4(b).

Suppose for contradiction that priors $\pi_a \preceq_{\text{FOSD}} \pi_b$ FOSD-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 any $t < t_1^a$:

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

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

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 4(b), region I).

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 4(b), region V).

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 4(b)): 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.

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 $\text{supp}(\pi_a) \subseteq (-\infty, t_2^b]$.

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 $[t_1^a, t_2^b]$, regions II and III (Figure 4(b), the outlined bars).

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. In the drawn configuration of Figure 4(b), the clash is visible directly: Step 2 requires π_a-mass in region V while Steps 1 and 4 confine $\text{supp}(\pi_a)$ to regions II and III. □

A.2 Proof of Proposition 1

Proposition 1. *Let $\tilde{L}(x \mid \cdot)$ be multimodal. For any prior π with $Z_\pi := \mathbb{E}_\pi[\tilde{L}] > 0$, let t_1^π and t_2^π denote the leftmost and rightmost points of $\{\theta : \tilde{L}(x \mid \theta) \geq Z_\pi\}$.*

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

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

(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) is immediate from the necessity argument for Theorem 1. Step 1 there shows that a strict downward FOSD shift $\hat{\pi}_a \prec_{\text{FOSD}} \pi_a$ forces $\text{supp}(\pi_a) \subseteq [t_1^a, \infty)$: below t_1^a we have $\tilde{L} < Z_a$, so the Bayes factor $\tilde{L}/Z_a < 1$ erodes any mass there and the posterior CDF cannot dominate the prior CDF, contradicting the downward shift. The mirror of Step 1 gives $\text{supp}(\pi_b) \subseteq (-\infty, t_2^b]$. Neither uses unimodality of $\tilde{L}$, which enters the Theorem 1 necessity proof only at its later steps.

Part (ii) is a perturbation argument. Fix a target tolerance $\varepsilon > 0$. We construct full-support priors with FOSD violations below this tolerance by mixing an FOSD-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 an FOSD-polarizing pair $\pi_a^\star \preceq_{\text{FOSD}} \pi_b^\star$; the sufficiency construction in Appendix A.1 produces such a pair with bounded support. Choose a full-support distribution V on Θ with finite first moment, choose $\theta^\circ \in \Theta$ with $\tilde{L}(x \mid \theta^\circ) > 0$, and set $U = (1-\eta)V + \eta\delta_{\theta^\circ}$ for any fixed $\eta \in (0,1)$. Then U has full support and finite first moment, while

$$Z_U := \mathbb{E}_U[\tilde{L}] \geq \eta \tilde{L}(x \mid \theta^\circ) > 0.$$

Since $\tilde{L}(x \mid \cdot)$ is bounded, the posterior $\hat{U}$ also has a finite first moment. For a mixing weight $\delta \in (0,1)$ define

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

Each π_i^δ has full support since U does.

Step 2: FOSD ordering is preserved. Mixture preserves the FOSD ordering between the priors:

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

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

Step 3: Posteriors and means are continuous in δ. The marginal likelihood normalizes the mixture as

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

where $Z_\pi := \mathbb{E}_\pi[\tilde{L}]$. Since $w_i(\delta) \to 0$ as $\delta \to 0$, the posterior means satisfy $\hat{\mu}_i^\delta = (1 - w_i(\delta))\hat{\mu}_i^\star + w_i(\delta)\hat{\mu}_U \to \hat{\mu}_i^\star$. Prior means are linear in δ: $\mu_i^\delta = (1-\delta)\mu_i^\star + \delta\mathbb{E}_U[\theta]$, so $\mu_i^\delta \to \mu_i^\star$ as well.

Step 4: Mean polarization persists for small δ. The two strict outward inequalities $\hat{\mu}_a^\star < \mu_a^\star$ and $\mu_b^\star < \hat{\mu}_b^\star$ persist for all sufficiently small δ by Step 3. The middle ordering is preserved exactly because the same U is mixed into both priors:

$$\mu_b^\delta - \mu_a^\delta = (1-\delta)(\mu_b^\star - \mu_a^\star) \geq 0.$$

Hence $\hat{\mu}_a^\delta < \mu_a^\delta \leq \mu_b^\delta < \hat{\mu}_b^\delta$ for all sufficiently small δ.

Step 5: FOSD violations vanish with δ. At $\delta = 0$, $\hat{\pi}_a^\star \prec_{\text{FOSD}} \pi_a^\star$ and $\pi_b^\star \prec_{\text{FOSD}} \hat{\pi}_b^\star$, so

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

Moreover,

$$\sup_t |F_{\pi_i^\delta}(t) - F_{\pi_i^\star}(t)| \leq \delta, \qquad \sup_t |F_{\hat{\pi}_i^\delta}(t) - F_{\hat{\pi}_i^\star}(t)| \leq w_i(\delta) = O(\delta).$$

The triangle inequality therefore bounds each FOSD-violation supremum by $\delta + w_i(\delta)$, which tends to zero with δ. Choose δ small enough that the mean inequalities in Step 4 hold and both violation suprema are below the prescribed tolerance ε, and denote the resulting priors by $\pi_a^\varepsilon, \pi_b^\varepsilon$. $\square$

A.3 Proof of Theorem 2

Theorem 2. *Mean polarization is possible under signal structure (Θ, Λ, L, F) with Normal priors of shared variance, $\pi_a = \mathcal{N}(\mu_a, \sigma^2)$ and $\pi_b = \mathcal{N}(\mu_b, \sigma^2)$, 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.

Let $f_{\sigma^2}(s) = \mathcal{N}(s; 0, \sigma^2)$ denote the Gaussian density. Substituting $\theta = s + \mu$ in the posterior mean, the posterior mean shift for Normal prior $\mathcal{N}(\mu, \sigma^2)$ is $g(\mu) := \hat{\mu} - \mu = \varphi(\mu)/Z(\mu)$ where

$$\varphi(\mu) = \int s\, f_{\sigma^2}(s)\, \tilde{L}(x \mid s + \mu)\, ds, \qquad Z(\mu) = \int f_{\sigma^2}(s)\, \tilde{L}(x \mid s + \mu)\, ds > 0.$$

Using the Gaussian identity $s f_{\sigma^2}(s) = -\sigma^2 f'_{\sigma^2}(s)$ and integrating by parts,

$$\varphi(\mu) = -\sigma^2 \int f'_{\sigma^2}(s)\, \tilde{L}(x \mid s + \mu)\, ds = \sigma^2 \int f_{\sigma^2}(s)\, \tilde{L}'(x \mid s + \mu)\, ds = \sigma^2\, (f_{\sigma^2} * \tilde{L}')(\mu),$$

where $\tilde{L}'$ denotes the a.e. derivative with respect to θ; the boundary term $\big[-\sigma^2 f_{\sigma^2}(s)\, \tilde{L}(x \mid s + \mu) \big]_{-\infty}^{\infty}$ vanishes because $\tilde{L}$ is bounded and f_{σ^2} has Gaussian decay, and the last equality uses the symmetry of f_{σ^2}. Since $Z > 0$, $\operatorname{sign}(g) = \operatorname{sign}(\varphi)$, and mean polarization at x given $\mathcal{N}(\mu_a, \sigma^2)$ and $\mathcal{N}(\mu_b, \sigma^2)$ requires $\varphi(\mu_a) < 0 < \varphi(\mu_b)$ with $\mu_a < \mu_b$, that is, a $(-, +)$ sign change of $f_{\sigma^2} * \tilde{L}'$.

Note that the Gaussian density is a Pólya frequency density of order ∞, so by the variation-diminishing theorem of Karlin (1968), $f_{\sigma^2} * h$ has at most as many sign changes as h for any integrable function h.

Sufficiency. Suppose $\tilde{L}(x \mid \cdot)$ is multimodal, so by Definition 3 there are $\theta_1 < \theta_2 < \theta_3$ with $\tilde{L}(x \mid \theta_2) < \tilde{L}(x \mid \theta_1)$ and $\tilde{L}(x \mid \theta_2) < \tilde{L}(x \mid \theta_3)$. By absolute continuity, $\int_{\theta_1}^{\theta_2} \tilde{L}'\, d\theta = \tilde{L}(x \mid \theta_2) - \tilde{L}(x \mid \theta_1) < 0$, so $\tilde{L}' < 0$ on a subset of (θ_1, θ_2) of positive Lebesgue measure; since almost every point is a Lebesgue point of the integrable function $\tilde{L}'$, that subset contains a Lebesgue point μ_a, with $\tilde{L}'(\mu_a) < 0$. Symmetrically, (θ_2, θ_3) contains a Lebesgue point μ_b of $\tilde{L}'$ with $\tilde{L}'(\mu_b) > 0$. Then $\mu_a < \theta_2 < \mu_b$. As $\sigma \to 0$, the Gaussian kernels f_{σ^2} form an approximate identity with radially decreasing integrable majorant, so $(f_{\sigma^2} * \tilde{L}')(\mu) \to \tilde{L}'(\mu)$ at every Lebesgue point μ of $\tilde{L}'$. For σ small enough, therefore, $\varphi(\mu_a) < 0 < \varphi(\mu_b)$, and the pair $\mathcal{N}(\mu_a, \sigma^2), \mathcal{N}(\mu_b, \sigma^2)$ mean-polarizes.

Necessity. Suppose $\tilde{L}(x \mid \cdot)$ is unimodal. By Definition 3 this means $\tilde{L}(x \mid \theta_2) \geq \min\{\tilde{L}(x \mid \theta_1), \tilde{L}(x \mid \theta_3)\}$ for all $\theta_1 < \theta_2 < \theta_3$, that is, $\tilde{L}(x \mid \cdot)$ is quasiconcave. By the maintained regularity its maximum is attained at some θ^*, and quasiconcavity then makes it non-decreasing on $(-\infty, \theta^*]$ and non-increasing on $[\theta^*, \infty)$, so $\tilde{L}'$ has at most one sign change, from $+$ to $-$. By the variation-

diminishing theorem, $f_{\sigma^2} * \tilde{L}'$ has at most one sign change. Since $g_\sigma := f_{\sigma^2} * \tilde{L}$ is nonnegative, not identically zero, and vanishes at both tails, its derivative $g'_\sigma = f_{\sigma^2} * \tilde{L}'$ cannot have a single $(-,+)$ sign change; any sign change must therefore have the $(+,-)$ orientation. Hence φ has no $(-,+)$ crossing, and no pair $\mu_a < \mu_b$ satisfies $\varphi(\mu_a) < 0 < \varphi(\mu_b)$. So no pair of Normal priors with shared variance mean-polarizes at $\tilde{L}(x \mid \cdot)$. □

A.4 Proof of Theorem 3

We first isolate the eventual-unimodality property behind the finiteness claim in part (iii).

Lemma 1. *Let h be a nonnegative integrable function on $\mathbb{R}$, not almost everywhere zero, and suppose either*

(i) h vanishes outside a bounded interval $[m, M]$, or

(ii) $h = \sum_{i=1}^{k} w_i f_{v_i}(\cdot - m_i)$ is a finite mixture of Gaussian densities with weights $w_i > 0$, variances $v_i > 0$, and means $m_i \in [m, M]$.

Then for every $\sigma \geq M - m$, the convolution $f_{\sigma^2} * h$ is strictly increasing up to a unique mode and strictly decreasing beyond it; in particular it is unimodal.

Proof. Write $g := f_{\sigma^2} * h$ and $\psi := \sigma^2 g'$.

In case (i), differentiation under the integral (justified by dominated convergence, since $h \in L^1$ and the derivatives of f_{σ^2} are bounded) gives

$$\psi(\theta) = \int (u - \theta)\, f_{\sigma^2}(\theta - u)\, h(u)\, du, \qquad \psi'(\theta) = \int \Big[\frac{(u-\theta)^2}{\sigma^2} - 1\Big] f_{\sigma^2}(\theta - u)\, h(u)\, du.$$

For $\theta \leq m$ we have $u - \theta \geq 0$ for all $u \in [m, M]$, strictly on a set of positive h-measure, so $\psi(\theta) > 0$; symmetrically $\psi(\theta) < 0$ for $\theta \geq M$. For $\theta \in [m, M]$ and $u \in [m, M]$ we have $|u - \theta| \leq M - m \leq \sigma$, so the bracket is nonpositive, vanishing only where $|u - \theta| = \sigma$, a set of u of Lebesgue measure zero; hence $\psi'(\theta) < 0$ on $[m, M]$. So ψ is strictly positive on $(-\infty, m]$, strictly decreasing on $[m, M]$, and strictly negative on $[M, \infty)$: it has a unique zero $\hat{\theta} \in (m, M)$, with $\psi > 0$ to its left and $\psi < 0$ to its right, which is the claim for g.

In case (ii), take without loss of generality $m = \min_i m_i$ and $M = \max_i m_i$, which can only shrink $M - m$ and so preserves $\sigma \geq M - m$. The Gaussian semigroup gives $g = \sum_i w_i f_{v_i + \sigma^2}(\cdot - m_i)$, so

$$\psi(\theta) = \sigma^2 \sum_i \frac{w_i (m_i - \theta)}{v_i + \sigma^2} f_{v_i + \sigma^2}(\theta - m_i), \qquad \psi'(\theta) = \sigma^2 \sum_i \frac{w_i}{v_i + \sigma^2} \Big[\frac{(m_i - \theta)^2}{v_i + \sigma^2} - 1\Big] f_{v_i + \sigma^2}(\theta - m_i).$$

If $m = M$, then g is a mixture of Gaussian densities with a common center and is trivially of the claimed shape. If $m < M$, then for $\theta \leq m$ every summand of $\psi(\theta)$ is nonnegative and any summand with $m_i > m$ is strictly positive, so $\psi(\theta) > 0$; symmetrically $\psi < 0$ on $[M, \infty)$. For $\theta \in [m, M]$, $(m_i - \theta)^2 \leq (M - m)^2 \leq \sigma^2 < v_i + \sigma^2$, so every bracket is strictly negative and $\psi'(\theta) < 0$ there. The conclusion follows as in case (i). □

Theorem 3. *Fix x with $\tilde{L}(x \mid \cdot)$ multimodal, and Normal priors $\pi_a = \mathcal{N}(\mu_a, \sigma^2)$, $\pi_b = \mathcal{N}(\mu_b, \sigma^2)$ with $\mu_a < \mu_b$.*

(i) *Mean polarization occurs at x given π_a, π_b if and only if $(\mu_a, \mu_b, \sigma^2) \in P(\tilde{L})$.*

(ii) *$P(\tilde{L})$ is open and non-empty.*

(iii) *$P(\tilde{L}) \subseteq \{\sigma^2 < \sigma^2_{\text{crit}}(\tilde{L})\}$, and if $\tilde{L}(x \mid \cdot)$ has compact support or is a finite mixture of Gaussian densities, then $\sigma^2_{\text{crit}}(\tilde{L}) < \infty$.*

Criterion. The identity $\varphi(\mu) = \sigma^2(f_{\sigma^2} * \tilde{L}')(\mu)$ and $\text{sign}(g) = \text{sign}(\varphi)$ from the proof of Theorem 2 give that mean polarization at x given π_a, π_b holds iff $\varphi(\mu_a) < 0 < \varphi(\mu_b)$, iff $(\mu_a, \mu_b, \sigma^2) \in P(\tilde{L})$ (using $\sigma^2 > 0$ and $\mu_a < \mu_b$). This is part (i).

Openness and non-emptiness. The map $(\mu, \sigma^2) \mapsto (f_{\sigma^2} * \tilde{L}')(\mu)$ is continuous, so $P(\tilde{L})$, cut out by strict inequalities, is open. Non-emptiness is the sufficiency argument for Theorem 2.

The inclusion $P(\tilde{L}) \subseteq \{\sigma^2 < \sigma^2_{\text{crit}}\}$. Write $\varphi_\sigma := f_{\sigma^2} * \tilde{L}'$. If $(\mu_a, \mu_b, \sigma_0^2) \in P(\tilde{L})$, then $\varphi_{\sigma_0}(\mu_a) < 0 < \varphi_{\sigma_0}(\mu_b)$ with $\mu_a < \mu_b$, so φ_{σ_0} has a $(-,+)$ sign change and $\sigma_0^2 \le \sigma^2_{\text{crit}}$ by the definition of the supremum. The inequality is strict: the map $(\mu, \sigma^2) \mapsto \varphi_\sigma(\mu)$ is continuous, so both strict inequalities at μ_a and μ_b persist for all variances in a neighborhood of σ_0^2; variances slightly above σ_0^2 therefore also carry a $(-,+)$ sign change, and the supremum strictly exceeds σ_0^2. (The same argument shows the supremum is never attained by a variance carrying a $(-,+)$ crossing.) This proves the inclusion, with no hypothesis on $\tilde{L}(x \mid \cdot)$ beyond the maintained regularity and smoothness conditions.

Finiteness of σ^2_{crit} under the added hypothesis. By the Gaussian semigroup $f_{\sigma_1^2+\sigma_2^2} = f_{\sigma_1^2} * f_{\sigma_2^2}$, for $\sigma'^2 > \sigma^2$ we have $\varphi_{\sigma'} = f_{\sigma'^2-\sigma^2} * \varphi_\sigma$, so by the variation-diminishing theorem the number of sign changes of φ_σ is non-increasing in σ^2. Suppose $\tilde{L}(x \mid \cdot)$ has compact support or is a finite mixture of Gaussian densities. Lemma 1 applies to $h = \tilde{L}(x \mid \cdot)$, in the compactly supported case directly and in the mixture case with $[m, M]$ the convex hull of the component means, and yields $\bar{\sigma}^2 < \infty$ at which $g_{\bar{\sigma}} = f_{\bar{\sigma}^2} * \tilde{L}$ is strictly increasing up to a unique mode and strictly decreasing beyond it. Its derivative $\varphi_{\bar{\sigma}}$ (the identity $g'_\sigma = f_{\sigma^2} * \tilde{L}'$ is from the proof of Theorem 2) then has the single sign pattern $(+,-)$. For every $\sigma^2 \ge \bar{\sigma}^2$, variation diminishing therefore leaves φ_σ with at most one sign change. Since g_σ is nonnegative and vanishes at both tails, that sign change cannot have the $(-,+)$ orientation. Hence $\sigma^2_{\text{crit}} \le \bar{\sigma}^2 < \infty$. Small variances do carry a $(-,+)$ crossing, by the sufficiency argument for Theorem 2, so $\sigma^2_{\text{crit}} \in (0, \infty)$. □

A.5 Proof of Proposition 2

Proposition 2. *Fix x with $\tilde{L}(x \mid \cdot)$ bimodal, with peaks $\theta_1 < \theta_3$ and valley θ_2: strictly increasing on $(-\infty, \theta_1]$, strictly decreasing on $[\theta_1, \theta_2]$, strictly increasing on $[\theta_2, \theta_3]$, and strictly decreasing on $[\theta_3, \infty)$. Consider Normal priors with shared variance σ^2 and means $\mu_a < \mu_b$, and write $g_\sigma := f_{\sigma^2} * \tilde{L}(x \mid \cdot)$ for the smoothed likelihood.*

(i) *Some pair mean-polarizes at variance σ^2 if and only if $\sigma^2 < \sigma^2_{\text{crit}}(\tilde{L})$.*

(ii) *For every $\sigma^2 < \sigma^2_{\text{crit}}(\tilde{L})$, the smoothed likelihood g_σ is bimodal, with peaks $a_\sigma < c_\sigma$ and valley b_σ, and, off the isolated zeros of g'_σ, the pair mean-polarizes at variance σ^2 if and only if $a_\sigma < \mu_a < b_\sigma < \mu_b < c_\sigma$. For $\sigma^2 \geq \sigma^2_{\text{crit}}(\tilde{L})$, g_σ is unimodal.*

(iii) *If $\theta_1 < \mu_a < \theta_2 < \mu_b < \theta_3$, the pair mean-polarizes at all sufficiently small σ^2; if $\mu_a \notin [\theta_1, \theta_2]$ or $\mu_b \notin [\theta_2, \theta_3]$, it does not mean-polarize at all sufficiently small σ^2.*

Throughout, fix x, abbreviate $\tilde{L} := \tilde{L}(x \mid \cdot)$, and write $\varphi_\sigma := f_{\sigma^2} * \tilde{L}'$, so that $g'_\sigma = \varphi_\sigma$ and, by the criterion of Theorem 3(i), the pair $\mathcal{N}(\mu_a, \sigma^2), \mathcal{N}(\mu_b, \sigma^2)$ with $\mu_a < \mu_b$ mean-polarizes at x if and only if $\varphi_\sigma(\mu_a) < 0 < \varphi_\sigma(\mu_b)$. Under the maintained smoothness conditions, φ_σ is real-analytic (a Gaussian convolution) and not identically zero, so its zeros are isolated, and g_σ is nonnegative, not identically zero, and vanishes at $\pm\infty$.

Step 1: the sign pattern of φ_σ is $(+,-)$ or $(+,-,+,-)$. The strict bimodality of $\tilde{L}$ gives $\tilde{L}'$ exactly three sign changes, in the order $(+,-,+,-)$, so by the variation-diminishing theorem φ_σ has at most three. φ_σ has at least one sign change, since $\int \varphi_\sigma = g_\sigma(\infty) - g_\sigma(-\infty) = 0$ while $\varphi_\sigma \not\equiv 0$. If the first sign of φ_σ were $-$, then g_σ would strictly decrease from its limit of 0 at $-\infty$ and turn negative, contradicting $g_\sigma \geq 0$; symmetrically, if the last sign were $+$, then g_σ would be negative near $+\infty$. So the sign sequence begins with $+$, ends with $-$, and has an odd number of changes: the pattern is $(+,-)$ or $(+,-,+,-)$.

Step 2: the bimodal pattern holds exactly below σ^2_{crit}. A $(-,+)$ sign change occurs precisely in the $(+,-,+,-)$ pattern, so, given Step 1, $\sigma^2_{\text{crit}}(\tilde{L})$ is the supremum of the variances at which φ_σ has the pattern $(+,-,+,-)$, finite or not. By the Gaussian semigroup and the variation-diminishing theorem, as in the proof of Theorem 3, the number of sign changes of φ_σ is non-increasing in σ^2, so the set of variances with the $(+,-,+,-)$ pattern is an interval with left endpoint 0; it is non-empty because small σ retains the valley's $(-,+)$ crossing (sufficiency argument for Theorem 2). When σ^2_{crit} is finite, the right endpoint is excluded: if the pattern held at σ^2_{crit}, there would be $\mu_a < \mu_b$ with $\varphi_\sigma(\mu_a) < 0 < \varphi_\sigma(\mu_b)$ at $\sigma^2 = \sigma^2_{\text{crit}}$, and the continuity of $(\mu, \sigma^2) \mapsto \varphi_\sigma(\mu)$ would preserve both strict inequalities at slightly larger variances, contradicting the supremum. Hence φ_σ has the pattern $(+,-,+,-)$ for $\sigma^2 < \sigma^2_{\text{crit}}$ and the pattern $(+,-)$ for $\sigma^2 \geq \sigma^2_{\text{crit}}$.

Step 3: parts (i) and (ii). For $\sigma^2 \geq \sigma^2_{\text{crit}}$, the $(+,-)$ pattern puts every negative value of φ_σ to the right of every positive value, so no pair $\mu_a < \mu_b$ satisfies the criterion, and g_σ increases to its single flip point and decreases thereafter, hence unimodal. For $\sigma^2 < \sigma^2_{\text{crit}}$, the $(+,-,+,-)$ pattern realizes the criterion (take μ_a in the first negative region and μ_b in the second positive region), proving (i). Its three sign-changing zeros occur at points $a_\sigma < b_\sigma < c_\sigma$; any additional isolated zeros are tangencies and do not alter the sign pattern. Thus, off its zeros, $\varphi_\sigma > 0$ on $(-\infty, a_\sigma)$, $\varphi_\sigma < 0$ on (a_σ, b_σ), $\varphi_\sigma > 0$ on (b_σ, c_σ), and $\varphi_\sigma < 0$ on (c_σ, ∞). Hence g_σ strictly increases on $(-\infty, a_\sigma]$, strictly decreases on $[a_\sigma, b_\sigma]$, strictly increases on $[b_\sigma, c_\sigma]$, and strictly decreases on $[c_\sigma, \infty)$, isolated interior zeros of φ_σ not interrupting strict monotonicity: g_σ is bimodal with peaks a_σ, c_σ and valley b_σ.

The criterion $\varphi_\sigma(\mu_a) < 0 < \varphi_\sigma(\mu_b)$ requires $\mu_a \in (a_\sigma, b_\sigma) \cup (c_\sigma, \infty)$ and $\mu_b \in (-\infty, a_\sigma) \cup (b_\sigma, c_\sigma)$; the ordering $\mu_a < \mu_b$ eliminates every combination except $\mu_a \in (a_\sigma, b_\sigma)$ and $\mu_b \in (b_\sigma, c_\sigma)$, and conversely every such pair off the zeros of φ_σ satisfies the criterion. This is the box in (ii).

Step 4: part (iii). We first show that for any μ strictly inside one of the four monotone segments of $\tilde{L}$, the sign of $\varphi_\sigma(\mu)$ at all sufficiently small σ is that of the segment. Let μ lie in the open segment I on which $\tilde{L}$ is strictly decreasing (the argument on increasing segments is the mirror image), and pick $\delta > 0$ with $[\mu - \delta, \mu + \delta] \subset I$. Strict monotonicity gives $\tilde{L}' \leq 0$ almost everywhere on I and $\int_{\mu-\delta}^{\mu+\delta} \tilde{L}'\, d\theta = \tilde{L}(\mu+\delta) - \tilde{L}(\mu-\delta) =: -c < 0$. Split $\varphi_\sigma(\mu) = \int f_{\sigma^2}(\mu - u)\, \tilde{L}'(u)\, du$ into the window $|u - \mu| \leq \delta$, the remainder of I, and the complement of I. On the window, $f_{\sigma^2}(\mu - u) \geq f_{\sigma^2}(\delta)$ and the integrand is almost everywhere nonpositive, so the window term is at most $-c\, f_{\sigma^2}(\delta)$; the remainder term is nonpositive; and the complement term is at most $f_{\sigma^2}(d)\, \|\tilde{L}'\|_{L^1}$ in absolute value, where $d := \operatorname{dist}(\mu, \mathbb{R} \setminus I) > \delta$, since $f_{\sigma^2}(\mu - u) \leq f_{\sigma^2}(d)$ off I. Hence

$$\varphi_\sigma(\mu) \leq -c\, f_{\sigma^2}(\delta) + f_{\sigma^2}(d)\, \|\tilde{L}'\|_{L^1},$$

and since $f_{\sigma^2}(d)/f_{\sigma^2}(\delta) = e^{-(d^2-\delta^2)/2\sigma^2} \to 0$ as $\sigma \to 0$, we get $\varphi_\sigma(\mu) < 0$ for all sufficiently small σ.

If $\theta_1 < \mu_a < \theta_2 < \mu_b < \theta_3$, then μ_a sits strictly inside the decreasing segment (θ_1, θ_2) and μ_b strictly inside the increasing segment (θ_2, θ_3), so $\varphi_\sigma(\mu_a) < 0 < \varphi_\sigma(\mu_b)$ for all sufficiently small σ, and the pair mean-polarizes by the criterion.

Conversely, suppose $\mu_a \notin [\theta_1, \theta_2]$ or $\mu_b \notin [\theta_2, \theta_3]$; we check in each case that one of the two criterion inequalities fails for all sufficiently small σ. If $\mu_a < \theta_1$ or $\mu_a \in (\theta_2, \theta_3)$, then μ_a lies strictly inside an increasing segment and $\varphi_\sigma(\mu_a) > 0$ eventually. If $\mu_a \geq \theta_3$, then $\mu_b > \theta_3$ lies strictly inside the decreasing segment (θ_3, ∞) and $\varphi_\sigma(\mu_b) < 0$ eventually. If $\mu_b \in (\theta_1, \theta_2)$ or $\mu_b > \theta_3$, then μ_b lies strictly inside a decreasing segment and $\varphi_\sigma(\mu_b) < 0$ eventually; and if $\mu_b \leq \theta_1$, then $\mu_a < \theta_1$, covered above. These cases exhaust the complement of the closed box, completing part (iii). $\square$

References

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

Sandeep Baliga, Eran Hanany, and Peter Klibanoff. Polarization and ambiguity. *American Economic Review*, 103(7):3071–3083, 2013.

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.

Adam Corner, Lorraine Whitmarsh, and Dimitrios Xenias. Uncertainty, scepticism and attitudes towards climate change: Biased assimilation and attitude polarisation. *Climatic Change*, 114 (3–4):463–478, 2012.

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

Ward Edwards. Conservatism in human information processing. In Benjamin Kleinmuntz, editor, *Formal Representation of Human Judgment*, pages 17–52. Wiley, 1968.

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.

Samuel Karlin. *Total Positivity, Vol. I.* Stanford University Press, 1968.

Deanna Kuhn and Joseph Lao. Effects of evidence on attitudes: Is polarization the norm? *Psychological Science*, 7(2):115–120, 1996.

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.

Arthur G. Miller, John W. McHoskey, Cynthia M. Bane, and Timothy G. Dowd. 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, 1993.

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.

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

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.

Charles S. Taber and Milton Lodge. Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3):755–769, 2006.