

Oberlin

Digital Commons at Oberlin

Oberlin College Kasper Economics and
Business Working Paper Series

Economics

8-2026

HKC07 - Bayesian conspiracy theorists? Equivocal signals and persistent polarization

Javier A. Granados Samayoa
Texas Christian University

Timothy Hyde
Department of Economics, Oberlin College

Follow this and additional works at: https://digitalcommons.oberlin.edu/economics_wps



Part of the [Economics Commons](#)

Recommended Citation

Granados Samayoa, Javier A. and Hyde, Timothy, "HKC07 - Bayesian conspiracy theorists? Equivocal signals and persistent polarization" (2026). *Oberlin College Kasper Economics and Business Working Paper Series*. 7.

https://digitalcommons.oberlin.edu/economics_wps/7

This Working Paper is brought to you for free and open access by the Economics at Digital Commons at Oberlin. It has been accepted for inclusion in Oberlin College Kasper Economics and Business Working Paper Series by an authorized administrator of Digital Commons at Oberlin. For more information, please contact mmcfarli@oberlin.edu.



OBERLIN COLLEGE KASPER ECONOMICS AND BUSINESS
WORKING PAPER SERIES

HKC Working Paper No 07
August 2026

Bayesian conspiracy theorists? Equivocal signals and persistent polarization

Javier A. Granados Samayoa, Timothy Hyde

Bayesian conspiracy theorists? Equivocal signals and persistent polarization*

Javier A. Granados Samayoa

Timothy Hyde

August 2026

Abstract

Conspiracy beliefs are the subject of a rich literature in psychology, which explains them through personal dysfunction. The corresponding economics literature is scant. We offer a demand-side model of belief in specific conspiracy theories with no biases, misspecification, or strategic manipulation. A conspiracy, by its nature, generates equivocal evidence about its own existence, because the same strength that produces more to detect also produces better concealment. We model belief in a specific conspiracy theory as learning about the strength q of a cabal, a stand-in for the secretly coordinating actors of any conspiracy theory. In the baseline the cabal influences events with probability q and conceals that influence with probability q , so public evidence arrives at rate $q(1 - q)$ and cannot distinguish a weak cabal from a strong, well-concealed one. Bayesian agents converge to two-point beliefs on the pair of observationally equivalent states, with weights given by the ratio of their prior densities there, and their optimal actions diverge accordingly. Divergence requires only that concealment improve elastically with strength over some range, not exact symmetry. It is also invisible to standard belief elicitation: agents with indistinguishable measured priors can be driven to opposite camps by identical public information. Polarization, opposite responses to common information, and stable adherence follow from Bayes' rule alone. The model provides a complementary account to the one offered by the psychology literature.

Keywords: belief polarization, Bayesian learning, partial identification, merging of opinions, conspiracy theories

JEL Codes: D83, D84, D91

*Javier A. Granados Samayoa: Texas Christian University, j.granadossamayoa@tcu.edu. Timothy Hyde: Oberlin College, thyde@oberlin.edu. We thank participants at the Northeast Ohio Economics Workshop for their attention and feedback. All remaining errors are our own.

1 Introduction

Conspiracy beliefs – beliefs that “two or more actors have coordinated in secret to achieve an outcome and that their conspiracy is of public interest but not public knowledge” (Douglas and Sutton 2023) – command growing attention. Across English-language books, the *New York Times*, and cable news, attention to conspiracy theories rose sharply after 2015, peaked around the COVID-19 pandemic and the 2020 United States election, and remains roughly three times its level of the early 2010s (Figure 1). Simultaneously, rates of belief in individual conspiracy theories tend to be relatively stable over time, at both the population level (Uscinski et al. 2022) and the individual level (Williams et al. 2024), with the population rarely coming to a unanimous conclusion that a theory is right or wrong. This lack of resolution is in tension with the basic intuition about Bayesian updating that a population should converge to the same opinion as more information becomes available (Blackwell and Dubins 1962). A substantial research literature explains conspiracy belief as dysfunctional, rooted in frustrated psychological needs, cognitive deficits, or manipulation by propagandists, and documents its correlation with support for political violence (Enders et al. 2024). Beliefs themselves are rarely modeled as the output of rational inference from the distinctive information environment that a conspiracy creates.

Our object is belief in a specific conspiracy theory, as distinct from the generic conspiracy mentality that psychologists measure separately (Trella et al. 2024). Particular conspiracy theories vary widely in content and valence, but we abstract from that variety and model belief in any one of them along a single continuous dimension, the strength $q \in [0, 1]$ of a “cabal” trying to influence events. When q is near zero, everything is as it seems; when q is near one, the most florid conspiracy theories are correct. Strength has two implications: a stronger cabal does more, which makes it more visible, and hides better, which makes it less visible. We argue that a conspiracy, by its nature, generates equivocal evidence about its own existence, because the same strength that produces more to detect also produces less detection.

Each period, the cabal influences events with probability q and, conditional on doing so, conceals its influence with probability q , so evidence surfaces publicly in a fraction $q(1 - q)$ of periods. This detection rate is symmetric about $q = \frac{1}{2}$. A marginal, mostly inert cabal ($q = 0.1$) and a powerful, disciplined one ($q = 0.9$) generate identical distributions of public evidence. As evidence accumulates, every agent’s posterior collapses onto the *mirror pair* $\{q^*, 1 - q^*\}$, and the weight she places on the conspiratorial branch is the ratio of her prior densities at those two points. Agents agree about the content of the public evidence, and disagree permanently about what that evidence means.

Two further results follow. First, the model yields an elicitation-impossibility corollary that appears to be new to the disagreement literature. Limiting beliefs are governed by prior density

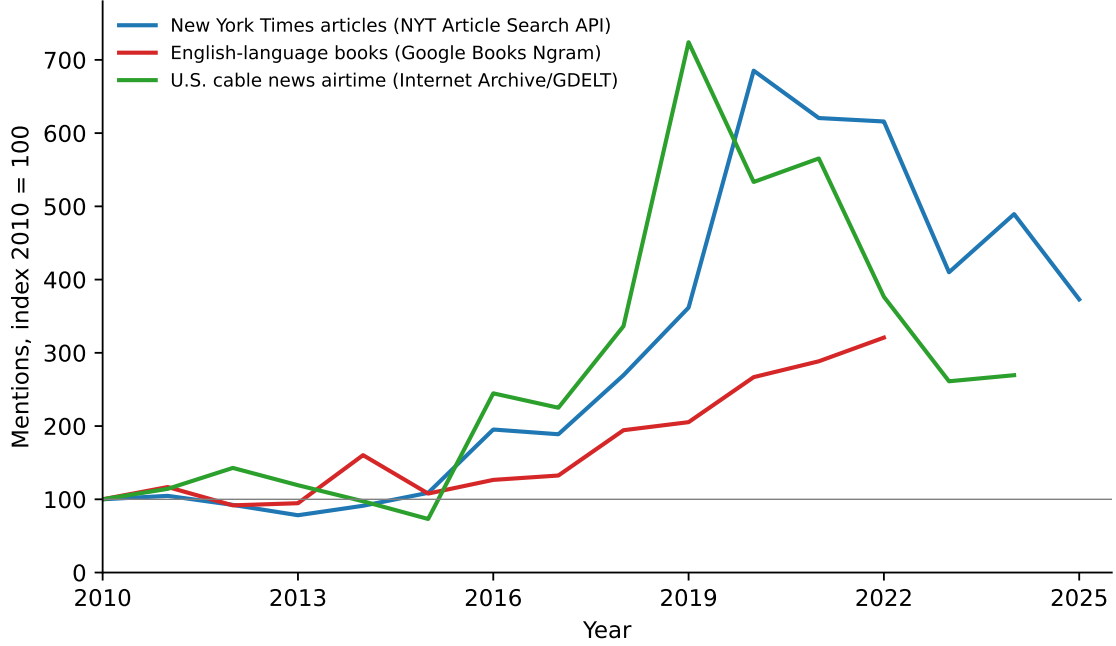


Figure 1: Media attention to conspiracy theories, 2010–2025

Each series counts exact-phrase mentions of “conspiracy theory” or “conspiracy theories” and is indexed to its 2010 value. Books: combined frequency of the two phrases per million words in the Google Books English corpus (Michel et al. 2011). *New York Times*: articles per year containing either phrase, from the NYT Article Search API. U.S. cable news: share of fifteen-second caption blocks on CNN, Fox News, and MSNBC mentioning “conspiracy theory,” authors’ calculations from the Internet Archive TV News Archive via the GDELT Television API.

ratios at ex ante negligible states, so the statistic that determines who polarizes is invisible to any elicitation that measures the prior where its mass lies. Two agents whose priors are indistinguishable to standard instruments can be driven by the same public evidence to opposite camps. Second, we show that the mechanism does not depend on the baseline’s symmetry. With an arbitrary concealment technology, the detection map is non-injective if and only if concealment improves elastically with strength somewhere, and polarization then obtains in cases when the detection level implied by a cabal’s strength is attained at more than one state (Proposition 4).

The contribution of this paper is that observation and the model built on it, a demand-side theory of conspiracy belief from standard ingredients, which the economics literature has lacked. The formal building blocks are established results in Bayesian econometrics and the theory of disagreement, reviewed in Section 2. What is new is the equivocal signal, derived rather than assumed, and derived from the defining property of a conspiracy. Relative to behavioral and strategic accounts, the model is more parsimonious, with no bias, misspecification, or propagandist needed. Relative to the Bayesian-disagreement literature, it is more specific. Deriving the identification failure from concealment is what yields the mirror pair, the tail-ratio weights, and the elicitation result.

The remainder of the paper proceeds as follows. Section 2 situates the paper in the literature and documents the stylized facts that discipline the model. Section 3 presents the model, and Section 4 presents the results and maps them to the stylized facts. Section 5 discusses limitations, policy implications, and extensions, and Section 6 concludes.

2 Related literature and stylized facts

2.1 The psychology of conspiracy belief

A substantial literature in psychology treats conspiracy belief as a symptom to be explained by individual-level dysfunction: threatened needs for safety, certainty, and self-esteem (Douglas et al. 2017, 2019), lower cognitive ability and actively open-minded thinking (Pennycook et al. 2020), schizotypy and paranoia proneness (Bowes et al. 2023), and a battery of cognitive biases in belief formation and updating (Gagliardi 2025). More recently, a growing macro-level literature relates conspiracy belief to societal conditions such as economic inequality (Jetten et al. 2021) and corruption (Alper 2023), relocating the dysfunction from the individual to the society but retaining the symptom framing.

Four empirical regularities organize this literature and discipline the model below. First, beliefs about specific conspiracy theories are polarized: each is endorsed by a substantial minority and rejected by most others (Oliver and Wood 2014; Uscinski et al. 2022). Second, common information can move different people’s beliefs in opposite directions (Lord et al. 1979; Nyhan and Reifler 2010).¹ The same news is read by adherents as confirmation and by skeptics as further reason for dismissal. Third, aggregate rates of adherence are remarkably stable over time. Figure 2 displays survey evidence from Uscinski et al. (2022): the population rate of adherence to a given conspiracy theory, polled years apart, moves down about as often as it moves up. Survey measures are, moreover, inherently limited here. The people whose beliefs are at issue are characterized by distrust, and those who distrust institutions are presumably the least likely to respond when institutions survey them. Fourth, the beliefs are behaviorally consequential, particularly in health domains, where conspiracy beliefs predict lower uptake of treatment and prevention (Thorburn and Bogart 2005; Grebe and Natrass 2012; Bierwiazzonek et al. 2022; Stasielowicz 2022).

2.2 Strategic and behavioral accounts in economics

Economists have studied phenomena adjacent to conspiracy belief, such as the spread and consumption of fake news (Allcott and Gentzkow 2017) and the political effects of the internet and

¹Such backfire effects are far from universal; see Wood and Porter (2019).

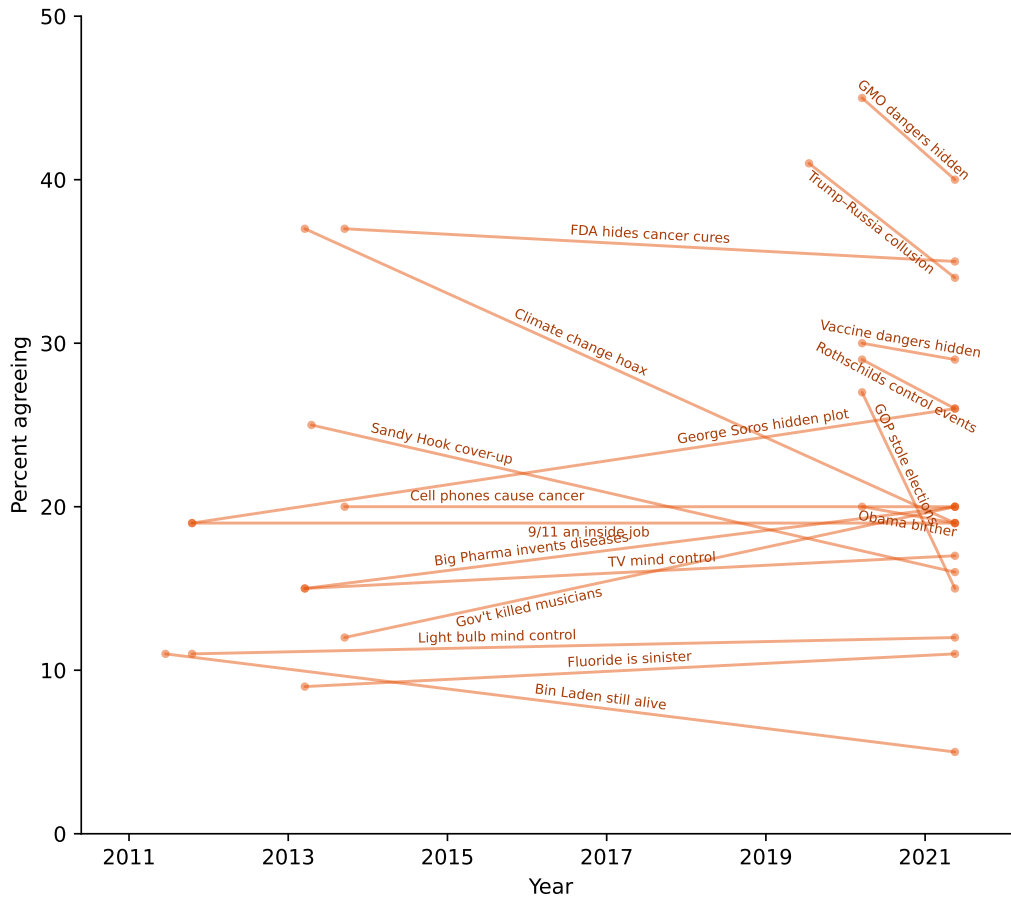


Figure 2: Belief in specific conspiracy theories does not converge to 100% or 0% over time

Each segment plots beliefs in one specific conspiracy theory among U.S. adults, measured at two different points in time. Data are from Table 3 of Uscinski et al. (2022), restricted to the 18 theories polled in May 2021 and at least once earlier, between 2010 and 2020. One additional belief, that the richest one percent control the government and the economy, is omitted for axis scale.

social media (Zhuravskaya et al. 2020), and Sunstein and Vermeule (2009) attribute conspiracy belief informally to a “crippled epistemology” of informational isolation. Formal treatments of conspiracy belief itself are recent and few, and each incorporates a behavioral bias. On the supply side, Szeidl and Szucs (2025) model populism as the strategic dissemination of a conspiratorial “alternative reality” that has zero objective probability, spreads through voters assumed receptive to propaganda, and is amplified by correlation neglect. On the demand side, Kishishita (2025) embeds conspiracy belief in an electoral-competition model in which voters trade accuracy against self-image, so that rising inequality leads the poor to endorse conspiracy theories more strongly, and Aytimur and Kishishita (2025) model conspiracy thinking as a misspecified worldview – one that treats experts as possibly conspiratorial and never mistaken – whose population share evolves under motivated-reasoning payoffs (for related misspecified-worldview and empirical accounts see Levy et al. 2022; Becchetti et al. 2023). Other examples in the Bayesian disagreement literature feature

similar biases: confirmatory bias (Rabin and Schrag 1999), coarse interpretation of ambiguous evidence (Fryer et al. 2019), boundedly rational multidimensional learning (Loh and Phelan 2019), pre-screening of expert credibility (Cheng and Hsiaw 2021), competing misspecified narratives (Eliaz and Spiegler 2020), or small reasoning biases amplified through learned trust (Gentzkow et al. 2025) – mechanisms that locate disagreement in how agents interpret evidence rather than in the structure of the evidence itself. What the economics literature lacks, and this paper supplies, is a demand-side account of what a Bayesian agent, facing the information environment that a specific conspiracy by its nature creates, should come to believe about it.

2.3 Bayesian disagreement and partial identification

The Bayesian econometrics of non-identified and partially-identified models is well studied. Poirier (1998) shows that in a nonidentified model the conditional prior given the identified reduced-form parameter is never revised, and Moon and Schorfheide (2012) show that data concentrate the posterior on the identified set but cannot reallocate mass within it. In the theory of disagreement, Acemoglu et al. (2016) establish that identification problems can generate asymptotic disagreement among Bayesian agents and that agreement is fragile to small perturbations of priors; Dixit and Weibull (2007) show that common information can move rational agents apart; and Danenberg (2025) characterizes the public signals under which Bayesians polarize across many dimensions. Propositions 1 and 2 below are specializations of these established results to the conspiracy environment; what is new is the concealment microfoundation that generates the identification failure, together with the elicitation-impossibility result of Proposition 3. Closest to our interpretive claim, Benoit and Dubra (2019) show that the attitude polarization observed in classic experiments is consistent with unbiased Bayesian updating, so that observed polarization is not by itself proof of bias; we extend that point from experimental snapshots to the long-run limit of learning in the environment a conspiracy itself creates.

A parallel literature in cognitive science and philosophy shows that belief polarization can be normatively Bayesian and that conspiracy belief specifically can be rational (e.g. McCauley and Jacques 1979; Madsen et al. 2017; Doyle 2021; Poth and Dolega 2023; Tsapos 2024; Young et al. 2025), but in each case the disagreement enters through what agents bring to the evidence – trust, hidden variables, auxiliary beliefs, or network position – whereas here agents share a common likelihood and fully public information, and bring only their priors. Closest to our mechanism, Herrick (2026) argues informally that a successful conspiracy drives the likelihood ratio toward one, so that updating stalls; Proposition 1 derives that neutralization from the concealment technology, and Proposition 2 shows it produces divergence, not merely deadlock.

3 Model

Time is discrete, $t = 1, 2, \dots$. A cabal has strength $q \in [0, 1]$, fixed by nature. Agents indexed by i hold heterogeneous prior densities π_i over q (Assumption 2 below). Priors are subjective primitives; no common prior is imposed (Morris 1995). Each period every agent chooses an action $a \in [0, 1]$ and receives utility $-(a - q)^2$, so the optimal action is the posterior mean, $a_{it}^* = E_i[q \mid h_t]$, where h_t is the public history.

Each period the cabal, to the extent it exists, attempts to influence global events and succeeds with probability q . Conditional on influencing events, it attempts to cover its tracks and succeeds with an independent probability q . The world then reveals a public binary signal $s_t \in \{0, 1\}$: evidence of cabal influence surfaces ($s_t = 1$) if and only if the cabal acted and failed to conceal it:

$$\Pr(s_t = 1 \mid q) = q(1 - q) \equiv p(q). \quad (1)$$

Draws are independent across periods. Two assumptions are maintained throughout.

Assumption 1 (Signals). *Conditional on q , the signals $s_1, s_2, \dots$ are independent and identically distributed, with $\Pr(s_t = 1 \mid q) = p(q)$.*

Assumption 2 (Priors). *Each prior π_i is a continuous, strictly positive density on $[0, 1]$.*

We write P_{q^*} for the law of the signal sequence when the true strength is q^* ; “almost surely” always refers to P_{q^*} . Assumption 2 makes Bayes’ rule well defined after every history and is the only restriction placed on priors.

The essential feature of (1) is that detection is not monotone in q : stronger cabals do more, but they hide better. Detection peaks at $q = \frac{1}{2}$ and vanishes at both endpoints: an idle cabal generates no evidence, and so does an omnipotent one. Agents have common knowledge of (1) and update by Bayes’ rule; this is explicitly not a misspecified-models story. The difficulty is that the map from states to signal distributions is not invertible: $p(q) = p(1 - q)$, so the states q and $1 - q$ are observationally equivalent. The identified object is the detection rate $m \equiv q(1 - q) \in [0, \frac{1}{4}]$; the structural parameter q is identified only up to the unordered pair $\{q, 1 - q\}$. Throughout, we refer to this pair—the identified set for q , whose elements are reflections of each other about $q = \frac{1}{2}$ —as the *mirror pair*, and to its elements as *mirror states*.

4 Results

4.1 Prior-odds invariance

Proposition 1 (Prior-odds invariance). *Under Assumptions 1–2, for every history $h_t = (s_1, \dots, s_t)$ and every $q \in [0, 1]$,*

$$\pi_i(q \mid h_t) \pi_i(1 - q) = \pi_i(1 - q \mid h_t) \pi_i(q);$$

that is, for every $q \in (0, 1)$ the posterior odds between the mirror states q and $1 - q$ equal the prior odds, $\pi_i(q \mid h_t) / \pi_i(1 - q \mid h_t) = \pi_i(q) / \pi_i(1 - q)$.

All proofs are in the Appendix. The argument here is one line: the likelihood of any history, $L(h_t \mid q) = p(q)^{k_t} [1 - p(q)]^{t-k_t}$ with $k_t = \sum_{\tau \leq t} s_\tau$, depends on q only through $p(q)$, and $p(q) = p(1 - q)$, so the likelihood ratio between mirror states is one after every history and the normalizing constant cancels.

Proposition 1 is the specialization to this environment of a general fact about nonidentified models: the conditional prior over the structural parameter, given the identified reduced form, is never updated (Poirier 1998; Moon and Schorfheide 2012).

Mechanically, the proposition is the label-switching problem of Bayesian mixture models (Stephens 2000). A likelihood that is invariant under a permutation of its parameters produces symmetric posterior modes among which the data can never adjudicate. What mixture estimation treats as a nuisance to be corrected is, in the conspiracy environment, the phenomenon of interest.

4.2 Limit beliefs and the tail ratio

Proposition 2 (Limit beliefs and the tail ratio). *Let Assumptions 1–2 hold and let the true strength be $q^* \in (0, 1)$ with $q^* \neq \frac{1}{2}$. Then, almost surely, agent i 's posterior converges weakly to the two-point distribution*

$$w_i \delta_{q^*} + (1 - w_i) \delta_{1-q^*}, \quad w_i = \frac{\pi_i(q^*)}{\pi_i(q^*) + \pi_i(1 - q^*)};$$

equivalently, for every sufficiently small $\zeta > 0$, $\Pr_i(|q - q^*| < \zeta \mid h_t) \rightarrow w_i$ almost surely. Her action converges to $a_i^\infty = w_i q^* + (1 - w_i)(1 - q^*)$.

We call w_i agent i 's *limiting branch weight* on q^* . The proof has two steps. First, by the strong law the empirical detection frequency converges to $m^* = p(q^*)$, and the posterior concentrates on the identified set $\{q : p(q) = m^*\} = \{q^*, 1 - q^*\}$. Second, the relative mass on the two roots is settled by a Laplace-type comparison: the likelihood depends on q only through $p(q)$, so after the change of variables $u = p(q)$ near each root the two branch integrals share a common exponential

factor and differ only through the prior density and the local slope $|p'(q_j)|$ at the root. In the baseline the slopes are equal, $|p'(q^*)| = |p'(1 - q^*)| = |1 - 2q^*|$, so the branch weights are exactly $\pi_i(q^*) : \pi_i(1 - q^*)$.

A natural conjecture is that the limiting posteriors, while bimodal, should be approximately symmetric: since agents cannot distinguish $q = 0.1$ from $q = 0.9$, an agent without strong prior views might be expected to assign the two branches roughly equal weight, so that initially disparate priors would converge to common, balanced two-humped posteriors—a poor match for the stylized facts. Proposition 2 shows that this conjecture fails: balanced weights require a tail ratio of exactly one, and generic priors—including mild, unimodal, agnostic ones—are not symmetric in their tails.

Figures 3 and 4 illustrate with two agents whose beta priors have means 0.52 and 0.48. By any conventional description the agents hold very similar beliefs: both are centered near agnosticism, both assign the overwhelming majority of mass to the interior, and both regard $q \leq 0.15$ and $q \geq 0.85$ as remote. The true strength is $q^* = 0.1$, so the true detection level is $m^* = 0.09$: public evidence arrives 9 percent of the time. After 25 periods (Figure 3) the posteriors have flattened and pulled modestly apart. After 600 periods (Figure 4) each posterior has collapsed onto sharp spikes at $q \approx 0.1$ and $q \approx 0.9$ with mirror-image weights: agent 1 places roughly three-quarters of her mass on the conspiratorial branch ($E[q] \approx 0.70$), agent 2 the same mass on the innocent branch ($E[q] \approx 0.30$). A prior gap of 0.04 in means has become a permanent posterior gap of 0.40, dictated entirely by tail densities at states both agents began by dismissing. And both agents now agree, with near certainty, that $m^* \approx 0.09$: they cite the same facts. Nothing in the figures depends on which element of the mirror pair is true—the same signal distribution, posteriors, and camps arise under $q^* = 0.9$ —and the stipulation $q^* = 0.1$ determines only which agent’s limiting action is, unknowably to her, closer to the truth.

This polarization has real-world consequences. Because each agent’s optimal action is her posterior mean, the limiting actions are $a_i^\infty = w_i q^* + (1 - w_i)(1 - q^*)$, so two agents who have watched the same history forever act as if they live in different worlds: in the example, agent 1 behaves as though $q \approx 0.70$ and agent 2 as though $q \approx 0.30$ and neither will ever revise. The gap between their actions, $|w_1 - w_2| |1 - 2q^*|$, is permanent, and it is widest precisely when the truth is least ambiguous: as q^* approaches zero, a world with essentially no cabal, the believer’s action approaches one. The environment thus delivers the fourth stylized fact together with the first three: persistent disagreement in beliefs translates one-for-one into persistent disagreement in behavior, and the most innocent worlds produce the widest behavioral gulf.

The same limit that splits agents on q unites them on everything the data identify: every agent’s belief about the detection rate m converges to m^* , so adherents and skeptics come to agree, with arbitrary precision, about how often evidence of the cabal surfaces. This is the precise sense in which the model complements rather than contradicts Blackwell and Dubins (1962): merging

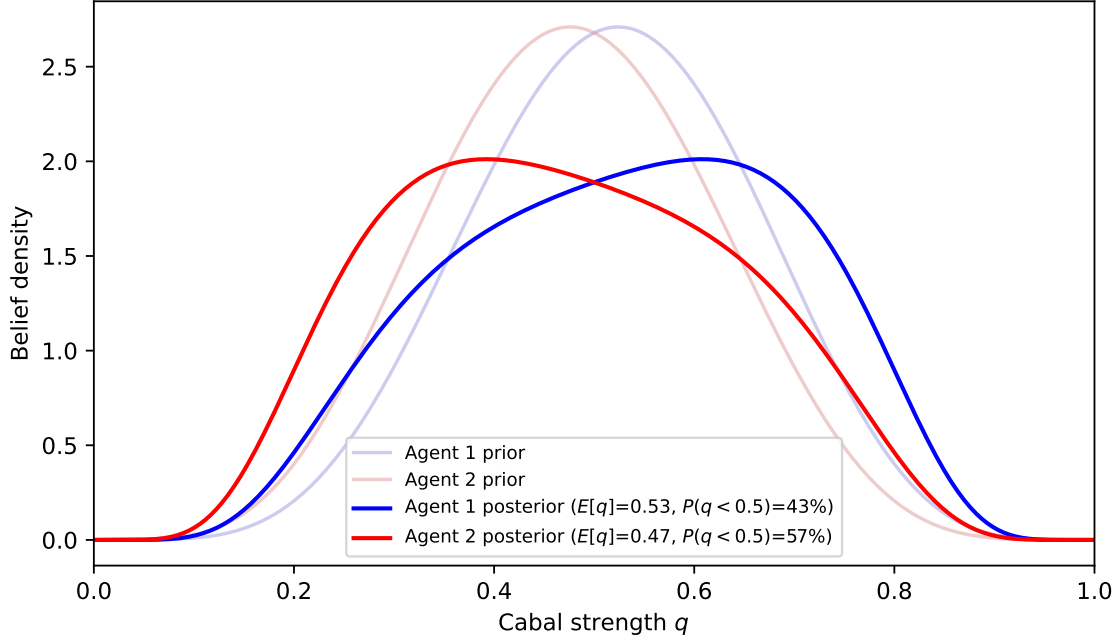


Figure 3: Priors and posteriors after 25 periods, true $q^* = 0.1$

Two agents with beta priors with means 0.52 (agent 1) and 0.48 (agent 2) observe an identical public signal history generated by Equation 1 with $q^* = 0.1$. Posteriors are computed by Bayes' rule on a grid. The two priors are close by any standard elicited summary — mean, median, credible intervals — and Proposition 3 shows they can be made arbitrarily close without changing the limiting outcome.

obtains on the observables, and the disagreement that survives lives entirely on the identified set. Nor does watching one another help: because priors are subjective primitives rather than private information about q , another agent's actions are a function of the public history and her own prior, so observing them reveals at most that prior—which, by Proposition 1, carries no information about the state.

4.3 Polarization is invisible ex ante

The tail ratio is a feature of the prior that conventional belief elicitation does not measure. Standard elicitation instruments collect means, medians, probabilities of intervals, and confidence statements (Manski 2004)—functionals dominated by where the prior's mass sits. Proposition 3 states the resulting impossibility: agents indistinguishable to any such elicitation can be driven arbitrarily far apart by identical public data. Throughout, $\|\pi - \pi'\|_{TV} = \frac{1}{2} \int_0^1 |\pi(q) - \pi'(q)| dq$ denotes total variation distance, and an *elicitation functional* is any real-valued functional of the prior that is continuous with respect to $\|\cdot\|_{TV}$; means, interval probabilities, and quantiles at points where the density is bounded below all qualify.

Proposition 3 (Elicitation impossibility). *Let Assumption 1 hold. Fix $q^* \in (0, 1)$ with $q^* \neq \frac{1}{2}$,*

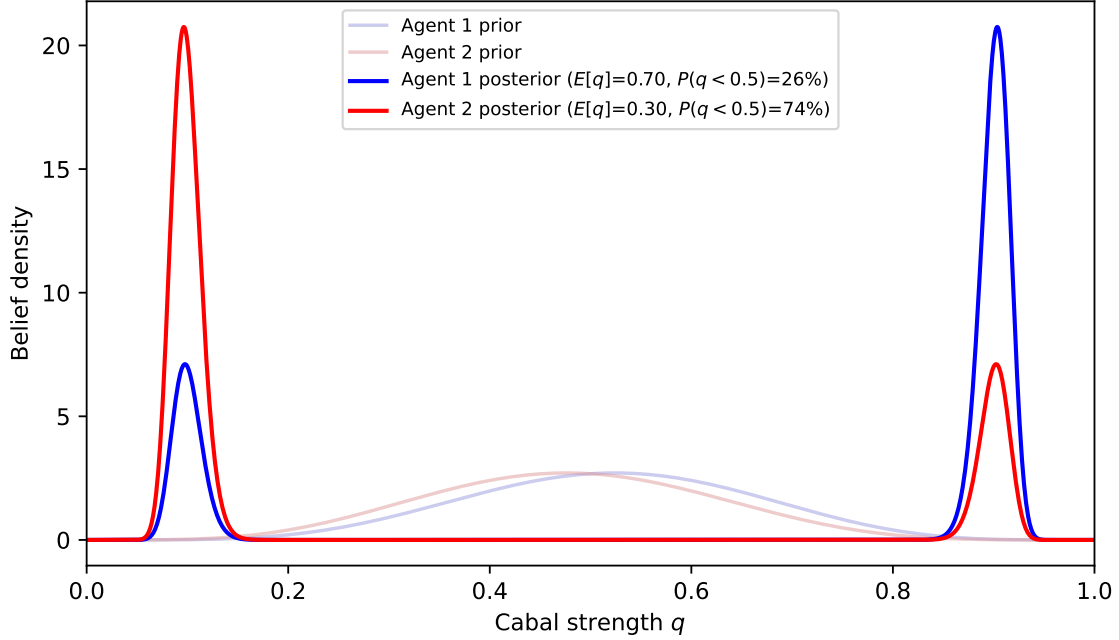


Figure 4: Priors and posteriors after 600 periods, true $q^* = 0.1$

Same two agents and signal history as Figure 3, extended to 600 periods. Posteriors concentrate on the mirror pair $\{0.1, 0.9\}$ with mirror-image branch weights equal to each agent's prior tail ratio (Proposition 2).

any $\delta > 0$, and any target branch weights $w_1, w_2 \in (0, 1)$. There exist priors π_1, π_2 satisfying Assumption 2 with $\|\pi_1 - \pi_2\|_{TV} < \delta$ such that agent i 's limiting branch weight on q^* (Proposition 2) is w_i , $i = 1, 2$. Consequently, for any finite collection of elicitation functionals $F_1, \dots, F_n$ and any $\gamma > 0$, the priors can be chosen so that $|F_\ell(\pi_1) - F_\ell(\pi_2)| < \gamma$ for every ℓ , while the agents' limiting actions differ by $|w_1 - w_2| |1 - 2q^*|$, which can be made arbitrarily close to $|1 - 2q^*|$.

The construction adds to a common base prior a narrow bump at q^* or at $1 - q^*$ whose height sets the density ratio at the mirror pair to $w_i : (1 - w_i)$. The bump's total mass is proportional to its width, so shrinking the width drives the two priors together in total variation while, by Proposition 2, leaving the limiting weights untouched.

The interpretation is stark. Who radicalizes is determined by relative openness to two scenarios that both agents may sincerely dismiss—densities of order 10^{-6} against 10^{-8} settle the question as decisively as 0.6 against 0.4. No elicitation that measures the prior where its mass is can predict the outcome, and respondents asked directly would deny any meaningful difference between themselves. In this model, the empirical difficulty of predicting radicalization from observables is not noise; it is an implication of Proposition 3.

Proposition 3 sharpens the fragility result of Acemoglu et al. (2016) into a statement about observability. In their environment, agents face uncertainty about the conditional distributions of signals, and asymptotic agreement fails discontinuously in the prior. Here the signal distribution is

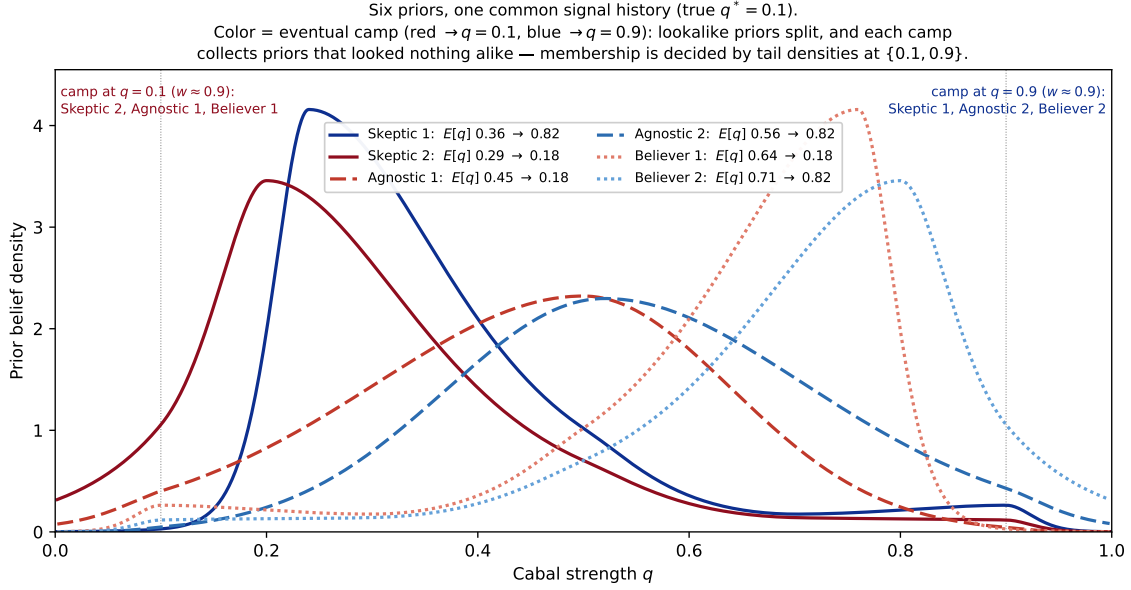


Figure 5: Six priors, one common signal history, true $q^* = 0.1$

Each prior is a smooth freehand density (a shape-preserving interpolation of hand-placed knots in log-density space); within each of the three shape pairs, the two densities share a similar bulk shape and differ decisively in the ratio of their values at the mirror states $\{0.1, 0.9\}$, set to 9:1 in opposite directions. Color indicates the eventual camp: red densities converge to limiting weight ≈ 0.9 on the $q = 0.1$ branch and blue densities to weight ≈ 0.9 on the $q = 0.9$ branch, after a common history of 25,000 signals. Legend entries report each agent's prior and limiting posterior means.

common knowledge, and the same discontinuity arises instead from the geometry of the identified set. Limiting beliefs are continuous in the prior everywhere except in the direction of the density ratio at the mirror pair, and that is precisely the direction that total variation distance—and any elicitation of the prior's bulk—cannot detect.

Figure 5 illustrates the result with six agents rather than two, and with priors that belong to no parametric family. Each prior is a smooth density drawn freehand—a shape-preserving interpolation of hand-placed knots—in one of three shapes: two skeptic-leaning densities peaked near $q = 0.2$, two agnostic densities peaked near one-half, and two believer-leaning densities peaked near $q = 0.75$. Within each pair, the two densities share a similar bulk shape; the decisive difference is the ratio of densities at the mirror states, set to nine-to-one in opposite directions. Exposed to a single common signal history, the six agents sort into two camps of three, each camp placing roughly ninety percent of limiting mass on its branch—and the camps cut across the visible shapes: one skeptic-leaning agent ends with $E[q] \approx 0.82$ while one believer-leaning agent ends with $E[q] \approx 0.18$. Priors that look alike split, and each camp collects priors that looked nothing alike; membership is determined only by the tails.

Two remarks delimit the result. First, its scope is unconditional elicitation. Proposition 3 covers functionals of the prior that are continuous in total variation, which includes the summaries that

standard instruments collect—means, medians, interval probabilities, confidence statements—but not conditional elicitation: a question of the form “if the truth is extreme, which extreme?” targets the invariant conditional odds directly and is, in principle, askable. Second, the result creates an apparent falsifiability tension that is worth making explicit: the parameter that determines who polarizes is, by the theory’s own account, unobservable *ex ante* in standard field data. The theory remains testable on its other margins—the directional predictions for null and positive signals (Section 4.4) and the identified-signal comparative static of Section 5 do not involve the tails—and in the laboratory the tension dissolves entirely, since induced priors are known to the experimenter by construction, so the tail-ratio prediction itself becomes directly testable.

4.4 Direction of updating

Two within-sample properties map onto observed behavior. Write $\ell_0(q) = 1 - p(q)$ and $\ell_1(q) = p(q)$ for the likelihoods of a null and a positive signal. A signal $s_{t+1} = s$ moves agent i ’s posterior mean by

$$E_i[q \mid h_t, s_{t+1} = s] - E_i[q \mid h_t] = \frac{\text{Cov}_i(q, \ell_s(q) \mid h_t)}{E_i[\ell_s(q) \mid h_t]},$$

so the direction of updating is the sign of the posterior covariance between the state and the likelihood. When $s_t = 0$ —no evidence surfaces— $\ell_0(q) = 1 - q(1 - q)$ is U-shaped, decreasing below $\frac{1}{2}$ and increasing above it. For an agent whose posterior is concentrated above $\frac{1}{2}$ the covariance is positive and she shifts *further toward* $q = 1$: through a conspiratorial lens, silence is what a competent cabal produces. The signature inferential move of conspiracy communities—absence of evidence as evidence of the cover-up (Sunstein and Vermeule 2009; Boudry and Hofhuis 2024)—is correct Bayesian reasoning under (1). Because an agent concentrated below $\frac{1}{2}$ simultaneously shifts down, a null signal polarizes: the same public observation moves the two camps apart. When $s_t = 1$, $\ell_1(q) = q(1 - q)$ peaks at $\frac{1}{2}$ and posterior mass in the interior gains at the expense of the tails—a detected plot is evidence against an omnipotent cabal. Both effects operate on mass *within* a branch, and both therefore attenuate as learning proceeds: an agent whose posterior has collapsed onto the mirror pair is unmoved by either signal, since the mirror states generate both signals at the same rate (Proposition 1). That positive signals should moderate partially convinced adherents is the model’s least plausible prediction; Section 5 returns to it.

4.5 General concealment technologies

The baseline model ties influence and concealment to the same parameter, so that detection occurs at rate $q(1 - q)$ and the identified set is symmetric. Neither feature is necessary for the mechanism. Let the cabal conceal its influence with probability $c(q)$ and write $s(q) = 1 - c(q)$ for the slip

probability, so that detection occurs at rate $p(q) = q s(q)$.

Assumption 3 (Concealment technology). $c : [0, 1] \rightarrow [0, 1]$ is continuously differentiable and increasing, with $c(q) < 1$ for $q < 1$, and $p(q) = q [1 - c(q)]$ has finitely many critical points.

Stronger cabals hide better, though not necessarily at the same rate at which they influence better. The condition $c(q) < 1$ for $q < 1$ rules out perfect concealment short of full strength, so that $p > 0$ on $(0, 1)$, and finiteness of the critical set excludes technologies with infinitely many detection plateaus; the baseline $c(q) = q$ satisfies both. Define the *stealth elasticity* on $(0, 1)$

$$\varepsilon(q) = -\frac{q s'(q)}{s(q)} \geq 0,$$

the percentage improvement in concealment per percentage increase in strength. Differentiating, $p'(q) = s(q) [1 - \varepsilon(q)]$: detection rises with strength precisely where the cabal's marginal capability outruns its marginal stealth. Where the elasticity exceeds one, detection falls with strength and p is non-monotone—and hence non-injective. Let $\mathcal{M} \equiv \{m : \# p^{-1}(m) \geq 2\}$ denote the set of detection levels attained at more than one state, and call a level m *regular* if $p'(q) \neq 0$ at every q with $p(q) = m$.

Proposition 4 (Concealment technologies and polarization). *Let Assumptions 1–3 hold.*

- (i) *If $\varepsilon(q) < 1$ for all $q \in (0, 1)$, then p is strictly increasing on $[0, 1]$ and q is point-identified: for every true $q^* \in (0, 1)$, every agent's posterior converges weakly to the point mass δ_{q^*} almost surely, and all disagreement vanishes.*
- (ii) *If $\varepsilon(q) > 1$ on some open interval, then p is non-monotone and $\mathcal{M}$ is a nonempty set of positive measure. Let the true $q^* \in (0, 1)$ be such that $m^* = p(q^*)$ is a regular level with $m^* \neq p(1)$. Then the identified set $p^{-1}(m^*) = \{q_1, \dots, q_K\}$ is a finite subset of $(0, 1)$, with $K \geq 2$ exactly when $m^* \in \mathcal{M}$; posterior odds between any two of its elements are never updated; and almost surely agent i 's posterior converges weakly to $\sum_{j=1}^K w_{ij} \delta_{q_j}$, where*

$$w_{ij} = \frac{\pi_i(q_j)/|p'(q_j)|}{\sum_{k=1}^K \pi_i(q_k)/|p'(q_k)|}.$$

Part (i) follows from $p'(q) = s(q)[1 - \varepsilon(q)] > 0$: an invertible detection map makes the environment identified, and the concentration argument behind Proposition 2 applies with a single root. In part (ii) the likelihood depends on q only through $p(q)$, so states with equal detection rates have identical likelihoods after every history, which generalizes Proposition 1 verbatim. The Laplace comparison behind Proposition 2 applies at each root, but the local slopes $|p'(q_j)|$ now

differ across roots, and a steeper slope shrinks the neighborhood of a root that the data leave plausible: hence the deflator $|p'(q_j)|^{-1}$. The two excluded coincidences— m^* a critical level, or $m^* = p(1)$ —are knife-edge cases in which a root sits at a plateau of p or at the boundary of the state space; the posterior still concentrates on the identified set, but the relative weights take a different form.²

Part (ii) contains two lessons. First, the polarization mechanism requires only that the detection map be non-injective somewhere—that concealment improve elastically with strength over some range—and not the knife-edge of the baseline. The scope of the mechanism is then delimited by $\mathcal{M}$ rather than by the technology as a whole: multiplicity, and hence polarization, obtains for truths whose detection level lies in $\mathcal{M}$, while truths outside it remain point-identified even though the technology is elastic elsewhere. The baseline is the special case $c(q) = q$, for which $\varepsilon(q) = q/(1 - q)$ crosses one exactly at $q = \frac{1}{2}$ and $\mathcal{M}$ comprises every interior detection level: each level below the peak is attained twice, so every truth except the knife-edge $q^* = \frac{1}{2}$ polarizes. The measure of $\mathcal{M}$ is accordingly a comparative static of independent interest: it is the share of possible worlds that is polarization-prone, and it grows with the interval on which the stealth elasticity exceeds one. Figure 6 illustrates the three regimes.

Second, the limiting weights acquire a slope deflator: branches where detection varies steeply with the state are easier to rule in or out locally and receive correspondingly less residual mass. Because the deflator $|p'(q_j)|$ is a feature of the commonly known technology rather than of any agent’s prior, the tail-ratio logic and the elicitation-impossibility construction of Proposition 3 carry over unchanged: who polarizes is still governed by prior densities at the identified states, and those densities remain invisible to elicitation. Part (i) is the mirror-image policy lesson, taken up in Section 5: an information environment whose stealth elasticity stays below one everywhere does not merely reduce polarization—it removes the identification failure that produces it.

A caveat on common knowledge is in order. Throughout, agents know the concealment technology $c(\cdot)$ —and hence the detection map $p(\cdot)$ —exactly; disagreement is a pure prior phenomenon. Uncertainty about the technology itself changes the environment in ways this paper does not analyze: it can restore injectivity over part of the range, partially restoring identification, or it can generate disagreement through a distinct channel in which agents never learn how to interpret signals, the environment studied by Acemoglu et al. (2016). What survives any technology uncertainty is the invariance logic itself: for whatever set of states the *believed* technology class renders observationally equivalent, posterior odds within that set are never updated. The results here should accordingly be read as characterizing the benchmark in which the information structure

²At a root where $p' = 0$ the Laplace mass is of larger order in t , so in the limit all mass migrates to the roots of highest-order contact; at a boundary root $q_j = 1$ the one-sided neighborhood halves that root’s weight relative to the formula in (ii).

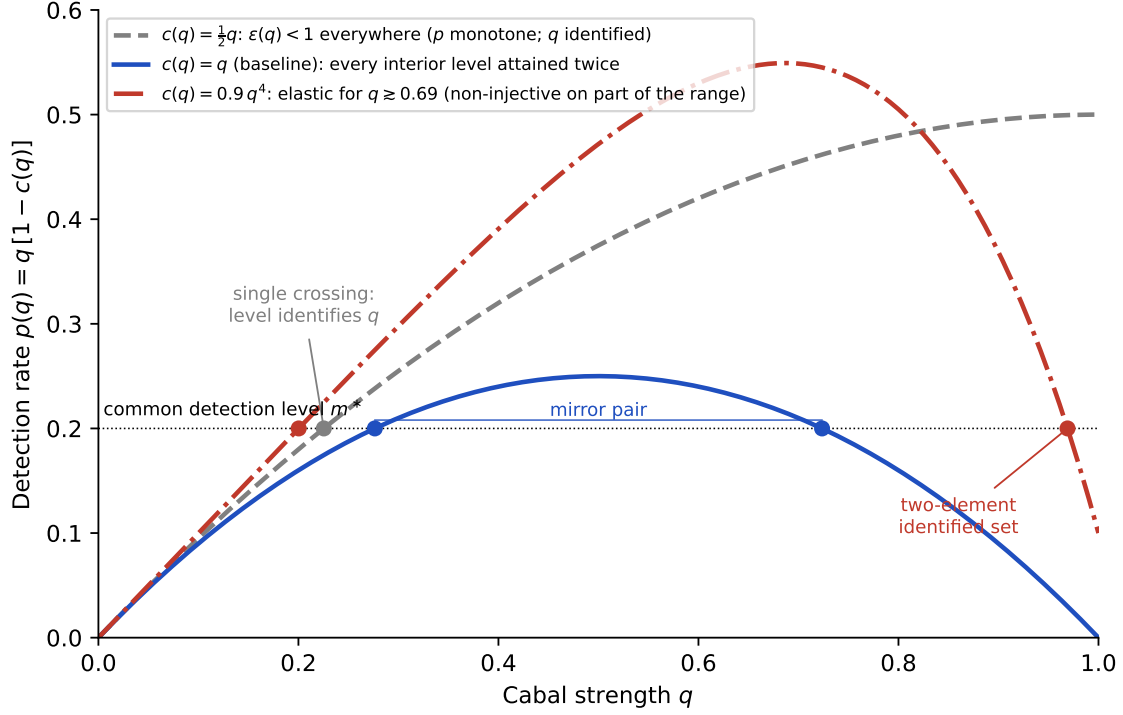


Figure 6: Detection maps under three concealment technologies

Detection rates $p(q) = q[1 - c(q)]$ induced by three concealment technologies. With $c(q) = \frac{1}{2}q$ (gray, dashed), the stealth elasticity is below one everywhere, p is strictly increasing, and any detection level identifies q : the horizontal line at the common level m^* crosses once. With the baseline $c(q) = q$ (blue, solid), every interior detection level is attained at two states, and the level m^* is attained at a symmetric mirror pair. With $c(q) = 0.9q^4$ (red, dash-dotted), the elasticity exceeds one only for $q \gtrsim 0.69$: levels in $\mathcal{M}$, including m^* , are attained at a two-element identified set that is no longer symmetric about $\frac{1}{2}$, while levels below $p(1) = 0.1$ remain identified.

is transparent and only the state is unknown; the interaction of state uncertainty with technology uncertainty is a natural subject for future work.

Table 1 summarizes the correspondence of the stylized facts identified in Section 2.1 to the features of the proposed environment that generate them. All four facts are reproduced with just three ingredients: Bayes' rule under a commonly known likelihood, heterogeneous priors, and a detection rate under which a stronger cabal produces more to detect but conceals it better.

5 Limitations and extensions

5.1 The response to positive evidence

One prediction of the model is difficult to reconcile with observed behavior. When evidence of cabal activity *does* surface, the likelihood $q(1 - q)$ pulls interior posterior mass toward the center, so

Table 1: Stylized facts and model mechanisms

Stylized fact	Model mechanism
Beliefs about specific theories polarized: endorsed by a substantial minority, rejected by most	Every posterior concentrates on the mirror pair $\{q^*, 1 - q^*\}$ (Prop. 2)
Common information shifts beliefs in opposite directions	Likelihood of a null signal, $1 - q(1 - q)$, is U-shaped: mass above $\frac{1}{2}$ moves up, mass below moves down
Population adherence rates stable over time; little convergence	Posterior odds between the mirror states equal prior odds after every history (Prop. 1)
People act on conspiracy beliefs	Optimal action is the posterior mean; branch weights translate directly into behavior

The table maps stylized facts about specific conspiracy beliefs to the feature of the equivocal-signal environment that generates each. All results obtain for fully Bayesian agents with common knowledge of the signal structure in equation (1).

adherents who retain uncertainty within a branch should moderate: every detected plot is evidence that the cabal is fallible. Evidence suggests the opposite: revelations of official misconduct tend to energize rather than temper distrust. In one famous instance, the 1972 disclosure that the U.S. Public Health Service had, in the Tuskegee syphilis experiment, knowingly withheld treatment from Black men for forty years measurably increased medical mistrust and reduced health-care use among Black men (Alsan and Wanamaker 2018). In richer specifications with uncertainty over the signal technology, the proof of existence provided by such a revelation may dominate.

5.2 Policy and extensions

The policy implication is an impossibility result. Interventions that supply more common public evidence – fact checks, investigations, transparency initiatives – move beliefs about the identified detection rate but cannot move the odds between branches (Proposition 1), and investigations that come up empty rationally deepen adherents’ convictions. Closing the gap requires changing what is identifiable—signals whose distributions differ across the mirror states, such as credible institutional audits whose informativeness does not scale with the cabal’s own concealment.

Several families of extensions preserve the identification failure while enriching its implications.

A first family introduces intermediaries. With multiple media outlets whose detection technologies differ, the symmetry of (1) breaks only if the technologies are common knowledge: an outlet known to detect at a rate $g(q)$ with $g(q) \neq g(1 - q)$ separates the mirror states, so cross-outlet comparison is informative precisely where any single outlet is not. If instead the outlets’ technologies are themselves uncertain—or are chosen by outlets with preferences over audience beliefs—equivocality is reintroduced endogenously: an outlet that slants coverage toward one branch is, from the audience’s standpoint, simply another mapping from states to reports whose

inverse is set-valued. The supply-side machinery of Szeidl and Szucs (2025) could be nested in this environment, with the strategic actor choosing the concealment technology rather than the narrative. More pointedly, the cabal is passive in the baseline: strength is fixed by nature, and concealment does not respond to being learned about. A strategic cabal that anticipates public learning might choose its concealment effort to maximize equivocality in the spirit of signal jamming (Fudenberg and Tirole 1986; Holmström 1999). If inducing non-injectivity is optimal for the cabal, then the identification failure is not an assumption about the environment but an equilibrium outcome of the cabal’s own problem.

A second family operates on the prior side. Private shocks to beliefs—a personal encounter, a trusted acquaintance’s testimony, an algorithmic recommendation—perturb tail densities directly and therefore generate discontinuous jumps in limiting branch weights: because the decisive statistic is a ratio of densities at ex ante negligible states, a shock that moves negligible total mass can move the ratio by orders of magnitude. This is a natural formalization of “going down the rabbit hole”—and, since a shock can move the ratio in either direction, equally of climbing back out—and it carries a distinctive prediction: radicalization, and likewise recovery from it, should appear sudden in beliefs and behavior while remaining nearly invisible in elicited summaries until it is complete.

Finally, the benchmark invites a formal comparison with behavioral accounts of polarization. Confirmatory bias (Rabin and Schrag 1999) and coarse interpretation of ambiguous evidence (Fryer et al. 2019) generate polarization through the updating rule; the present model generates it through the environment. The two classes are distinguishable in principle on two margins: the response to positive signals, where the rational benchmark predicts compression (Section 5.1) while confirmation-based accounts predict reinforcement, and the effect of replacing the equivocal signal with an identified one – a detection rate satisfying the point-identification condition of Proposition 4(i) – which removes environment-driven polarization but leaves bias-driven polarization intact.

6 Conclusion

Conspiracy theories are usually treated as symptoms of defective reasoning, whether motivated (Douglas et al. 2017, 2019; Kishishita 2025), cognitive (Pennycook et al. 2020; Bowes et al. 2023; Gagliardi 2025), misspecified (Aytimur and Kishishita 2025), or supplied by propagandists (Szeidl and Szucs 2025). This paper offers a demand-side model in which, once the defining property of a conspiracy – concealment that scales with capability – is written into the signal structure, Bayesian learning and optimization reproduce stylized facts with no behavioral ingredients. Persistent, polarized, behavior-relevant disagreement among people who agree about the facts is therefore not, by itself, evidence that anyone is updating badly. A sobering corollary follows: additional common evidence cannot close the divide.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Claude (Anthropic) to assist with literature search and citation verification, drafting and editing of text, checking and tightening the statements and proofs of the propositions, and writing the code that retrieves the media-attention data and produces the figures. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

- Acemoglu, Daron, Victor Chernozhukov, and Muhamet Yildiz (2016) “Fragility of Asymptotic Agreement under Bayesian Learning,” *Theoretical Economics*, 11 (1), 187–225.
- Allcott, Hunt and Matthew Gentzkow (2017) “Social Media and Fake News in the 2016 Election,” *Journal of Economic Perspectives*, 31 (2), 211–236.
- Alper, Sinan (2023) “There Are Higher Levels of Conspiracy Beliefs in More Corrupt Countries,” *European Journal of Social Psychology*, 53 (3), 503–517.
- Alsan, Marcella and Marianne Wanamaker (2018) “Tuskegee and the Health of Black Men,” *Quarterly Journal of Economics*, 133 (1), 407–455.
- Aytimur, R. Emre and Daiki Kishishita (2025) “Population Dynamics of Conspiracy Thinking,” Working paper, SSRN 5800942, November 2025.
- Becchetti, Leonardo, Davide Bellucci, and Fabio Pisani (2023) “Inequality and Conspiracy Beliefs,” SSRN 4652001.

- Benoit, Jean-Pierre and Juan Dubra (2019) “Apparent Bias: What Does Attitude Polarization Show?” *International Economic Review*, 60 (4), 1675–1703.
- Bierwiazzonek, Kinga, Aleksander B. Gundersen, and Jonas R. Kunst (2022) “The Role of Conspiracy Beliefs for COVID-19 Health Responses: A Meta-Analysis,” *Current Opinion in Psychology*, 46, 101346.
- Blackwell, David and Lester Dubins (1962) “Merging of Opinions with Increasing Information,” *Annals of Mathematical Statistics*, 33 (3), 882–886.
- Boudry, Maarten and Steije Hofhuis (2024) “On Epistemic Black Holes: How Self-Sealing Belief Systems Develop and Evolve,” *Theoria*, 90 (4), 429–447.
- Bowes, Shauna M., Thomas H. Costello, and Arber Tasimi (2023) “The Conspiratorial Mind: A Meta-Analytic Review of Motivational and Personological Correlates,” *Psychological Bulletin*, 149 (5–6), 259–293.
- Cheng, Ing-Haw and Alice Hsiaw (2021) “Distrust in Experts and the Origins of Disagreement,” *Journal of Economic Theory*, 200, 105401.
- Danenberg, Tuval (2025) “Bayesian Polarization,” arXiv:2509.02513.
- Dixit, Avinash K. and Jörgen W. Weibull (2007) “Political Polarization,” *Proceedings of the National Academy of Sciences*, 104 (18), 7351–7356.
- Douglas, Karen M. and Robbie M. Sutton (2023) “What Are Conspiracy Theories? A Definitional Approach to Their Correlates, Consequences, and Communication,” *Annual Review of Psychology*, 74, 271–298.
- Douglas, Karen M., Robbie M. Sutton, and Aleksandra Cichocka (2017) “The Psychology of Conspiracy Theories,” *Current Directions in Psychological Science*, 26 (6), 538–542.
- Douglas, Karen M., Joseph E. Uscinski, Robbie M. Sutton, Aleksandra Cichocka, Turkay Nefes, Chee Siang Ang, and Farzin Deravi (2019) “Understanding Conspiracy Theories,” *Political Psychology*, 40 (S1), 3–35.
- Doyle, Gabriel (2021) “A Bayesian-Rational Framework for Conspiracy Theories,” PsyArXiv preprint, doi:10.31234/osf.io/fupq3.
- Eliaz, Kfir and Ran Spiegler (2020) “A Model of Competing Narratives,” *American Economic Review*, 110 (12), 3786–3816.

- Enders, Adam, Casey Klofstad, and Joseph Uscinski (2024) “The Relationship Between Conspiracy Theory Beliefs and Political Violence,” *Harvard Kennedy School Misinformation Review*.
- Fryer, Roland G., Jr., Philipp Harms, and Matthew O. Jackson (2019) “Updating Beliefs when Evidence Is Open to Interpretation: Implications for Bias and Polarization,” *Journal of the European Economic Association*, 17 (5), 1470–1501.
- Fudenberg, Drew and Jean Tirole (1986) “A “Signal-Jamming” Theory of Predation,” *RAND Journal of Economics*, 17 (3), 366–376.
- Gagliardi, Lorenzo (2025) “The Role of Cognitive Biases in Conspiracy Beliefs: A Literature Review,” *Journal of Economic Surveys*, 39 (1), 32–65.
- Gentzkow, Matthew, Michael B. Wong, and Allen T. Zhang (2025) “Ideological Bias and Trust in Information Sources,” *American Economic Journal: Microeconomics*, 17 (2), 162–213.
- Grebe, Eduard and Nicoli Nattrass (2012) “AIDS Conspiracy Beliefs and Unsafe Sex in Cape Town,” *AIDS and Behavior*, 16 (3), 761–773.
- Herrick, Merlin (2026) “Conspiracy Theorising Is a Bayesian Nightmare,” *Social Epistemology*, forthcoming.
- Holmström, Bengt (1999) “Managerial Incentive Problems: A Dynamic Perspective,” *Review of Economic Studies*, 66 (1), 169–182.
- Jetten, Jolanda, Kim Peters, Belén Álvarez et al. (2021) “Consequences of Economic Inequality for the Social and Political Vitality of Society: A Social Identity Analysis,” *Political Psychology*, 42 (S1), 241–266.
- Kishishita, Daiki (2025) “Inequality, Conspiracy Theories, and Redistribution,” *European Journal of Political Economy*, 88, 102698.
- Levy, Gilat, Ronny Razin, and Alwyn Young (2022) “Misspecified Politics and the Recurrence of Populism,” *American Economic Review*, 112 (3), 928–962.
- Loh, Isaac and Gregory Phelan (2019) “Dimensionality and Disagreement: Asymptotic Belief Divergence in Response to Common Information,” *International Economic Review*, 60 (4), 1861–1876.
- Lord, Charles G., Lee Ross, and Mark R. Lepper (1979) “Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence,” *Journal of Personality and Social Psychology*, 37 (11), 2098–2109.

- Madsen, Jens Koed, Richard Bailey, and Toby D. Pilditch (2017) “Growing a Bayesian Conspiracy Theorist: An Agent-Based Model,” in *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, 2657–2662.
- Manski, Charles F. (2004) “Measuring Expectations,” *Econometrica*, 72 (5), 1329–1376.
- McCauley, Clark and Susan Jacques (1979) “The Popularity of Conspiracy Theories of Presidential Assassination: A Bayesian Analysis,” *Journal of Personality and Social Psychology*, 37 (5), 637–644.
- Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden et al. (2011) “Quantitative Analysis of Culture Using Millions of Digitized Books,” *Science*, 331 (6014), 176–182.
- Moon, Hyungsik Roger and Frank Schorfheide (2012) “Bayesian and Frequentist Inference in Partially Identified Models,” *Econometrica*, 80 (2), 755–782.
- Morris, Stephen (1995) “The Common Prior Assumption in Economic Theory,” *Economics and Philosophy*, 11 (2), 227–253.
- Nyhan, Brendan and Jason Reifler (2010) “When Corrections Fail: The Persistence of Political Misperceptions,” *Political Behavior*, 32 (2), 303–330.
- Oliver, J. Eric and Thomas J. Wood (2014) “Conspiracy Theories and the Paranoid Style(s) of Mass Opinion,” *American Journal of Political Science*, 58 (4), 952–966.
- Pennycook, Gordon, James Allan Cheyne, Derek J. Koehler, and Jonathan A. Fugelsang (2020) “On the Belief that Beliefs Should Change According to Evidence: Implications for Conspiratorial, Moral, Paranormal, Political, Religious, and Science Beliefs,” *Judgment and Decision Making*, 15 (4), 476–498.
- Poirier, Dale J. (1998) “Revising Beliefs in Nonidentified Models,” *Econometric Theory*, 14 (4), 483–509.
- Poth, Nina and Krzysztof Dolega (2023) “Bayesian Belief Protection: A Study of Belief in Conspiracy Theories,” *Philosophical Psychology*, 36 (6), 1182–1207.
- Rabin, Matthew and Joel L. Schrag (1999) “First Impressions Matter: A Model of Confirmatory Bias,” *Quarterly Journal of Economics*, 114 (1), 37–82.
- Stasielowicz, Lukasz (2022) “A Continuous Time Meta-Analysis of the Relationship Between Conspiracy Beliefs and Individual Preventive Behavior During the COVID-19 Pandemic,” *Scientific Reports*, 12, 11508.

- Stephens, Matthew (2000) “Dealing with Label Switching in Mixture Models,” *Journal of the Royal Statistical Society: Series B*, 62 (4), 795–809.
- Sunstein, Cass R. and Adrian Vermeule (2009) “Conspiracy Theories: Causes and Cures,” *Journal of Political Philosophy*, 17 (2), 202–227.
- Szeidl, Adam and Ferenc Szucs (2025) “A Model of Populism as a Conspiracy Theory,” *American Economic Review*, 115 (9), 3214–3247.
- Thorburn, Sheryl and Laura M. Bogart (2005) “Conspiracy Beliefs about Birth Control: Barriers to Pregnancy Prevention among African Americans of Reproductive Age,” *Health Education & Behavior*, 32 (4), 474–487.
- Trella, Carolina, Robbie M. Sutton, and Karen M. Douglas (2024) “Semantic and Causal Relations Between the Conspiracy Mentality and Belief in Conspiracy Theories,” *Zeitschrift für Psychologie*, 232 (1), 7–17.
- Tsapos, Melina (2024) “Betting on Conspiracy: A Decision Theoretic Account of the Rationality of Conspiracy Theory Belief,” *Erkenntnis*, 90 (5), 1913–1931.
- Uscinski, Joseph, Adam Enders, Casey Klofstad, Michelle Seelig, Hugo Drochon, Kamal Premaratne, and Manohar Murthi (2022) “Have Beliefs in Conspiracy Theories Increased over Time?” *PLOS ONE*, 17 (7), e0270429.
- Williams, Matt N., Mathew Ling, John R. Kerr, Stephen R. Hill, Mathew D. Marques, Hollie Mawson, and Edward J. R. Clarke (2024) “People Do Change Their Beliefs About Conspiracy Theories—But Not Often,” *Scientific Reports*, 14, 3836.
- Wood, Thomas and Ethan Porter (2019) “The Elusive Backfire Effect: Mass Attitudes’ Steadfast Factual Adherence,” *Political Behavior*, 41 (1), 135–163.
- Young, David J., Jens Koed Madsen, and Lee de Wit (2025) “Belief Polarization Can Be Caused by Disagreements over Source Independence: Computational Modelling, Experimental Evidence, and Applicability to Real-World Politics,” *Cognition*, 259, 106126.
- Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov (2020) “Political Effects of the Internet and Social Media,” *Annual Review of Economics*, 12, 415–438.

Appendix – intended for online publication

Throughout the appendix q^* denotes the true strength, $m^* = p(q^*)$, $k_t = \sum_{\tau \leq t} s_\tau$, $r_t = k_t/t$, and all almost-sure statements are with respect to P_{q^*} . By Assumption 1 and the strong law of large numbers, $r_t \rightarrow m^*$ almost surely. For q with $p(q) \in (0, 1)$ the log posterior density is, up to a constant that does not depend on q ,

$$\log \pi_i(q \mid h_t) = \log \pi_i(q) + t f_t(q), \quad f_t \equiv \varphi_{r_t} \circ p, \quad \varphi_r(u) \equiv r \log u + (1-r) \log(1-u),$$

and $f \equiv \varphi_{m^*} \circ p$ denotes the limit. For $m^* \in (0, 1)$ the function φ_{m^*} is strictly concave on $(0, 1)$ and uniquely maximized at $u = m^*$, so f is maximized on $[0, 1]$ exactly on the identified set $Q^* \equiv p^{-1}(m^*)$, with common value $f^* = \varphi_{m^*}(m^*)$. Under Assumption 3, p is strictly monotone on each of finitely many intervals, so every level set is finite; we write $Q^* = \{q_1, \dots, q_K\}$. The baseline $p(q) = q(1-q)$ is the case $c(q) = q$ of Assumption 3, with the single critical point $q = \frac{1}{2}$.

Proof of Proposition 1

By Assumption 1 the likelihood of h_t is $L(h_t \mid q) = p(q)^{k_t} [1 - p(q)]^{t-k_t}$, and Bayes' rule gives the posterior density $\pi_i(q \mid h_t) = \pi_i(q) L(h_t \mid q) / Z_t$ with $Z_t = \int_0^1 \pi_i(q) L(h_t \mid q) dq$. The normalizer is strictly positive because π_i is continuous and strictly positive (Assumption 2) and $L(h_t \mid \cdot) > 0$ on $(0, 1)$, so the density is well defined after every history. Since $p(q) = p(1-q)$, $L(h_t \mid q) = L(h_t \mid 1-q)$ for every $q \in [0, 1]$, and therefore

$$\pi_i(q \mid h_t) \pi_i(1-q) = \frac{\pi_i(q) \pi_i(1-q) L(h_t \mid q)}{Z_t} = \pi_i(1-q \mid h_t) \pi_i(q).$$

For $q \in (0, 1)$ both posterior densities are strictly positive, and the odds form follows by division. $\square$

Two lemmas

Lemma 1 (Concentration). *Let Assumptions 1–3 hold and let $q^* \in (0, 1)$, so that $m^* \in (0, 1)$. For $\zeta > 0$ small enough that the intervals $B_j = (q_j - \zeta, q_j + \zeta)$, $j = 1, \dots, K$, are disjoint and contained in $(0, 1)$,*

$$\Pr_i \left(q \notin \bigcup_j B_j \mid h_t \right) \longrightarrow 0 \quad \text{almost surely.}$$

Proof. Let $K_\zeta = [0, 1] \setminus \bigcup_j B_j$. Since p is continuous and K_ζ is compact, $p(K_\zeta)$ is a compact set not containing m^* , so by strict concavity of φ_{m^*} ,

$$\gamma \equiv f^* - \sup_{K_\zeta} f > 0,$$

where the supremum includes points with $p \in \{0, 1\}$, at which $f = -\infty$. Fix $\delta \in (0, \frac{1}{2})$ small enough that $\frac{m^*}{2} \log \delta < f^* - 2\gamma$ and $\frac{1-m^*}{2} \log \delta < f^* - 2\gamma$, and split K_ζ into the collar $D_\delta = \{q \in K_\zeta : p(q) \leq \delta \text{ or } p(q) \geq 1 - \delta\}$ and the core $C_\delta = K_\zeta \setminus D_\delta$.

Collar. Almost surely $r_t \in [m^*/2, (1+m^*)/2]$ for all large t . Both terms of φ_r are nonpositive, so where $p(q) \leq \delta$, $f_t(q) \leq r_t \log p(q) \leq \frac{m^*}{2} \log \delta$, and where $p(q) \geq 1 - \delta$, $f_t(q) \leq (1-r_t) \log[1 - p(q)] \leq \frac{1-m^*}{2} \log \delta$. Hence $\sup_{D_\delta} f_t \leq f^* - 2\gamma$ for all large t .

Core. On C_δ the functions $\log p$ and $\log(1-p)$ are bounded, so $|f_t - f| \leq |r_t - m^*| \cdot \sup_{C_\delta} |\log p - \log(1-p)| \rightarrow 0$ uniformly, and $\sup_{C_\delta} f_t \leq f^* - \gamma + o(1)$ almost surely.

Combining,

$$\int_{K_\zeta} \pi_i(q) e^{t f_t(q)} dq \leq \left(\sup_{[0,1]} \pi_i \right) \left[e^{t(f^* - \gamma + o(1))} + e^{t(f^* - 2\gamma)} \right].$$

For a lower bound, $\bar{B}_1$ is a compact subset of $(0, 1)$, on which $p \in (0, 1)$ and hence $f_t \rightarrow f$ uniformly; by continuity of f there is a subinterval $B' \subset B_1$ of positive length $|B'|$ on which $f \geq f^* - \gamma/2$, so

$$\int_{B_1} \pi_i(q) e^{t f_t(q)} dq \geq \left(\inf_{[0,1]} \pi_i \right) |B'| e^{t(f^* - \gamma/2 - o(1))}.$$

The ratio of the upper to the lower bound tends to zero almost surely, which is the claim. $\square$

Lemma 2 (Relative mass). *In the setting of Lemma 1, suppose in addition that m^* is a regular level of p . Then for ζ small enough and every j, k , almost surely*

$$\frac{\int_{B_j} \pi_i(q) e^{t f_t(q)} dq}{\int_{B_k} \pi_i(q) e^{t f_t(q)} dq} \longrightarrow \frac{\pi_i(q_j)/|p'(q_j)|}{\pi_i(q_k)/|p'(q_k)|}.$$

Proof. Since p' is continuous and $p'(q_j) \neq 0$, for ζ small p is strictly monotone on each $\bar{B}_j$ with $|p'|$ bounded away from zero there, so $U_j \equiv p(B_j)$ is an open interval containing m^* and p has a continuously differentiable inverse $\psi_j : U_j \rightarrow B_j$. The change of variables $u = p(q)$ gives

$$\int_{B_j} \pi_i(q) e^{t f_t(q)} dq = \int_{U_j} g_j(u) e^{t \varphi_{r_t}(u)} du, \quad g_j(u) \equiv \frac{\pi_i(\psi_j(u))}{|p'(\psi_j(u))|},$$

where g_j is continuous and strictly positive on U_j and—this is the point of the substitution—the exponent φ_{r_t} is the same function for every j . Let $U = \bigcap_j U_j$, an open interval containing m^* , and fix a compact interval $V \subset U$ with m^* in its interior. Almost surely $r_t \in V$ for all large t . For $r \in V$, φ_r is strictly concave on $(0, 1)$ with its unique maximum at $u = r$, and $\varphi_r''(u) = -r/u^2 - (1-r)/(1-u)^2$ is bounded away from zero uniformly for $u \in U$ and $r \in V$. Laplace's method with the maximizer as a parameter—the weight g_j continuous and positive, the curvature bounded away from zero,

the maximizer ranging over the compact V interior to U_j , and $\varphi_r(u) \leq \varphi_r(r) - c$ for some $c > 0$ uniformly in $r \in V$ whenever $u \in U_j$ lies outside a fixed neighborhood of V —yields, uniformly in $r_t \in V$,

$$\int_{U_j} g_j(u) e^{t\varphi_{r_t}(u)} du = g_j(r_t) e^{t\varphi_{r_t}(r_t)} \sqrt{\frac{2\pi}{t|\varphi_{r_t}''(r_t)|}} [1 + o(1)].$$

Taking the ratio for two roots j, k , the factors $e^{t\varphi_{r_t}(r_t)}$ and $\sqrt{2\pi/(t|\varphi_{r_t}''(r_t)|)}$ are identical and cancel exactly, leaving $g_j(r_t)/g_k(r_t) \rightarrow g_j(m^*)/g_k(m^*)$ by continuity and $r_t \rightarrow m^*$. Since $\psi_j(m^*) = q_j$, this is the claim. $\square$

Two comments on Lemma 2. First, the substitution is not merely convenient. Expanding f_t to second order around the fixed roots q_j leaves a first-order term $t f_t'(q_j)(q - q_j)$ with $f_t'(q_j) \propto (r_t - m^*)$; completing the square shows it contributes the factor $\exp\{t(r_t - m^*)^2/[2m^*(1 - m^*)]\}$ to each branch integral, which by the law of the iterated logarithm does not converge to one almost surely. That factor is common to every root and cancels in the ratio, which the substitution makes automatic. Second, the lemma uses only that p is continuously differentiable with $p'(q_j) \neq 0$; no second derivative of p is needed, because all curvature is carried by φ_r .

Proof of Proposition 2

The baseline technology satisfies Assumption 3. For $q^* \in (0, 1)$ with $q^* \neq \frac{1}{2}$, $m^* = q^*(1 - q^*) \in (0, \frac{1}{4})$, $Q^* = \{q^*, 1 - q^*\}$, and m^* is a regular level since $p'(q) = 1 - 2q \neq 0$ at both roots. Write μ_t for the posterior and B_1, B_2 for the ζ -neighborhoods of q^* and $1 - q^*$. By Lemma 1, $\mu_t(K_\zeta) \rightarrow 0$; by Lemma 2 and $|p'(q^*)| = |p'(1 - q^*)| = |1 - 2q^*|$, $\mu_t(B_1)/\mu_t(B_2) \rightarrow \pi_i(q^*)/\pi_i(1 - q^*)$; hence $\mu_t(B_1) \rightarrow w_i$ and $\mu_t(B_2) \rightarrow 1 - w_i$ almost surely, for every small ζ . For any bounded continuous ϕ on $[0, 1]$, with modulus of continuity ω_ϕ ,

$$\left| \int \phi d\mu_t - [\mu_t(B_1)\phi(q^*) + \mu_t(B_2)\phi(1 - q^*)] \right| \leq \omega_\phi(\zeta) + \|\phi\|_\infty \mu_t(K_\zeta).$$

Letting $t \rightarrow \infty$ and then $\zeta \rightarrow 0$ along a countable sequence (each step almost surely) gives $\int \phi d\mu_t \rightarrow w_i\phi(q^*) + (1 - w_i)\phi(1 - q^*)$, which is weak convergence to $w_i\delta_{q^*} + (1 - w_i)\delta_{1-q^*}$. The statement about $\Pr_i(|q - q^*| < \zeta \mid h_t)$ is $\mu_t(B_1) \rightarrow w_i$, and a_i^∞ follows by taking $\phi(q) = q$. $\square$

Proof of Proposition 3

Fix $q^* \in (0, 1)$ with $q^* \neq \frac{1}{2}$, targets $w_1, w_2 \in (0, 1)$, and $\delta > 0$. Write $q_1 = q^*$, $q_2 = 1 - q^*$, and let $\rho_i = w_i/(1 - w_i)$ denote the target density ratios.

Construction. Let π_0 be any continuous density on $[0, 1]$ bounded below by $\underline{\pi} > 0$. Choose $\eta > 0$ with $\eta < \frac{1}{4} \min\{q^*, 1 - q^*, |1 - 2q^*|\}$, so that the intervals $(q_j - \eta, q_j + \eta)$, $j = 1, 2$, are disjoint and contained in $(0, 1)$. Let ϕ_η^j denote the continuous triangular bump of height one supported on $(q_j - \eta, q_j + \eta)$ with $\phi_\eta^j(q_j) = 1$, and define the unnormalized densities

$$\tilde{\pi}_i(q) = \pi_0(q) + H_i^1 \phi_\eta^1(q) + H_i^2 \phi_\eta^2(q), \quad i = 1, 2.$$

Nonnegative heights achieving the target ratios exist for any $\rho_i > 0$: if $\rho_i \geq \pi_0(q_1)/\pi_0(q_2)$, set $H_i^2 = 0$ and $H_i^1 = \rho_i \pi_0(q_2) - \pi_0(q_1) \geq 0$; otherwise set $H_i^1 = 0$ and $H_i^2 = \pi_0(q_1)/\rho_i - \pi_0(q_2) > 0$. In either case $\tilde{\pi}_i(q_1)/\tilde{\pi}_i(q_2) = \rho_i$, and the heights are bounded by a constant $\bar{H}$ that depends only on π_0 and the targets, not on η . Let $\pi_i = \tilde{\pi}_i/Z_i$ with $Z_i = 1 + \eta(H_i^1 + H_i^2)$, since each triangular bump integrates to η times its height. Each π_i is continuous and strictly positive on $[0, 1]$, so it satisfies Assumption 2, and normalization scales the densities at q_1 and q_2 by the common factor Z_i^{-1} , leaving the ratio ρ_i unchanged.

Limit weights. By Proposition 2, agent i 's limiting branch weight on q_1 is $\pi_i(q_1)/[\pi_i(q_1) + \pi_i(q_2)] = \rho_i/(1 + \rho_i) = w_i$, as required, and the limiting actions differ by

$$|a_1^\infty - a_2^\infty| = |w_1 - w_2| |1 - 2q^*|,$$

which is fixed by the targets and independent of η .

Total variation. For each i ,

$$\|\pi_i - \pi_0\|_{L^1} \leq \frac{1}{Z_i} \|\tilde{\pi}_i - \pi_0\|_{L^1} + \left(1 - \frac{1}{Z_i}\right) \int_0^1 \pi_0 \leq \eta(H_i^1 + H_i^2) + \eta(H_i^1 + H_i^2) \leq 4\bar{H}\eta,$$

so by the triangle inequality $\|\pi_1 - \pi_2\|_{TV} = \frac{1}{2} \|\pi_1 - \pi_2\|_{L^1} \leq \frac{1}{2} (\|\pi_1 - \pi_0\|_{L^1} + \|\pi_0 - \pi_2\|_{L^1}) \leq 4\bar{H}\eta$. Choosing $\eta < \delta/(4\bar{H})$ delivers $\|\pi_1 - \pi_2\|_{TV} < \delta$.

Elicitation functionals. Let $F_1, \dots, F_n$ be functionals of the prior that are continuous with respect to $\|\cdot\|_{TV}$, and fix $\gamma > 0$. Since $\|\pi_i - \pi_0\|_{TV} \leq 2\bar{H}\eta \rightarrow 0$ as $\eta \rightarrow 0$, continuity of each F_ℓ at π_0 gives $|F_\ell(\pi_i) - F_\ell(\pi_0)| < \gamma/2$ for $i = 1, 2$ once η is small enough, and hence $|F_\ell(\pi_1) - F_\ell(\pi_2)| < \gamma$ for every ℓ , while the limiting actions remain $|w_1 - w_2| |1 - 2q^*|$ apart. Means and interval probabilities are Lipschitz in total variation on $[0, 1]$, and a quantile is continuous at π_0 whenever π_0 is bounded below near it, so each qualifies. $\square$

Proof of Proposition 4

Part (i). Under Assumption 3, $s(q) > 0$ for $q < 1$, so ε is well defined on $(0, 1)$ and $p'(q) = s(q)[1 - \varepsilon(q)] > 0$ there; by continuity p is strictly increasing on $[0, 1]$. For a true $q^* \in (0, 1)$, $m^* =$

$q^*s(q^*) \in (0, 1)$ and $Q^* = \{q^*\}$, so Lemma 1 applies with a single interval $B_1 = (q^* - \zeta, q^* + \zeta)$: $\mu_t(B_1) \rightarrow 1$ almost surely for every small ζ , which is weak convergence to δ_{q^*} , and every agent's limiting action equals q^* .

Part (ii). Suppose $\varepsilon > 1$ on an open interval $(a, b) \subset (0, 1)$, so that $p' < 0$ there, while $p'(0) = s(0) > 0$; hence p is non-monotone. Pick $a < q_a < q_b < b$, so $p(q_b) < p(q_a)$. Since $p(0) = 0 < p(q_b)$ and p is continuous, every level in $(p(q_b), p(q_a))$ is attained on $(0, q_a)$ by the intermediate value theorem and again on (q_a, q_b) ; thus $(p(q_b), p(q_a)) \subset \mathcal{M}$, and $\mathcal{M}$ has positive measure. Now fix $q^* \in (0, 1)$ with $m^* = p(q^*)$ a regular level and $m^* \neq p(1)$. The identified set $Q^* = p^{-1}(m^*)$ is finite (preliminaries), and its elements lie in $(0, 1)$ because $p(0) = 0 < m^*$ and $p(1) \neq m^*$; $K \geq 2$ exactly when $m^* \in \mathcal{M}$ by the definition of $\mathcal{M}$. Because the likelihood of any history depends on q only through $p(q)$, $L(h_t \mid q_j) = L(h_t \mid q_k)$ for all j, k and all histories, so posterior odds within Q^* equal prior odds, which generalizes Proposition 1. For the limit, Lemma 1 gives $\mu_t(K_\zeta) \rightarrow 0$, and Lemma 2, which applies because m^* is regular, gives $\mu_t(B_j)/\mu_t(B_k) \rightarrow [\pi_i(q_j)/|p'(q_j)|]/[\pi_i(q_k)/|p'(q_k)|]$ for all j, k ; together these yield $\mu_t(B_j) \rightarrow w_{ij}$ with w_{ij} as displayed in the proposition. Weak convergence to $\sum_j w_{ij} \delta_{q_j}$ then follows exactly as in the proof of Proposition 2, with K atoms in place of two. $\square$