Rationalizations and political polarization*

Yves Le Yaouanq, Peter Schwardmann, and Joël J. van der Weele

September 10, 2026

Abstract

We present a signaling model formalizing the view that people base their moral and political judgments on emotion and private motives, and then rationalize to appear reasonable. Rationalizations are strategic complements: when others rationalize, their actions reveal less about inconvenient truths and their rationales may provide cover for one's own actions. This produces groupthink among co-partisans and anti-groupthink among counter-partisans, so that agents are most realistic when facing a moderate. Agents prefer echo chambers populated by extreme and skilled rationalizers. Naiveté about one's own rationalizations yields ideological and affective polarization, and explains why cross-partisan contact reduces both, yet is shunned.

JEL Classification codes: D72, D83, D91, P16.
Keywords: esteem, moral behavior, self-deception, group decisions, polarization

*Le Yaouanq: CREST–École Polytechnique–Institut Polytechnique de Paris; Schwardmann: Carnegie Mellon University; Van der Weele: Amsterdam School of Economics, Tinbergen Institute. We are grateful to Suzanne Bellue, Luca Braghieri, Tilman Fries, Russell Golman, Yucheng Liang, Andrew Little, Reed Orchinik, Davide Pace, Roland Rathelot, Alyssa Rusonik, Klaus Schmidt and Egon Tripodi for helpful comments. We also thank seminar audiences at Berlin Behavioral Economics Seminar, CESifo Behavioral Economics Conference, CREST, EEA/ESEM 2024, EWMES 2024, EWBEE 2026, INRAE Montpellier, LMU Munich, Université Paris-Dauphine PSL, Université Louvain, University of Amsterdam, University of Bonn, University of Nancy, and University of Saint-Denis.

1 Introduction

Political views are formed and defended in the company of others. With the advent of the internet and social media, it has never been easier to choose that company. Given the choice, people avoid counter-partisans and moderates, and gravitate toward the most extreme and persuasive voices on their own side (Frimer et al., 2017; Goldenberg et al., 2023; Rathje et al., 2024; Braghieri et al., 2024a). Because deliberation in like-minded groups is prone to move to extremes (Sunstein, 2002), the resulting divisions run deep, both in policy positions and in the moral regard each side holds for the other (Iyengar et al., 2019). Since democratic policymaking requires a shared perception of reality and some measure of mutual regard between opponents, the stakes are considerable (Kingzette et al., 2021).

A leading theory in political psychology attributes our political divisions to a clash of moral intuitions. According to this perspective, moral and political judgments are not primarily the product of reason, but instead arise from deep-seated intuitions and emotions (Haidt, 2001, 2012; Graham et al., 2013), which predict political preferences across a range of policy domains (Enke, 2020; Enke et al., 2023). Reasoning merely serves to rationalize these moral intuitions ex post, in order to produce the appearance of being reasonable and impartial. In this paper, we formalize this "social intuitionist" model of moral decision making and show that it captures several of the defining features of politics in the modern information environment.

In the model, two agents observe a perfectly informative public signal about which of two actions is reasonable or socially appropriate. Each then chooses an action and tries to come up with a rationale for it, before the public signal is collectively forgotten. In doing so, strategic agents trade off their "core motives"—intuitions or interests that favor one action regardless of the evidence—against the image utility of being perceived as a reasonable, evidence-based decision-maker. Coming up with rationales that cut against the public signal, which we call "rationalizations", may fail and leave the agent morally dumbfounded (Haidt and Hersh, 2001; Haidt, 2001). Instead, successful rationales enter public discourse, where they may constrain or amplify the other agent's tendency to rationalize.

Many political disputes have this structure. The two sides have opposing core motives — interests, group attachments, or moral intuitions — yet both profess the same standard of reasonable, evidence-based judgment. What is contested is which action ultimately is socially appropriate and reasonable given the evidence, which is where rationalizations come into play. Consider the conflict over discriminatory hiring practices playing out in hiring committees and in politics at large. Evaluators' or voters' motives differ by whether their own group gains and by their solidarity with those who would benefit (Enke et al., 2023),

while both sides claim to want a fair process that selects the most deserving candidate. Studies find that evaluators resolve this tension by redefining merit to fit the candidate they already favor (Norton et al., 2004; Uhlmann and Cohen, 2005). Similar patterns appear in other domains. While core motives like moral intuitions and self-interest are predictive of policy preferences on gun control, immigration, climate change, and redistribution, the debate on these topics is conducted in evidentiary terms and frequently ends in disagreement about the facts themselves.[1]

The first insight from our model is that rationalizations are strategic complements. Among co-partisans, whose motives are aligned, complementarities arise through the moral "cover" generated by rationalizations. A partner who rationalizes supplies arguments that shelter the agent's own action, while a realistic agent who follows the original signal may reveal those actions as inappropriate. The result is groupthink: rationalizations by one agent beget rationalizations by the other agent. The model thus predicts that like-minded groups move to extremes, as they are better able to generate good narratives (Sunstein, 2002), and clarifies how public rhetoric licenses inappropriate expressions and behavior (Bursztyn et al., 2020, 2023; Müller and Schwarz, 2023). Despite the complementarity, equilibrium is unique, because an agent who often rationalizes discounts her own rationales.

The complementarity survives among counter-partisans, but through a different mechanism. When opposing partisan rationales enter public discourse, esteem accrues to the more credible speaker. In equilibrium, the weaker the opponent's commitment to the truth, the more credibly the agent can blame the disagreement on the other side's rationalizations. Thus, rationalizing opponents weaken epistemic discipline and encourage rationalizations, a form of polarization or anti-groupthink. Combining the co- and counter-partisan cases yields a sharp prediction: an agent's realism is hill-shaped in her partner's core motive. She is most truthful facing a moderate—who stays realistic and thus disciplines discourse—and least truthful facing an extremist of either stripe.

Next, we ask which discourse environments agents seek out, comparing solitary reasoning to an echo chamber (a co-partisan partner) and an agora (a counter-partisan partner). We find that agents prefer environments that allow for more rationalizations. Agents select into echo chambers, but only when the co-partisan is both sufficiently extreme and skilled at rationalizing, so that the prospect of obtaining cover for a convenient action outweighs the

[1] In particular, moral universalism, or how far someone's altruism extends to outgroup members, predicts policy preferences over immigration, redistribution, affirmative action, and climate change mitigation (Enke et al., 2023; Stötzer and Zimmermann, 2026). At the same time, perceptions of the corresponding facts are politically divided: partisans draw opposite conclusions from identical crime data on gun control (Kahan et al., 2017); and beliefs about the number and economic contribution of immigrants (Alesina et al., 2023), about social mobility (Alesina et al., 2018), and about the size and causes of racial gaps (Alesina et al., 2021) all vary systematically with political position.

risk of being confronted with an inconvenient truth. This mixed finding aligns with the literature, which shows that generic exposure to like-minded content does little to change attitudes and actions, while skilled influencers move attitudes a lot (Nyhan et al., 2023; Allcott et al., 2024; Rathje et al., 2024). The agora always disciplines rationalizations through the injection of contrary rationales, and agents avoid it for that reason (Frimer et al., 2017; Chopra et al., 2022). Finally, the agent's preference over the core motive of her interlocutor exhibits "acrophily", the documented preference for co-partisans more extreme than oneself (Goldenberg et al., 2023; Zimmerman et al., 2024).

Because agents in the baseline model are Bayesian, the model does not generate systematic disagreement in beliefs. The final section assumes that agents are insufficiently skeptical of their own rationales while remaining alert to everyone else's, in line with empirical evidence (Kappes and Sharot, 2019; Melnikoff and Strohminger, 2020; Kurzban, 2011; Hagenbach and Saucet, 2025). This naiveté assumption jointly produces ideological polarization, self-righteousness, and affective polarization as each side tilts toward its convenient facts, overrates its own reasonableness, and underrates the integrity of the other camp. It also predicts that forced counter-partisan contact depolarizes, because naive agents are positively surprised by the quality of their opponents' arguments—as documented across field settings (Santoro and Broockman, 2022; Braghieri et al., 2024a; Broockman and Kalla, 2025; Hobolt et al., 2024; Blattner and Koenen, 2023). The model thus resolves a puzzle in this literature: counter-partisan contact reduces prejudice, yet it is avoided, because the price of depolarization is paid in esteem.

Our paper contributes to literatures in economics, psychology, and political science. In psychology, we build on the social intuitionist model (Haidt, 2001) and moral foundations theory (Graham et al., 2013), which stress the primacy of intuition in moral judgment and cast reasoning as a rationalizing force. By providing a formal model of rationalization and the psychological and social incentives behind it, we clarify how these psychological assumptions translate into social and political outcomes. Compared to the original formulation in Haidt (2001), we are able to spell out the distinct roles of image concerns, complementarities, and naiveté, and enlarge the set of empirical claims that can be used to evaluate and test the theory.

Our theoretical approach draws on economic models of motivated cognition (Bénabou and Tirole, 2016). Polarization in our model results from rationalization, not from non-motivated updating errors (Baliga et al., 2013; Fryer Jr. et al., 2019), heterogeneous mental models (Levy et al., 2022), or media competition (Perego and Yuksel, 2022). The model differs from existing accounts of motivated cognition in its mechanism and application: the cost of biased beliefs and behavior is damage to image rather than worse decisions (Bénabou and

Tirole, 2002; Bodner and Prelec, 2002; Brunnermeier and Parker, 2005; Bénabou, 2013), and the benefit of rationalization is an image of reasonableness rather than reduced cognitive dissonance (Yariv, 2005; Acharya et al., 2018; Eyster et al., 2021) or anticipatory utility (Brunnermeier and Parker, 2005; Bénabou, 2013; Le Yaouanq, 2023).[2]

A novel aspect of our model is that image concerns generate complementarities in belief distortions, because the disciplining effect of social interaction depends on others' realism. A small set of papers obtains complementarities in beliefs through interdependent anticipatory payoffs (Bénabou, 2013; Le Yaouanq, 2023) or inference from others' actions (Bénabou and Tirole, 2011; Hestermann et al., 2020). Because these models rest on different assumptions about the nature, costs, and benefits of biased beliefs, they do not extend easily to cultural conflict or to the drivers of political polarization.

We also relate to a nascent literature on narratives in economic theory. Bénabou et al. (2019) and Foerster and van der Weele (2021) study the strategic transmission of exculpatory narratives. In work on political correctness, Golman (2023) shows that the willingness to voice dissent depends on how many people's true opinions align with the socially acceptable one. We provide two key novelties in this literature. First, we model the production of narratives through a novel rationalization and sharing technology, and identify a new channel for complementarities in rationalization. Second, we model interdependencies in the propensity to rationalize among senders with aligned or opposite motives. This allows us to study selection into informational environments (autarky, echo chambers, and agoras).

Finally, our paper offers a unified framework for several hitherto separate accounts of polarization in political science (Boxell et al., 2024): moral foundations (Graham et al., 2013), which we take as a primitive; rationalization, self-deception, and partisan bias (Taber and Lodge, 2006; Taber et al., 2009; Little, 2025), which are our central mechanisms; and counter-partisan contact (Santoro and Broockman, 2022; Hobolt et al., 2024) and social media (Kubin and von Sikorski, 2021), which act as mediators.

2 Model

We model a two-player game in which agents choose an action and try to articulate a rationale supporting it. The central tension is between a core motive that favors a particular action and the reputational cost of being seen to rationalize rather than reason.

[2]Some self-signaling models presume that agents value an image of altruism (Bodner and Prelec, 2002; Bénabou and Tirole, 2006; Grossman and van der Weele, 2017; Reichel, 2024). Agents in our model instead signal a commitment to evidence-based reasoning; they may disagree about what is appropriate, and the norm that confers esteem is endogenous.

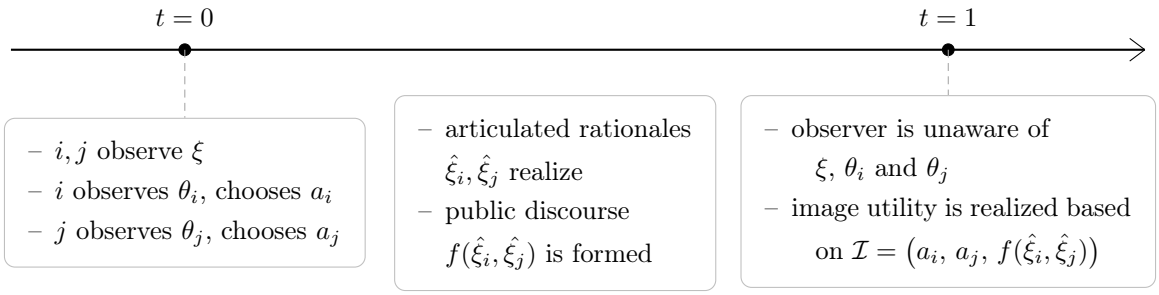

Figure 1: Timeline of the game.

Players and actions. Figure 1 summarizes the timing of the game. There are two agents, i and j, each of whom chooses an action $a \in \{0, 1\}$ at $t = 0$. These actions may represent speech acts, like vocally staking out a (policy) position regarding controversial topics, or represent morally contentious behavior, like discriminating against outgroup members, engaging in affirmative action, voting for higher redistribution, or polluting the environment. Below we describe the preferences of agent i; those of agent j are defined analogously.

State and appropriateness. Which action is socially appropriate is governed by a binary state $\xi \in \{0, 1\}$ that is drawn with $\Pr[\xi = 1] = \lambda \in (0, 1)$ and is publicly observed by both players at $t = 0$. In the subsequent period, $t = 1$, this signal is no longer available, and can only be inferred from actions and public discourse, as we describe below. We say that $a = 1$ is the *socially appropriate* action when $\xi = 1$, and $a = 0$ is socially appropriate when $\xi = 0$. For instance, when committee members are making hiring decisions, the application package may generate an unambiguous signal of who the most deserving candidate is. We deliberately sketch an extreme situation where all parties initially agree about what is appropriate, in order to then study how such agreement could be subverted.

Types. The population is heterogeneous in its willingness to follow unbiased moral reasoning (Saccardo and Serra-Garcia, 2023). We model a fundamental distinction between those who pursue the Socratic ideal of following the logic wherever it leads, and those willing to engage in sophistry to justify a predetermined conclusion. With probability $\rho \in (0, 1)$, agent i is a *reasonable* (or Socratic) type ($\theta_i = 1$): she accepts the normative facts or moral logic embodied in ξ and always follows the normatively appropriate action. With probability $1-\rho$, agent i is a *strategic* (or sophist) type ($\theta_i = 0$), who chooses her action based on her core motive and the expected consequences for her image. As we describe below, we model this by letting θ_i multiply a utility term for the appropriate action. Player j's type θ_j is drawn similarly and independently of θ_i.

Core motive and image utility. Strategic agents are pulled in two directions. On one side, an agent's *core motive* is a parameter $b_i \in \mathbb{R}$ that multiplies the action a_i. When $b_i > 0$, agent i favors action $a = 1$, and when $b_i < 0$, she favors $a = 0$. Core motives may be based on moral intuitions and emotions like disgust, compassion, or fairness, which have been shown to differ systematically across the political spectrum (Haidt, 2012; Graham et al., 2009). They may also reflect private material benefits, as when richer agents favor lower taxes or firms favor lighter regulation. In the hiring decision example, core motives may reflect ingroup loyalty or, more generally, how universally one's altruism extends to groups with greater social or ethnic distance (Enke et al., 2023).

On the other side, agents value a reputation for being reasonable and responsive to evidence, a Socratic ideal that is pervasive in many cultures. Experimental evidence shows that people want to appear honest and unbiased to others (Abeler et al., 2019) and need to be paid to share fake news, even when politically congruent (Altay et al., 2022). Moreover, people spend more effort constructing unbiased arguments (Lerner and Tetlock, 1999) and consume a more balanced information diet (Moehring and Molina, 2023) when they expect a critical audience. We model this *image utility* from being a reasonable decision-maker as

$$\mu_i \Pr\big[\theta_i = 1 \,\big|\, \mathcal{I}\big],$$

where $\mu_i > 0$ is a scaling parameter that multiplies the probability an observer assigns to the agent being a reasonable type, given the observer's information set $\mathcal{I}$, described below.

The utility accruing to agent i if she chooses action a_i is

$$\underbrace{\theta_i \gamma_i \mathbf{1}\{a_i = \xi\}}_{\text{Socratic utility}} + \underbrace{b_i a_i}_{\text{core utility}} + \underbrace{\mu_i \Pr\big[\theta_i = 1 \,\big|\, \mathcal{I}\big]}_{\text{image}}. \tag{1}$$

Thus, relative to the strategic type ($\theta_i = 0$), the reasonable type ($\theta_i = 1$) has an additional Socratic utility component that puts a premium on taking the appropriate action. The individual scaling parameter γ_i is such that $\gamma_i > |b_i| + \mu_i$, so that the Socratic utility dominates: reasonable types always play the normatively appropriate action ($a = \xi$).[3] All parameters are common knowledge. Types are private information, and ξ is publicly observed at $t = 0$ but unavailable at $t = 1$. We next describe the elements of $\mathcal{I}$ and the nature of the observer.

[3]The reasonable type could thus be modeled with a much simpler utility function. We model this type in a fashion that is comparable to the strategic type only to facilitate the analysis of preferences over interaction environments (see Section 4).

Rationales and rationalizations. Central to the model are instances of conflict between the core motive b and the socially appropriate action indicated by ξ. For instance, agents' core motives may favor an ingroup job applicant, but the applicant may score poorly on socially acceptable criteria like skills or credentials, captured by ξ. To avoid a conflict between core motive and image, the agent may engage in rationalization. Through the use of selective attention or creative reinterpretation, she may reframe the decision context to make her preferred action appear more reasonable. An employer may thus favor the ingroup applicant by selectively highlighting her value to the organization, or by questioning the validity of certain diplomas or test scores of the outgroup applicant.

After observing ξ, each agent chooses an action $a \in \{0, 1\}$ that determines her core utility and the type of rationale she tries to articulate in favor of her position. Let $\hat{\xi}_i \in \{0, 1, \varnothing\}$ denote agent i's *articulated rationale*, i.e., the rationale that actually enters discourse. When $a_i = \xi$, the rationale merely restates the public signal and is straightforwardly available, so $\hat{\xi}_i = \xi$. When $a_i \neq \xi$, the agent attempts a *rationalization* — an argument that goes against the original signal. We think of rationales as facts, pieces of evidence, or logical arguments that are credible and compelling enough to prescribe behavior to any reasonable person. The binary nature of our model makes the distinction between truthful rationales and rationalizations particularly stark. In reality, rationalizations may range from outright falsehoods to subtle misrepresentations of the truth, and, depending on the domain, may be more or less compelling than arguments in favor of the truth.

Following Williams (2023), we assume that such rationales are hard to construct: with probability $\tau_i \in (0, 1)$ the attempt succeeds and $\hat{\xi}_i = a_i$ enters discourse; with probability $1 - \tau_i$ it fails and the agent is *morally dumbfounded* (Haidt and Hersh, 2001; Haidt, 2001), unable to articulate any rationale at all, and $\hat{\xi}_i = \varnothing$. We assume that τ_i and τ_j are common knowledge, and rationalization outcomes are independent across agents conditional on attempting.[4]

The limit case with $\tau_i, \tau_j \to 1$ features perfect co-movement between actions and articulated rationales. In this case, the presence of rationales does not generate predictions about actions that do not follow from a model without rationales. When rationalization can fail ($\tau_i, \tau_j < 1$), the existence of rationales carries information about the state ξ, and can affect how actions are interpreted, and therefore, the actions themselves.

[4]We assume that the choice of an action automatically generates an attempt to rationalize it. An alternative, richer model is one where the agent selects an action and independently decides whether to try to rationalize it. We formulate and analyze such a model in the Supplemental Appendix and establish that, under a mild tie-breaking assumption, this model predicts that players always try to come up with a convenient narrative in equilibrium, as doing so (weakly) improves their image. Thus, the two models predict the same equilibrium actions, rationales, on-path beliefs, and discourse.

Information set and discourse. At $t = 1$, an observer makes inferences about the types of the players. Since the original signal ξ is no longer available, inference is based on the players' actions and "public discourse". Once articulated, a rationale enters a public discourse, i.e. the pool of rationales available at the moment of judgment. A dumbfounded agent contributes nothing to this pool: her articulated rationale is $\varnothing$. Rationales are *non-rival*: a successful rationalization becomes a piece of common discourse available to support any aligned action by any player. The observer registers only which arguments are represented in discourse, not their multiplicity, attribution, or origin. In particular, when one agent is dumbfounded and the other has articulated a rationale $\hat{\xi} \in \{0, 1\}$, the observer sees only that $\hat{\xi}$ is present; she cannot tell whether both agents forwarded this rationale, or one agent was dumbfounded. Formally, the observer's view of discourse is summarized by a coarsening f of the articulated rationales with the following properties:

$$\begin{cases} f(\varnothing,\, \varnothing) = \varnothing, \\ f(\hat{\xi},\, \varnothing) = f(\varnothing,\, \hat{\xi}) = f(\hat{\xi},\, \hat{\xi}) = \{\hat{\xi}\} \quad \text{for } \hat{\xi} \in \{0, 1\}, \\ f(0,\, 1) = f(1,\, 0) = \{0, 1\}. \end{cases}$$

The observer's information set is then $\mathcal{I} = \big(a_i,\, a_j,\, f(\hat{\xi}_i,\, \hat{\xi}_j)\big)$, which she uses to confer image utility on each agent.[5]

Three interpretations of the observer. We do not commit to a particular identity for the observer. Instead, the image term in Equation (1) admits three interpretations, which are mathematically equivalent because they share the same information set $\mathcal{I}$:

1. *Self-image.* The observer is i's own date-1 self. By the time image utility is realized she has lost direct access to the original signal and her own type θ_i, and so infers her morality from the observed actions and from public discourse, $\mathcal{I}$. The model is then a model of self-signaling, or self-deception, in which the agent actively manages her own beliefs (Bodner and Prelec, 2002; Bénabou and Tirole, 2011, 2016). This line of theorizing is supported by evidence on excuse seeking in anonymous contexts (Dana et al., 2007; Exley, 2015; Dana and van der Weele, 2025; Bénabou and Henkel, 2025).

2. *Social image.* The observer is the other agent j, who has also lost access to ξ and her own type θ_j. Player i at $t = 0$ cares about j's actual perception of her type at $t = 1$,

[5]The modeling of rationalization is a formalization of the key insights of the social intuitionist model in Haidt (2001). The assumption that narratives can be shared and used between agents is not explicit in Haidt, but does align with some of his writing, e.g. that "moral judgment is not just a single act that occurs in a single person's mind but is an ongoing process, often spread out over time and over multiple people. Reasons and arguments can circulate and affect people..." (p. 829).

which j forms from the same publicly available record $\mathcal{I}$.[6] The model then captures how the agent persuades herself inadvertently, as a by-product of her attempts to persuade agent j, consistent with evidence described in Mercier and Sperber (2011) and Schwardmann et al. (2022).

3. *External image.* The observer is an uninformed third party (e.g., an audience) who never had access to ξ, θ_i or θ_j, and makes inferences from $\mathcal{I}$. This interpretation of the image term maps naturally into the business of politics as well as the business of political influence online, two settings that frequently feature discourse between individuals with private interests in pursuit of votes or clout.

The three interpretations differ in the source of asymmetric information — what the agent knows or has forgotten about herself and about the state — but yield the same equilibrium analysis. Image concerns may also reflect a mixture of the three foundations.

Equilibrium concept. We look for Perfect Bayesian Equilibria of the game. Since reasonable types have a strictly dominant strategy, we only have to predict the behavior of strategic agents. If i is strategic, we represent her strategy as a pair $(\sigma_i^{\xi=0}, \sigma_i^{\xi=1})$, where $\sigma_i^{\xi} \in [0,1]$ is the probability that i plays the action prescribed by the signal ξ ($a_i = \xi$). With the complementary probability $1-\sigma_i^{\xi}$, i plays the opposite action and attempts to rationalize it ($a_i = 1-\xi$). We use a star superscript to denote equilibrium strategies.

Off-path beliefs are disciplined by two refinements, stated in full in Appendix A.1. The first rules out permissive attributions of off-path dumbfoundedness; the second imposes a global consistency condition on the observer's beliefs across agents that is a consequence of Bayes' rule for on-path beliefs.

1. *Reasonable types never rationalize.* If an agent deviates from a fully realistic equilibrium strategy and has no supporting rationale ex post, the deviation is attributed to the strategic type.

2. *Consistency.* On counter-partisan disagreement events that are off path under full realism by both players, the observer's posteriors about θ_i and θ_j sum to the prior ρ.

3 Equilibrium analysis

We start by defining some terms that will be useful in our analysis. First, we will refer to actions and rationales that are aligned with the core motive as "convenient", and the

[6] We note that social image here may map into how pleasant a conversation feels, an object that is measured in many empirical studies on political conversations.

other actions/rationales as "inconvenient". Second, following the terminology in Bursztyn et al. (2023), we say an action receives "cover" when public discourse contains a supportive rationale for that action. Cover will play a key role in the equilibrium analysis, as a lack of cover identifies a failed rationalization and leads to an image of zero.

We investigate equilibrium behavior in three environments. We first provide a benchmark in which the observer sees only i's action a_i and articulated rationale $\hat{\xi}_i$, a case that we will refer to as *autarky*. We then analyze social interactions between i and j, where the observer sees both actions and the public discourse constructed from the articulated rationales. The compatibility of i and j's core motives is crucial to the results, as the signs of b_i and b_j determine the directions in which individuals are tempted to rationalize. Without loss of generality, we focus on a player with $b_i > 0$. We will refer to interactions with a co-partisan, who has aligned interests ($b_j > 0$), as an *echo chamber*, and interactions with a counter-partisan ($b_j < 0$) as an *agora*, after the classical Greek public square in which the exchange of ideas took place.

A useful preliminary result is that we can characterize the strategy in the good-news state in any equilibrium and interaction environment:

Lemma 1. *Suppose $b_i > 0$. In every equilibrium, player i stays realistic when state and core motive align: $\sigma_i^{\xi=1,*} = 1$.*

When news is good, there is no trade-off: the agent takes the action prescribed by her core motives, and truthfully advances the rationale that supports that action. Lemma 1 applies symmetrically when the agent prefers state $\xi = 0$. We can therefore describe any equilibrium by i's bad-news realism $\sigma_i^{\xi=0,*}$ and the analogous bad-news realism of j, namely $\sigma_j^{\xi=0,*}$ for a co-partisan ($b_j > 0$) and $\sigma_j^{\xi=1,*}$ for a counter-partisan ($b_j < 0$). For convenience we drop the ξ superscript and write the objects simply as (σ_i^*, σ_j^*). We write $U_i^{\xi}(a; \sigma_i, \sigma_j)$ for strategic i's interim utility from playing action a in state ξ given the strategy profile (σ_i, σ_j). In autarky, we omit the absent σ_j argument and write $U_i^{\xi}(a; \sigma_i)$.

3.1 Autarky

To illustrate the fundamental trade-offs in the model, and to develop a benchmark for the social setting, we study the case of solitary reasoning, or autarky. In autarky, the agent's image $\Pr[\theta_i = 1 \,|\, \mathcal{I}]$ is based only on her own action a_i and her articulated rationale $\hat{\xi}_i$.

Fix a tentative level of realism σ_i for player i. Upon observing inconvenient news ($\xi = 0$), agent i may remain realistic and play $a_i = 0$. Her utility from this action is

$$U_i^0(0; \sigma_i) = \mu_i \frac{\rho}{\rho + (1-\rho)\sigma_i}. \tag{2}$$

Playing $a_i = 0$ reveals that $\xi = 0$, because no agent, reasonable or strategic, plays $a_i = 0$ after $\xi = 1$. This leads to a positive image, the magnitude of which is decreasing in σ_i: when the strategic type is more realistic, her strategy is closer to that of the reasonable type, so the inconvenient action is less informative.

If, instead, i chooses to rationalize and plays the convenient action $a_i = 1$, she reaps utility b_i from indulging her core motives. To obtain a positive image, she needs a public rationale that supports this action ("cover"), which she generates with probability τ_i. By Bayes' rule, the payoff from rationalizations is given by

$$U_i^0(1;\sigma_i) = b_i + \mu_i\tau_i\Big[\frac{\lambda\rho}{\lambda + (1-\lambda)(1-\rho)\tau_i(1-\sigma_i)}\Big]. \tag{3}$$

The agent's image is now increasing in σ_i, as more realism by the strategic type makes the convenient action less diagnostic of a rationalization.

Equations (2) and (3) illustrate the trade-off that the agent faces when confronted with bad news ($\xi = 0$). Advancing the truthful rationale and playing the inconvenient action $a_i = 0$ improves her image, but results in forgoing $b_i > 0$. The net benefit from rationalization $U_i^0(1;\sigma_i) - U_i^0(0;\sigma_i)$ determines the agent's best response. Since this expression is strictly increasing in σ_i, there is a unique equilibrium characterized as follows.[7]

Proposition 1 (Equilibrium in autarky)**.** *Suppose $b_i > 0$. There exists a unique equilibrium of the game, in which the agent's level of realism σ_i^* is nondecreasing in μ_i (image concerns) and nonincreasing in b_i (core motive) and τ_i (rationalization ability).*

The individual equilibrium features a number of results that will carry over to our analysis of social interactions. First, the model generates motivated cognition, since the agent's core motives and characteristics causally determine the agent's ex-post beliefs about the state and the morality of the different actions. This criterion for identifying motivated beliefs is the one proposed in Little (2025). Second, the agent is more likely to rationalize when her core motive is large, rationalizations are easier to come by, and image concerns are smaller. Note that the fixed point may be interior and the equilibrium is unique, because a higher equilibrium probability of rationalization undermines the effectiveness of the rationalization, as a Bayesian agent discounts it more heavily.

Finally, rationales are not just epiphenomenal justifications but actually drive behavior. In particular, if rationalization becomes easier (τ_i increases), a strategic type is more likely to engage in the convenient action. The reason is that rationales allow the agent to obfuscate her true motives, and cushion the image consequences of ignoring the original signal.

[7] An interior σ_i^* equates (2) and (3), while $\sigma_i^* = 1$ is an equilibrium if and only if $U_i^0(0;1) \geq U_i^0(1;1)$, and $\sigma_i^* = 0$ is an equilibrium if and only if $U_i^0(0;0) \leq U_i^0(1;0)$.

3.2 Echo chambers

Consider two co-partisans with $b_i > 0$ and $b_j > 0$. Players i and j both prefer action $a = 1$ and will be tempted to seek rationalizations for it. We fix the strategy σ_j of j and consider a tentative best response σ_i by strategic agent i upon seeing bad news $\xi = 0$.

Once again, we compare utilities from realism and rationalization. If i is realistic and plays $a_i = 0$, her utility is the same as in the individual case, and given by (2). Since her action reveals the state, the opponent's strategy does not matter. If i plays $a_i = 1$ instead, she reaps utility b_i from playing her favored action. To obtain a positive image, she needs the original state to remain concealed, which requires there to be cover in public discourse. In contrast to the individual case, cover can be generated by either agent, since rationales are non-rival. Moreover, since motives are aligned, both agents have an incentive to find such a rationale. Conditional on both agents trying to rationalize, the probability that either i or j generates cover is

$$K_{ij} \coloneqq 1 - (1 - \tau_i)(1 - \tau_j) = \tau_i + \tau_j - \tau_i\tau_j. \tag{4}$$

Just obtaining cover is not sufficient for a positive image, however: the echo chamber requires that public discourse contains no contradicting action or rationale, as this would identify the true state. Thus, agent j must also deny the original signal, which happens with probability $(1 - \rho)(1 - \sigma_j)$. With the complementary probability, j remains realistic and reveals the inconvenient state, so any cover generated by i is rendered useless.

For $\sigma_i < 1$, collecting these arguments and computing the Bayesian posteriors,[8] we obtain the following expression for i's expected utility from playing $a_i = 1$:

$$U_i^0(1; \sigma_i, \sigma_j) = b_i + \mu_i \underbrace{K_{ij}\,(1 - \rho)(1 - \sigma_j)}_{\substack{\text{prob. of } a_j=1 \\ \text{and } f(\hat{\xi}_i,\hat{\xi}_j)=\{1\}}} \underbrace{\frac{\lambda\rho}{\lambda + (1 - \lambda)(1 - \rho)^2(1 - \sigma_i)(1 - \sigma_j)\,K_{ij}}}_{\text{image when } (a_i,a_j,f(\hat{\xi}_i,\hat{\xi}_j))=(1,1,\{1\})}. \tag{5}$$

Equation (5) shows how the image term generates strategic complementarities. If player j is a probable realist (high σ_j), she is likely to articulate the truthful rationale $\hat{\xi}_j = 0$, destroying cover for $a = 1$ along with the image of the agent playing it.

As before, the net benefit from rationalization $U_i^0(1; \sigma_i, \sigma_j) - U_i^0(0; \sigma_i)$ determines player i's best response. This expression is increasing in σ_i, yielding a unique best response, and

[8] The posterior beliefs are pinned down by Bayes' rule, except in the off-path scenarios where $\sigma_i = 1$ and i deviates, and either: (i) both i and j's rationalization attempts fail, leaving i dumbfounded; or (ii) i's rationalization attempt succeeds but j stays realistic and reveals $\xi = 0$ through $a_j = 0, \hat{\xi}_j = 0$. The former case is covered by Refinement 1. We do not restrict off-path beliefs in the latter case, but note that the least charitable interpretation of the deviation for i (attributing it to a strategic type) is the one that supports realism on the largest range of parameters. See the proof in the appendix for more details.

decreasing in σ_j, reflecting strategic complementarity. The next result describes the unique equilibrium[9] in the echo chamber, defined as the intersection of i's and j's best-response functions.

Proposition 2 (Equilibrium in echo chambers)**.** *Suppose $b_i > 0$ and $b_j > 0$. There exists a unique equilibrium of the game. In this equilibrium, agent i's level of realism σ_i^* is non-increasing in b_i, b_j (core motives) and τ_i, τ_j (rationalization abilities), and nondecreasing in μ_i, μ_j (image concerns).*

Several things are noteworthy about this equilibrium. The first is that rationalizations are complements. For one agent's rationale to work, there cannot be counter-rationales or incongruent actions, which requires the other agent to rationalize as well. Echo chambers thus generate a form of "groupthink", by which agents' denial and behavior reinforce each other. Sunstein (2002) discusses evidence on when group deliberation among like-minded individuals leads to more extreme outcomes, emphasizing the important role of "argument pools", i.e. the generation of multiple arguments for a given position. This mechanism is formalized by our model, which makes the intuitive prediction that echo chambers lead to more rationalizations and actions that go counter to the state, as rationalizers provide cover by contributing arguments to the pool. The key role of cover is borne out in Bursztyn et al. (2023), where experimentally increasing the public (as opposed to private) supply of rationales increases the number of subjects who are willing to support contentious police reform and anti-immigrant sentiments on Twitter.

Second, the equilibrium is unique despite strategic complementarity because rationalization undermines itself: the more likely a player is to rationalize in equilibrium, the more she discounts her rationales, putting a brake on rationalizations and the strength of the complementarity. Third, comparative statics in parameters relating to agent i follow the same logic as Proposition 1: realism requires a low level of emotional or material attachment (low b_i), low rationalization ability (low τ_i), and high levels of image concerns (high μ_i).

Finally, because of the complementarity in rationalizations, agent i's equilibrium denial is increasing in agent j's core motive and rationalization ability, and decreasing in j's desire to appear reasonable. The dependence of agent i's cognition and ultimate action on agent j's preferences and skills is a new, testable prediction from the model. This implies that an intervention that narrows one agent's scope for rationalization should spill over to her conversation partner, raising the partner's realism as well.

[9]Throughout the paper, we say that the equilibrium is unique when the equilibrium strategy profile (σ_i^*, σ_j^*) is uniquely determined, even if it can be supported by multiple off-path beliefs.

3.3 Agoras

Suppose now that $b_i > 0$ but $b_j < 0$, so that the two players are looking for opposite rationales. The signals $\xi = 1$ and $\xi = 0$ are good and bad news for i, respectively, and vice versa for j.

To understand the strategic incentives, we once more compare the expected utility for i from realism and from rationalization upon seeing bad news, conditional on the strategies (σ_i, σ_j). The utility from realism is the same as in Equation (2). When we turn to the utility from rationalization, there are two key differences from the echo chamber. First, conditional on observing $\xi = 0$, i knows that j faces good news and will always play $a_j = 0$, with $\hat{\xi}_j = 0$. Since j will not provide cover for $a_i = 1$, i will need to do so herself, or obtain an image of zero. Second, if neither player stays fully realistic, the action profile $\{(a_i = 1, \hat{\xi}_i = 1), (a_j = 0, \hat{\xi}_j = 0)\}$ does not reveal ξ: in contrast to the echo chamber, disagreement implies that both players are playing their convenient actions, and either player may be rationalizing.

Despite these differences, we again find that rationalizations are complements, as i's expected utility after rationalizing equals

$$U_i^0(1; \sigma_i, \sigma_j) = b_i + \mu_i \underbrace{\tau_i}_{\text{prob. of } \hat{\xi}_i = 1} \underbrace{\frac{\lambda\rho\tau_j(1-\sigma_j)}{\lambda\tau_j(1-\sigma_j) + (1-\lambda)\tau_i(1-\sigma_i)}}_{\text{image when } (a_i, a_j, f(\hat{\xi}_i, \hat{\xi}_j)) = (1,0,\{0,1\})}. \tag{6}$$

This expression is well-defined whenever $\min(\sigma_i, \sigma_j) < 1$, in which case it is decreasing in σ_j and exhibits complementarity in rationalizations. While the result is similar, the intuition differs from that in the echo chamber. If i's rationalization succeeds ($\hat{\xi}_i = 1$) it enters discourse alongside j's truthful $\hat{\xi}_j = 0$. The observer must then adjudicate between two stories: $\xi = 0$ with i rationalizing, or $\xi = 1$ with j rationalizing. She bases her judgment on the equilibrium strategies: the more realistic j is (the higher σ_j), the more credible her rationale, and the lower i's image. By contrast, the higher j's rationalization probability is, the more credible i's rationale appears, raising i's image.

The following proposition characterizes the unique equilibrium.

Proposition 3 (Equilibrium in the agora)**.** *Suppose that $b_i > 0$ and $b_j < 0$. There exists a unique equilibrium of the game. In this equilibrium, agent i's level of realism σ_i^* is nonincreasing in b_i (i's core motive) and τ_i, τ_j (rationalization abilities), and nondecreasing in b_j (j's core motive) as well as μ_i, μ_j (image concerns).*

Proposition 3 shows that complementarities in rationalizations lead to anti-groupthink, or *polarization*, whereby core motives b lead the two camps to diverge in their actions and the

accompanying rationalizations. As was the case in the echo chamber, polarization is driven by the agent's own core motive b, her efficiency of rationalization τ, and the strength of her image concerns μ, as well as those of her partner. However, unlike in the echo chamber, complementarities are not driven by joint investments in the same rationale. Instead, they arise because the opponent's tendency to rationalize decreases her credibility as a source of epistemic discipline. The model thus predicts that if one side weakens its commitment to the truth — e.g., through an increase in τ or $|b|$ — the other side will strategically increase its own rationalizations, as it can now more credibly blame any disagreement on the other side's denial.

Returning to the leading example, the same logic operates within a hiring committee. An evaluator known to advocate for her preferred candidate whatever the evidence is easy to discount, and her presence licenses rather than restrains her opponent, who can attribute any disagreement to her partisanship. The prediction also applies to political movements. The model can explain instances where progressive and nativist positions on questions of race and immigration both drift from the evidence and polarize, each licensed by the other's excesses. To our knowledge, these are novel predictions that should be tested by empirical research.

In both the echo chamber and the agora, comparative statics obtain under each of the three interpretations of the image term introduced in Section 2. A more realistic partner makes it harder for the agent to believe her own rationales (self-image). A more realistic partner is also more critical of i's behavior and rationales (social image), and acts to make them less credible to an outside audience (external image). As a result, the complementarity we uncover is likely to be robust to different audiences and, through self-image, can operate even when the agent's own behavior is not observed by anyone else.

3.4 Extremism breeds extremism

The combined comparative statics of Propositions 2 and 3 yield a prediction that we lay out in the following corollary.

Corollary 1. *The equilibrium level of realism σ_i^* of individual i is hill-shaped in b_j.*

Figure 2 illustrates Corollary 1 by plotting i's equilibrium realism against the opponent's core motive b_j. We make a number of observations. First, the hill shape confirms that agent i is most realistic when interacting with a moderate opponent (small $|b_j|$). If $|b_j|$ is close to zero, j is fully realistic, and the exact magnitude of j's core motive is irrelevant to i's behavior, which explains the flat top of the hill. Similarly, when $|b_j|$ is large, j is already maximally in denial, and σ_i^* is flat. The hill shape shows that political moderates enforce

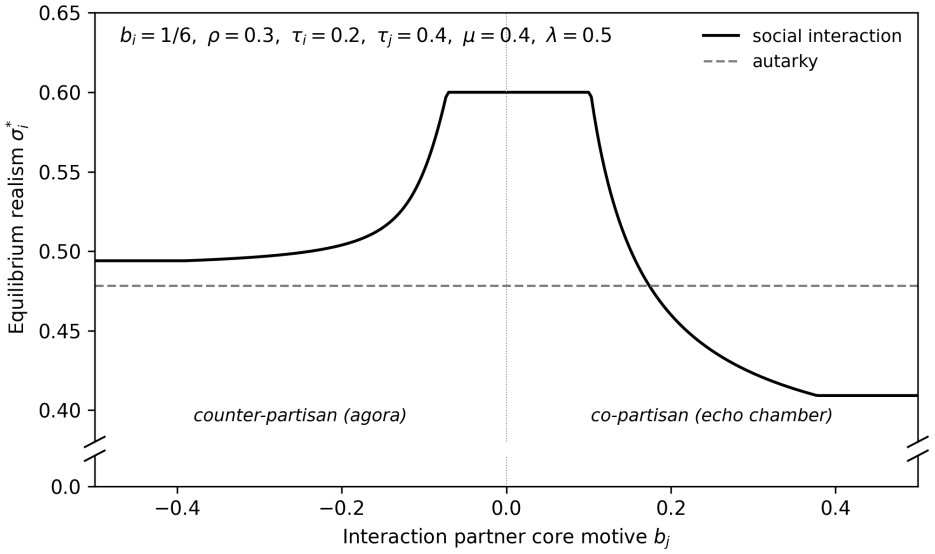

Figure 2: Equilibrium realism σ_i^* as a function of the partner's core motive b_j. Parameters: $b_i = 1/6$, $\mu_i = \mu_j = 0.4$, $\rho = 0.3$, $\lambda = 0.5$, $\tau_i = 0.2$, $\tau_j = 0.4$.

epistemic discipline: they provide less cover to co-partisan rationalizations and are more credible opponents to counter-partisans. Conversely, an extremist opponent, whether allied or opposed, is less informative and less credible, and hence weakens the image penalty on rationalizations.

Thus, "extremism breeds extremism," as the radicalization of one player (through an exogenous increase in $|b_j|$) has an epistemic multiplier effect which increases rationalizations by both players. Similarly, party leaders with more extreme preferences provide cover for their supporters who are tempted to rationalize. This aligns with results from Bursztyn et al. (2020), who showed how the 2016 election of Donald Trump, who sometimes referred to (Latin American) immigrants as "rapists" and "bad hombres", increased the willingness to publicly express xenophobic views. In an experiment, Buschinger et al. (2026) find that exposure to extreme justifications for discrimination increases polarization and the prevalence of extreme reasoning.

Second, conditional on $|b_j|$, whether a co- or counter-partisan partner has a larger effect on i's realism is ambiguous and depends on the relative size of two effects. On the one hand, co-partisans can provide cover for the favored action by producing viable rationales even when i's attempt at rationalizing fails. This reduces the inherent risk of rationalizing and playing the convenient action. On the other hand, if a co-partisan disagrees, for instance because she is a reasonable type, this perfectly reveals the inconvenient state, thereby making i's rationalizing more risky. This effect is smaller for counter-partisans, because their dissent

can be chalked up to insincerity.[10]

Finally, the dashed horizontal line in Figure 2 displays the equilibrium realism under autarky. With the parameter values considered here, realism in autarky is lower than in counter-partisan interactions, but higher than in interactions with very ideological co-partisans. In the next section we will make these comparisons more precise, and discuss how agents value autarky in comparison to the different types of social interactions.

4 Preferences for interaction partners

Up to this point, we have investigated equilibrium strategies conditional on the state, the type of the agent and the type of her opponent. In this section, we ask whether an agent may prefer a particular interaction, given the resulting equilibria. More specifically, we ask whether an agent would prefer an echo chamber or an agora, rather than reasoning in isolation. This allows the model to speak to a literature on assortative matching that has attracted considerable attention in the social sciences, and has generated renewed interest with the arrival of social media. We first discuss the criteria by which an agent may evaluate these interactions. We then derive results on preferences for particular interactions and discuss the relation to the empirical literature.

4.1 The value of denial

We define two ways in which the agent may evaluate the equilibrium outcomes of different interactions. First, we consider an ex-ante perspective where agents do not yet know their type or the state, and average over both to construct and compare expected utilities (W_i). This reflects the idea that agents cannot choose interactions based on their type, to which they may not have introspective access, or on the (stochastic) arrival of public signals. Second, we characterize expected utility conditional on the type (W_i^R and W_i^S, respectively), while still averaging over the state. This allows us to compare the preferences of reasonable and strategic types.

When we apply these expected utility terms to equilibrium outcomes, it turns out that agents prefer environments that are conducive to more rationalization. The following proposition and discussion make this precise.

[10]Note that the asymmetry provides an incentive to mischaracterize a co-partisan who plays a misaligned action/narrative as a counter-partisan in disguise, in order to allow dissent to be characterized as self-serving. While such mischaracterization is not a formal feature of the model, it does explain why people are particularly sensitive to deviations by ingroup members.

Proposition 4 (The value of denial)**.** *Suppose that $b_i > 0$ and consider a game played by i in autarky, or against a partner with core motive b_j. Let σ_i^* be i's equilibrium strategy in that game.*

i) Agent i's ex-ante expected utility is

$$W_i \;=\; C_i \;+\; b_i\,(1-\rho)(1-\lambda)\left(1-\sigma_i^*\right), \tag{7}$$

with $C_i \coloneqq \rho\gamma_i \;\; + \; \rho\,\mu_i \; + \; \lambda\,b_i$.

ii) Suppose $\sigma_i^* > 0$. Then, expected utility for reasonable and strategic types is given by

$$W_i^R = W_i + (1-\rho)\gamma_i \text{ and} \tag{8}$$

$$W_i^S = W_i - \rho\gamma_i, \tag{9}$$

respectively. Thus, if $\sigma_i^* > 0$ in autarky, echo chamber and agora, strategic and reasonable types rank environments identically, and prefer those that induce less realism.

iii) Suppose instead that $\sigma_i^* = 0$ in two environments. The strategic and reasonable types then have the opposite preference over these environments.

Proposition 4.i) shows that ex-ante utility is made up of two components. The first is a constant C_i, which depends on expected payoffs from truthfulness ($\rho\gamma_i$), the expected image ($\rho\mu_i$) and the expected payoff from satisfying the core motive in the good state (λb_i). The second term reflects the frequency with which the strategic type satisfies her core motive in the inconvenient state, which is directly proportional to the equilibrium degree of rationalization ($1-\sigma_i^*$). We will refer to this component as the *indulgence effect*. Note that any image effect of rationalizations (across states, or types) washes out of ex-ante utility due to the martingale property of beliefs. Expected utility depends on i's rationalization ability τ_i and on j's characteristics (b_j, μ_j, τ_j) only through σ_i^*, and hence displays the opposite comparative statics to equilibrium realism, as pinned down in Propositions 2 and 3.

Proposition 4.ii) considers expected utility for the two types separately. The key finding is that even though they have different utility functions, both types rank environments in the same way, unless at least one environment allows full denial. To understand this, consider that more rationalizations reduce the image associated with the convenient action, which is played disproportionally by the strategic type. This implies a transfer of image from the strategic to the reasonable type, as the Socratic virtues of the latter become more visible. This *virtue salience effect* explains why the reasonable type benefits from lower realism. For

the strategic type, the virtue salience effect is dominated by the indulgence effect whenever $\sigma_i^* > 0$, explaining why both types prefer environments with less realism.

Conversely, Proposition 4.iii) shows that if $\sigma_i^* = 0$ in each of the environments, the two types have opposing preferences. In the corner solution the environment does not discipline behavior, the indulgence effect disappears, and only the zero-sum image transfer remains. In that case, the expected utility of both types starts to depend on the cognitive parameters that determine ex-post image (τ_i, and σ_j^*, τ_j in interactions) directly, and not just through their effects on σ_i^*.

We highlight a number of take-aways from this exercise. First, agents display a "madder-is-better" ethos, whereby the appeal of different interactions is fully determined by the degree of denial that they facilitate. The strategic types seek liberation from a moralizing public discourse that keeps them from indulging their core motives. The reasonable types appreciate an environment where they can flaunt their epistemic virtues most saliently. Second, outside of the corner solution where $\sigma_i^* = 0$, it does not matter whether we evaluate environments by ex-ante expected utility, or the expected utility of either type of agent. Since all types prefer the same environment, there is also no signaling value of explicitly choosing one environment over the other.

Finally, one should not mistake these evaluation criteria for the welfare of society as a whole. Overall welfare will depend on other elements that we have not modeled. For instance, a social planner might be concerned about the externalities associated with inappropriate actions, as in Bénabou and Tirole (2026). This would generate a concern for the overall information environment and a demand for policies to deter rationalizations.

4.2 Preferences for echo chambers

We now investigate when the agent prefers to enter an echo chamber or an agora, as opposed to reasoning alone. We will take the ex-ante perspective of Section 4.1, i.e. that of an agent who does not yet know whether she will be strategic or reasonable and who internalizes the utility of each type. We have seen in the previous section that an agent's expected utility is increasing in the amount of rationalizations in the environment. Characterizing preferences for interactions thus boils down to determining whether the degree of realism in the echo chamber or agora exceeds that in autarky. We can characterize (weak) preferences in the following way.

Proposition 5 (Selection into interactions)**.** *Suppose that* $b_i > 0$.

i) If $(1-\rho)\,K_{ij} > \tau_i$, *there exists a threshold* $\overline{b}_j > 0$ *such that agent* i *prefers an echo chamber to autarky if* $b_j \geq \overline{b}_j$, *and agent* i *prefers autarky to an echo chamber if* $b_j \leq \overline{b}_j$.

If $(1-\rho)\,K_{ij} \leq \tau_i$*, agent* i *always prefers autarky to an echo chamber.*

ii) Agent i *always prefers autarky to an agora.*

Proposition 5 illuminates the preferences over environments, which are illustrated in Figure 2 (since utility is decreasing in σ_i^*). Proposition 5.i) shows that agents prefer an echo chamber over autarky if two conditions are met, both of which hinge on the cover for i's preferred action. In autarky, cover exists only if i's own attempt succeeds, which happens with probability τ_i. In an echo chamber, cover exists whenever *either* attempt succeeds, which happens with the larger probability K_{ij}, defined in equation (4). However, the additional cover materializes only when j is a strategic type who is herself rationalizing, an event of probability $(1-\rho)(1-\sigma_j^*)$. So the agent selects into an echo chamber whenever $(1-\rho)(1-\sigma_j^*)K_{ij} > \tau_i$, which happens when $(1-\rho)K_{ij} > \tau_i$ and b_j is large enough.

The relative probabilities of cover, and hence the value of echo chambers, also depend crucially on the rationalization abilities of the agents. The condition $(1-\rho)K_{ij} > \tau_i$ is satisfied for agents who have little cognitive flexibility in constructing their own stories (small τ_i), and expect to meet speakers who are gifted at doing so (large τ_j). This highlights how leaders and politicians may influence voters' beliefs: by being a reliable source of cover, they encourage rationalizations by others.

Proposition 5.ii) states that the agora is unambiguously disciplining. Upon observing a bad public signal, i knows that her counter-partisan interaction partner faced good news and will play $a_j = 0$ with $\hat{\xi}_j = 0$ for sure. Discourse therefore always contains a rationale *against* i's preferred action. Even if agent i has reason to discount j's narrative due to her tendency to rationalize, the mere existence of a counter-rationale exerts discipline on i's own rationalizations, which, as we showed above, she does not like. Notably, agents in the agora prefer a counter-partisan with high rhetorical ability τ_j, as this makes it easier to discount her dissenting rationales.

Proposition 5 predicts a preference for co-partisan interactions over autarky and thereby provides a cognitive foundation for sorting into biased political information environments like slanted newspapers or peer conversations (Mullainathan and Shleifer, 2005; Nyhan et al., 2023; Robertson et al., 2023; Braghieri et al., 2024b), and, closest to our interpretation, in the context of political conversations (Braghieri et al., 2024a; Frimer et al., 2017; Bauer et al., 2026). However, a distinctive prediction of the model is that agents' preference is conditional. Echo chambers are most attractive and influential if they contain skilled and biased rationalizers, those who facilitate the agent's own efforts at rationalization. This separates the model from theories based on a pure taste for like-minded interactions.

This is also important to align the predictions of the model with the empirical evidence.

The literature indeed shows no clear effect of the generic use of social media on attitudes and beliefs, but does chronicle a clear impact of partisan influencers. In particular, disconnecting Facebook users prior to the 2020 U.S. presidential election had no effect on affective or issue polarization (Allcott et al., 2024), and there is no evidence that attitudes move with exposure to like-minded sources (Nyhan et al., 2023) or to reshared content (Guess et al., 2023). However, unfollowing hyperpartisan influencers on Twitter durably reduces out-party animosity (Rathje et al., 2024), and Donald Trump's tweets about Muslims causally increased xenophobic tweets and hate crimes by his followers (Müller and Schwarz, 2023).

4.3 Preferences for extremes

Finally, we highlight a direct corollary of Proposition 4 when combined with Propositions 2 and 3. Agents have a preference for extremism, as extremist interaction partners allow more rationalizations. This prediction contrasts with identity-based models of social interactions, which typically assume that individuals prefer partners whose preferences are more similar to their own. The following corollary, which mirrors Corollary 1, makes this precise.

Corollary 2 (Preference for extremism)**.** *Fix $b_i > 0$. Then W_i is nondecreasing in τ_i and τ_j (rationalization abilities), nonincreasing in μ_j (j's image concerns), and, by Corollary 1, valley-shaped in b_j: agent i is worst off when matched with a moderate partner and better off against an extremist of either stripe.*

The preference for extremes derives from the indulgence effect, the tendency of strategic types to avoid moralizing public discourse. This can explain the phenomenon of "political acrophily", a documented preference to connect to co-partisans who are more extreme than oneself (Goldenberg et al., 2023). This tendency is amplified on social media, where users disproportionately follow and retweet more radical ingroup voices (Zimmerman et al., 2024; Zarouali, 2026). The demand for extremes can also help explain why social media companies algorithmically favor divisive political content (Horwitz, 2023).

5 Naiveté and polarization

So far, our analysis has assumed that observers are fully sophisticated about the equilibrium rationalization probabilities. At least for the case where the agent observes herself (self-image interpretation), this contradicts psychological evidence on the nature of motivated reasoning, which shows that people are largely unaware of their tendency to engage in rationalizations and their selective attention to evidence (Simler and Hanson, 2017; Kappes and Sharot,

2019; Melnikoff and Strohminger, 2020). At the same time, people are highly skeptical when others put forward competing rationales, and they are often quick to point out self-serving biases in others' reasoning (Pronin et al., 2002; Kurzban, 2011; Mercier and Sperber, 2011; Hagenbach and Saucet, 2025).

The assumption of sophistication also leads to some counterfactual predictions. For instance, for any information set, both players are in perfect agreement about the probability that either side engaged in rationalization, the likely value of the original ξ, and the assessment of each other's moral type. This contradicts well-documented facts about political beliefs, such as the existence of political polarization, and the research on the contact hypothesis, which finds that people update their beliefs in a systematic direction when they talk to outgroup members.

In order to explain such phenomena, and belief biases more generally, we propose a parsimonious model of *naive* Bayesians. Motivated by the literature cited above, we model naiveté as insufficient skepticism about one's own rationalizations. At $t = 1$, agent i believes that the probability that her own rationalizations are successful is given by $\tau_i(1 - \chi_i)$, while remaining sophisticated about j's rationalization tendencies. Thus, $\chi_i \in [0, 1)$ measures the degree of naiveté: at $\chi_i = 0$ we recover the sophisticated benchmark, while as $\chi_i \to 1$ the agent takes her own rationale at face value, as in Brunnermeier and Parker (2005).[11] While we consider this model of naiveté to be particularly well-supported by evidence, the results below can also be obtained from slightly different modeling choices.[12]

Because different observers with different motives differ in the direction of their naiveté, the three types of observers (self, social and external) are no longer equivalent. For simplicity, we focus on the self-image case. Note that if naiveté is unanticipated, the equilibrium analysis of Section 3 remains exactly the same. If naiveté is anticipated, all qualitative equilibrium properties survive, but the level of realism might change: $\chi_i > 0$ pushes σ_i^* down, as naiveté cushions the image loss from rationalizations.[13]

As we now show, naiveté allows the model to fit stylized facts about political polarization and intergroup contact. The results in this section hold for any level of $\sigma_i, \sigma_j < 1$, even off-equilibrium. Of course, in equilibrium, the equilibrium comparative statics hold, e.g. an

[11] We do not discuss the case where $\chi_i = 1$. A complete lack of awareness may confront the agent with incongruous results. For instance, she might rationalize into the convenient belief yet learn the true, inconvenient state from the action of the other player. This case might be resolved either way by making additional assumptions.

[12] For instance, one might assume that agents are insufficiently skeptical about any convenient rationale, including those generated by their co-partisans, that they update more after good news about θ_i than after bad news, or that they have an inflated (and self-serving) prior about θ_i.

[13] The effect of j's naiveté χ_j on i's equilibrium realism, assuming i anticipates it, goes in the same direction: a naive partner denies inconvenient signals with a larger probability, which pushes σ_i^* down.

increase in $|b_j|$ will increase the level of polarization by affecting the strategies of both parties, as shown in Propositions 2 and 3.

5.1 Ideological and affective polarization

Political polarization is sometimes broken down into *ideological polarization*, where different sides of the political spectrum disagree on policy positions, and *affective polarization*, where each side sees the other political camp as less moral than its own. The model with naiveté can accommodate these findings. Let i and j play with each other, and observe the same information set. It turns out that the introduction of naiveté generates both types of polarization, as well as a form of moral superiority. We write p_i and p_j for i and j's subjective (naive) posterior beliefs, distinct from the correct Bayesian posterior considered in the previous sections.

Proposition 6 (Polarization under naiveté)**.** *Suppose $b_i > 0$, $\chi_i > 0$, $\chi_j > 0$ and $\sigma_i < 1$, $\sigma_j < 1$. Then the posterior beliefs display*

1. **Ideological polarization:** *If $b_j < 0$, then $\mathbb{E}[p_i(\xi = 1)] > \lambda > \mathbb{E}[p_j(\xi = 1)]$: i and j on average disagree about the value of ξ, with each agent overestimating the likelihood of her convenient signal;*

2. **Self-righteousness:** *$\mathbb{E}[p_i(\theta_i = 1)] > \rho$: i's mean posterior about her own morality is biased upwards relative to the truth;*

3. **Affective polarization:** *If $b_j > 0$ (resp. $b_j < 0$), then $\mathbb{E}[p_i(\theta_j = 1)] > \rho$ (resp. $\mathbb{E}[p_i(\theta_j = 1)] < \rho$): i's mean posterior about j's morality is biased upwards (resp. downwards) relative to the truth.*

Proposition 6 shows how naiveté affects the agent's belief system. When an agent retrieves a convenient signal at $t = 1$, naiveté leads her to underestimate the likelihood of successful rationalization and overestimate the likelihood that it reflects the original signal. This tilts her beliefs towards the convenient state, which generates *ideological polarization* between counter-partisans, as average beliefs about ξ, i.e. about the appropriate action, diverge between i and j. Naiveté does not generate systematic disagreements between co-partisans, although the size of the individual bias will vary with individual parameters χ_i and χ_j.

Second, naiveté generates a form of overconfidence we call *self-righteousness*: even though she knows the true strategy profile, the agent overestimates the probability that the recovered rationale reflects the original signal, and hence the probability that she is a reasonable

type. By the same mechanism, naiveté also leads her to overrate the morality of her co-partisans. The flip side of this is *affective polarization*: as i overestimates the probability of the convenient signal ξ, she also overestimates the likelihood of rationalization by j, who plays the opposite action (i.e. with oppositely signed b). Thus she is too confident the other side is wrong and unprincipled.[14]

The model thus offers a cognitive explanation for political polarization. This contrasts with explanations based on social identity theory, which rely on identification with partisan ingroups (Iyengar et al., 2012, 2019). Our model does not assume any ingroup favoritism or outgroup animus, but highlights selection of convenient rationales and naiveté about that selection.[15] This suggests that the rise in polarization, particularly affective polarization in the United States (Gentzkow, 2016; Iyengar et al., 2019), may have to do with changes to the sharing of and search for rationales. For instance, technology like social media that increases access to convenient rationales by lowering the cost of following skilled rationalizers may lead to increased polarization.[16]

5.2 Forced contact with counter-partisans

There exists a large literature on the "contact hypothesis", the idea that contact with an outgroup member can reduce prejudice and hostility. This literature has produced a number of robust findings. First, people avoid contact with outgroups, and are even willing to pay to do so (Rao, 2019; Chetty et al., 2022; Braghieri et al., 2024a; Bauer et al., 2026). Second, if they do interact with outgroups, this often reduces prior prejudice and animosity (Allport, 1954; Rao, 2019; Lowe, 2021; Schindler and Westcott, 2021; Achard et al., 2025). This pattern also holds in politics, where forced interactions between counter-partisans have been found to decrease both affective and factual polarization (Santoro and Broockman, 2022; Blattner and Koenen, 2023; Hobolt et al., 2024; Braghieri et al., 2024a). Moreover, exposing Facebook users to counter-attitudinal news (Levy, 2021) and paying Fox News viewers to watch CNN for a month (Broockman and Kalla, 2025) have similar depolarizing effects.

[14] This notion of affective polarization is more specific than that often used in the empirical literature, which relies on more general measurements, such as a "feeling thermometer", to measure attitudes towards the outgroup (Iyengar et al., 2012).

[15] A distinctive prediction of this model is that the extent of affective polarization depends on the parameters governing the interaction, in particular the core motives, the aptitude for rationalization, and the level of naiveté of both players.

[16] A literature on meta-perceptions and polarization shows that people overestimate how negative outgroups are about them, driving misattributions and a desire for social distance (Lees and Cikara, 2020, 2021; Moore-Berg et al., 2020). The model can deliver such results if people overestimate naiveté in others, in line with Pronin et al. (2002). If i overestimates j's naiveté, then i will overestimate j's bias in all beliefs. This is true for fixed σ_j, but also if σ_j is endogenous, because i then underestimates j's realism.

The pattern is puzzling: if counter-partisan contact builds consensus and reduces animosity, why do people shun it? To study contact, we now consider a pre-contact information structure in which i initially observes $\{a_i, a_j, \hat{\xi}_i\}$ and subsequently observes $\hat{\xi}_j$. The logic of Proposition 6 implies that i's view of j is initially prejudiced, and we ask what happens upon "forced contact", i.e. the (expected or unexpected) observation of the counter-partisan's articulated narrative.

Proposition 7 (Forced contact)**.** *Suppose $b_i > 0 > b_j$, $\chi_i > 0$, i and j have formed their beliefs using $\sigma_i < 1, \sigma_j < 1$, and i has received information set $\{a_i, a_j, \hat{\xi}_i\}$. Suppose that i then observes $\hat{\xi}_j$. Then, i's mean posteriors about ξ, θ_i and θ_j become less biased.*

Proposition 7 shows that the model can match the empirical record. Since the agent is naive about her own rationalizations, she is on average surprised by the quality of the counter-partisan's rationale (i.e. the lack of dumbfoundedness) and updates her beliefs towards the truth. Forced contact with counter-partisans thus reduces disagreement about the state ξ and raises the esteem that agent i bestows on her interlocutor. Since this effect works in opposite directions for both camps, contact reduces polarization.[17]

The model also explains why agents nevertheless shun the outgroup. The price of depolarization is paid in (self-)esteem: learning that the other side is more likely to be right reduces the agent's confidence in her own morality. The upshot is that policy makers who want to foster social cohesion through counter-partisan interactions may have to incentivize or force such contact.

6 Discussion

We investigate the co-production of rationalizations and (im)moral behavior, modeling the trade-off between core motives and esteem as a reasonable decision-maker. Agents produce rationalizations for inappropriate actions that are then available to other agents. Unlike in other models of communication, the exchange of rationales in our model does not serve to inform receivers or persuade them to take certain actions. Instead, communication serves to increase esteem for the speaker and make her actions seem more appropriate.

The model predicts a range of group behaviors, like ingroup preferences, groupthink, prejudice towards the outgroup, and a demand for extremes. It does so without assuming any

[17] The effect of contact is hill-shaped in agent i's naiveté χ_i. Indeed, the average beliefs of fully sophisticated agents ($\chi_i = 0$) do not react to contact because they are well-calibrated either way, while fully naive agents ($\chi_i \approx 1$) are not de-biased either because they interpret disagreement, including a countervailing narrative articulated by a counter-partisan, in a maximally self-serving manner. Thus, the effect of forced contact is maximal for a moderate level of naiveté.

ingroup love, outgroup hostility, or preference for conformity or belief consonance. Instead, all results are driven by the desire to obtain esteem, thus providing a new micro-foundation for models of group identity, such as Akerlof and Kranton (2000).

The model delivers nuanced predictions about belief formation and selection into groups. For instance, the preference for echo chambers over autarky is conditional and depends on agents' relative advantage in rationalization skills. Because the dislike of outgroups is not hardwired, it can lessen upon contact. Further predictions highlight the role of discourse, such as the licensing role of moral cover, and the resulting preference for more extreme ingroup partners. The broad focus on core motives also means that the model can be applied to cultural and identity-driven debates, such as gun control, affirmative action and pandemic responses; or historical interpretations, like justifications of colonialism, or the storming of the U.S. Capitol building on January 6, 2021.

A natural direction for future work is to empirically identify and quantify the interpersonal spillovers of rationalizations predicted by the model. Another direction is to better understand the nature of rationalizations themselves. Because we model a univariate state, all rationalization is directed at a single belief. In richer environments, an agent pursuing a desired conclusion can choose where to direct her rationalization: she may contest the facts or reinterpret them. She may also concede them while reassigning blame, as when partisans converge on casualty counts in the Iraq War while diverging on the interpretation of those numbers (Gaines et al., 2007), or when they agree on economic conditions while disagreeing about which party is responsible (Bisgaard, 2015, 2019). A better understanding of which beliefs make better targets for rationalization, and why, will help sharpen predictions about the incidence of polarization.

We can envision at least two fruitful extensions of our model. First, our model only has two agents; extending the model to larger groups and evaluating how the effect and desirability of information environments move with the number of agents within them would be interesting. Second, one might investigate the externalities associated with rationalizations. As we have shown, rationalizations license others to indulge their core motives. From the perspective of a planner who cares about moral actions being played, rationalizations thus constitute a negative externality that pollutes the information environment. This may generate demand for policies to deter them, like censorship and other speech regulation.

References

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond, "Preferences for Truth-Telling," *Econometrica*, 2019, *87* (4), 1115–1153.

Achard, Pascal, Sabina Albrecht, Riccardo Ghidoni, Elena Cettolin, and Sigrid Suetens, "Local Exposure to Refugees Changed Attitudes to Ethnic Minorities in the Netherlands," *The Economic Journal*, 2025, *135* (667), 808–837.

Acharya, Avidit, Matthew Blackwell, and Maya Sen, "Explaining Preferences from Behavior: A Cognitive Dissonance Approach," *Journal of Politics*, 2018, *80* (2), 400–411.

Akerlof, George and Rachel Kranton, "Economics and Identity," *Quarterly Journal of Economics*, 2000, *115* (3), 715–753.

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva, "Immigration and Redistribution," *Review of Economic Studies*, 2023, *90* (1), 1–39.

_ , Matteo F. Ferroni, and Stefanie Stantcheva, *Perceptions of Racial Gaps, Their Causes and Ways to Reduce Them*, Vol. 29245, National Bureau of Economic Research Cambridge, MA, 2021.

_ , Stefanie Stantcheva, and Edoardo Teso, "Intergenerational Mobility and Preferences for Redistribution," *American Economic Review*, 2018, *108* (2), 521–554.

Allcott, Hunt, Matthew Gentzkow, Winter Mason, Arjun Wilkins, Pablo Barberá, Taylor Brown, Juan Carlos Cisneros, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon et al., "The Effects of Facebook and Instagram on the 2020 Election: A Deactivation Experiment," *Proceedings of the National Academy of Sciences*, 2024, *121* (21), e2321584121.

Allport, Gordon Willard, *The Nature of Prejudice*, Addison-Wesley, 1954.

Altay, Sacha, Anne-Sophie Hacquin, and Hugo Mercier, "Why Do So Few People Share Fake News? It Hurts Their Reputation," *New Media & Society*, 2022, *24* (6), 1303–1324.

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

Bauer, Kevin, Yan Chen, Florian Hett, and Michael Kosfeld, "Group Identity and Belief Formation: A Microfoundation of Political Polarization," *The Economic Journal*, 2026, p. ueag105.

Bénabou, Roland, "Groupthink: Collective Delusions in Organizations and Markets," *Review of Economic Studies*, 2013, *80* (2), 429–462.

_ and Jean Tirole, "Self-Confidence and Personal Motivation," *Quarterly Journal of Economics*, 2002, *117* (3), 871–915.

_ and _ , "Incentives and Prosocial Behavior," *American Economic Review*, 2006, *96* (5), 1652–1678.

_ **and** _ , "Identity, Morals, and Taboos: Beliefs as Assets," *Quarterly Journal of Economics*, 2011, *126* (2), 805–855.

_ **and** _ , "Mindful Economics: The Production, Consumption, and Value of Beliefs," *Journal of Economic Perspectives*, 2016, *30* (3), 141–64.

_ **and** _ , "Laws and Norms," *Journal of Political Economy*, 2026, *134* (2), 731–772.

_ **and Luca Henkel**, "Identity as Self-Image," 2025. Working paper.

_ **, Armin Falk, and Jean Tirole**, "Narratives, Imperatives and Moral Reasoning," 2019. Working paper.

Bisgaard, Martin, "Bias will Find a Way: Economic Perceptions, Attributions of Blame, and Partisan-Motivated Reasoning during Crisis," *The Journal of Politics*, 2015, *77* (3), 849–860.

_ , "How Getting the Facts Right Can Fuel Partisan-Motivated Reasoning," *American Journal of Political Science*, 2019, *63* (4), 824–839.

Blattner, Adrian and Martin Koenen, "Does Contact Reduce Affective Polarization? Field Evidence from Germany," 2023. Working paper.

Bodner, Ronit and Drazen Prelec, "Self-Signaling and Diagnostic Utility in Everyday Decision Making," in Isabelle Brocas and Juan D. Carillo, eds., *Collected Essays in Psychology and Economics*, Oxford University Press, 2002.

Boxell, Levi, Matthew Gentzkow, and Jesse M. Shapiro, "Cross-Country Trends in Affective Polarization," *Review of Economics and Statistics*, 2024, *106* (2), 557–565.

Braghieri, Luca, Peter Schwardmann, and Egon Tripodi, "Talking across the Aisle," 2024. Working paper.

_ **, Sarah Eichmeyer, Ro'ee Levy, Markus Möbius, Jacob Steinhardt, and Ruiqi Zhong**, "Article-Level Slant and Polarization of News Consumption on Social Media," 2024. Working paper.

Broockman, David E. and Joshua L. Kalla, "Consuming Cross-Cutting Media Causes Learning and Moderates Attitudes: A Field Experiment with Fox News Viewers," *Journal of Politics*, 2025, *87* (1).

Brunnermeier, Markus K. and Jonathan A. Parker, "Optimal Expectations," *American Economic Review*, 2005, *95* (4), 1092–1118.

Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin, "From Extreme to Mainstream: The Erosion of Social Norms," *American Economic Review*, 2020, *110* (11), 3522–3548.

_ , _ **, Ingar Haaland, Aakaash Rao, and Christopher Roth**, "Justifying Dissent," *Quarterly Journal of Economics*, 2023, *138* (3), 1403–1451.

Buschinger, Christiane, Markus Eyting, Florian Hett, and Judd Kessler, "Extreme Justifications Fuel Polarisation," *Economic Journal*, 2026, p. ueag062.

Chetty, Raj, Matthew O. Jackson, Theresa Kuchler, Johannes Stroebel, Nathaniel Hendren, Robert B. Fluegge, Sara Gong, Federico Gonzalez, Armelle Grondin, Matthew Jacob et al., "Social Capital II: Determinants of Economic Connectedness," *Nature*, 2022, *608* (7921), 122–134.

Chopra, Felix, Ingar Haaland, and Christopher Roth, "Do People Demand Fact-Checked News? Evidence from US Democrats," *Journal of Public Economics*, 2022, *205*, 104549.

Dana, Jason and Joël J. van der Weele, "Willful Ignorance: A Perspective from Economics," *Current Opinion in Psychology*, 2025, p. 102210.

_ , Roberto A. Weber, and Jason Xi Kuang, "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness," *Economic Theory*, 2007, *33* (1), 67–80.

Enke, Benjamin, "Moral Values and Voting," *Journal of Political Economy*, 2020, *128* (10), 3679–3729.

_ , Ricardo Rodríguez-Padilla, and Florian Zimmermann, "Moral Universalism and the Structure of Ideology," *Review of Economic Studies*, 2023, *90* (4), 1934–1962.

Exley, Christine L., "Excusing Selfishness in Charitable Giving: The Role of Risk," *Review of Economic Studies*, 2015, *83* (2), 587–628.

Eyster, Erik, Shengwu Li, and Sarah Ridout, "A Theory of Ex Post Rationalization," 2021. Working paper.

Foerster, Manuel and Joël J. van der Weele, "Casting Doubt: Image Concerns and the Communication of Social Impact," *Economic Journal*, 2021, *131* (639), 2887–2919.

Frimer, Jeremy A., Linda J. Skitka, and Matt Motyl, "Liberals and Conservatives Are Similarly Motivated to Avoid Exposure to One Another's Opinions," *Journal of Experimental Social Psychology*, 2017, *72*, 1–12.

Fryer Jr., Roland G., 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*, 2019, *17* (5), 1470–1501.

Gaines, Brian J., James H. Kuklinski, Paul J. Quirk, Buddy Peyton, and Jay Verkuilen, "Same Facts, Different Interpretations: Partisan Motivation and Opinion on Iraq," *The Journal of Politics*, 2007, *69* (4), 957–974.

Gentzkow, Matthew, "Polarization in 2016," 2016. Working paper.

Goldenberg, Amit, Joseph M. Abruzzo, Zi Huang, Jonas Schöne, David Bailey, Robb Willer, Eran Halperin, and James J. Gross, "Homophily and Acrophily as Drivers of Political Segregation," *Nature Human Behaviour*, 2023, *7* (2), 219–230.

Golman, Russell, "Acceptable Discourse: Social Norms of Beliefs and Opinions," *European Economic Review*, 2023, *160*, 104588.

Graham, Jesse, Jonathan Haidt, and Brian A. Nosek, "Liberals and Conservatives Rely on Different Sets of Moral Foundations," *Journal of Personality and Social Psychology*, 2009, *96* (5), 1029.

_ , _ , **Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto**, "Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism," in "Advances in experimental social psychology," Vol. 47, Elsevier, 2013, pp. 55–130.

Grossman, Zachary and Joël J. van der Weele, "Self-Image and Willful Ignorance in Social Decisions," *Journal of the European Economic Association*, 2017, *15* (1), 173–217.

Guess, Andrew M., Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow et al., "Reshares on Social Media Amplify Political News but Do Not Detectably Affect Beliefs or Opinions," *Science*, 2023, *381* (6656), 404–408.

Hagenbach, Jeanne and Charlotte Saucet, "Motivated Skepticism," *Review of Economic Studies*, 2025, *92* (3), 1882–1919.

Haidt, Jonathan, "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment," *Psychological Review*, 2001, *108* (4), 814.

_ , *The Righteous Mind: Why Good People Are Divided by Politics and Religion*, Vintage, 2012.

_ **and Matthew A. Hersh**, "Sexual Morality: The Cultures and Emotions of Conservatives and Liberals 1," *Journal of Applied Social Psychology*, 2001, *31* (1), 191–221.

Hestermann, Nina, Yves Le Yaouanq, and Nicolas Treich, "An Economic Model of the Meat Paradox," *European Economic Review*, 2020, *129*, 103569.

Hobolt, Sara B., Katharina Lawall, and James Tilley, "The Polarizing Effect of Partisan Echo Chambers," *American Political Science Review*, 2024, *118* (3), 1464–1479.

Horwitz, Jeff, *Broken Code: Inside Facebook and the Fight to Expose Its Harmful Secrets*, Doubleday, 2023.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes, "Affect, Not Ideology: A Social Identity Perspective on Polarization," *Public Opinion Quarterly*, January 2012, *76* (3), 405–431.

_ **, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood**, "The Origins and Consequences of Affective Polarization in the United States," *Annual Review of Political Science*, 2019, *22*, 129–146.

Kahan, Dan M., Ellen Peters, Erica Cantrell Dawson, and Paul Slovic, "Motivated Numeracy and Enlightened Self-Government," *Behavioural Public Policy*, 2017, *1* (1), 54–86.

Kappes, Andreas and Tali Sharot, "The Automatic Nature of Motivated Belief Updating," *Behavioural Public Policy*, 2019, *3* (1), 87–103.

Kingzette, Jon, James N. Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan, "How Affective Polarization Undermines Support for Democratic Norms," *Public Opinion Quarterly*, 2021, *85* (2), 663–677.

Kubin, Emily and Christian von Sikorski, "The Role of (Social) Media in Political Polarization: A Systematic Review," *Annals of the International Communication Association*, 2021, *45* (3), 188–206.

Kurzban, Robert, *Why Everyone (Else) Is a Hypocrite: Evolution and the Modular Mind*, Princeton University Press, 2011.

Le Yaouanq, Yves, "A Model of Voting with Motivated Beliefs," *Journal of Economic Behavior & Organization*, 2023, *213*, 394–408.

Lees, Jeffrey and Mina Cikara, "Inaccurate Group Meta-Perceptions Drive Negative Out-Group Attributions in Competitive Contexts," *Nature Human Behaviour*, 2020, *4* (3), 279–286.

_ and _, "Understanding and Combating Misperceived Polarization," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2021, *376* (1822), 20200143.

Lerner, Jennifer S. and Philip E. Tetlock, "Accounting for the Effects of Accountability," *Psychological Bulletin*, 1999, *125* (2), 255.

Levy, Gilat, Ronny Razin, and Alwyn Young, "Misspecified Politics and the Recurrence of Populism," *American Economic Review*, 2022, *112* (3), 928–962.

Levy, Ro'ee, "Social Media, News Consumption, and Polarization: Evidence from a Field Experiment," *American Economic Review*, 2021, *111* (3), 831–870.

Little, Andrew T., "How to Distinguish Motivated Reasoning from Bayesian Updating," *Political Behavior*, 2025, pp. 1–25.

Lowe, Matt, "Types of Contact: A Field Experiment on Collaborative and Adversarial Caste Integration," *American Economic Review*, 2021, *111* (6), 1807–1844.

Melnikoff, David E. and Nina Strohminger, "The Automatic Influence of Advocacy on Lawyers and Novices," *Nature Human Behaviour*, 2020, *4* (12), 1258–1264.

Mercier, Hugo and Dan Sperber, "Why Do Humans Reason? Arguments for an Argumentative Theory," *Behavioral and Brain Sciences*, 2011, *34* (2), 57–74.

Milgrom, Paul and John Roberts, "Rationalizability, Learning, and Equilibrium in Games with Strategic Complementarities," *Econometrica*, 1990, pp. 1255–1277.

Moehring, Alex and Carlos Molina, "Social Influence and News Consumption," 2023. Working paper.

Moore-Berg, Samantha L., Boaz Hameiri, and Emile Bruneau, "The Prime Psychological Suspects of Toxic Political Polarization," *Current Opinion in Behavioral Sciences*, 2020, *34*, 199–204.

Mullainathan, Sendhil and Andrei Shleifer, "The Market for News," *American Economic Review*, 2005, *95* (4), 1031–1053.

Müller, Karsten and Carlo Schwarz, "From Hashtag to Hate Crime: Twitter and Antiminority Sentiment," *American Economic Journal: Applied Economics*, 2023, *15* (3), 270–312.

Norton, Michael I., Joseph A. Vandello, and John M. Darley, "Casuistry and Social Category Bias," *Journal of Personality and Social Psychology*, 2004, *87* (6), 817.

Nyhan, Brendan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y. Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery et al., "Like-Minded Sources on Facebook Are Prevalent but Not Polarizing," *Nature*, 2023, *620* (7972), 137–144.

Perego, Jacopo and Sevgi Yuksel, "Media Competition and Social Disagreement," *Econometrica*, 2022, *90* (1), 223–265.

Pronin, Emily, Daniel Y. Lin, and Lee Ross, "The Bias Blind Spot: Perceptions of Bias in Self versus Others," *Personality and Social Psychology Bulletin*, 2002, *28* (3), 369–381.

Rao, Gautam, "Familiarity Does Not Breed Contempt: Generosity, Discrimination, and Diversity in Delhi Schools," *American Economic Review*, March 2019, *109* (3), 774–809.

Rathje, Steve, Clara Pretus, James Kunling He, Trisha Harjani, Jon Roozenbeek, Kurt Gray, Sander van der Linden, and Jay J. Van Bavel, "Unfollowing Hyperpartisan Social Media Influencers Durably Reduces Out-Party Animosity," 2024. Working paper.

Reichel, Friederike, "What's in a Name? The Breadth of Moral Labels," 2024. Working paper.

Robertson, Ronald E., Jon Green, Damian J. Ruck, Katherine Ognyanova, Christo Wilson, and David Lazer, "Users Choose to Engage with More Partisan News than They Are Exposed to on Google Search," *Nature*, 2023, *618* (7964), 342–348.

Saccardo, Silvia and Marta Serra-Garcia, "Enabling or Limiting Cognitive Flexibility? Evidence of Demand for Moral Commitment," *American Economic Review*, 2023, *113* (2), 396–429.

Santoro, Erik and David E. Broockman, "The Promise and Pitfalls of Cross-Partisan Conversations for Reducing Affective Polarization: Evidence from Randomized Experiments," *Science Advances*, 2022, *8* (25), eabn5515.

Schindler, David and Mark Westcott, "Shocking Racial Attitudes: Black G.I.s in Europe," *The Review of Economic Studies*, January 2021, *88* (1), 489–520.

Schwardmann, Peter, Egon Tripodi, and Joël J. van der Weele, "Self-Persuasion: Evidence from Field Experiments at International Debating Competitions," *American Economic Review*, 2022, *112* (4), 1118–46.

Simler, Kevin and Robin Hanson, *The Elephant in the Brain: Hidden Motives in Everyday Life*, Oxford University Press, 2017.

Stötzer, Lasse Simon and Florian Zimmermann, "Morality and Climate Policy Attitudes," 2026. Working paper.

Sunstein, Cass R, "The Law of Group Polarization," *Journal of Political Philosophy*, 2002, *10*, 175–195.

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

_ , Damon Cann, and Simona Kucsova, "The Motivated Processing of Political Arguments," *Political Behavior*, 2009, *31* (2), 137–155.

Uhlmann, Eric Luis and Geoffrey L. Cohen, "Constructed Criteria: Redefining Merit to Justify Discrimination," *Psychological Science*, 2005, *16* (6), 474–480.

Vives, Xavier, "Nash Equilibrium with Strategic Complementarities," *Journal of Mathematical Economics*, 1990, *19* (3), 305–321.

Williams, Daniel, "The Marketplace of Rationalizations," *Economics & Philosophy*, 2023, *39* (1), 99–123.

Yariv, Leeat, "I'll See It When I Believe It—A Simple Model of Cognitive Consistency," 2005. Working paper.

Zarouali, Brahim, "Pushed to the Edge: When Radical Political Messages Persuade More on Social Media," *Technology, Mind, and Behavior*, 2026.

Zimmerman, Federico, David D. Bailey, Goran Muric, Emilio Ferrara, Jonas Schöne, Robb Willer, Eran Halperin, Joaquín Navajas, James J. Gross, and Amit Goldenberg, "Attraction to Politically Extreme Users on Social Media," *PNAS Nexus*, 2024, *3* (10), pgae395.

Appendix A: Preliminaries

A.1 Refinement definitions

The first refinement pertains to the inferences about an agent who is expected to be fully realistic, but deviates and has no convenient narrative to articulate. When a player who denies inconvenient signals with positive probability has no supporting rationale, Bayes' rule prescribes inferring that this player is strategic with probability one. We extend this inference to out-of-equilibrium situations, in line with the idea that reasonable types are flawless moral reasoners. Formally, let us write $\mathcal{A} = f(\hat{\xi}_i, \hat{\xi}_j)$ for the set of narratives articulated ex post.

Refinement 1 (Reasonable types never rationalize)**.** *Let* $\tilde{\xi} \in \{0,1\}$*. If* $\mathcal{I} = (a_i = \tilde{\xi}, a_j, \mathcal{A})$ *is such that* $\tilde{\xi} \notin \mathcal{A}$*, then* $\Pr[\theta_i = 1 \mid \mathcal{I}] = 0$*. The same refinement holds symmetrically for player* j*.*

The next refinement applies to interactions between two counter-partisans with $b_i > 0 > b_j$. If both players are realistic after their inconvenient signal, they must play the same action. Thus, the information set $(a_i = 1, a_j = 0, \{0,1\})$ is off the equilibrium path, and we discipline the observer's beliefs as follows.

Refinement 2 (Consistency)**.** *Suppose* $b_i > 0 > b_j$*. If* $\sigma_i^{\xi=0,*} = \sigma_j^{\xi=1,*} = 1$*, then*

$$\Pr[\theta_i = 1 \mid (a_i = 1, a_j = 0, \{0,1\})] + \Pr[\theta_j = 1 \mid (a_i = 1, a_j = 0, \{0,1\})] = \rho.$$

Refinement 2 imposes a consistency condition on the observer's beliefs about the types of the two players following the off-path occurrence in which each player plays their convenient action and has a convenient narrative. This identity is implied by Bayes' rule whenever the disagreement information set is on the equilibrium path. We extend it to rule out unnatural out-of-equilibrium beliefs where the observer would, for instance, believe that i and j are both reasonable with probability one (or both with probability zero).

A.2 Proof of Lemma 1

We start by proving that good news is never denied. We provide a proof that covers the one-player game and both versions of the two-player game (with co- and counter-partisan interactions).

Suppose without loss of generality that $b_i > 0$ and, towards a contradiction, that $\sigma_i^{\xi=1,*} < 1$. Throughout the proof, we suppress the strategy profile from the notation. We write

$$R_i^{\xi}(a) \coloneqq \mathbb{E}[\Pr[\theta_i = 1 \mid \mathcal{I}] \mid \xi, \theta_i = 0, i \text{ chooses } a]$$

for a strategic i's expected image in the eyes of the observer whenever i reacts to the signal ξ by playing action a (and trying to rationalize it if $a \neq \xi$). Note that this includes potential deviations.

The expected utility for a strategic type playing a after signal ξ then equals

$$U_i^{\xi}(a) = b_i a + \mu_i R_i^{\xi}(a). \tag{A.1}$$

First case: $\sigma_i^{\xi=0,*} = 1$**.** Suppose i plays $a_i = 1$. This implies that the observer learns for sure that $\xi = 1$, and hence

$$R_i^1(1) = \frac{\rho}{\rho + (1-\rho)\sigma_i^{\xi=1,*}} > \rho.$$

By contrast, after $a_i = 0$, the ex-post image satisfies $\Pr[\theta_i = 1 \mid \mathcal{I}] = 0$ if $\Pr[\xi = 0 \mid \mathcal{I}] = 0$, and otherwise

$$\begin{aligned}\Pr[\theta_i = 1 \mid \mathcal{I}] &= \Pr[\theta_i = 1, \xi = 0 \mid \mathcal{I}] \\ &= \Pr[\theta_i = 1 \mid \xi = 0, \mathcal{I}] \Pr[\xi = 0 \mid \mathcal{I}] \\ &= \rho \Pr[\xi = 0 \mid \mathcal{I}] \\ &\leq \rho,\end{aligned}$$

since the strategic and the reasonable types behave identically conditional on $\xi = 0$. Thus, in both cases we have $\Pr[\theta_i = 1 \mid \mathcal{I}] \leq \rho$.

This implies that $R_i^1(1) > \rho \geq R_i^1(0)$. This observation, together with $b_i > 0$, implies that $U_i^1(1) > U_i^1(0)$, which contradicts $\sigma_i^{\xi=1,*} < 1$.

Second case: $\sigma_i^{\xi=0,*} < 1$. Since i denies both signals with positive probability, both actions $a_i = 0$ and $a_i = 1$ are played with positive probability in each state ξ. We will use the following lemma without proof, as it is a direct consequence of Bayes' rule.

Lemma A.1. *Suppose that $a_i = 1$ occurs with positive probability in both states. Then, for any on-path information set $\mathcal{I}$ containing $a_i = 1$,*

$$\Pr[\theta_i = 1 \mid \mathcal{I}] = \frac{\rho}{\rho + (1-\rho)\sigma_i^{\xi=1,*}} \Pr[\xi = 1 \mid \mathcal{I}].$$

Symmetrically, if $a_i = 0$ occurs with positive probability in both states,

$$\Pr[\theta_i = 1 \mid \mathcal{I}] = \frac{\rho}{\rho + (1-\rho)\sigma_i^{\xi=0,*}} \Pr[\xi = 0 \mid \mathcal{I}]$$

for any on-path information set $\mathcal{I}$ containing $a_i = 0$.

We will then prove that $R_i^1(1) > R_i^0(1)$: that is, i's expected image when playing $a_i = 1$ is higher when this action is played in the congruent state $\xi = 1$ than in the non-congruent state $\xi = 0$.

To establish this result, we use the following elementary property of Bayesian posteriors.

Lemma A.2. *Let $a \in \{0, 1\}$, and suppose that the event $a_i = a$ occurs with positive probability under both $\xi = 0$ and $\xi = 1$. Then*

$$\mathbb{E}[\Pr(\xi = 1 \mid \mathcal{I}) \mid \xi = 1, a_i = a] \geq \mathbb{E}[\Pr(\xi = 1 \mid \mathcal{I}) \mid \xi = 0, a_i = a].$$

The inequality is strict if $\Pr(\xi = 1 \mid \mathcal{I})$ is nondegenerate conditional on $a_i = a$.

Proof. Let $\pi \coloneqq \Pr[\xi = 1 \mid a_i = a] \in (0, 1)$. We have

$$\begin{aligned}\pi\, \mathbb{E}[\Pr(\xi = 1 \mid \mathcal{I}) \mid \xi = 1, a_i = a] &= \mathbb{E}[\Pr(\xi = 1 \mid \mathcal{I}) \mathbb{1}_{\{\xi=1\}} \mid a_i = a] \\ &= \mathbb{E}[\Pr(\xi = 1 \mid \mathcal{I}) \mathbb{E}[\mathbb{1}_{\{\xi=1\}} \mid \mathcal{I}] \mid a_i = a] \\ &= \mathbb{E}[\Pr(\xi = 1 \mid \mathcal{I})^2 \mid a_i = a],\end{aligned}$$

where the second equality is due to the tower property, since a_i is observed as part of $\mathcal{I}$ and $\Pr(\xi = 1 \mid \mathcal{I})$ is $\mathcal{I}$-measurable.

Similarly,

$$(1-\pi)\,\mathbb{E}[\Pr(\xi=1\mid\mathcal{I})\mid\xi=0,a_i=a]=\mathbb{E}[\Pr(\xi=1\mid\mathcal{I})[1-\Pr(\xi=1\mid\mathcal{I})]\mid a_i=a].$$

Combining these two identities and using $\pi=\mathbb{E}[\Pr(\xi=1\mid\mathcal{I})\mid a_i=a]=\Pr(\xi=1\mid a_i=a)$, we get

$$\mathbb{E}[\Pr(\xi=1\mid\mathcal{I})\mid\xi=1,a_i=a]-\mathbb{E}[\Pr(\xi=1\mid\mathcal{I})\mid\xi=0,a_i=a]=\frac{\mathrm{Var}(\Pr(\xi=1\mid\mathcal{I})\mid a_i=a)}{\pi(1-\pi)}\geq 0.$$

The inequality is strict whenever $\Pr(\xi=1\mid\mathcal{I})$ is not constant conditional on $a_i=a$. □

Notice that the conditioning in Lemma A.2 is consistent with the definition of $R_i^{\xi}(a)$. When $a\neq\xi$, observing $a_i=a$ identifies i as strategic. When $a=\xi$, reasonable and strategic types who choose a generate the same distribution of $\mathcal{I}$, since both play the congruent action–rationale pair. Hence, conditional on $(\xi,a_i=a)$, the distribution of $\mathcal{I}$ is the same as conditional on $(\xi,\theta_i=0,i\text{ chooses }a)$.

We now combine Lemmas A.1 and A.2 to show that

$$R_i^1(1)>R_i^0(1). \tag{A.2}$$

Consider first the case $a=1$. By Lemma A.1,

$$\Pr[\theta_i=1\mid\mathcal{I}]=\frac{\rho}{\rho+(1-\rho)\sigma_i^{\xi=1,*}}\Pr[\xi=1\mid\mathcal{I}].$$

Since the multiplicative term is positive, Lemma A.2 applied to $a=1$ implies $R_i^1(1)\geq R_i^0(1)$. Moreover, the inequality is strict because $\Pr(\xi=1\mid\mathcal{I})$ is not constant conditional on $a_i=1$. To make this point, we now distinguish between the one-player and two-player versions of the game.

In the one-player game, conditional on $\xi=0$ the agent is attempting to rationalize, which fails with probability $1-\tau_i$. The resulting information set contains no rationale supporting action 1, yielding $\Pr(\xi=1\mid\mathcal{I})=0$, whereas under the information set $\mathcal{I}=(a_i=1,\hat{\xi}_i=1)$, which happens with positive probability if $\xi=1$, we have $\Pr(\xi=1\mid\mathcal{I})>0$.

In the two-player game, an agent who attempts to rationalize also reveals the state $\xi=0$ for sure if this player's rationalization fails and the other agent does not provide cover, which happens with positive probability (e.g., if the other player is reasonable). Conversely, in state $\xi=1$ it is clear that some positive-probability information sets $\mathcal{I}$ are associated with a positive $\Pr[\xi=1\mid\mathcal{I}]$. The inequality in Lemma A.2 is therefore strict, which proves (A.2).

A symmetric argument applies conditional on $a_i=0$, and gives

$$R_i^0(0)>R_i^1(0). \tag{A.3}$$

We can now compare the relative attractiveness of action $a_i=1$ across the two states. By Equation (A.1),

$$[U_i^1(1)-U_i^1(0)]-[U_i^0(1)-U_i^0(0)]=\mu_i\Big[\underbrace{R_i^1(1)-R_i^0(1)}_{>0\text{ by (A.2)}}+\underbrace{R_i^0(0)-R_i^1(0)}_{>0\text{ by (A.3)}}\Big]>0. \tag{A.4}$$

But $\sigma_i^{\xi=1,*}<1$ implies that a strategic i chooses $a_i=0$ with positive probability after $\xi=1$. Hence $U_i^1(0)\geq U_i^1(1)$. Similarly, $\sigma_i^{\xi=0,*}<1$ implies that a strategic i chooses $a_i=1$ with positive probability after $\xi=0$, and therefore $U_i^0(1)\geq U_i^0(0)$. These two inequalities contradict (A.4). This completes the proof. □

Appendix B: Proofs

B.1 Proofs of Section 3

B.1.1 Proof of Proposition 1

Lemma 1 implies that $\sigma_i^{\xi=1,*} = 1$ for a player with $b_i > 0$, so we simply write and characterize $\sigma_i^* = \sigma_i^{\xi=0,*}$. Let

$$\Delta_i(\sigma_i) \coloneqq U_i^0(1;\sigma_i) - U_i^0(0;\sigma_i)$$

be the net benefit from denial at σ_i, where the expressions for $U_i^0(1;\sigma_i)$ and $U_i^0(0;\sigma_i)$ are given in the main text. The function Δ_i is continuous and strictly increasing in σ_i. Thus, there is a unique equilibrium characterized by the following conditions:

$$\begin{cases} \sigma_i^* = 1 \text{ if } \Delta_i(1) \leq 0; \\ \sigma_i^* = \Delta_i^{-1}(0) \in (0,1) \text{ if } \Delta_i(0) < 0 < \Delta_i(1); \\ \sigma_i^* = 0 \text{ if } \Delta_i(0) \geq 0. \end{cases}$$

The resulting value of σ_i^* is nondecreasing in μ_i and nonincreasing in b_i and τ_i. This proves Proposition 1.

B.1.2 Equilibrium analysis in echo chambers

B.1.2.1 Analysis of the best response

The best-response behavior of i is characterized in the following lemma.[18]

Lemma B.1. *Given some σ_j, there exists a unique best response $\sigma_i \in [0,1]$ for i. This best-response level of realism is continuous and nondecreasing in σ_j (strategic complementarity) and μ_i (image concerns), and is nonincreasing in b_i (core motive), τ_i and τ_j (rationalization abilities).*

Proof. In the following, we fix $\sigma_j \in [0,1]$. Let

$$V^i_{a_i,a_j,\mathcal{A}}(\sigma_i,\sigma_j) \coloneqq \Pr[\theta_i = 1 \mid a_i, a_j, \mathcal{A}]$$

be i's ex-post image conditional on the information set $(a_i, a_j, \mathcal{A})$ given the strategies (σ_i, σ_j).

Suppose that $\xi = 0$. Then i's expected utility from playing $a_i = 1$ equals

$$U_i^0(1;\sigma_i,\sigma_j) = b_i + \mu_i\big[(1-\rho)(1-\sigma_j)K_{ij}V^i_{1,1,\{1\}}(\sigma_i,\sigma_j) + (\rho + (1-\rho)\sigma_j)\tau_i V^i_{1,0,\{0,1\}}(\sigma_i,\sigma_j)\big],$$

where

$$V^i_{1,1,\{1\}}(\sigma_i,\sigma_j) = \frac{\lambda\rho}{\lambda + (1-\lambda)(1-\rho)^2K_{ij}(1-\sigma_i)(1-\sigma_j)},$$

[18] When the lemma identifies $\sigma_i = 1$ as the unique best response to some σ_j, we mean that $\sigma_i = 1$ is the unique strategy that can be supported as a best response to σ_j by some appropriate off-path belief. The supporting off-path belief itself need not be unique (see Footnote 9).

and $V^i_{1,0,\{0,1\}}(\sigma_i, \sigma_j) = 0$ on path, that is, if $\sigma_i < 1$. Otherwise, it is an out-of-equilibrium belief.

Indeed, i receives positive image in two scenarios: (i) if j also denies, and one of the rationalization attempts is successful, which happens with probability K_{ij}; (ii) if j complies with the signal but i deviates from a realism equilibrium and produces a narrative, which happens with probability τ_i.

Conversely, if i plays $a_i = 0$, this perfectly identifies the state $\xi = 0$. Thus, i's expected utility from playing $a_i = 0$ equals

$$U_i^0(0; \sigma_i, \sigma_j) = \mu_i \frac{\rho}{\rho + (1-\rho)\sigma_i}.$$

Let us define the function

$$\Delta_i(\sigma_i, \sigma_j) \coloneqq b_i + \mu_i \Big[\frac{K_{ij}(1-\rho)(1-\sigma_j)\lambda\rho}{\lambda + (1-\lambda)(1-\rho)^2 K_{ij}(1-\sigma_i)(1-\sigma_j)} - \frac{\rho}{\rho + (1-\rho)\sigma_i} \Big]. \tag{B.1}$$

The function $\Delta_i(\sigma_i, \sigma_j)$ captures the net expected utility gain from denial relative to realism given the strategy profile (σ_i, σ_j), ignoring out-of-equilibrium image payoffs if $\sigma_i = 1$. The function Δ_i is continuous and increasing in σ_i. As a result, there is a unique best response characterized by the following conditions:

$$\begin{cases} \sigma_i = 1 \text{ if } \Delta_i(1, \sigma_j) \leq 0; \\ \sigma_i = (\Delta_i(\cdot, \sigma_j))^{-1}(0) \in (0,1) \text{ if } \Delta_i(0, \sigma_j) < 0 < \Delta_i(1, \sigma_j); \\ \sigma_i = 0 \text{ if } \Delta_i(0, \sigma_j) \geq 0. \end{cases}$$

When $\sigma_i = 1$, denial is off path. Its net payoff is $\Delta_i(1, \sigma_j) + \mu_i(\rho + (1-\rho)\sigma_j)\tau_i V^i_{1,0,\{0,1\}}(1, \sigma_j)$. Since image is nonnegative, the belief $V^i_{1,0,\{0,1\}}(1, \sigma_j) = 0$ minimizes the payoff from denial and is therefore the most permissive belief for sustaining $\sigma_i = 1$. Hence $\sigma_i = 1$ can be supported if and only if $\Delta_i(1, \sigma_j) \leq 0$.

This establishes existence and uniqueness of the best response, which, in addition, is continuous in all parameter values. The comparative statics then follow from the variations of the function $\Delta_i(\sigma_i, \sigma_j)$ with the corresponding parameter values. This proves Lemma B.1. □

B.1.2.2 Proof of Proposition 2

Existence Lemma B.1 establishes that the best response σ_i is a continuous function of σ_j (and vice versa, by symmetry). Brouwer's fixed-point theorem then establishes the existence of an equilibrium.

Uniqueness We prove equilibrium uniqueness by showing that the best-response functions have a slope strictly smaller than one. We prove it for player i. The result holds for j as well by symmetry, which implies that the best-response functions cannot intersect multiple times, and hence the equilibrium is unique.

Let $\mathrm{BR}_i(\sigma_j)$ be i's best-response function. Consider first an interior best response. Recall the definition of Δ_i in (B.1). Differentiating Δ_i gives

$$\frac{\partial \Delta_i}{\partial \sigma_i} = \mu_i \rho (1-\rho) \left[\frac{K_{ij}^2 \lambda (1-\lambda)(1-\rho)^2 (1-\sigma_j)^2}{\left[\lambda + (1-\lambda)(1-\rho)^2 K_{ij}(1-\sigma_i)(1-\sigma_j)\right]^2} + \frac{1}{\left[\rho + (1-\rho)\sigma_i\right]^2} \right],$$

while

$$\frac{\partial \Delta_i}{\partial \sigma_j} = -\mu_i \rho (1-\rho) \frac{K_{ij}\lambda^2}{\left[\lambda + (1-\lambda)(1-\rho)^2 K_{ij}(1-\sigma_i)(1-\sigma_j)\right]^2}.$$

Moreover,

$$\frac{K_{ij}\lambda^2}{\left[\lambda + (1-\lambda)(1-\rho)^2 K_{ij}(1-\sigma_i)(1-\sigma_j)\right]^2} < \frac{1}{\left[\rho + (1-\rho)\sigma_i\right]^2},$$

since $\lambda + (1-\lambda)(1-\rho)^2 K_{ij}(1-\sigma_i)(1-\sigma_j) \geq \lambda$, $K_{ij} < 1$, and $\rho + (1-\rho)\sigma_i \leq 1$. Hence

$$-\frac{\partial \Delta_i}{\partial \sigma_j} < \frac{\partial \Delta_i}{\partial \sigma_i}.$$

By the implicit-function theorem, whenever the best response is interior,

$$0 < \frac{d\mathrm{BR}_i(\sigma_j)}{d\sigma_j} = -\frac{\partial \Delta_i / \partial \sigma_j}{\partial \Delta_i / \partial \sigma_i} < 1.$$

It follows that for any $\widetilde{\sigma}_j > \sigma_j$ such that BR_i is interior on $[\sigma_j, \widetilde{\sigma}_j]$,

$$\mathrm{BR}_i(\widetilde{\sigma}_j) - \mathrm{BR}_i(\sigma_j) < \widetilde{\sigma}_j - \sigma_j.$$

We now extend this result to arbitrary $\widetilde{\sigma}_j > \sigma_j$, including points at which the best response is at a boundary. If $\mathrm{BR}_i(\widetilde{\sigma}_j) = \mathrm{BR}_i(\sigma_j)$, the desired inequality is immediate. Otherwise, by continuity and monotonicity of BR_i, there exist $\sigma_j \leq x < y \leq \widetilde{\sigma}_j$ such that

$$\mathrm{BR}_i(x) = \mathrm{BR}_i(\sigma_j), \qquad \mathrm{BR}_i(y) = \mathrm{BR}_i(\widetilde{\sigma}_j),$$

and $\mathrm{BR}_i(z) \in (0,1)$ for all $z \in (x,y)$. The mean value theorem then implies that, for some $z^* \in (x,y)$,

$$\mathrm{BR}_i(\widetilde{\sigma}_j) - \mathrm{BR}_i(\sigma_j) = \mathrm{BR}_i'(z^*)(y-x) < y - x \leq \widetilde{\sigma}_j - \sigma_j.$$

Hence, for any $\widetilde{\sigma}_j > \sigma_j$,

$$\mathrm{BR}_i(\widetilde{\sigma}_j) - \mathrm{BR}_i(\sigma_j) < \widetilde{\sigma}_j - \sigma_j. \tag{B.2}$$

To complete the proof of uniqueness, suppose towards a contradiction that there exist two distinct equilibria (σ_i^*, σ_j^*) and $(\widetilde{\sigma}_i^*, \widetilde{\sigma}_j^*)$. Suppose without loss of generality that $\widetilde{\sigma}_j^* > \sigma_j^*$. Since i's best response is nondecreasing, we must have $\widetilde{\sigma}_i^* > \sigma_i^*$ with strict inequality, otherwise the best response by j would be the same in both cases. By (B.2), we have $\widetilde{\sigma}_i^* - \sigma_i^* < \widetilde{\sigma}_j^* - \sigma_j^*$, and applying the same inequality to player j gives $\widetilde{\sigma}_j^* - \sigma_j^* < \widetilde{\sigma}_i^* - \sigma_i^*$, which is a contradiction. Hence the equilibrium is unique.

Comparative statics Comparative statics follow from classic arguments from the literature on supermodular games (e.g., Milgrom and Roberts, 1990; Vives, 1990). We lay out the argument for the comparative statics in b_i. The other results can be obtained with similar arguments, noting that the best-response functions co-vary with all parameters.

We now index i's best response $\mathrm{BR}_i(\sigma_j, b_i)$ by b_i. Conversely, $\mathrm{BR}_j(\sigma_i)$ is independent of b_i. The analysis of the best-response behavior in the previous section shows that BR_i is nondecreasing in σ_j and nonincreasing in b_i, whereas BR_j is nondecreasing in σ_i.

Consider then two values $b_i < \widetilde{b}_i$, and let $(\sigma_i^*(b_i), \sigma_j^*(b_i))$ be the equilibrium when the parameter

is b_i. We have

$$\forall \sigma_j \le \sigma_j^*(b_i), \mathrm{BR}_i(\sigma_j, \widetilde{b}_i) \le \mathrm{BR}_i(\sigma_j, b_i) \le \mathrm{BR}_i(\sigma_j^*(b_i), b_i) = \sigma_i^*(b_i),$$

and

$$\forall \sigma_i \le \sigma_i^*(b_i), \mathrm{BR}_j(\sigma_i) \le \mathrm{BR}_j(\sigma_i^*(b_i)) = \sigma_j^*(b_i).$$

Therefore the best-response correspondence $(\sigma_i, \sigma_j) \mapsto \big(\mathrm{BR}_i(\sigma_j, \widetilde{b}_i), \mathrm{BR}_j(\sigma_i)\big)$ maps the rectangle $[0, \sigma_i^*(b_i)] \times [0, \sigma_j^*(b_i)]$ into itself. By Brouwer's theorem, there exists a fixed point of this restricted map. Since the equilibrium is unique, the unique equilibrium with parameter $\widetilde{b}_i$ must therefore satisfy $\sigma_i^*(\widetilde{b}_i) \le \sigma_i^*(b_i)$ and $\sigma_j^*(\widetilde{b}_i) \le \sigma_j^*(b_i)$.

B.1.3 Equilibrium analysis in agoras

B.1.3.1 Analysis of the best response

Here we provide a lemma analogous to the one we have for echo chambers.[19]

Lemma B.2. *Given some $\sigma_j \in [0,1]$, there exists a unique best-response level of realism $\sigma_i \in [0,1]$ for i. This best response is continuous and nondecreasing in σ_j and μ_i, and nonincreasing in b_i, τ_i, and τ_j.*

Proof. Fix $\sigma_j \in [0,1]$. Recall that σ_i denotes i's probability of realism after the inconvenient state $\xi = 0$, whereas σ_j denotes j's probability of realism after her inconvenient state $\xi = 1$.

Suppose first that $\min\{\sigma_i, \sigma_j\} < 1$. Conditional on the information set $(a_i, a_j, \mathcal{A}) = (1, 0, \{0,1\})$, Bayes' rule gives

$$V^i_{1,0,\{0,1\}}(\sigma_i, \sigma_j) = \frac{\lambda\rho\tau_j(1-\sigma_j)}{\lambda\tau_j(1-\sigma_j) + (1-\lambda)\tau_i(1-\sigma_i)}. \tag{B.3}$$

Indeed, if $\xi = 1$, player i truthfully plays $a_i = 1$ while the information set can arise only if a strategic j denies the signal and successfully rationalizes. If $\xi = 0$, player j truthfully plays $a_j = 0$ while the information set can arise only if a strategic i denies the signal and successfully rationalizes.

Upon observing $\xi = 0$, player i knows that j faces good news and, by Lemma 1 (applied symmetrically to $b_j < 0$), plays $a_j = 0$. If i denies the signal and plays $a_i = 1$, she obtains a positive image only when her rationalization succeeds. Her expected utility from denial is therefore

$$U_i^0(1; \sigma_i, \sigma_j) = b_i + \mu_i\tau_i \frac{\lambda\rho\tau_j(1-\sigma_j)}{\lambda\tau_j(1-\sigma_j) + (1-\lambda)\tau_i(1-\sigma_i)}.$$

Conversely, playing $a_i = 0$ reveals the state to be $\xi = 0$, and hence yields

$$U_i^0(0; \sigma_i, \sigma_j) = \mu_i \frac{\rho}{\rho + (1-\rho)\sigma_i}.$$

Let

$$\Delta_i(\sigma_i, \sigma_j) \coloneqq b_i + \mu_i \left[\tau_i \frac{\lambda\rho\tau_j(1-\sigma_j)}{\lambda\tau_j(1-\sigma_j) + (1-\lambda)\tau_i(1-\sigma_i)} - \frac{\rho}{\rho + (1-\rho)\sigma_i}\right] \tag{B.4}$$

denote the net utility gain from denial relative to realism. Whenever $\min\{\sigma_i, \sigma_j\} < 1$, this expression is pinned down by Bayes' rule.

[19] Again, uniqueness refers to the strategy that can be supported as a best-response to σ_j, but the off-path belief (if $\sigma_i = \sigma_j = 1$) need not be unique.

For any fixed $\sigma_j < 1$, the function $\Delta_i(\sigma_i, \sigma_j)$ is continuous and strictly increasing in σ_i. Therefore the best response is uniquely characterized by

$$\begin{cases} \sigma_i = 1 \quad \text{if} \quad \Delta_i(1, \sigma_j) \leq 0; \\ \sigma_i = (\Delta_i(\cdot, \sigma_j))^{-1}(0) \in (0,1) \quad \text{if} \quad \Delta_i(0, \sigma_j) < 0 < \Delta_i(1, \sigma_j); \\ \sigma_i = 0 \quad \text{if} \quad \Delta_i(0, \sigma_j) \geq 0. \end{cases}$$

Consider now $\sigma_j = 1$. For every $\sigma_i < 1$, Equation (B.4) reduces to

$$\Delta_i(\sigma_i, 1) = b_i - \mu_i \frac{\rho}{\rho + (1-\rho)\sigma_i}.$$

If $\sigma_i = 1$ as well, the information set $(1, 0, \{0,1\})$ is off path. Let $V^i_{1,0,\{0,1\}}(1,1)$ denote the corresponding off-equilibrium image. Full realism by i can then be supported whenever

$$b_i + \mu_i \tau_i V^i_{1,0,\{0,1\}}(1,1) \leq \mu_i \rho.$$

In particular, if $b_i \leq \mu_i \rho$, it can be supported by setting $V^i_{1,0,\{0,1\}}(1,1) = 0$. This value is admissible under Refinement 2, provided the observer assigns image ρ to j at the same information set. Whether the beliefs supporting the best responses of i and j can simultaneously satisfy their equilibrium incentive constraints is a joint issue that we address in the proof of Proposition 3. If $b_i > \mu_i \rho$, full realism cannot be supported under any belief, and the unique best response satisfies $\sigma_i < 1$ and is characterized by the preceding expression. This analysis also shows that $\lim_{\sigma_j \uparrow 1} \mathrm{BR}^i(\sigma_j) = \mathrm{BR}^i(1)$, so BR^i is continuous everywhere.

It remains to establish the comparative statics. Direct differentiation of (B.4) shows that it is nonincreasing in σ_j and increasing in b_i, τ_i, and τ_j. Moreover, the expected image from denial is strictly smaller than the image from realism, since

$$\tau_i V^i_{1,0,\{0,1\}}(\sigma_i, \sigma_j) \leq \tau_i \rho < \rho \leq \frac{\rho}{\rho + (1-\rho)\sigma_i}.$$

Hence Δ_i is decreasing in μ_i. Since Δ_i is strictly increasing in σ_i, the unique best-response level of realism is therefore nondecreasing in σ_j and μ_i, and nonincreasing in b_i, τ_i, and τ_j. A similar argument applies symmetrically to player j. □

B.1.3.2 Proof of Proposition 3

Existence. Consider the best-response functions $\mathrm{BR}_i(\sigma_j)$ and $\mathrm{BR}_j(\sigma_i)$ found in Lemma B.2. The map $(\sigma_i, \sigma_j) \mapsto \big(\mathrm{BR}_i(\sigma_j), \mathrm{BR}_j(\sigma_i)\big)$ is continuous and maps $[0,1]^2$ to itself, so it admits a fixed point by Brouwer's theorem, i.e. the two best-response curves intersect. If an intersection exists where $\min(\sigma_i, \sigma_j) < 1$, it is an equilibrium of the game, so existence is established.

If the only intersection is $(1,1)$, we need to find supporting off-path beliefs for each best response that satisfy Refinement 2. To do so, consider the function

$$H(\sigma_i) = \mathrm{BR}^i(\mathrm{BR}^j(\sigma_i))$$

defined on $[0,1]$.

Since the only fixed point of $(\sigma_i, \sigma_j) \mapsto (\mathrm{BR}^i(\sigma_j), \mathrm{BR}^j(\sigma_i))$ is $(1,1)$, the only fixed point of H is $\sigma_i = 1$. Moreover, $H(0) \geq 0$ and $H(0) = 0$ would give another fixed point with $\sigma_i = 0$, which is a

contradiction. Therefore, $H(0) > 0$, which then implies (together with the continuity of H) that

$$H(\sigma_i) > \sigma_i \text{ for all } \sigma_i \in [0, 1). \tag{B.5}$$

Now consider an increasing sequence $\sigma_i^n \uparrow 1$, $\sigma_i^n < 1$. Set $\sigma_j^n = \text{BR}^j(\sigma_i^n)$. By monotonicity and continuity of BR^j, $\sigma_j^n \uparrow 1$.

Since $\sigma_i^n < 1$ for all n, every pair (σ_i^n, σ_j^n) is such that the information set $(a_i = 1, a_j = 0, \{0, 1\})$ is on path and associated with the following posterior beliefs:

$$V^i_{1,0,\{0,1\}}(\sigma_i^n, \sigma_j^n) = \frac{\lambda\rho\tau_j(1-\sigma_j^n)}{\lambda\tau_j(1-\sigma_j^n) + (1-\lambda)\tau_i(1-\sigma_i^n)}$$

and

$$V^j_{1,0,\{0,1\}}(\sigma_i^n, \sigma_j^n) = \frac{(1-\lambda)\rho\tau_i(1-\sigma_i^n)}{(1-\lambda)\tau_i(1-\sigma_i^n) + \lambda\tau_j(1-\sigma_j^n)}.$$

Note that $V^i_{1,0,\{0,1\}}(\sigma_i^n, \sigma_j^n) + V^j_{1,0,\{0,1\}}(\sigma_i^n, \sigma_j^n) = \rho$ for all n.

The sequence $\left(V^i_{1,0,\{0,1\}}(\sigma_i^n, \sigma_j^n), V^j_{1,0,\{0,1\}}(\sigma_i^n, \sigma_j^n)\right)$ lies in the compact set $[0, \rho]^2$, so it admits a convergent subsequence, whose limit we write (V_*^i, V_*^j). We now work with this subsequence instead of the original sequence, and note that

$$V_*^i + V_*^j = \rho.$$

Thus, the off-path beliefs (V_*^i, V_*^j) satisfy Refinement 2. It remains to show that V_*^i supports $\sigma_i = 1$ as a best response to $\sigma_j = 1$, and similarly for j. To do so, note that, since $\sigma_i^n > 0$ for large n, and $H(\sigma_i^n) > \sigma_i^n$ by Equation (B.5), we have $\text{BR}_i(\sigma_j^n) > \sigma_i^n$. Since $\Delta_i(\cdot, \sigma_j^n)$ is strictly increasing and such that $\Delta_i(\text{BR}_i(\sigma_j^n), \sigma_j^n) \leq 0$ by definition of the best response (with equality for an interior best response), this implies that $\Delta_i(\sigma_i^n, \sigma_j^n) < 0$, and hence player i prefers realism to denial at (σ_i^n, σ_j^n). Formally,

$$b_i + \mu_i\tau_i V^i_{1,0,\{0,1\}}(\sigma_i^n, \sigma_j^n) \leq \mu_i \frac{\rho}{\rho + (1-\rho)\sigma_i^n}.$$

Taking the limit implies $b_i + \mu_i\tau_i V_*^i \leq \mu_i\rho$, which is exactly the condition for $\sigma_i = 1$ to be a best response (together with off-path belief V_*^i) to $\sigma_j = 1$.

Last, we have defined $\sigma_j^n = \text{BR}_j(\sigma_i^n)$, so for large n we have $\sigma_j^n > 0$ and the payoff from realism cannot be strictly smaller than the payoff from denial since realism is played with positive probability. A continuity argument similar to the one used for i completes the proof. This proves existence of the equilibrium.

Uniqueness The previous steps imply that, at any information set $(1, 0, \{0, 1\})$, any equilibrium (σ_i^*, σ_j^*) is such that

$$V^i_{1,0,\{0,1\}}(\sigma_i^*, \sigma_j^*) + V^j_{1,0,\{0,1\}}(\sigma_i^*, \sigma_j^*) = \rho, \tag{B.6}$$

where this identity is either given by Bayes' rule (on path) or by Refinement 2 (off path).

Suppose, towards a contradiction, that there exist two distinct equilibria (σ_i^*, σ_j^*) and $(\widetilde{\sigma}_i^*, \widetilde{\sigma}_j^*)$. Since best responses are nondecreasing and unique, we can suppose without loss of generality that $\sigma_i^* < \widetilde{\sigma}_i^*$ and $\sigma_j^* < \widetilde{\sigma}_j^*$.

Since $\sigma_i^* < 1$, denial is played with positive probability by a strategic i in the first equilibrium,

and hence

$$b_i + \mu_i \tau_i V^i_{1,0,\{0,1\}}(\sigma_i^*, \sigma_j^*) \geq \mu_i \frac{\rho}{\rho + (1-\rho)\sigma_i^*}.$$

Since $\widetilde{\sigma}_i^* > 0$, realism is played with positive probability in the second equilibrium, and hence

$$b_i + \mu_i \tau_i V^i_{1,0,\{0,1\}}(\widetilde{\sigma}_i^*, \widetilde{\sigma}_j^*) \leq \mu_i \frac{\rho}{\rho + (1-\rho)\widetilde{\sigma}_i^*}.$$

If the second equilibrium is $(1,1)$, the image in the latter inequality is the off-path belief specified by the equilibrium. Combining the two inequalities, and using $\sigma_i^* < \widetilde{\sigma}_i^*$, gives

$$V^i_{1,0,\{0,1\}}(\sigma_i^*, \sigma_j^*) > V^i_{1,0,\{0,1\}}(\widetilde{\sigma}_i^*, \widetilde{\sigma}_j^*).$$

The symmetric argument for player j gives

$$V^j_{1,0,\{0,1\}}(\sigma_i^*, \sigma_j^*) > V^j_{1,0,\{0,1\}}(\widetilde{\sigma}_i^*, \widetilde{\sigma}_j^*).$$

These two inequalities contradict

$$V^i_{1,0,\{0,1\}}(\sigma_i^*, \sigma_j^*) + V^j_{1,0,\{0,1\}}(\sigma_i^*, \sigma_j^*) = V^i_{1,0,\{0,1\}}(\widetilde{\sigma}_i^*, \widetilde{\sigma}_j^*) + V^j_{1,0,\{0,1\}}(\widetilde{\sigma}_i^*, \widetilde{\sigma}_j^*) = \rho,$$

which is given by (B.6). This proves equilibrium uniqueness.

Comparative statics Lemma B.2 establishes strategic complementarity and the direction in which each parameter shifts the best-response functions. The comparative statics of the equilibrium strategy then follow from the same arguments as in the echo chamber. We omit the details for brevity. □

B.1.4 Proof of Corollary 1

The variation of σ_i^* on each side of zero is given by Propositions 2 and 3. It only remains to establish continuity at zero. This is given directly by the fact that $\sigma_j^* \to 1$ when $b_j \to 0$, and the function $\Delta_i(\cdot, \sigma_j^*)$ converges to the same limit when $b_j \to 0^+$ and $b_j \to 0^-$.

B.2 Proofs of Section 4

B.2.1 Proof of Proposition 4

Part i) Since the observer is Bayesian, the martingale property of beliefs implies

$$\mathbb{E}\left[\Pr[\theta_i = 1 \mid \mathcal{I}]\right] = \Pr[\theta_i = 1] = \rho.$$

Thus, ex-ante expected image utility is equal to $\rho\mu_i$ in any environment.

By Lemma 1, all types play $a_i = 1$ when $\xi = 1$. When $\xi = 0$, a reasonable type plays $a_i = 0$, while a strategic type plays $a_i = 1$ with probability $1 - \sigma_i^*$. Hence

$$\mathbb{E}[a_i] = \lambda + (1-\rho)(1-\lambda)(1-\sigma_i^*).$$

Finally, the reasonable type receives Socratic utility with probability one, while the strategic type does not receive this utility. Ex-ante expected Socratic utility is therefore $\rho\gamma_i$. Collecting these

terms gives

$$\begin{aligned} W_i &= \rho\gamma_i + \rho\mu_i + b_i\left[\lambda + (1-\rho)(1-\lambda)(1-\sigma_i^*)\right] \\ &= C_i + b_i(1-\rho)(1-\lambda)(1-\sigma_i^*), \end{aligned}$$

where

$$C_i \coloneqq \rho\gamma_i + \rho\mu_i + \lambda b_i.$$

This establishes part i.

Part ii) Recall the notation $R_i^{\xi}(a)$ introduced in the proof of Lemma 1. Conditional on a given state, a reasonable type generates the same distribution of observables as a strategic type choosing the realistic action. Hence

$$W_i^R = \gamma_i + \lambda\left[b_i + \mu_i R_i^1(1)\right] + (1-\lambda)\mu_i R_i^0(0),$$

whereas

$$W_i^S = \lambda\left[b_i + \mu_i R_i^1(1)\right] + (1-\lambda)\left[\sigma_i^*\mu_i R_i^0(0) + (1-\sigma_i^*)\left(b_i + \mu_i R_i^0(1)\right)\right].$$

It follows that

$$W_i^R - W_i^S = \gamma_i + (1-\lambda)(1-\sigma_i^*)\left[\mu_i R_i^0(0) - b_i - \mu_i R_i^0(1)\right]. \tag{B.7}$$

Suppose $\sigma_i^* > 0$. If $\sigma_i^* = 1$, the second term in (B.7) is equal to zero. If instead $\sigma_i^* \in (0,1)$, both realism and denial are played with positive probability by the strategic type after $\xi = 0$. The equilibrium indifference condition therefore gives

$$\mu_i R_i^0(0) = b_i + \mu_i R_i^0(1).$$

Hence, in either case, $W_i^R - W_i^S = \gamma_i$. Since $W_i = \rho W_i^R + (1-\rho)W_i^S$, we obtain

$$W_i^R = W_i + (1-\rho)\gamma_i \text{ and } W_i^S = W_i - \rho\gamma_i.$$

Together with part i), this implies that whenever $\sigma_i^* > 0$, both types rank environments in the same way and prefer environments with lower equilibrium realism.

Part iii) Consider two environments, denoted $\mathcal{E}$ and $\mathcal{E}'$, in which $\sigma_i^* = 0$. By part i), $W_i^{\mathcal{E}} = W_i^{\mathcal{E}'}$ since ex-ante utility depends on the environment only through σ_i^*. Moreover, $W_i^{\mathcal{E}} = \rho W_i^{R,\mathcal{E}} + (1-\rho)W_i^{S,\mathcal{E}}$ and analogously for $\mathcal{E}'$. Hence (rearranging),

$$\rho\left(W_i^{R,\mathcal{E}} - W_i^{R,\mathcal{E}'}\right) = -(1-\rho)\left(W_i^{S,\mathcal{E}} - W_i^{S,\mathcal{E}'}\right).$$

Thus the reasonable and strategic types have opposite preferences over the two environments: whenever one type strictly prefers one environment, the other type strictly prefers the other. This proves Proposition 4. □

B.2.2 Proof of Proposition 5

By Proposition 4, agent i's ex-ante expected utility is nonincreasing in her equilibrium level of realism σ_i^*. We can therefore compare environments by comparing the equilibrium values of σ_i^* that they induce.

Echo chamber versus autarky Recall that the net utility gain from denial relative to realism in autarky is

$$\Delta_i^A(\sigma_i) = b_i + \mu_i \left[\frac{\tau_i \lambda \rho}{\lambda + (1-\lambda)(1-\rho)\tau_i(1-\sigma_i)} - \frac{\rho}{\rho + (1-\rho)\sigma_i} \right].$$

In an echo chamber, conditional on σ_j, the corresponding expression is

$$\Delta_i^E(\sigma_i, \sigma_j) = b_i + \mu_i \left[\frac{K_{ij}(1-\rho)(1-\sigma_j)\lambda\rho}{\lambda + (1-\lambda)(1-\rho)^2 K_{ij}(1-\sigma_i)(1-\sigma_j)} - \frac{\rho}{\rho + (1-\rho)\sigma_i} \right].$$

Some algebra shows that

$$\Delta_i^E(\sigma_i, \sigma_j) \geq \Delta_i^A(\sigma_i) \quad \iff \quad (1-\rho)(1-\sigma_j)K_{ij} \geq \tau_i.$$

Since both functions are strictly increasing in σ_i, it follows that

$$\sigma_i^{E,*} \leq \sigma_i^{A,*} \quad \text{whenever} \quad (1-\rho)(1-\sigma_j^*)K_{ij} \geq \tau_i.$$

Suppose first that $(1-\rho)K_{ij} \leq \tau_i$. This implies $(1-\rho)(1-\sigma_j^*)K_{ij} \leq \tau_i$ for every equilibrium value of σ_j^*. Thus $\sigma_i^{E,*} \geq \sigma_i^{A,*}$ for every co-partisan j, and Proposition 4 implies that agent i prefers autarky to the echo chamber.

Suppose now that $(1-\rho)K_{ij} > \tau_i$. The proof of Proposition 2 shows that equilibrium realism σ_j^* in the echo chamber is continuous and nonincreasing in b_j. Moreover, $\sigma_j^* = 1$ when $b_j \approx 0$ and $\sigma_j^* = 0$ when b_j is sufficiently large, e.g., $b_j > \mu_j$. Therefore there exists a threshold $\bar{b}_j$ such that $(1-\rho)K_{ij}(1-\sigma_j^*) = \tau_i$ when $b_j = \bar{b}_j$, and this threshold is unique since σ_j^* is strictly decreasing in b_j in the region where j plays a mixed strategy. We then have $\sigma_i^{E,*} \leq \sigma_i^{A,*}$ if $b_j \geq \bar{b}_j$ and $\sigma_i^{E,*} \geq \sigma_i^{A,*}$ if $b_j \leq \bar{b}_j$, which proves part i).

Agora versus autarky If the equilibrium in the agora is $(\sigma_i^{G,*} = 1, \sigma_j^{G,*} = 1)$, then it is immediate that $\sigma_i^{G,*} \geq \sigma_i^{A,*}$. Otherwise, conditional on σ_j, the net utility gain from denial in the counter-partisan interaction is

$$\Delta_i^G(\sigma_i, \sigma_j) = b_i + \mu_i \left[\tau_i \frac{\lambda\rho\tau_j(1-\sigma_j)}{\lambda\tau_j(1-\sigma_j) + (1-\lambda)\tau_i(1-\sigma_i)} - \frac{\rho}{\rho + (1-\rho)\sigma_i} \right],$$

where the expression is well-defined since $\min(\sigma_i, \sigma_j) < 1$. Some algebra shows that $\Delta_i^G(\sigma_i, \sigma_j) \leq \Delta_i^A(\sigma_i)$ whenever $\min(\sigma_i, \sigma_j) < 1$. This further implies that $\sigma_i^{G,*} \geq \sigma_i^{A,*}$. This proves part ii). □

B.3 Proofs of Section 5

B.3.1 Proof of Proposition 6

We first establish the bias in i's belief about the state. We then show that the biases in beliefs about the players' types follow directly from this state bias.

Consider first a co-partisan interaction, with $b_j > 0$. The only information set at which i assigns

positive probability to both states is

$$\mathcal{I} = (a_i = 1, a_j = 1, \{1\}).$$

If $\xi = 1$, this information set occurs with probability one. If $\xi = 0$, both players must be strategic and deny the signal, and at least one of their rationalization attempts must succeed. Its true probability conditional on $\xi = 0$ is therefore

$$(1-\rho)^2(1-\sigma_i)(1-\sigma_j)K_{ij}.$$

By contrast, agent i underestimates the probability that her own rationalization succeeds and perceives the corresponding probability as

$$(1-\rho)^2(1-\sigma_i)(1-\sigma_j)\left[\tau_j + \tau_i(1-\chi_i)(1-\tau_j)\right].$$

Hence

$$p_i(\xi = 1 \mid \mathcal{I}) = \frac{\lambda}{\lambda + (1-\lambda)(1-\rho)^2(1-\sigma_i)(1-\sigma_j)\left[\tau_j + \tau_i(1-\chi_i)(1-\tau_j)\right]}.$$

All other information sets reveal the state. Averaging i's posterior under the true distribution therefore gives

$$\mathbb{E}[p_i(\xi = 1)] = \lambda + \beta_i^E,$$

where

$$\beta_i^E \coloneqq \frac{\lambda(1-\lambda)(1-\rho)^2\tau_i\chi_i(1-\tau_j)(1-\sigma_i)(1-\sigma_j)}{\lambda + (1-\lambda)(1-\rho)^2(1-\sigma_i)(1-\sigma_j)\left[\tau_j + \tau_i(1-\chi_i)(1-\tau_j)\right]} > 0.$$

Consider next a counter-partisan interaction, with $b_j < 0$. In this case, the only information set at which both states remain possible is

$$\mathcal{I} = (a_i = 1, a_j = 0, \{0, 1\}).$$

If $\xi = 1$, this information set requires j to be strategic, deny her inconvenient signal, and successfully rationalize. If $\xi = 0$, it requires the analogous events for i. Thus,

$$p_i(\xi = 1 \mid \mathcal{I}) = \frac{\lambda\tau_j(1-\sigma_j)}{\lambda\tau_j(1-\sigma_j) + (1-\lambda)\tau_i(1-\chi_i)(1-\sigma_i)}.$$

Again all other information sets reveal the state. Averaging under the true distribution yields

$$\mathbb{E}[p_i(\xi = 1)] = \lambda + \beta_i^G,$$

where

$$\beta_i^G \coloneqq \frac{\lambda(1-\lambda)(1-\rho)\tau_i\tau_j\chi_i(1-\sigma_i)(1-\sigma_j)}{\lambda\tau_j(1-\sigma_j) + (1-\lambda)\tau_i(1-\chi_i)(1-\sigma_i)} > 0. \tag{B.8}$$

Thus, in either environment, we can write

$$\mathbb{E}[p_i(\xi = 1)] = \lambda + \beta_i \tag{B.9}$$

for some $\beta_i > 0$. Agent i therefore overestimates, on average, the probability of her convenient state $\xi = 1$. In a counter-partisan interaction, the symmetric argument for player j gives

$$\mathbb{E}[p_j(\xi = 1)] = \lambda - \beta_j,$$

with $\beta_j > 0$ whenever $\chi_j > 0$. Hence each counter-partisan overestimates the probability of her own convenient state, establishing ideological polarization.

We now show that the remaining results follow immediately from (B.9). Consider first i's belief about her own type. Whenever $a_i = 1$, a reasonable i can only have generated the observation in state $\xi = 1$, and conditional on that state reasonable and strategic types generate the same action and rationale. Hence

$$p_i(\theta_i = 1 \mid \mathcal{I}) = \rho\, p_i(\xi = 1 \mid \mathcal{I}) \qquad \text{if } (a_i = 1) \in \mathcal{I}.$$

When $a_i = 0$, the state is revealed to be $\xi = 0$, and

$$p_i(\theta_i = 1 \mid a_i = 0) = \frac{\rho}{\rho + (1-\rho)\sigma_i}.$$

Since $\Pr[a_i = 0] = (1-\lambda)\left[\rho + (1-\rho)\sigma_i\right]$, we obtain

$$\begin{aligned}\mathbb{E}[p_i(\theta_i = 1)] &= \rho\, \mathbb{E}[p_i(\xi = 1)] + (1-\lambda)\rho \\ &= \rho(1+\beta_i) > \rho.\end{aligned}$$

This establishes self-righteousness.

If j is a co-partisan, the same argument applies to j, since j also always plays $a_j = 1$ in state $\xi = 1$. Therefore

$$\mathbb{E}[p_i(\theta_j = 1)] = \rho(1+\beta_i) > \rho.$$

If instead j is a counter-partisan, she always plays $a_j = 0$ in state $\xi = 0$. Whenever $a_j = 0$,

$$p_i(\theta_j = 1 \mid \mathcal{I}) = \rho\, p_i(\xi = 0 \mid \mathcal{I}),$$

whereas $a_j = 1$ reveals $\xi = 1$ and contributes $\lambda\rho$ to the average posterior image. Hence

$$\begin{aligned}\mathbb{E}[p_i(\theta_j = 1)] &= \rho\left[1 - \mathbb{E}[p_i(\xi = 1)]\right] + \lambda\rho \\ &= \rho(1-\beta_i) < \rho.\end{aligned}$$

Thus i overestimates the morality of a co-partisan and underestimates that of a counter-partisan. This proves Proposition 6. □

B.3.2 Proof of Proposition 7

Suppose that $b_i > 0 > b_j$. Agent i has formed her beliefs p_i^a with the strategies σ_i, σ_j and conditional on the information set $\{a_i, a_j, \hat{\xi}_i\}$. We compare these beliefs to p_i^G, those obtained after she additionally observes j's articulated rationale $\hat{\xi}_j$.

Before observing $\hat{\xi}_j$, the only information set at which the state remains uncertain is the one in which i has retrieved her convenient rationale and j has chosen her convenient action,

$$(a_i = 1, \hat{\xi}_i = 1, a_j = 0).$$

Conditional on $\xi = 1$, this event requires j to be strategic and deny her inconvenient signal, which occurs with probability $(1-\rho)(1-\sigma_j)$. Conditional on $\xi = 0$, it requires i to be strategic, deny her inconvenient signal, and successfully rationalize, which occurs with probability $(1-\rho)\tau_i(1-\sigma_i)$. Agent i perceives the latter probability as $(1-\rho)\tau_i(1-\chi_i)(1-\sigma_i)$.

Averaging i's posterior under the true distribution therefore gives

$$\mathbb{E}[p_i^a(\xi = 1)] = \lambda + \beta_i^a,$$

where

$$\beta_i^a := \frac{\lambda(1-\lambda)(1-\rho)\tau_i\chi_i(1-\sigma_i)(1-\sigma_j)}{\lambda(1-\sigma_j) + (1-\lambda)\tau_i(1-\chi_i)(1-\sigma_i)}.$$

If i additionally observes $\hat{\xi}_j$, her average posterior belief is given by

$$\mathbb{E}[p_i^G(\xi = 1)] = \lambda + \beta_i^G,$$

where β_i^G is defined in (B.8).

Since $\tau_j < 1$, some algebra shows that $\beta_i^G < \beta_i^a$. Hence, observing the counter-partisan's rationale moves i's average belief about the state strictly towards its true value λ.

The effects on beliefs about types then follow from relations similar to those established in the proof of Proposition 6: in particular,

$$\begin{cases} \mathbb{E}[p_i^a(\theta_i = 1)] = \rho(1 + \beta_i^a), \ \mathbb{E}[p_i^a(\theta_j = 1)] = \rho(1 - \beta_i^a), \\ \mathbb{E}[p_i^G(\theta_i = 1)] = \rho(1 + \beta_i^G), \ \mathbb{E}[p_i^G(\theta_j = 1)] = \rho(1 - \beta_i^G). \end{cases}$$

□

Supplemental Appendix

In this Supplemental Appendix we formulate and analyze a version of our model where players independently choose an action a_i and whether to attempt to articulate a rationale supporting that action at $t = 0$. Formally, after choosing a_i, player i chooses $r_i \in \{a_i, \varnothing\}$. If $r_i = \varnothing$, the player does not attempt to articulate a rationale and $\hat{\xi}_i = \varnothing$. If $r_i = a_i$, the player attempts to forge a rationale for a_i. If $a_i = \xi$, the truthful rationale ($\hat{\xi}_i = \xi$) is found with probability one; if $a_i = 1 - \xi$, it is found ($\hat{\xi}_i = a_i$) with probability τ_i, and otherwise the agent is dumbfounded ($\hat{\xi}_i = \varnothing$).

We will show that, under some mild conditions, this game produces exactly the same equilibria as the ones derived in the main text. The two conditions we impose are the following.

First, in Equation (1) the Socratic utility is replaced by $\theta_i \gamma_i \mathbf{1}\{a_i = \xi, r_i = a_i\}$. This guarantees that reasonable types always produce a truthful rationale on top of playing the action prescribed by the signal.

Second, we break ties in favor of attempting to articulate a rationale supporting one's action. This condition is used only in situations where a player expects the same esteem whether she comes up with a rationale or not (e.g., when the other player already provides a rationale supporting her action), and is thus indifferent between trying or not.

Assumption S.1 (Tie-breaking)**.** *If a player is indifferent between playing* $(a_i, r_i = a_i)$ *and* $(a_i, r_i = \varnothing)$*, we select the former strategy.*

The next proposition establishes the equivalence between the two games, and applies to all environments considered in the text (autarky, echo chambers and agoras).

Proposition S.1. *The equilibrium of this game is unique, features* $r_i = a_i$ *for both players, and has the same equilibrium strategies as the model analyzed in the main text.*

Proof. We start by proving that, given the choice of an action a_i, a player cannot strictly benefit from giving up rather than attempting to articulate a rationale supporting her action. For reasonable types, this is obvious given the form of Socratic utility. For strategic types, since a_i is fixed the core utility is unchanged, and we only need to compare esteem.

Let us distinguish cases as a function of $\hat{\xi}_j$. If $\hat{\xi}_j = a_i$, then $\mathcal{A} = \{a_i\}$ regardless of whether i attempts to articulate a rationale of her own or not. Thus, i receives exactly the same esteem under $r_i = a_i$ and $r_i = \varnothing$.

Suppose now that $\hat{\xi}_j = 1 - a_i$. If i chooses $r_i = \varnothing$, then $\mathcal{A} = \{1 - a_i\}$, so that $a_i \notin \mathcal{A}$ and i receives zero image by Refinement 1. If instead i chooses $r_i = a_i$, failure to articulate the rationale leads to the same outcome, while a successful attempt yields $\mathcal{A} = \{a_i, 1 - a_i\}$ and therefore weakly positive image. Hence attempting is weakly better.

Finally, suppose that $\hat{\xi}_j = \varnothing$, which also covers the version of the game in autarky. If i chooses $r_i = \varnothing$, then $\mathcal{A} = \varnothing$, so that i receives zero image by Refinement 1. If instead i chooses $r_i = a_i$, failure again leads to zero image, whereas a successful attempt yields $\mathcal{A} = \{a_i\}$ and weakly positive image. Hence attempting is again weakly better.

Thus, conditional on any action a_i and any realization of the other player's articulated rationale, choosing $r_i = a_i$ yields weakly higher utility than choosing $r_i = \varnothing$. By Assumption S.1, all players therefore choose $r_i = a_i$ in equilibrium.

Conditional on $r_i = a_i$, the rationale technology is exactly the one analyzed in the main text. The alternative game therefore reduces to the game analyzed in the main text.

Conversely, any equilibrium characterized in the main text can be supported in this alternative game by setting $r_i = a_i$ for both players. Deviating to $r_i = \varnothing$, holding the action fixed, cannot be

profitable by the preceding argument, while deviations with $r_i = a_i$ are exactly the action deviations already available in the main model. Hence the two games have the same equilibrium strategies. This concludes the proof. □