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

Working Paper

Rationalizations and Political Polarization

CESifo Working Paper, No. 11897

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Le Yaouanq, Yves; Schwardmann, Peter; van der Weele, Joël J. (2025) : Rationalizations and Political Polarization, CESifo Working Paper, No. 11897, CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/320118

CESifo WORKING PAPERS

11897
2025

Mai 2025

Rationalizations and Political Polarization

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

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.ifo.de/en/cesifo/publications/cesifo-working-papers
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.ifo.de/en/cesifo/publications/cesifo-working-papers

Rationalizations and political polarization*

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

May 15, 2025

Abstract

We present a self- and social-signaling model formalizing findings in political psychology that moral and political judgments stem primarily from intuition and emotion, while reasoning serves to rationalize these intuitions to maintain an image of impartiality. In social interactions, agents' rationalizations are strategic complements: others' rationalizations weaken their ability to judge critically and make their actions less revealing of (inconvenient) truths. When agents are naive about their own rationalizations, our model predicts ideological and affective polarization, with each side assigning inappropriate motives to the other. Cross-partisan exchanges of narratives reduce polarization but are avoided by the agents. In within-group exchanges agents favor skilled speakers, whose narratives worsen polarization. Our model explains partisan disagreements over policy consequences, aligns with empirical polarization trends, and offers insights into efforts to disrupt echo chambers.

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, 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, INRAE Montpellier, LMU Munich, Université Paris-Dauphine PSL, University of Amsterdam, University of Bonn, University of Nancy, and University of Saint-Denis.

1 Introduction

Politics is increasingly defined by both ideological and affective polarization between partisan groups (Gentzkow, 2016; Sunstein, 2017; Iyengar et al., 2019; Braghieri et al., 2024b). Paradoxically, the advent of the internet and other advancements in communication technology, which have reduced the cost of information exchange, appear to have done little to reverse these trends. Since effective policymaking depends on social cohesion as well as a shared perception of reality, these developments pose serious challenges to the functioning of democracies (Kingzette et al., 2021).

A leading theory in political psychology attributes polarization 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 moral intuitions and emotions with evolutionary origins (Haidt, 2001, 2012; Graham et al., 2013). Reasoning merely serves to rationalize moral intuitions ex-post, in order to produce the appearance of rationality and impartiality – a Socratic ideal that is prominent in many cultures. Indeed, there is strong empirical evidence that links political preferences to moral intuitions. For instance, the distinction between communal and universalist moral intuitions predicts political behavior and preferences in several policy domains (Enke, 2020; Enke et al., 2023).

However, by itself, the idea that politics is rooted in fixed moral intuitions fails to capture the role of social influence and communication in shaping political ideology. In this paper, we develop a formal framework to explore these issues. In our model, two agents receive a public signal about the impact of a binary policy. They then publicly decide which policy to endorse, and whether to privately engage in rationalizations of the signal. The agents are caught between several forces. On the one hand, their moral intuitions or "core motives" lead them to favor a particular policy regardless of the evidence. On the other hand, agents wish to appear reasonable, i.e. responsive to evidence and logic. Agents can maintain a self-image of reasonableness *and* stick to their preferred action by selectively rationalizing away evidence that contradicts their moral intuitions. However, this tendency may be checked by their interaction with the other agent, whose esteem they also value. To understand the outcomes of these interactions, we analyze the best responses and equilibria of the resulting signaling game.

Our first result is that agents' rationalizations are strategic complements, operating through both a self and a social image channel. The self-image channel arises because rationalizations make behavior less informative of the actual evidence, allowing the other agent to feel better about her own delusions. The social image channel arises because rationalizations soften critical judgment of the other agent's actions. Thus, rationalizations

degrade the information environment by making agents both less informative to others, and less critical of them. This result holds both when the agents' moral intuitions are similar ("co-partisans") and when they diverge ("counter-partisans"). In co-partisan groups, the complementarity fosters belief conformity or groupthink. In cross-partisan encounters, it causes polarization: as one side abandons the truth, the other side will also lean into their moral intuitions, as they can now credibly attributing disagreement to their opponent's delusions. As we show in Section 5, the model explains empirically documented channels of social influence, such as accountability to (un)informed audiences (Lerner and Tetlock, 1999), and the strong reactions to in-group deviants (Marques et al., 1988).

Our second set of results emphasizes a key role for *naiveté*: agents' lack of awareness about their tendency to rationalize. The assumption of naiveté is in line with empirical evidence (Haidt, 2001; Kurzban, 2011; Kappes and Sharot, 2019; Melnikoff and Strohminger, 2020), and we show that it is necessary to align the predictions of the model with the stylized facts of political polarization. In particular, naiveté implies ideological polarization as counter-partisans (but not co-partisans) disagree about the evidence and moral value of different actions. These (mis)perceptions then infect the agents' entire belief system, causing them to overestimate their own commitment to truthful reasoning ("self-righteousness"), and underestimate it among counter-partisans. The result is affective polarization: a mutual mistrust about the moral reasoning of the other side. Our model thus suggests the existence of a common primitive driving these various belief distortions.

Our third set of results concerns the communication of narratives, the opportunity for which has drastically increased with the advent of social media. In addition to the mutual observation of behavior, we now allow one agent (the speaker) to disclose her "narrative": a moral argument or empirical fact that can either be the truthful recollection of the public signal, or a self-serving rationalization of it. Since producing good rationalizations is hard, the quality of the disclosed narrative may reveal the true state of the world. We show that the impact of communication depends crucially on the alignment of moral intuitions between speaker and listener. Co-partisan conversations improve the listener's self-image and further bias her beliefs, but only if the speaker is skilled at crafting rationalizations and the listener is naive enough to accept them. In addition, if the listener anticipates encountering a skilled co-partisan speaker, she becomes more prone to rationalizing. Thus, politicians, public intellectuals and influencers affect beliefs not just through persuasion, but also by providing cover for their follower's own rationalizations.

By contrast, a counter-partisan speaker unambiguously decreases factual and affective polarization since, on average, a naive listener is surprised by the strength of the speaker's arguments. Nevertheless, in an effort to protect their self-image, listeners prefer echo cham-

bers over being exposed to counter-partisan narratives. As we describe in Section 5, our model can explain why co-partisan content on social media drives polarization only when it is created by skilled influencers (Allcott et al., 2024; Rathje et al., 2024), while encounters with counter-partisan content and individuals generally decrease polarization (Levy, 2021; Blattner and Koenen, 2023; Fang et al., 2023; Hobolt et al., 2024; Braghieri et al., 2024a). The model then also explains why people self-select into echo chambers (Sunstein, 2001, 2009, 2017; Gentzkow and Shapiro, 2011; Nelson and Webster, 2017; Flaxman et al., 2016; González-Bailón et al., 2023; Guess et al., 2018; Guess, 2021; Braghieri et al., 2024a,b; Brown et al., 2024) and implies that an analyst attempting a welfare analysis of forced inter-group exposure must decide how to weigh the (systematic) loss of moral self-esteem implied by such an intervention.

Immigration provides a clear example of how our model applies. The issue activates conflicting moral intuitions—compassion for refugees and other immigrants on the left, and protective instincts along with loyalty to the ingroup on the right (Hoewe et al., 2022; Nath et al., 2022). Self-image concerns then lead individuals to frame their positions in universalist terms, ultimately causing conservatives and liberals to diverge in their perceptions of immigrants' behavior and impact on welfare (Alesina et al., 2023), resulting in mutual suspicion and affective polarization. Highlighting how social image concerns constrain the expression of socially inappropriate views, Bursztyn et al. (2023) find that providing scientific cover for anti-immigrant sentiments causally increases their public expression. At the same time, as documented by Hobolt et al. (2024), exposure to counter-partisan arguments reduces both policy and affective polarization around immigration. Nevertheless, our model warns that contact with the political outgroup may not be a policy panacea, as piercing comforting narratives can provoke backlash when individuals are eager protect the utility derived from self-image.

Our paper contributes to literatures in economics, psychology, and political science. In psychology, we build on ideas from the social intuitionist model (Haidt, 2001) and moral foundations theory (Graham et al., 2013). Both frameworks highlight the primacy of emotions and intuitions in moral decision-making and view moral reasoning primarily as a rationalizing force. By developing a formal model of rationalization and the underlying psychological and social incentives, we clarify how psychological assumptions translate into social and political implications. Our model elucidates the relationship between image concerns and complementarities in rationalizations and identifies the crucial role of naiveté in driving polarization and self-righteousness.

Our theoretical approach draws on economic models of motivated cognition that emphasize rationalization and self-deception in belief formation (Bénabou and Tirole, 2016).

Polarization in our model results specifically from rationalizations, rather than mechanical updating errors (Baliga et al., 2013; Fryer Jr. et al., 2019), heterogeneity in mental models (Levy et al., 2022), or media competition (Perego and Yuksel, 2022). However, our model differs from other economic models of motivated cognition in terms of its underlying mechanisms, its political application, and its focus on communication. Specifically, the cost of biased beliefs in our model arises from damage to social image rather than decision-making errors (Bénabou and Tirole, 2002; Bodner and Prelec, 2002; Brunnermeier and Parker, 2005; Bénabou, 2013). The benefits of rationalization stem from maintaining a self-image as a reasonable decision-maker despite following core motives, rather than from cognitive dissonance reduction (Yariv, 2005; Acharya et al., 2018; Eyster et al., 2021) or anticipatory utility (Brunnermeier and Parker, 2005; Bénabou, 2013; Le Yaouanq, 2023).[1]

A novel aspect of our model is that image concerns produce complementarities in belief distortions, because the disciplining effect of social interactions depends on others' own factual beliefs. A small set of other papers explores complementarities in beliefs within contexts like group work or consumption—either through interdependent anticipatory payoffs (Bénabou, 2013; Le Yaouanq, 2023) or inference from others' actions similar to our self-image channel (Bénabou and Tirole, 2011; Hestermann et al., 2020). Because these models are based on different assumptions about the nature, costs and benefits of biased beliefs, they do not easily extend to cultural issues or capture the drivers of polarization. Relatedly, Bénabou et al. (2019) and Foerster and van der Weele (2021) study strategic narrative transmission in the presence of externalities, without allowing for self-deception. Focused on a different set of questions surrounding political correctness, Golman (2023) shows that people's willingness to express dissenting opinions depends positively on how many people's true opinions align with what is politically correct.

Finally, our paper provides a unified framework for several previously separate explanations of polarization within political science (Boxell et al., 2024). These include moral foundations theory (Graham et al., 2013), which we incorporate as a primitive; rationalization, self-deception, and partisan bias (Taber and Lodge, 2006; Taber et al., 2009; Little, 2025), which are central mechanisms; and cross-partisan contact (Santoro and Broockman, 2022; Hobolt et al., 2024) and social media usage (Kubin and Von Sikorski, 2021), which

[1]Some self-signaling models focused on individual decision making presume that agents value a self-image as altruistic agents (Bodner and Prelec, 2002; Bénabou and Tirole, 2006; Grossman and van der Weele, 2017; Reichel, 2024). By contrast, agents in our model signal a commitment to evidence-based reasoning. They may disagree about what is appropriate, and the social norm that confers esteem is endogenous. Beyond the novel mechanisms we emphasize, we also make a conceptual contribution to the broader signaling literature: by extending notions of naiveté to off-path observations, our model can accommodate scenarios where heterogeneous agents make differential inferences from the same data (see equilibrium refinements in Appendix A).

serve as important mediators.

2 Model

We consider two agents, i and j. At time period $t = 0$, both agents independently select an action $a \in \{0, 1\}$. At $t = 1$, actions are observed by the other player. These actions represent a morally contentious behavior, like discriminating against outgroup members, engaging in affirmative action, voting for higher redistribution, polluting the environment, etc. It may also apply to speech acts, like vocally staking out a (policy) position regarding controversial topics. Our model distinguishes between two types of drivers behind these actions: core motives (or intuitions) and socially appropriate motives, that reflect the tension between the individual and social spheres. We discuss these drivers in turn, and then describe how individuals manage a potential conflict between these motives.

Core motives. The primary psychological drivers of the agents are given by their "core motives". These motives derive from our evolutionary history, our genetic make-up, or our personal experience, all of which are outside of the model. We denote core motives by the parameter b that multiplies the utility of the action a. Thus, core motives will lead an agent to favor action $a = 1$ (resp. $a = 0$) if $b > 0$ (resp. $b < 0$). Core motives have two possible foundations:

Moral intuitions. Moral intuitions and emotions like disgust, compassion or anger are the primary drivers of moral judgments and behavior. Moral emotions are inherently subjective and differ across the political spectrum (Haidt, 2012). In particular, liberals score relatively higher on emotions associated with care and fairness like compassion and guilt. Conservatives score higher on emotions associated with community, like respect and belonging (Graham et al., 2009). To the extent that these differences can be captured by or projected onto a single dimension, b can be interpreted as a political left-right distinction, or a distinction between universal and communal values (Enke, 2020; Enke et al., 2023).

Material benefits. Agents may favor actions that personally benefit them. This may drive their political activities through voting, lobbying or advocacy. For instance, richer people may favor lower taxes, and companies may favor lighter regulation.

Socially appropriate motives. While core motives capture *individual* drivers, society prescribes *socially* acceptable reasons for actions. We assume that agents use a common cul-

tural and moral framework to evaluate actions normatively, e.g. utilitarianism or a Christian morality. Such moral frameworks will typically produce general moral prescriptions that are uncontroversial and praiseworthy, such as "protect the children", "take care of the planet" or "reward hard work and merit". However, in a particular decision context, there is uncertainty about which actions best implement these moral rules, which require factual evidence and logical arguments. For instance, even though agents agree that it is important to promote the most qualified person for the job, it may not be obvious who that person is.

To capture this uncertainty, we introduce the variable ξ. This variable can be high ($\xi = \xi_H$) or low ($\xi = \xi_L$) with respective probabilities λ and $1 - \lambda$. In particular, we assume that all agents agree that $a = 1$ is the normatively appropriate action if $\xi = \xi_H$, whereas $a = 0$ is normatively appropriate if $\xi = \xi_L$. At $t = 0$ both agents receive a public signal that perfectly communicates the value of ξ. Thus, $\xi = \xi_H$ and $\xi = \xi_L$ can be interpreted as facts, pieces of evidence, or logical arguments that are credible and compelling enough to prescribe behavior to any reasonable person.

Conflicts and rationalizations. Central to our model are conflicts between the core motive b and the socially appropriate action indicated by ξ. For instance, agents' personal motives may favor an in-group job applicant, but the applicant may score poorly on socially acceptable criteria like skills and competences, captured by ξ. In terms of our model, we will refer to the realization of ξ as "good news" if it aligns with the agent's motives and interests, and as "bad news" if it does not.

To avoid a conflict behind private motives and public morality, the agent may engage in rationalizations. Through the use of selective attention and creative reinterpretation, she may reframe the decision context to make her preferred action appear more reasonable. Continuing our previous example, an employer may selectively highlight qualities of the in-group applicant, to present her as more valuable to the organization. Similarly, a capitalist may promote "trickle-down" economics to justify why his personal benefit is aligned with the general interest (Oreskes and Conway, 2023). Psychological evidence indicates that people often come to believe these ex-post rationalizations. Moreover, these rationalizations allow them to make decisions that would otherwise be socially unacceptable, such as expressing anti-immigrant sentiments (Bursztyn et al., 2023) engaging in selfish behavior (Dana et al., 2007; Exley, 2015), or indulging in guilty pleasures (Woolley and Risen, 2021).

Following Bénabou and Tirole (2002), we formalize this process via an intra-personal communication game between two incarnations of each agent, Self 0 and Self 1. At $t = 0$, upon seeing the public signal ξ, each agent's Self 0 simultaneously chooses an action $a \in \{0, 1\}$ and a *narrative* $\tilde{\xi} \in \{\xi_L, \xi_H\}$ that she wants to transmit to her future self (Self 1). The

encoded narrative can either be truthful and correspond to the signal ($\tilde{\xi} = \xi$), or the agent can try to transmit an untruthful counter-narrative corresponding to the opposite signal ($\tilde{\xi} \neq \xi$). We will refer to the untruthful narrative as a "rationalization", and to narratives that are congruent with the speaker's convenient action (whether they are rationalizations or not) as "justifications" or "convenient" narratives.

We assume that the process of finding a rationalization may fail, reflecting the idea that constructing credible justifications is hard (Williams, 2023). We capture this by a parameter $\tau \in (0, 1)$, the probability with which the narrative by Self 1, which we write as $\hat{\xi}$, is indeed the rationalization sent by Self 0 ($\hat{\xi} = \tilde{\xi}$). With probability $1-\tau$, rationalization fails and Self 1 instead retrieves the true signal ($\hat{\xi} = \xi$). In this case, following Haidt and Graham (2007), we say an agent is "morally dumbfounded", as she cannot produce a socially approved moral reason for her actions. One can interpret τ as the accessibility of rationalizations, which depends on the degree of situational moral wiggle room and the creativity of the agent.

Metacognition. A key issue in the analysis is the degree to which individuals trust their own rationalizations. At one extreme, *sophisticated* individuals are fully aware of their (equilibrium) tendency to self-deceive, as in Bénabou and Tirole (2002). At the other extreme, *naive* individuals take convenient narratives at face value, as if it were the original signal ξ, as in Brunnermeier and Parker (2005). The psychological evidence supports a more naive interpretation, and shows that individuals are more skeptical and suspicious of inconvenient narratives than convenient ones (Ditto and Lopez, 1992; Taber and Lodge, 2006; Mercier and Sperber, 2011; Simler and Hanson, 2017; Hagenbach and Saucet, 2024).

Following the evidence, we allow for a flexible degree of naiveté to capture all possible cases. At $t = 1$, upon internally retrieving a convenient narrative, Self 1 believes that the probability that her own rationalizations are successful is given by $\tau(1-\chi)$. Here, $\chi \in [0, 1)$ measures the degree of naiveté: $\chi \approx 1$ indicates near-perfect naiveté, where the agent believes that her attempts at rationalization always fail, and her current narrative $\hat{\xi}$ must therefore reflect reality.[2] The case $\chi = 0$ indicates full sophistication where the agent is aware of the actual rationalization technology. Because motivated cognition is self-serving, we assume that the agent is sophisticated about the rationalizations of the other player. Indeed, the psychological evidence indicates that people are often quick to point out self-serving biases in *others*, especially when they advance a different narrative (Kurzban, 2011; Mercier and Sperber, 2011; Trivers, 2011).

[2]We do not discuss the case where $\chi = 1$. 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 which way by making additional assumptions.

To close the model, we must also specify the agent's meta-metacognition, that is, Self 0's belief about Self 1's χ. The main text exposes our results under the assumption that Self 0 has correct beliefs about χ (i.e. sophistication about future naiveté). This assumption turns out to be inessential for our main results, at least qualitatively.[3] In addition, we assume awareness of others' naiveté.

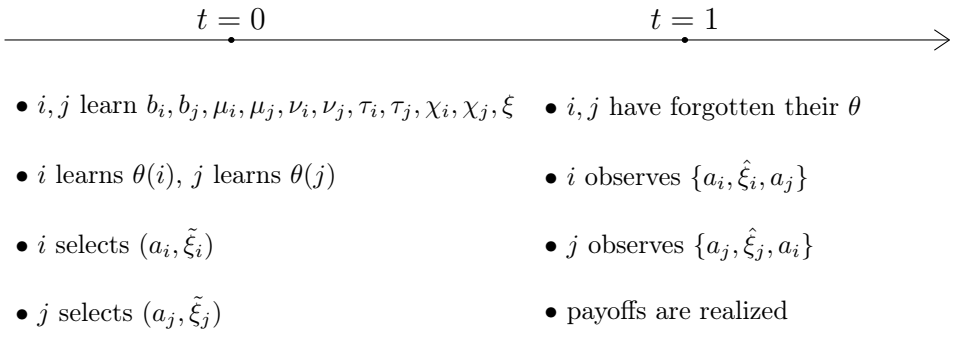
Moral reasoning and types. An ideal moral reasoner uses impartial logic and follows the evidence wherever it leads, a Socratic ideal that is prevalent in most philosophical traditions. In reality, as described above, people are heterogeneous in their ability and willingness to follow unbiased moral reasoning (Saccardo and Serra-Garcia, 2023). We capture this heterogeneity by assuming two different types. With probability ρ the agent is a *Socratic* or *reasonable* type ($\theta = R$) and accepts the normative facts or moral logic embodied in the signal ξ. She transmits the true value of ξ with probability 1, and always follows the normatively appropriate action: $a = 1$ if $\xi = \xi_H$, and $a = 0$ if $\xi = \xi_L$. We do not model the reasonable type as an explicit optimizer: she has a personal motive b, but her intellectual honesty is strong enough not to let it cloud her judgment. With probability $1 - \rho$, the agent is a *sophist* or *strategic type* ($\theta = S$): she chooses her belief and action as the result of an explicit cost-benefit analysis guided by her core motives. If opportune, she will engage in rationalizations to cast her behavior in a better light. Our analysis focuses on the behavior of strategic individuals.

Self-image concerns. We assume that people want to be perceived as a reasonable type, as outlined in our introduction. Image concerns in our model come in two flavors. First, we assume agents care about how they perceive themselves. To model this, we enrich the model with a self-signaling technology, as in Bénabou and Tirole (2011). At $t = 0$, the agent's Self 0 observes her type θ through the experiences and emotions that accompany the decision. At time $t = 1$, the agent's Self 1 no longer has access to her type, and introspection alone cannot elucidate her past motives. Instead, she makes a Bayesian inference about her type based on some concrete observations: her retrieved narrative $\hat{\xi}$, the action she played, and the action of the other player. She then reaps a self-image payoff proportional to the probability she attaches to being reasonable.

Social image concerns. In addition to self-image, agents care about being perceived as a reasonable person by the other player. This is line with Moehring and Molina (2023), who

[3] Behavior under the alternative assumption of naiveté about future naiveté can be studied by substituting $\chi = 1$ for the actual value of χ in all equilibrium conditions below.

Figure 1: Timeline of the game



show that people select more balanced news once their news diet is under public scrutiny. At $t = 1$, each agent uses the information set (a_i, a_j) and her private recollection $\hat{\xi}$ to form posterior beliefs about her own type (reasonable or strategic) and the type of the other player. Crucially, at $t = 0$, agent i cares about agent j's true perception of i, not about i's Self 1 beliefs over j's image of her. That is, the agent actually seeks a good reputation, not the erroneous belief that she enjoys a good reputation. This preference may be motivated by the (material) benefits accruing to the genuinely esteemed, and distinguishes social image concerns from self-image concerns. In Section 4, we also allow agents to share their narrative with the other player in order to improve their reputation.

Timing and utility. The timing of the game outlined above is summarized in Figure 1. The utility accruing to a strategic agent i who chooses action a_i and who retrieves $\hat{\xi}_i$ at date 1 is:

$$\underbrace{b_i a_i}_{\text{core utility}} + \mu_i \underbrace{p_i[\theta(i) = R \mid (a_i, \hat{\xi}_i), a_j]}_{\text{self-image}} + \nu_i \underbrace{p_j[\theta(i) = R \mid (a_j, \hat{\xi}_j), a_i]}_{\text{social image}}. \tag{1}$$

The utility of player j (if strategic) is defined symmetrically. The first part of the utility function captures the core motives of the agent.[4] The second and third part capture self-image and social image concerns, which are scaled by parameters μ_i and ν_i respectively. Image concerns reflect the inferences of agent i (self-image) and j (social image) about the type of agent i. They depend on the action profile of both agents, and the observer's private recollection of the public signal $\hat{\xi}$. Apart from the types θ, which are privately known at $t = 0$ and forgotten before $t = 1$, all other preference and cognitive parameters (b, μ, ν, τ,

[4]In this representation, strategic agents are intrinsically insensitive to moral arguments, as ξ does not appear in their utility function—in contrast with Socratic agents, who are maximally sensitive to moral arguments. One could alternatively replace the material payoff of strategic types by $(b_i + \xi)a_i$ without affecting the main results.

χ) are common knowledge.

Equilibrium concept. We look for Perfect Bayesian Equilibria that satisfy a number of criteria. First, we disregard equilibria in weakly dominated strategies. This allows us to focus on *congruent* action-recollection pairs ($a = 1, \tilde{\xi} = \xi_H$) or ($a = 0, \tilde{\xi} = \xi_L$).[5] Thus, we can represent agent i's strategy as a pair (σ_L^i, σ_H^i), where each σ^i measures the degree of realism after the corresponding signal: σ_H^i is the probability with which a strategic i plays $(a_i = 1, \tilde{\xi}_i = \xi_H)$ conditional on $\xi = \xi_H$, and σ_L^i is the probability with which she plays $(a_i = 0, \tilde{\xi}_i = \xi_L)$ after $\xi = \xi_L$.

Second, we restrict attention to a class of equilibria we call *motonone*, in which the equilibrium probability with which a player selects an action is larger under its congruent signal than under the opposite signal:

Definition 1. *A Perfect Bayesian Equilibrium is monotone if* $\rho+(1-\rho)\sigma_H^i \geq (1-\rho)(1-\sigma_L^i)$ *and* $\rho + (1-\rho)\sigma_H^j \geq (1-\rho)(1-\sigma_L^j)$.

In a monotone equilibrium, actions respond to the normative arguments embedded in ξ in the expected direction. Our analysis will show that there always exists a (unique) monotone equilibrium.[6]

Third, we impose three restrictions on off-path beliefs. We believe these refinements are natural, as they extend the logic of (naive) Bayesian belief formation to off-path beliefs. They may also be of more general use in Bayesian games with some form of naiveté that feature potential disagreements between players. We briefly sketch the logic of the three refinements, and defer formal definitions to Appendix A.

1. **Reasonable types never rationalize.** An agent who is expected to be fully realistic may deviate and find himself dumbfounded. This refinement ensures that such deviations are attributed to the strategic type.

2. **Intra-personal consistency.** Consider an interaction between two counter-partisans (e.g. $b_i > 0$ and $b_j < 0$). In an equilibrium where both agents are expected to be fully

[5] On the equilibrium path, any non-congruent action-recollection pair must be attributed to a strategic type, and is therefore weakly dominated by playing the corresponding action but attempting to send the congruent narrative. Since non-congruent pairs are not strictly dominated, there exists equilibria where these actions are played. But an agent is indifferent between playing $(a, \hat{\xi} = \xi_H)$ and $(a, \hat{\xi} = \xi_L)$ if and only if her expected self-image equals zero in both cases, which only happens when social interactions are perfectly informative about ξ. In such an equilibrium, the pairs $(a, \hat{\xi} = \xi_H)$ and $(a, \hat{\xi} = \xi_L)$ give rise to the same ex post observables (actions and beliefs). We therefore do not lose generality by ruling out noncongruent pairs.

[6] Every equilibrium is monotone if ρ is large enough ($\rho \geq 1/2$ suffices), if players are sophisticated enough (small χ_i, χ_j) or in the absence of social interactions (see Section 3.1). In Section B.4.3 of the Appendix, we show that if these conditions are violated, non-monotone equilibria where players coordinate on the "wrong" action may exist.

realistic, actions should always agree. We impose that, upon observing off-path self-serving disagreement ($a_i = 1$ and $a_j = 0$), each player's beliefs about herself and the other agent are consistent (that is, they sum to the prior ρ). This rules out pathological inferences such as believing that both players are reasonable with probability one.

3. **Inter-personal consistency.** Consider again full realism and off-path disagreement in actions, as in the previous case. This refinement posits a consistency between the interpretations of agents with symmetric information sets, ensuring that their beliefs diverge only due to differences in their personal parameters χ and τ.

Applications. The theory has a rich set of applications: it can capture any situation in which diverging moral intuitions generate opposing beliefs and narratives. In addition to the immigration example from the introduction, other divisive topics fit this description. For instance, on abortion, b captures diverging moral foundations of freedom (for mothers) vs. sanctity (of the unborn life) (Lockhart et al., 2023). Opposing narratives $\hat{\xi}$ reflect motivated reasoning in this area, which has been documented both in evaluating deductive arguments about abortion (Čavojová et al., 2018) and in interpreting U.S. legislative documents surrounding the repeal of Roe vs. Wade (Geller et al., 2025).

As another example of cultural division in the U.S., the gun control debate pits moral foundations related to liberty and authority against foundations of care (Cook, 2015), which can be represented by oppositely signed b's. These drivers then generate motivated reasoning about the relationship between gun control and social safety (Kahan et al., 2013), epitomized by universalist narratives ($\hat{\xi}$) that stress deterrence ("the only thing that stops a bad guy with a gun is a good guy with a gun") or victimhood ("protect children, not guns").[7] More generally, values related to individualism and hierarchy correlate not just with moral views about issues like abortion, global warming and homosexuality, but also with a perception of their risks (Kahan et al., 2007, 2008).

3 Equilibrium analysis

We now turn to our analysis and results. In Section 5 we discuss how these results help organize existing findings on social influence, polarization and political discussion.

[7]Note that applications need not be symmetric. When their motives align, members of a society may subscribe to the same rationalizations, as in the case of factory farming and animal welfare (Hestermann et al., 2020), or the use of "natural law" to justify European colonialism (Waswo, 1996). Applications also extend to situations where the norm used to define reasonable behavior is not necessarily moral: e.g. the degree to which smoking is perceived as (un)reasonable depends on evidence about its health impact.

3.1 Preliminary individual analysis

Before moving to our main analysis, we focus on a simple case without social interactions to illustrate the logic of rationalizations and to develop some terminology. That is, we ignore the social image terms in Equation (1), reflecting the idea that a_i is not publicly observable, and we let the self-image $p_i[\theta(i) = R \mid a_i, \hat{\xi}_i]$ be based only on the action of agent i and her private recollection of the signal. For clarity, we drop the subscript i in this section.

Proposition 1 establishes existence and uniqueness of an equilibrium when $b > 0$.[8]

Proposition 1 (Individual behavior)**.** *Suppose $b > 0$. Then there exists a unique equilibrium of the game, in which:*

1. *$\sigma_H = 1$ (good news is always transmitted).*
2. *σ_L (the propensity to transmit inconvenient arguments) is a nondecreasing function of μ (image concerns) and a nonincreasing function of b (personal motive), τ (ease of finding rationalizations) and χ (naiveté).*

This equilibrium features a number of results that will carry over to our analyses below. First, when news is good, there is no trade-off: the agent takes the action prescribed by her core motives, and truthfully transmits the narrative that supports that action. The agent may rationalize only bad news, which implies a bias towards the self-serving action $a = 1$. Second, the model generates motivated cognition, since the personal motives and characteristics causally determine the agent's beliefs about the morality of the different actions.

Third, the agent is more likely to rationalize when her motive b is larger, her naiveté is larger, rationalizations are easier to come by, and self- image concerns are smaller. Note that the fixed point may be internal, because a higher equilibrium probability to rationalize undermines the effectiveness of the rationalization, as a sophisticated agent discounts it more heavily.

Finally, narratives are not just epiphenomenal justifications but actually drive behavior. In particular, if rationalization becomes easier (τ increases), a sophist is more likely to engage in the convenient action. The reason is that narratives allow the agent to obfuscate her true motives, and cushion the self-image consequences of ignoring the original signal (see also Grossman and van der Weele, 2017).

[8] There do not exist non-monotone equilibria in that case. The situation where $b < 0$ is the mirror image. If the agent's core motives are neutral ($b = 0$), then the only equilibrium is $\sigma_H = \sigma_L = 1$ (no bias).

3.2 Social influence in rationalizations

We are interested in how j's rationalizations affect i's incentives to rationalize and vice versa. Therefore, we now analyze social interactions between i and j, where both players observe each other's action. In Section 4, we extend the model with the option to reveal the retrieved narrative, which better captures in-person interactions, or communication on social media.

Social interactions affect utility through both self and social image. First, i's self-image will depend on observing the action a_j, which might reveal information about ξ and therefore about i's type. Second, i cares about j's perception of her type, which in turn depends on j's decision to stay realistic or to rationalize. Moreover, it turns out that the compatibility of i and j's personal 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. Thus, in the remainder, we will fix $b_i > 0$ and distinguish between two cases: co-partisans, with $b_j > 0$, or counterpartisans, with $b_j < 0$ (the cases with $b_i < 0$ are symmetric).

Given our focus on monotone equilibria, we can prove that players always accept the signal realizations that are convenient for them, in both co-partisan and counter-partisan case.

Lemma 1. *In every monotone equilibrium, good news is accepted:* $\sigma_H^i = 1$.

We can therefore describe any monotone equilibrium by the level of realism σ_L^i of i following $\xi = \xi_L$, and the level of realism of j, equal to σ_L^j for a co-partisan ($b_j > 0$) and σ_H^j for a counter-partisan ($b_j < 0$). We drop the subscripts and write the objects simply (σ^i, σ^j) for convenience.

3.2.1 Equilibrium analysis with co-partisans

We first analyze the case of 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 start by fixing the level of realism of j at some σ^j, and consider a tentative best-response σ^i by i upon seeing bad news $\xi = \xi_L$. If i plays $(a_i = 0, \tilde{\xi}_i = \xi_L)$, her utility is given by

$$U_{0,i}^{\xi_L}(\sigma^i) = (\mu_i + \nu_i) \underbrace{\frac{\rho}{\rho + (1-\rho)\sigma^i}}_{\text{reputation for } a_i=0}. \tag{2}$$

Playing $a_i = 0$ reveals that $\xi = \xi_L$ because, as we discussed above, no agent (reasonable or strategic) plays $a_i = 0$ after $\xi = \xi_H$. This leads to a positive self- and social image, which depends on the equilibrium probability, as the only uncertainty left is whether i played the virtuous action because i is a reasonable player, or for signaling purposes.

If i plays $(a_i = 1, \tilde{\xi}_i = \xi_H)$ instead, she reaps utility b_i from playing her favorite action. However, her image (both self and social) will depend on a successful rationalization of $\xi = \xi_H$ by herself and player j. To see this, first consider self-image: i's self-image is positive only if she is not dumbfounded, i.e. $\hat{\xi}_i = \xi_H$, which happens with probability τ_i. Moreover, since $a_j = 0$ reveals the signal to be $\xi = \xi_L$, player j must play the convenient action $a_j = 1$, which happens with probability $(1-\rho)(1-\sigma^j)$. If these conditions are met, i's self-image will be positive, but discounted by the probability that both players are strategic and engaged in rationalization (corrected by naiveté). Next, consider i's reputation vis-à-vis j. Image utility is positive if and only if j plays $a_j = 1$ and successfully rationalizes this action, i.e. $\hat{\xi}_j = \xi_H$, which happens with probability $\tau_j(1-\rho)(1-\sigma^j)$. Again, image will be discounted by the probability that both players are strategic and engaged in rationalization.

Collecting these terms and computing the Bayesian posteriors, i's expected utility from playing $a_i = 1$ equals[9]

$$U_{1,i}^{\xi_L}(\sigma^i, \sigma^j) = b_i + \mu_i \underbrace{\tau_i(1-\rho)(1-\sigma^j)}_{\text{proba of } a_j=1,\ \hat{\xi}_i=\xi_H} \underbrace{\frac{\lambda\rho}{\lambda + (1-\lambda)\tau_i(1-\chi_i)(1-\rho)^2(1-\sigma^i)(1-\sigma^j)}}_{\text{self-image for } (a_i=1,\ \hat{\xi}_i=\xi_H),\ a_j=1}$$
$$+ \nu_i \underbrace{\tau_j(1-\rho)(1-\sigma^j)}_{\text{proba of } a_j=1,\ \hat{\xi}_j=\xi_H} \underbrace{\frac{\lambda\rho}{\lambda + (1-\lambda)\tau_j(1-\chi_j)(1-\rho)^2(1-\sigma^i)(1-\sigma^j)}}_{\text{reputation for } (a_j=1,\ \hat{\xi}_j=\xi_H),\ a_i=1}. \quad (3)$$

Equations (2) and (3) illustrate the trade-off that agent i faces when confronted with bad news ($\xi = \xi_L$). Encoding the normative arguments and playing the inconvenient action $a_i = 0$ improves her self- and social image, but results in forgoing $b_i > 0$. The equations also illustrate how complementarities arise. If player j is a probable realist, i.e. has a high σ^j, then j enhances i's realism via two channels. The first is a *self-image channel*: if j is more realistic, i is more likely to be confronted with j playing an inconvenient action, which undermines i's rationalizations and self-image. The second is a *social image channel*: if j is more realistic, i receives a lower reputation when she rationalizes and plays her convenient action, as j will judge her more harshly.

To show these complementarities formally, note that the net benefit from self-deception $U_{1,i}^{\xi_L}(\sigma^i, \sigma^j) - U_{0,i}^{\xi_L}(\sigma^i)$ determines player i's best response.[10] This expression is increasing in σ^i, yielding a unique best response, and decreasing in σ^j, reflecting strategic complementarity.

[9] The posterior beliefs are all pinned down by Bayes' rule, except in the out-of-equilibrium scenarios where $\sigma^i = 1$ and $[(a_i = 1, \hat{\xi}_i = \xi_H), a_j = 0]$ (for i's beliefs) or $[(a_j = 0, \hat{\xi}_j = \xi_L), a_i = 1]$ (for j's beliefs). See the appendix for more details.

[10] The candidate strategy σ^i is a best-response if and only if one of the following conditions holds: (i) $\sigma^i = 1$ and $U_{0,i}^{\xi_L}(1) \geq U_{1,i}^{\xi_L}(1, \sigma^j)$; (ii) $\sigma^i \in (0,1)$ and $U_{0,i}^{\xi_L}(\sigma^i) = U_{1,i}^{\xi_L}(\sigma^i, \sigma^j)$; (iii) $\sigma^i = 0$ and $U_{0,i}^{\xi_L}(0) \leq U_{1,i}^{\xi_L}(0, \sigma^j)$.

Lemma 2 (Complementarity in best response)**.** *Given some σ^j, there exists a unique best response $\sigma^i \in [0,1]$ for i. This best-response level of compliance with inconvenient signals is nondecreasing in σ^j (strategic complementarity), in μ_i and in ν_i (image concerns), and nonincreasing in b_i (personal motive), τ_i and τ_j (rationalization efficiency), and χ_i and χ_j (naiveté).*

The next result describes the unique monotone equilibrium of the game, defined as an intersection of i's and j's best-response functions.

Proposition 2 (Equilibrium with co-partisans)**.** *There exists a unique equilibrium of the game. In this equilibrium, i's level of realism σ^i is nonincreasing in b_i, b_j (personal motives), τ_i, τ_j (wiggle room) and χ_i, χ_j (naiveté), and nondecreasing in $\mu_i, \nu_i, \mu_j, \nu_j$ (image concerns).*

Several things are noteworthy about this equilibrium. First, the comparative statics with respect to $b_i, \mu_i, \nu_i, \tau_i$ follows the same logic as Lemma 2 above: realism requires a low level of emotional or material attachment (low b), low situational wiggle room or personal creativity (low τ) and high levels of image concerns (high μ and ν).

Second, our model generates a form of "groupthink" by predicting that motivated beliefs are mutually reinforcing. Thus, motivated beliefs do not only depend on an agent's own preferences and abilities, but also on those of her interaction partners. Finally, note that despite players' strategies being strategic complements, the equilibrium is unique. A key reason for this is that rationalizations undermine themselves: the more likely a player is to rationalize in equilibrium, the more she discounts her rationalizations. This puts a brake on the rationalizations of both players i and j, and limits the strength of the complementarity.

3.2.2 Equilibrium analysis with counterpartisans

Suppose now that $b_i > 0$ but $b_j < 0$, so that both players are looking for opposite justifications. The signals $\xi = \xi_H$ and $\xi = \xi_L$ are respectively good and bad news for i, and vice versa for j. Again, both players will always transmit good news truthfully in equilibrium, so we write (σ^i, σ^j) to indicate the (equilibrium) probabilities with which each agent truthfully transmits what is inconvenient for them.

To understand the strategic incentives, we again compare the expected utility for player i from choosing either realism or rationalization upon seeing bad news. The utility from playing realism is the same as given in Equation (2). As before, i fully reveals the state through her actions, so $j's$ actions or rationalizations do not matter. We now turn to the utility from rationalizing. Upon observing $\xi = \xi_L$, i knows that j will observe the same signal and play $(a_j = 0, \tilde{\xi}_j = \xi_L)$. In contrast to the co-partisan case, conflict does not necessarily

reveal the true ξ, as both counter-partisans play their convenient actions and may have engaged in rationalization. To adjudicate this conflict, both agents use the equilibrium strategies to update beliefs about ξ.

Despite this difference between the two settings, similar complementarities arise in the counter-partisan case as well: increased realism by j will make her a more credible party in the disagreement, to the detriment of i's self- and social image. First, through the self-image channel, a higher σ^j will increase i's tendency to attribute disagreement to her own rationalizations and lower her self-image. Second, through the social image channel, it reinforces j's confidence in her own narrative and causes her to update more negatively about i's type. All in all, i's expected utility after rationalizing equals

$$U_{1,i}^{\xi_L}(\sigma^i,\sigma^j) = b_i + \mu_i\tau_i \underbrace{\frac{\lambda\rho(1-\sigma^j)}{\lambda(1-\sigma^j)+(1-\lambda)\tau_i(1-\chi_i)(1-\sigma^i)}}_{\text{self-image for } (a_i=1,\ \tilde{\xi}_i=\xi_H),\ a_j=0} + \nu_i \underbrace{\frac{\lambda\rho\tau_j(1-\chi_j)(1-\sigma^j)}{\lambda\tau_j(1-\chi_j)(1-\sigma^j)+(1-\lambda)(1-\sigma^i)}}_{\text{reputation for } (a_j=0,\ \tilde{\xi}_j=\xi_L),\ a_i=1}.$$

This expression is well-defined only if $\min(\sigma^i,\sigma^j) < 1$. In that case, the expression is decreasing in σ^j, demonstrating complementarity in rationalizations. In Section B.5 of the Appendix, we prove uniqueness of the equilibrium under our equilibrium refinements. We also exhibit a case of equilibrium multiplicity if these conditions are violated.

Proposition 3 (Equilibrium with counter-partisans)**.** *There exists a unique equilibrium of the game. In this equilibrium, i's level of realism σ^i is nonincreasing in b_i (i's personal motive), τ_i, τ_j (wiggle room) and χ_i (i's naiveté), and nondecreasing in b_j (j's personal motive) and $\mu_i, \nu_i, \mu_j, \nu_j$ (image concerns).*

Complementarities in rationalizations lead to anti-groupthink, or *polarization*. As in the co-partisan case, polarization is driven both by the agent's own emotional/material attachment b and efficiency of rationalization τ as well as those of the opposite side. Thus, our model 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 rationalization, as it can now more credibly blame any disagreement on the other side's delusions. The combined comparative statics in b_j from Propositions 2 and 3 also reveal that moderates (those with $|b_j|$ near zero) serve as the strongest deterrents to their peers' rationalizations. Specifically, σ_i is hill-shaped in b_j: it is minimized when agent j has a strong core motive (large $|b_j|$ of either sign) and maximized when $|b_j|$ is moderate.

3.2.3 Rationalizations and the nature of social influence

These findings relate to empirical results on social influence, which we discuss in Section 5, and to existing theoretical approaches. First, they are reminiscent of the Mutually Assured Delusion principle in Bénabou (2013), which also describes a form of groupthink. However, the mechanism is entirely different: the driving force of complementaries in Bénabou (2013) is anticipatory utility about payoff externalities. By contrast, in our model, strategic complementarity arises through self and social image. Moreover, unlike in Bénabou (2013), or herding models of social learning, complementarities may lead beliefs and actions to polarize rather than conform: we may observe groupthink or anti-groupthink depending on the alignment of core motives.

Second, one can view our theory in terms of an *informational public good*, as rationalizations produce externalities on other agents. From the perspective of a strategic agent, this externality is positive, as others' rationalizations license her to produce her own, which increases her utility, at least on average. From the perspective of a Socratic agent, or a planner who cares about reasonable actions being played, rationalizations constitute a negative externality that pollutes the information environment by rendering agents less morally critical, and their actions less informative of the true state.

Third, we can relate our equilibrium results to the theory of social norms (Bicchieri, 2006). Using the pertinent terminology, *descriptive norms* refer to behavior that is actually played, while *injunctive norms* describe behavior that is (perceived as) socially desirable. Rationalizations are associated with both types of social norms. For instance, an exogenous shift in parameters b or τ will shift behavior, both by changing i's own view of what is acceptable and her expectations of what j finds acceptable. Thus, our theory provides a "thick" description of social norms, which includes behavior, beliefs and narratives.

Finally, we can expect strategic complementarities to be empirically robust, seeing as they rely on two separate image mechanisms. In particular, mutual observability of actions is not a necessary condition. Indeed, i's tendency to rationalize is affected by j's behavior as soon as i observes j's action (via the self-image channel) or j observes i's action (via the social image channel). The insights of the model are therefore relevant for both symmetric personal interactions and for asymmetric ones, where one of the players is a public figure and the other an anonymous observer. In Section 5 we discuss evidence for both esteem channels. In addition, we note that the complementarity we identified in the counter-partisan case also extends to a situation where a player is not concerned about her reputation vis-à-vis the other side but towards a passive audience that is uninformed about ξ and makes an inference about both players from observing their actions (a_i and a_j).

3.3 The effects of naiveté

All results thus far are robust to assuming perfect sophistication about rationalizations. However, sophistication implies that both agents are (on average) 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. These predictions contradict stylized facts of political polarization. In reality, we observe both *ideological polarization*, where different sides of the political spectrum disagree on policy positions, and *affective polarization*, where each side sees the other camp as less moral compared to their own. Moreover, the assumption of sophistication contradicts psychological evidence on the nature of motivated cognition, which is often immediate and occurs below awareness level (Kappes and Sharot, 2019; Melnikoff and Strohminger, 2020).

To address these points, Proposition 4 summarizes the effect of naiveté on equilibrium beliefs in both the co- and counter-partisan case. Here, all mentions of "beliefs" refer to the ex-ante expected posterior belief at $t = 1$, averaged over all strategies and realizations of ξ.

Proposition 4 (Naiveté)**.** *Suppose that $b_i > 0$, $\sigma^i < 1$, $\sigma^j < 1$ and $\chi_i > 0$, i.e. both players sometimes rationalize and i is less than perfectly sophisticated. Then the equilibrium involves*

1. **Ideological polarization***: if $b_j < 0$, then i and j disagree about the value of ξ, with each agent overestimating the likelihood of their "convenient" signal.*

2. **Self-righteousness***: i's ex-post belief about her own morality $p_i[\theta(i) = R]$ is biased upwards relative to the truth.*

3. **Affective polarization***: if $b_j > 0$ ($b_j < 0$), then i's ex-post belief $p_i[\theta(j) = R]$ about j's morality is biased upwards (downwards) relative to the truth.*

Proposition 4 shows a series of results that fit the stylized facts of political polarization. First, naiveté generates *ideological polarization* between counter-partisans. 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 was coded honestly and hence reflects the original signal. Since i takes her rationalizations too seriously, average beliefs about ξ diverge between i and j. Note that naiveté does not generate systematic disagreements between co-partisans, although the size of the individual bias will vary with individual parameters, e.g. the magnitude of χ_i and χ_j.

Second, naiveté predicts *self-righteousness*: as the agent underestimates the possibility of rationalization, she overestimates the probability of coding the signal correctly, and hence being a reasonable type. This happens even though she knows the true equilibrium

strategies. Finally, naiveté predicts *affective polarization*, as self-righteousness affects beliefs about the interaction partner. Since i overestimates the probability of the convenient signal ξ, she underestimates the likelihood of rationalization by j who plays the same convenient action, i.e. an agent with b of the same sign. Conversely, she overestimates the likelihood of rationalization by j who plays the opposite action (i.e. with opposite signed b).

Proposition 4 highlights two general points. First, naiveté about rationalizations is crucial in explaining systematic differences in opinion that could never arise under full Bayesian rationality. Second, the beliefs about the agent are interconnected, as belief biases about the underlying state ripple through her belief system and affect her moral valuation of both herself as well as her co- and counterpartisans.

4 Communication of narratives

So far, agents only observed each others' actions. In many contexts, people also have the opportunity to share arguments to justify their preferred position, either in face-to-face communication or via social media. Here, we model such communication and formally analyze the incentives for sharing and listening to arguments from different sides of the ideological spectrum. We are especially interested in studying the agents' incentives to self-select into "echo chambers" as well as the effect of echo chambers on ideological and affective polarization.

4.1 Modeling communication

To study communication between agents from different parts of the political spectrum, we consider a game with three players: a passive *listener* i ($b_i > 0$) and two *speakers* $\underline{i}$ and $\underline{k}$. Speaker $\underline{i}$ is a co-partisan to i (with $b_{\underline{i}} > 0$), and speaker $\underline{k}$ is a counter-partisan to i (with $b_{\underline{k}} < 0$).

The modified game consists of two stages. In the rationalization stage (stage 0), all players choose a level of realism σ and an action a, as in the previous sections. We assume that $a_{\underline{i}}$ and $a_{\underline{k}}$ are publicly observed, but that a_i and all internal narratives ($\hat{\xi}_i$, $\hat{\xi}_{\underline{i}}$ and $\hat{\xi}_{\underline{k}}$) are (by default) privately observed. This captures a setting where speakers' actions, e.g. their support for a political candidate, are known, but understanding the underlying justifications and arguments requires dedicated attention and communication channels like social media. The focus on uni-directional communication makes the analysis more tractable and facilitates the exposition of the mechanisms. It implies that i's behavior is not observed and, hence, that i is only concerned about her self-image, and not her social image.

In the communication stage (stage 1), a speaker has the choice between disclosing her retrieved narrative $\hat{\xi}$ to listener i ($m = \hat{\xi}$) or not communicating anything ($m = \varnothing$). Modeling communication as disclosure rather than cheap talk is in line with the assumption that $\tau < 1$: coming up with convincing rationalizations is hard and success is not guaranteed. To compare the effects of listening to various types of partisans, we assume that listener i is exogenously matched to speaker $\underline{i}$ or $\underline{k}$. The exogenous matching allows a clean comparison of exposure to different types of speakers, but we will also discuss the choice of a speaker by the listener. We will refer to communication between co-partisans i and $\underline{i}$ as an "echo-chamber" and between counter-partisans i and $\underline{k}$ as an "agora", after the classical Greek public square where the exchange of ideas took place. At the end of the game, the information set of listener i is $\{(a_i, \hat{\xi}_i), a_{\underline{i}}, a_{\underline{k}}, , m_z\}$ with $z \in \{\underline{i}, \underline{k}\}$.

We allow for listener's naiveté about speakers' narratives. Analogous to the framework of the previous section, where individuals are naive about the origin of their own internal narratives, agent i interprets $\underline{i}$'s narrative $\hat{\xi}_{\underline{i}}$ (if received) under the assumption that $\underline{i}$'s probability of successful rationalization (if attempted) equals $(1 - \chi_i^{\underline{i}})\tau_{\underline{i}}$. We define $\chi_i^{\underline{k}}$ separately to allow i to potentially apply different degrees of naiveté to the narratives of the two speakers. Unless otherwise stated, our results are true under standard Bayesian rationality at the communication stage, i.e. $\chi_i^{\underline{i}} = \chi_i^{\underline{k}} = 0$.

4.2 Equilibrium analysis

Communication of narratives can only be consequential to i's beliefs if all three agents have played their convenient action, as all other cases reveal the original signal with certainty. For simplicity, we assume that it is common knowledge that both speakers rationalize with probability one ($\sigma^{\underline{i}} = \sigma^{\underline{k}} = 0$), but qualitatively similar conclusions will obtain as long as $\sigma^{\underline{i}}, \sigma^{\underline{k}} < 1$. Thus, following our analysis in the previous section, we implicitly assume that $b_{\underline{i}}$ and $|b_{\underline{k}}| = -b_{\underline{k}}$ are (known to be) sufficiently large to produce rationalizations among speakers. We remain more flexible about i's strategy: we assume $\sigma^i < 1$, and make σ^i endogenous in Proposition 6.

Disclosure and unraveling. We first ask whether the speakers want to share their narrative ex-post. Key to the effects of communication is that a speaker who can communicate a congruent narrative (i.e. a strong argument) for her positions will improve the listener's opinion of her, as showing that she is not dumbfounded rules out a failed rationalization. Thus, speakers with a congruent narrative want to share it ex-post, given that their only incentive is to improve their reputation in the eye of the listener. In equilibrium, a lack of sharing therefore implies dumbfoundedness and identifies the speaker as a sophist. In

turn, this makes speakers without a congruent narrative indifferent between sharing or not. Thus, we obtain the familiar unraveling result from Milgrom (1981), whereby narratives are revealed in all equilibria. In addition, extending on the intuition from Proposition 4, shifting i's belief about ξ affects i's entire belief system, including i's self-image and opinion of other players. This is formalized in Lemma 3, which shows how the revelation of convenient narratives affects i's average posterior beliefs.[11]

Lemma 3 (Unraveling)**.** *If $a_{\underline{i}} = 1$ and $a_{\underline{k}} = 0$, there exists a unique (generic) equilibrium of the communication game where the speakers' narratives are fully revealed. In addition, the effect of a speaker communicating a congruent narrative (relative to a non-congruent narrative) is as follows:*

1. *it improves the speaker's reputation towards listener i;*
2. *a congruent narrative ($\hat{\xi}_{\underline{i}} = \xi_H$) shared by the co-partisan speaker $\underline{i}$ increases i's confidence in her preferred state ($\xi = \xi_H$), in her own morality ($\theta(i) = R$), in the co-partisan speaker's morality ($\theta(\underline{i}) = R$), and reduces i's confidence in the counter-partisan speaker's morality ($\theta(\underline{k}) = R$);*
3. *a congruent narrative ($\hat{\xi}_{\underline{k}} = \xi_L$) shared by the counter-partisan speaker $\underline{k}$ has the opposite effects.*

Lemma 3 clarifies why policy debates have implications for individuals' moral identity, even when these debates concern factual elements (e.g., the impact of policy decisions). Upon observing partisan disagreement ($a_i = 1, a_{\underline{i}} = 1, a_{\underline{k}} = 0$), i infers that one side of the divide has engaged in rationalization, whereas the other has maintained intellectual integrity. Any argument offered in favor of i's preferred policy therefore has the dual effect of improving her moral status and downgrading her opponent's moral status. The zero-sum nature of the allocation of esteem across the political spectrum also explains why partisans have incentives to disparage the other side and cast doubt on its morality. This happens in our model purely as a byproduct of reputation concerns, even though agents have no intrinsic preferences over others' reputations.

[11] If i (privately) played $a_i = 0$ at the rationalization stage or attempted to self-deceive but found herself morally dumbfounded, then communication is inconsequential. The average effect uncovered by Lemma 3 captures i's updating in the case where $a_i = 1$ and $\hat{\xi}_i = \xi_H$, which happens with positive probability conditional on $a_{\underline{i}} = 1$ and $a_{\underline{k}} = 0$ (both from the perspective of an unbiased observer and from that of the two speakers).

Agoras and echo chambers. To study and compare the average effect of communication by either co-partisan or counter-partisan speakers we now take an ex-ante perspective.[12] To compute i's average posterior beliefs with communication, we weigh the effects of the disclosure of congruent narratives and dumbfounded narratives on posterior beliefs with their (objective) likelihood. We fix i's probability of denial to some value σ^i, and contrast situations with and without communication. This case covers the situation where communication is not anticipated by i, or where i's core motive b_i is strong enough to imply $\sigma^i = 0$ with or without communication. Later, we will endogenize the effect of narrative transmission on i's behavior. Here, we obtain the following result.

Proposition 5. *Assume $\chi_i > 0$ (i is somewhat naive about her own rationalizations). Fix i's probability of denial to some $\sigma^i < 1$ and consider two forms of communication:*

1. **Echo chamber** *(communication from $\underline{i}$ to i): there exists a function $\iota(\tau_{\underline{i}}, \chi_i^{\underline{i}})$ increasing in both arguments with $\iota(1,0) < 0 < \iota(1,1)$ such that,*
   - *(a) if $\iota(\tau_{\underline{i}}, \chi_i^{\underline{i}}) < 0$, i's posterior beliefs are less biased towards the convenient state when i learns $\hat{\xi}_{\underline{i}}$ than without communication.*
   - *(b) if $\iota(\tau_{\underline{i}}, \chi_i^{\underline{i}}) > 0$, i's posterior beliefs are more biased towards the convenient state when i learns $\hat{\xi}_{\underline{i}}$ than without communication.*
2. **Agora** *(communication from $\underline{k}$ to i): for all parameter values, i's posterior beliefs are less biased when i learns $\hat{\xi}_{\underline{k}}$ than without communication.*

The first part of Proposition 5 shows that the effect of echo chambers on posterior beliefs is ambiguous. There are two countervailing effects. A *biasing effect* arises as listeners who are naive about the convenient narratives from co-partisans ($\chi_i^{\underline{i}} > 0$) overestimate their informativeness. A *debiasing effect* arises because observing a dumbfounded in-group member is highly damaging to the listener's self-image, as it reveals the inconvenient state to be true. This effect is moderated by χ_i and $\tau_{\underline{i}}$. If $\chi_i > 0$, i.e. the listener is somewhat naive about her own rationalizations in the first stage, she underestimates the frequency of dumbfoundedness, and will on average be disappointed by the quality of her co-partisan's arguments. However, the frequency of dumbfoundedness will be low as long as speaker is skilled at constructing convenient narratives (high $\tau_{\underline{i}}$).

Thus, if $\chi_i^{\underline{i}}$ and $\tau_{\underline{i}}$ are high, the first effect dominates, and echo chambers exacerbate belief biases about the original state, the level of self-righteousness, and mistrust about

[12] One might also consider a setting where the listener receives messages from both partisans. Here, the only situation with ex-post uncertainty is the one where both have a justifying narrative. As Proposition 5 suggests, the net effect of narrative communication in that case depends on the relative naiveté parameters $\chi_i^{\underline{i}}$ and $\chi_i^{\underline{k}}$.

the morality of the outgroup. Otherwise, and notably if $\chi_i^i = 0$ (unbiased social learning), echo chambers counteract the formation of self-serving beliefs. Note that a listener's strong reaction to encountering dumbfoundedness on behalf of a speaker is based on the assumption that speakers cannot use weak rationalizations to cover up their dumbfoundedness. In line with this, Mercier (2020) provides a wealth of evidence that people are rather apt at distinguishing between convincing and unconvincing evidence and narratives.

The second part of Proposition 5 shows that the agora — the communication between counterpartisans — always debiases self-serving beliefs. A naive individual overestimates the probability of the convenient state of the world, and hence is positively surprised by the average quality of counter-partisan arguments, i.e. their lack of dumbfoundedness. This leads her to revise her beliefs (about ξ, herself and the outgroup) towards a more accurate perception.

Self-selection into echo chambers. Proposition 5 shows that listening to out-group members reduces polarization. This raises the question of whether individuals will seek out these interactions by themselves. We can answer this question by comparing the expected utility (that is, the expected self-image) following communication by either kind of speaker.[13] We consider again a case with fixed σ^i. We then use the fact that, conditional on playing the convenient action, the utility of the strategic type is an increasing function of a belief in the convenient state, to extend Proposition 5 into the following Corollary. We focus on a case where speakers' abilities to rationalize are equal (for symmetry) and high, as otherwise the comparison is ambiguous (and i gets nearly equal expected utility from two unskilled speakers).

Corollary 1. *If both speakers are (equally) skilled at rationalizations ($\tau_{\underline{i}} = \tau_{\underline{k}} \geq \tau^*$ for some threshold τ^*), then listener i's ex-ante expected utility is higher in the echo chamber (listening to $\underline{i}$) than on the agora (listening to $\underline{k}$).*

The link between posterior beliefs and utility means that individuals prefer communication inside echo chambers, where convenient beliefs are less likely to be threatened.[14]

[13] Note that we do not explicitly model the choice of speaker (echo chamber vs. agora). This would lead to more complexities, as such a choice may itself become a signal of the type of the agent. However, our result is reminiscent of that in Grossman and van der Weele (2017), who use a signaling game to explicitly model the choice to avoid inconvenient information.

[14] The expected utilities in Corollary 1 are computed using the (unbiased) ex-ante beliefs. This speaks to situations in which agents can select a speaker before forming their own rationalizations. However, one can show that the result extends to situations where the expected utility is computed at the interim stage (after rationalization) from the perspective of an agent with (possibly) biased beliefs. Our results thus have two interpretations, one in which the expected utility corresponds to the actual, experienced utility (on average) in these interactions; the other is the individual's perceived expected utility, which might differ from the actual one if the agent has already formed wrong beliefs.

Anticipatory rationalizations. Our results in this section thus far are based on the assumption of a fixed, exogenous level of realism σ^i. However, anticipating communication with co- or counter-partisan may well affect incentives to engage in rationalizations in the first place. Our next result shows that the main qualitative insights obtained in this section survive if we endogenize σ^i as a function of the identity of the speaker in the second stage. We start from an equilibrium without communication that features $\sigma^i < 1$. We also assume sophistication about future naiveté, as in Section 3.[15]

Proposition 6. *Suppose that the equilibrium without communication features $\sigma^i < 1$. Then communication has the following impact:*

1. **Echo chamber** *(communication from $\underline{i}$ to i): there exists a function $\upsilon(\tau_{\underline{i}}, \chi_i^{\underline{i}})$ increasing in both arguments with $\upsilon(1,0) < 0 < \upsilon(1,1)$ such that,*

   (a) if $\upsilon(\tau_{\underline{i}}, \chi_i^{\underline{i}}) < 0$, σ^i is higher compared to a situation without communication.

   (b) if $\upsilon(\tau_{\underline{i}}, \chi_i^{\underline{i}}) > 0$, σ^i is lower compared to a situation without communication.

2. **Agora** *(communication from $\underline{k}$ to i): for all parameter values, σ^i is higher compared to a situation without communication.*

Proposition 6 proves that interactions with the outgroup spur more realism ex-ante, while echo chambers exacerbate self-deception under the same conditions that promote biases in posterior beliefs (large $\tau_{\underline{i}}$, large $\chi_i^{\underline{i}}$). The intuition behind this result and its relation to Proposition 5 is straightforward. Communication affects listener i only through self-image, which is itself an increasing function of the belief that the true state is convenient. Thus, the marginal return to rationalization depends on the likelihood that i's convenient beliefs survive the communication phase, creating a feedback loop from expected posterior beliefs, as specified in Proposition 5, to the level of realism in the rationalization stage.

Proposition 6 shows how initial rationalizations may depend on the anticipation of having one's beliefs validated by a speaker who is motivated and able enough to come up with a strong congruent narrative. Through this channel, politicians may influence voters' beliefs not merely by being persuasive, but by being a reliable source of cover that encourages rationalizations.

[15] In other words, we assume that the individual anticipates her future naiveté parameters ($\chi_i^{\underline{i}}$ and $\chi_i^{\underline{k}}$). This assumption is crucial, otherwise the player would not expect interactions to have a systematic effect on her self-image. We propose that the assumption of sophistication about future naiveté should not be taken too literally; instead, an individual with past experiences of encounters with both groups might have learned to associate co-partisan interactions with more pleasant affective states than cross-partisan conversations. She can then use that knowledge to predict her experience in future similar circumstances. Corollary 1 clarifies that, on average, individuals do have better hedonic experiences (as measured by their self-image) in echo chambers than on agoras.

5 Empirical evidence on social influence and polarization

In this section, we ask how the main predictions of our model help organize and interpret empirical evidence on the nature of social influence, groupthink, news consumption and political conversations.

5.1 Social influence and complementarity

Our model generally predicts conformity of cognition in like-minded groups. This is in line with a large literature, showing that people conform their opinions to those of an (in-group) audience (see Golman et al., 2016, for a review). More specifically, the model predicts that rationalizations and narratives are complements via both the self-image and a social image channel, which we discuss in turn.

Self-image channel. The self-image channel predicts that the prospect of observing a like-minded audience increases self-serving beliefs, as observing congruent behavior reinforces one's own rationalizations. At the same time, "deviations" from in-group members are potentially very damaging to in-group beliefs and self-image. Given the shared values and interests, one cannot explain such deviations away as self-serving delusions, as one might do with similar behavior from outgroup members.

This observation can explain why deviations by in-group members often attract strong reactions; cults typically punish defectors severely, and some religious groups even impose death penalties on apostates. Social psychologists have demonstrated such a "Black sheep effect" in the laboratory: judgments about in-group members are more extreme (both in a positive and negative dimension) than those about similarly behaved outgroup members (Marques et al., 1988), and the derogation of in-group deviations serves an identity-protective function (Marques and Paez, 1994; Lewis and Sherman, 2010).

Furthermore, according to our model, the degree to which outgroup members threaten self-image depends on their perceived commitment to realism. This explains an eagerness to expose self-serving reasoning and hypocrisy in outgroup members. Wolsky (2022) shows that Republicans indeed feel more positive about their own party after being informed about a hypocritical versus a non-hypocritical outgroup transgression.

Even though psychologists agree that self-esteem matters in social influence (Pool et al., 1998), direct and controlled evidence of the self-image channel remains scarce; a clean demonstration requires that decision makers can observe others, but are not observed themselves. We would therefore encourage future experiments that follow this structure.

The social image channel. The social image channel predicts that being observed by a critical audience induces more critical and realistic thinking. A literature in psychology on accountability investigates such audience effects on cognitive effort. Here "accountability" means that one has to justify one's views to an audience, and critical thinking is measured by the thoroughness and nuance of the justifications. In a review of both lab and field experiments, Lerner and Tetlock (1999) find that when the speaker knows the audience's opinion, they tend to provide low-quality, conformist justifications. Reasoning efforts are highest when the audience's views are unknown. Furthermore, effortful thinking is activated when the audience is known to be more informed and interested in accuracy. Beliefs about the audience's attitudes also affect how people express themselves politically. Bursztyn et al. (2020) shows how the rising popularity of Donald Trump reduced stigmatization and increased expression of xenophobic views.

5.2 Polarization and echo chambers

One strength of our approach is that it cannot only explain "belief consonance" in groups (Golman et al., 2016), but also polarization between different groups. In addition, it can capture both ideological and affective polarization. Empirically, polarization has increased in many countries. In particular, the U.S. has seen a stark rise in affective polarization (Gentzkow, 2016; Iyengar et al., 2019). The typical explanation is that social identification with partisan groups has increased (Iyengar et al., 2012, 2019). Our model shows that affective polarization may arise even without explicitly assuming in-group favoritism or out-group animus. Instead, it places the origins for polarization in the selection of self-serving narratives, and naiveté about the selection process. This implies that increasing the options to select one's own news sources online, for instance through social media, combined with naiveté about their content, generates heightened skepticism about the reasoning capabilities of the other side. We now discuss this evidence in more detail.

The effect of echo chambers. Our model predicts an ambiguous effect of echo chambers on polarization, that is more likely to be positive when speakers are skilled, i.e. $\tau_{\underline{i}}$ is high. This ambiguous prediction is consistent with mixed evidence on the effects of social media use on political attitudes. When researchers disconnected Facebook users prior to the 2020 US presidential election, this separation of users from their echo chamber or bubble had no effect on their affective and issue polarization (Allcott et al., 2024). In a similar vein, there is no evidence that Facebook users' exposure to like-minded sources (Nyhan et al., 2023) or exposure to shared content (Guess et al., 2023) affects polarization in attitudes and opinions.

At the same time, Rathje et al. (2024) find that unfollowing hyperpartisan influencers

on Twitter leads to a large and durable decrease in affective polarization. Similarly, Müller and Schwarz (2023) show a causal impact of Donald Trump's tweets about Muslims on xenophobic tweets by his followers, as well as hate crimes. In line with our model, these papers suggest that congruent narratives polarize opinions, attitudes and actions precisely when they are delivered by influential and persuasive figures. Our model then also explains the less obvious fact that overtly biased, but skilled political commentators are paid attention to, are highly valued by their audience, and, hence, can extract large rents. The reason for their success is that they provide their followers with valued boosts in moral self-esteem.

The effect of cross-partisan contact. Our second key prediction is that the agora, or cross-partisan contact, decreases polarization. In line with this, Levy (2021) finds that exposing Twitter users to counter-attitudinal news reduces their affective polarization. Similarly, Broockman and Kalla (2025) find that paying Fox News viewers to watch CNN for a month moderates their attitudes. Likewise, in the setting of political conversations, forced interactions between counter-partisans have generally been found to decrease both affective and factual polarization (Santoro and Broockman, 2022; Blattner and Koenen, 2023; Fang et al., 2023; Hobolt et al., 2024; Braghieri et al., 2024a).

Several authors in political science have hypothesized that confirmation bias and attachment to valued beliefs make people invulnerable to the debiasing effects of counter-attitudinal encounters and content (see Levendusky, 2013; and Coppock, 2023 for reviews of these arguments). Instead, consistent with the above evidence for the depolarizing effect of cross-partisan contact, our model supposes that agents' beliefs are affected by plausible narratives even if they are delivered by the other side.

Self-selection into echo chambers. Finally, our model predicts a preference for co-partisan contact and content over counter-partisan contact and content. Our model thus provides a cognitive foundation for the demand for slanted news (Mullainathan and Shleifer, 2005). Indeed, the fact that online news consumption is heavily slanted (Nyhan et al., 2023), especially at the article level (Braghieri et al., 2024b), is highly suggestive of this preference for echo chambers. The preference is also present in the setting of political conversations, where experimental participants prefer to have a conversation with a co- rather than counter-partisan (Braghieri et al., 2024a). Similarly, people are willing to pay not to listen to a counter-partisan (Frimer et al., 2017; Bauer et al., 2023), while the demand for fact checks is lower for ideologically aligned news sources (Chopra et al., 2022).

Since our model predicts that people misperceive the character of their counter-partisans, it is tempting to conclude that information or positive experiences with the other side could

persistently break people out of echo chambers. Unfortunately, our model tells a more pessimistic story, supported by the data. Agents stay away from cross-partisan contact to protect their ego, and it is precisely the compelling and reasonable interactions with outgroup members that produce the highest threat.

6 Conclusion

Our model studies the trade-off between core motives and image concerns over being perceived as reasonable agents. The equilibria capture key aspects of political cognition, including conformism within groups and ideological and affective polarization across groups. Communication within groups, especially by skilled speakers, can further bias beliefs and improve individual welfare, as the sharing of convenient narratives improves group members' self- and social image. By contrast, communication between groups has a zero-sum nature, as sharing narratives that are convenient for one group undermines the other group's self-image. As a result, our model generates self-selection into echo chambers and backlash against policies that force cross-partisan contact.

We show that signaling of cognitive traits like reasonableness can reproduce key stylized facts of not just polarization, but also group identity. In particular, the distortion of beliefs about the nature of reality affects the agent's entire belief system, including beliefs about herself and the morality of her in- and outgroup. Without assuming explicit animus towards particular people or groups, we thus provide a micro-foundation for models of group identity like Akerlof and Kranton (2000). This opens up (cultural) applications for economic analysis, like debates about 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 6th, 2020.

Our model helps explain a range of moral behaviors in groups, including the tendency of people to veer into groupthink. The problem is not just that people bend their beliefs to serve their passions. The very human quest for esteem implies that such tendencies are mutually reinforcing within groups: every time a group member rationalizes evidence away or avoids informative outgroup members, some information dissipates from the system, further weakening constraints on rationalizations and behavior. The question of how important these effects are in practice seems to us like an important area of future empirical investigation. Another urgent question is how to stem such a spiral. Here, our model may facilitate exploring the promise and pitfalls of regulating speech.

References

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.

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.

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 Decomposition of Political Polarization," 2023. Working Paper.

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), 141–164.

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

Bicchieri, Cristina, *The Grammar of Society: The Nature and Dynamics of Social Norms*, Cambridge University Press, 2006.

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).

Brown, Jacob, Enrico Cantoni, Ryan Enos, Vincent Pons, and Emilie Sartre, "The Increase in Partisan Segregation in the United States," 2024. Working Paper.

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.

Čavojová, Vladimíra, Jakub Šrol, and Magdalena Adamus, "My Point is Valid, Yours is not: Myside Bias in Reasoning about Abortion," *Journal of Cognitive Psychology*, 2018, *30* (7), 656–669.

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

Cook, Edgar Valentine, "Talking Past Each Other: The Diverging Moral Foundations of the Contemporary Gun Debate," *Berkeley Undergraduate Journal*, 2015, *28* (1).

Coppock, Alexander, *Persuasion in Parallel: How Information Changes Minds about Politics*, University of Chicago Press, 2023.

Dana, Jason, 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.

Ditto, Peter H. and David F. Lopez, "Motivated Skepticism: Use of Differential Decision Criteria for Preferred and Nonpreferred Conclusions," *Journal of Personality and Social Psychology*, 1992, *63* (4), 568.

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.

Fang, Ximeng, Sven Heuser, and Lasse S. Stötzer, "How In-Person Conversations Shape Political Polarization: Quasi-Experimental Evidence from a Nationwide Initiative," 2023. Working paper.

Flaxman, Seth, Sharad Goel, and Justin M. Rao, "Filter Bubbles, Echo Chambers, and Online News Consumption," *Public Opinion Quarterly*, 2016, *80*, 298–320.

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.

Geller, Rebecca C, Jamie D Gravell, Amy Richardson, and Stacy Ann Strang, ""My Thinking has Changed but Beliefs Have not": Motivated Reasoning in Learning to Teach Abortion," *Theory & Research in Social Education*, 2025, *53* (1), 1–26.

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

_ **and Jesse M. Shapiro**, "Ideological Segregation Online and Offline," *Quarterly Journal of Economics*, 11 2011, *126* (4), 1799–1839.

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

_ **, George Loewenstein, Karl Ove Moene, and Luca Zarri**, "The Preference for Belief Consonance," *Journal of Economic Perspectives*, September 2016, *30* (3), 165–188.

González-Bailón, Sandra, David Lazer, Pablo Barberá, Meiqing Zhang, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Deen Freelon, Matthew Gentzkow, Andrew M. Guess, Shanto Iyengar, Young Mie Kim, Neil Malhotra, Devra Moehler, Brendan Nyhan, Jennifer Pan, Carlos Velasco Rivera, Jaime Settle, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker, "Asymmetric Ideological Segregation in Exposure to Political News on Facebook," *Science*, 2023, *381* (6656), 392–398.

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, Benjamin Lyons, Brendan Nyhan, and Jason Reifler, "Avoiding the Echo Chamber about Echo Chambers: Why Selective Exposure to Like-Minded Political News is Less Prevalent than you Think," 2018. Working paper.

Guess, Andrew M., "Almost Everything in Moderation," *American Journal of Political Science*, 2021, *4* (65), 1007–1022.

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*, 2024. rdae058.

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 Jesse Graham**, "When Morality Opposes Justice: Conservatives have Moral Intuitions that Liberals may not Recognize," *Social Justice Research*, 2007, *20* (1), 98–116.

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.

Hoewe, Jennifer, Elliot Panek, Cynthia Peacock, Lindsey Sherrill, and Shannon Wheeler, "Using Moral Foundations to Assess Stereotypes: Americans' Perceptions of Immigrants and Refugees," *Journal of Immigrant & Refugee Studies*, 2022, *20* (4), 501–518.

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, David A Hoffman, and Donald Braman, "Whose Eyes Are You Going to Believe-Scott v. Harris and the Perils of Cognitive Illiberalism," *Harv. L. Rev.*, 2008, *122*, 837.

_ **, Donald Braman, John Gastil, Paul Slovic, and CK Mertz**, "Culture and Identity-Protective Cognition: Explaining the White-Male Effect in Risk Perception," *Journal of Empirical Legal Studies*, 2007, *4* (3), 465–505.

_ **, Ellen Peters, Erica Dawson, and Paul Slovic**, "Motivated Numeracy and Enlightened Self-Government," September 2013. Working paper.

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.

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

Levendusky, Matthew S., "Why do Partisan Media Polarize Viewers?," *American Journal of Political Science*, 2013, *57* (3), 611–623.

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.

Lewis, Amy C and Steven J Sherman, "Perceived Entitativity and the Black-Sheep Effect: When Yill we Denigrate Negative Ingroup Members?," *The Journal of social psychology*, 2010, *150* (2), 211–225.

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

Lockhart, Christopher, Carol HJ Lee, Chris G Sibley, and Danny Osborne, "The Sanctity of Life: The Role of Purity in Attitudes towards Abortion and Euthanasia," *International Journal of Psychology*, 2023, *58* (1), 16–29.

Marques, José M, Vincent Y Yzerbyt, and Jacques-Philippe Leyens, "The "Black Sheep Effect": Extremity of Judgments towards Ingroup Members as a Function of Group Identification," *European Journal of Social Psychology*, 1988, *18* (1), 1–16.

Marques, José M. and Dario Paez, "The 'Black Sheep Effect': Social Categorization, Rejection of Ingroup Deviates, and Perception of Group Variability," *European Review of Social Psychology*, 1994, *5* (1), 37–68.

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, *Not Born Yesterday: The Science of who we Trust and what we Believe*, Princeton University Press, 2020.

_ **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.

Milgrom, Paul R, "Good News and Bad News: Representation Theorems and Applications," *The Bell Journal of Economics*, 1981, pp. 380–391.

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

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.

Nath, Leda, Nicholas Pedriana, Christopher Gifford, James W McAuley, and Marta Fülöp, "Examining Moral Foundations Theory through Immigration Attitudes," *Athens Journal of Social Sciences*, 2022, *9* (1), 9–30.

Nelson, Jacob L. and James G. Webster, "The Myth of Partisan Selective Exposure: A Portrait of the Online Political News Audience," *Social Media + Society*, 2017, *3* (3).

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.

Oreskes, Naomi and Erik M Conway, *The Big Myth: How American Business Taught us to Loathe Government and Love the Free Market*, Bloomsbury Publishing USA, 2023.

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

Pool, Gregory J, Wendy Wood, and Kira Leck, "The Self-esteem Motive in Social Influence: Agreement with Valued Majorities and Disagreement with Derogated Minorities.," *Journal of Personality and Social Psychology*, 1998, *75* (4), 967.

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.

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.

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

Sunstein, Cass R, *Republic.com*, Princeton University press, 2001.

_ , *Going to Extremes: How Like Minds Unite and Divide*, Oxford University Press, 2009.

_ , *# Republic: Divided Democracy in the Age of Social Media*, Princeton University press, 2017.

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.

Trivers, Robert, *The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life*, Basic Books, 2011.

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

Waswo, Richard, "The Formation of Natural Law to Justify Colonialism, 1539-1689," *New Literary History*, 1996, *27* (4), 743–759.

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

Wolsky, Adam D, "Scandal, Hypocrisy, and Resignation: How Partisanship Shapes Evaluations of Politicians' Transgressions," *Journal of Experimental Political Science*, 2022, *9* (1), 74–87.

Woolley, Kaitlin and Jane L. Risen, "Hiding from the Truth: When and How Cover Enables Information Avoidance," *Journal of Consumer Research*, 2021, *47* (5), 675–697.

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

Appendix A: Refinement definitions

The first refinement pertains to the inferences of an agent who is expected to be fully realistic, but deviates and finds himself dumbfounded. While date 1-inferences by a morally dumbfounded player are given by Bayes' rule when the player denies a signal with a positive probability in equilibrium, they must be specified as out-of-equilibrium beliefs whenever the agent is expected to stay realistic. We extend the fact that a morally dumbfounded player infers being strategic with probability one, which is true on the path, to these out-of-equilibrium situations, in line with the idea that reasonable types are flawless moral reasoners. We add that the player's inferences about ξ and about the type of the partner must be consistent with this inference. Formally:

Refinement 1 (Reasonable people never rationalize)**.** *If $(a_i, \hat{\xi}_i)$ is an off-path non-congruent action-recollection pair, then for every information set $\mathcal{I} \supseteq \{a_i, \hat{\xi}_i\}$ and any a_j:*

i) $p_i[\xi = \hat{\xi}_i \mid \mathcal{I}] = 1$;

ii) $p_i[\theta(i) = R \mid \mathcal{I}] = 0$;

iii) $p_i[\theta(j) = R \mid (a_i, \hat{\xi}_i), a_j] = p_i[\theta(j) = R \mid \xi = \hat{\xi}_i, a_j]$.

Player j's beliefs satisfy the symmetric conditions.

To illustrate, consider a situation where i's equilibrium strategy is to stay realistic, but her date 1-information set is the off-path observation $[(a_i = 1, \hat{\xi}_i = \xi_L), a_j = 1]$. The three bullet points of Refinement 1 imply that i infers: *i)* $\xi = \xi_L$; *ii)* i is strategic; *iii)* j is strategic.

The next two refinements relate to interactions between two counter-partisans. If $\sigma^i = \sigma^j = 1$, i.e. both are realistic, they must play the same action. Thus, the information set $(a_i = 1, a_j = 0)$ is off the equilibrium path, and we cannot apply Bayes' rule to pin down posterior beliefs. We discipline off-path beliefs using two additional refinements.

Refinement 2 (Intra-personal consistency)**.** *If $\sigma^i = \sigma^j = 1$, then*

$$p_i[\theta(i) = R \mid (a_i = 1, \hat{\xi}_i = \xi_H), a_j = 0] + p_i[\theta(j) = R \mid (a_i = 1, \hat{\xi}_i = \xi_H), a_j = 0] = \rho,$$

and similarly for j's beliefs.

Refinement (2) imposes a consistency condition on i's beliefs about their own type and about j's type following the out-of-equilibrium observation $[(a_i = 1, \hat{\xi}_i = \xi_H), a_j = 0]$. This condition is implied by Bayes' rule on every equilibrium path: we extend it to rule out pathological out-of-equilibrium beliefs where a player would, for instance, believe that i and j are both reasonable with probability one (or with probability zero).

To explain the meaning of the next (and final) refinement, assume temporarily that $\chi_i = \chi_j = 0$ and $\tau_i = \tau_j = 1$: both players are sophisticated, and narratives are uninformative. Hence, the observation sets $[(a_i = 1, \hat{\xi}_i = \xi_H), a_j = 0]$ for i, and $[(a_j = 0, \hat{\xi}_j = \xi_L), a_i = 1]$ for j, are informationally equivalent. A standard refinement in signaling games is that, in such a situation players form the same out-of-equilibrium beliefs. Relaxing the hypotheses $\chi_i = \chi_j = 0$ and $\tau_i = \tau_j = 1$, we need to adapt this condition for two reasons: (i) players don't have the same information set, as narratives convey some information; (ii) naive individuals tend to form self-serving (and hence divergent) beliefs. We modify the condition as follows:

Refinement 3 (Inter-personal consistency). *If* $\sigma^i = \sigma^j = 1$, *then*

$$\frac{p_j[\theta(i) = R \mid (a_j = 0, \hat{\xi}_j = \xi_L), a_i = 1]}{p_j[\theta(j) = R \mid (a_j = 0, \hat{\xi}_j = \xi_L), a_i = 1]} = \tau_i(1-\chi_i)\tau_j(1-\chi_j)\frac{p_i[\theta(i) = R \mid (a_i = 1, \hat{\xi}_i = \xi_H), a_j = 0]}{p_i[\theta(j) = R \mid (a_i = 1, \hat{\xi}_i = \xi_H), a_j = 0]}.$$

To interpret this condition, note that the likelihood ratios on both sides measure how each player perceives i's morality relative to j's. Thus, these expressions pin down how players "allocate" esteem to i and j in case of disagreement. The equality is implied by Bayes' rule on the path. By extending it to off-path scenarios, Refinement (3) imposes that the level of self-servingness in ex-post attributions must be commensurate with players' degree of naiveté and with the informativeness of congruent narratives. It rules out situations where, for instance, two sophisticated players ($\chi_i = \chi_j = 0$), who always form the same beliefs on path, make very self-serving out-of-equilibrium inferences.

Appendix B: Proofs

B.1 Proof of Proposition 1

We start by proving formally that noncongruent pairs are never played in equilibrium.

Lemma 4. *In any equilibrium of the game, the probability of transmission of a noncongruent action-justification pair is 0.*

Proof. Let us show that transmitting a noncongruent action-justification pair is a strictly dominated strategy. Consider for instance the pair $(a = 1, \tilde{\xi} = \xi_L)$, the case $(a = 1, \tilde{\xi} = \xi_L)$ being symmetric. We distinguish two cases as a function of the value of ξ.

Suppose first that $\xi = \xi_L$. Then an agent playing $(a = 1, \tilde{\xi} = \xi_L)$ will retrieve $\hat{\xi} = \xi_L$ with probability one and reap utility b, as the recollection $(a = 1, \hat{\xi} = \xi_L)$ reveals a strategic type for sure, either by Bayes' rule or due to Refinement (1). By playing $(a = 1, \tilde{\xi} = \xi_H)$, the agent would obtain the same material utility but an expected self-image equal to $\tau p[\theta = R \mid a = 1, \hat{\xi} = \xi_H]$, which is positive regardless of the equilibrium strategies.

If $\xi = \xi_H$, then $\hat{\xi} = \xi_H$ with probability $1 - \tau$ and $\hat{\xi} = \xi_L$ with probability τ, yielding an expected self-image equal to $(1 - \tau)p[\theta = R \mid a = 1, \hat{\xi} = \xi_H]$, whereas the expected self-image from playing $(a = 1, \tilde{\xi} = \xi_H)$ equals $p[\theta = R \mid a = 1, \hat{\xi} = \xi_H]$, which is strictly larger. □

From Lemma 4 we can represent an agent's strategy by a pair (σ_L, σ_H) as defined in the main text. We then prove that good news is always accepted.

Lemma 5. *Every equilibrium is such that* $\sigma_H = 1$.

Proof. Consider a candidate equilibrium strategy (σ_L, σ_H), and let

$$U_0(\sigma_H, \sigma_L) = \mu\frac{(1-\lambda)\rho}{(1-\lambda)\rho + (1-\rho)[\lambda\tau(1-\chi)(1-\sigma_H) + (1-\lambda)\sigma_L]} \tag{B.1}$$

be the agent's ex-post utility conditional on the action $a = 0$ and the recollection $\hat{\xi} = \xi_L$. The ex-post utility conditional on $(a = 0, \hat{\xi} = \xi_H)$ is 0.

Similarly, the ex-post utility conditional on $(a = 1, \hat{\xi} = \xi_H)$ is

$$U_1(\sigma_H, \sigma_L) = b + \mu\frac{\lambda\rho}{\lambda\rho + (1-\rho)[\lambda\sigma_H + (1-\lambda)\tau(1-\chi)(1-\sigma_L)]}, \tag{B.2}$$

and the ex-post utility conditional on $(a = 1, \hat{\xi} = \xi_L)$ is b.

The profile of strategies (σ_L, σ_H) is an equilibrium if and only if:

$$\begin{cases} \sigma_H \in \underset{\sigma}{\operatorname{argmax}}\, \sigma U_1(\sigma_H, \sigma_L) + (1-\sigma)\tau U_0(\sigma_H, \sigma_L) \\ \sigma_L \in \underset{\sigma}{\operatorname{argmax}}\, \sigma U_0(\sigma_H, \sigma_L) + (1-\sigma)[\tau U_1(\sigma_H, \sigma_L) + (1-\tau)b]. \end{cases} \tag{B.3}$$

Let us assume towards a contradiction that $\sigma_H < 1$. This implies that $U_1(\sigma_H, \sigma_L) \leq \tau U_0(\sigma_H, \sigma_L)$, otherwise by (B.3) $\sigma_H = 1$ would be the unique best-response. But then

$$U_0(\sigma_H, \sigma_L) > \tau U_0(\sigma_H, \sigma_L) \geq U_1(\sigma_H, \sigma_L) \geq \tau U_1(\sigma_H, \sigma_L) + (1-\tau)b,$$

where the last inequality comes from the fact that $U_1(\sigma_H, \sigma_L) \geq b$. Hence, we have $U_0(\sigma_H, \sigma_L) > \tau U_1(\sigma_H, \sigma_L) + (1-\tau)b$, which by (B.3) implies that the equilibrium features $\sigma_L = 1$.

But substituting $\sigma_L = 1$ into expressions (B.1) and (B.2) reveals that $U_1(\sigma_H, 1) > U_0(\sigma_H, 1)$ for any σ_H, thereby contradicting the condition $U_1(\sigma_H, \sigma_L) < \tau U_0(\sigma_H, \sigma_L)$. This proves Lemma 5. □

To conclude the proof of Proposition 1, note that the profile $(\sigma_L, \sigma_H = 1)$ constitutes an equilibrium if and only if one of the following conditions holds:

$$\begin{cases} \sigma_L = 1 \text{ and } b \leq \mu\rho(1-\tau) \\ \sigma_L \in (0,1) \text{ and } b = \mu\Big[\dfrac{\rho}{\rho + (1-\rho)\sigma_L} - \tau\dfrac{\lambda\rho}{\lambda + (1-\lambda)(1-\rho)\tau(1-\chi)(1-\sigma_L)}\Big] \\ \sigma_L = 0 \text{ and } b \geq \mu\Big[1 - \tau\dfrac{\lambda\rho}{\lambda + (1-\lambda)(1-\rho)\tau(1-\chi)}\Big]. \end{cases}$$

For all parameter values this set of conditions defines a unique value of σ_L, which is nonincreasing in b, τ and χ, and nondecreasing in μ. This proves Proposition 1.

B.2 Equilibrium analysis: preliminaries

In this section we establish the equilibrium conditions. We cover both the co-partisan and the counter-partisan cases, as we assume that $b_i > 0$ but make no assumption about the sign of b_j.

Let us write $\Sigma = (\sigma_H^i, \sigma_L^i, \sigma_H^j, \sigma_L^j)$ for the profile of equilibrium strategies, and let us define:

$$\begin{cases} \mathcal{R}^{ii}_{(a_i,\hat{\xi}_i),a_j}(\Sigma) \coloneqq p_i[\theta(i) = R \mid (a_i, \hat{\xi}_i), a_j], \\ \mathcal{R}^{ji}_{(a_j,\tilde{\xi}_j),a_i}(\Sigma) \coloneqq p_j[\theta(i) = R \mid (a_j, \tilde{\xi}_j), a_i] \end{cases}$$

under strategies Σ. The expression $\mathcal{R}^{ii}_{(a_i,\hat{\xi}_i),a_j}(\Sigma)$ represents i's posterior self-reputation conditional on the equilibrium strategies Σ and on the dataset $[(a_i, \tilde{\xi}_i), a_j]$, while $\mathcal{R}^{ji}_{(a_j,\tilde{\xi}_j),a_i}(\Sigma)$ represents i's posterior reputation vis-à-vis j conditional on Σ and on the dataset $[(a_j, \tilde{\xi}_j), a_i]$. These expressions are obtained by Bayes' rule whenever possible, and are defined as out-of-equilibrium beliefs otherwise. For future reference, we collect here the values for all possible scenarios.

$$
\begin{cases}
\mathcal{R}^{ii}_{(1,\xi_H),1}(\Sigma) = \dfrac{\lambda\rho[\rho+(1-\rho)\sigma_H^j]}{\lambda[\rho+(1-\rho)\sigma_H^i][\rho+(1-\rho)\sigma_H^j]+(1-\lambda)(1-\rho)^2\tau_i(1-\chi_i)(1-\sigma_L^i)(1-\sigma_L^j)}, \\
\mathcal{R}^{ii}_{(1,\xi_H),0}(\Sigma) = \dfrac{\lambda\rho(1-\sigma_H^j)}{\lambda[\rho+(1-\rho)\sigma_H^i](1-\sigma_H^j)+(1-\lambda)\tau_i(1-\chi_i)(1-\sigma_L^i)[\rho+(1-\rho)\sigma_L^j]} \\
\qquad\qquad \text{if } \min(\sigma_H^j,\sigma_L^i)<1, \\
\mathcal{R}^{ii}_{(0,\xi_L),1}(\Sigma) = \dfrac{(1-\lambda)\rho(1-\sigma_L^j)}{\lambda\tau_i(1-\chi_i)(1-\sigma_H^i)[\rho+(1-\rho)\sigma_H^j]+(1-\lambda)[\rho+(1-\rho)\sigma_L^i](1-\sigma_L^j)} \\
\qquad\qquad \text{if } \min(\sigma_L^j,\sigma_H^i)<1, \\
\mathcal{R}^{ii}_{(0,\xi_L),0}(\Sigma) = \dfrac{(1-\lambda)\rho[\rho+(1-\rho)\sigma_L^j]}{\lambda(1-\rho)^2\tau_i(1-\chi_i)(1-\sigma_H^i)(1-\sigma_H^j)+(1-\lambda)[\rho+(1-\rho)\sigma_L^i][\rho+(1-\rho)\sigma_L^j]}.
\end{cases}
$$

All other scenarios involve a non-congruent action-justification pair for i and hence a self-image equal to zero. Some algebra shows that, in a monotone equilibrium Σ where the two expressions in the middle are well-defined,

$$
\mathcal{R}^{ii}_{(1,\xi_H),1}(\Sigma) \geq \mathcal{R}^{ii}_{(1,\xi_H),0}(\Sigma) \text{ and } \mathcal{R}^{ii}_{(0,\xi_L),1}(\Sigma) \leq \mathcal{R}^{ii}_{(0,\xi_L),0}(\Sigma). \tag{B.4}
$$

In other words, fixing her recollection $\hat{\xi}_i$, i has better self-image ex post when her action is aligned with j's action.

We similarly collect the values of $\mathcal{R}^{ji}_{(a_j,\tilde{\xi}_j),a_i}(\Sigma)$ for all possible scenarios:

$$\begin{cases}
\mathcal{R}^{ji}_{(1,\xi_H),1}(\Sigma) = \dfrac{\lambda\rho[\rho+(1-\rho)\sigma_H^j]}{\lambda[\rho+(1-\rho)\sigma_H^i][\rho+(1-\rho)\sigma_H^j]+(1-\lambda)(1-\rho)^2\tau_j(1-\chi_j)(1-\sigma_L^i)(1-\sigma_L^j)}, \\
\mathcal{R}^{ji}_{(1,\xi_L),1}(\Sigma) = 0, \\
\mathcal{R}^{ji}_{(0,\xi_H),1}(\Sigma) = \dfrac{\rho}{\rho+(1-\rho)\sigma_H^i}, \\
\mathcal{R}^{ji}_{(0,\xi_L),1}(\Sigma) = \dfrac{\lambda\rho\tau_j(1-\chi_j)(1-\sigma_H^j)}{\lambda[\rho+(1-\rho)\sigma_H^i]\tau_j(1-\chi_j)(1-\sigma_H^j)+(1-\lambda)(1-\sigma_L^i)[\rho+(1-\rho)\sigma_L^j]}, \\
\qquad\qquad \text{if } \min(\sigma_H^j,\sigma_L^i)<1, \\
\mathcal{R}^{ji}_{(0,\xi_L),0}(\Sigma) = \dfrac{(1-\lambda)\rho[\rho+(1-\rho)\sigma_L^j]}{\lambda(1-\rho)^2\tau_j(1-\chi_j)(1-\sigma_H^i)(1-\sigma_H^j)+(1-\lambda)[\rho+(1-\rho)\sigma_L^i][\rho+(1-\rho)\sigma_L^j]}, \\
\mathcal{R}^{ji}_{(0,\xi_H),0}(\Sigma) = 0, \\
\mathcal{R}^{ji}_{(0,\xi_L),1}(\Sigma) = \dfrac{\rho}{\rho+(1-\rho)\sigma_L^i}, \\
\mathcal{R}^{ji}_{(0,\xi_H),1}(\Sigma) = \dfrac{(1-\lambda)\rho\tau_j(1-\chi_j)(1-\sigma_L^j)}{\lambda(1-\sigma_H^i)[\rho+(1-\rho)\sigma_H^j]+(1-\lambda)[\rho+(1-\rho)\sigma_L^i]\tau_j(1-\chi_j)(1-\sigma_L^j)} \\
\qquad\qquad \text{if } \min(\sigma_L^j,\sigma_H^i)<1.
\end{cases}$$

The expressions for $\mathcal{R}^{ji}_{(1,\xi_L),1}(\Sigma)$, $\mathcal{R}^{ji}_{(0,\xi_H),1}(\Sigma)$, $\mathcal{R}^{ji}_{(1,\xi_L),0}(\Sigma)$ and $\mathcal{R}^{ji}_{(0,\xi_H),0}(\Sigma)$ all use Refinement (1) in the case where j's noncongruent recollection is off-path: in all these cases player j assigns probability one to their recollection being truthful.

We note again (after some algebra), using the monotonicity of the equilibrium, that

$$\mathcal{R}^{ji}_{(1,\xi_H),1}(\Sigma) \geq \mathcal{R}^{ji}_{(0,\xi_L),1}(\Sigma) \text{ and } \mathcal{R}^{ji}_{(0,\xi_L),0}(\Sigma) \geq \mathcal{R}^{ji}_{(1,\xi_H),0}(\Sigma) \tag{B.5}$$

whenever all these expressions are well-defined.

We explain how to obtain the equilibrium expected payoffs. Upon seeing $\xi=\xi_H$, i expects j to play $a_j=1$ with probability $\rho+(1-\rho)\sigma_H$, and $a_j=0$ with probability $(1-\rho)(1-\sigma_H^j)$. In the latter case, j manages to self-deceive (and thus obtains the recollection $\tilde{\xi}_j=\xi_L$) with probability τ_j, and fails ($\tilde{\xi}_j=\xi_H$) otherwise. If i tries to self-deceive, i expects this to succeed with probability τ_i, which is a necessary condition for a positive ex-post self-reputation, and j's self-deception to succeed with probability τ_j (if attempted), which is a necessary condition for a positive ex-post reputation vis-à-vis j. All in all, i's expected payoff from playing $a_i=1$ equals

$$\begin{aligned}
U^{\xi_H}_{1,i}(\Sigma) = b_i &+ [\rho+(1-\rho)\sigma_H^j][\mu_i\mathcal{R}^{ii}_{(1,\xi_H),1}(\Sigma)+\nu_i\mathcal{R}^{ji}_{(1,\xi_H),1}(\Sigma)] \\
&+ (1-\rho)(1-\sigma_H^j)[\mu_i\mathcal{R}^{ii}_{(1,\xi_H),0}(\Sigma)+\nu_i\tau_j\mathcal{R}^{ji}_{(0,\xi_L),1}(\Sigma)+\nu_i(1-\tau_j)\mathcal{R}^{ji}_{(0,\xi_H),1}(\Sigma)].
\end{aligned} \tag{B.6}$$

while the expected payoff from playing $a_i = 0$ equals

$$\begin{aligned} U_{0,i}^{\xi_H}(\Sigma) =&[\rho + (1-\rho)\sigma_H^j][\mu_i\tau_i\mathcal{R}^{ii}_{(0,\xi_L),1}(\Sigma) + \nu_i\mathcal{R}^{ji}_{(1,\xi_H),0}(\Sigma)] \\ &+ (1-\rho)(1-\sigma_H^j)[\mu_i\tau_i\mathcal{R}^{ii}_{(0,\xi_L),0}(\Sigma) + \nu_i\tau_j\mathcal{R}^{ji}_{(0,\xi_L),0}](\Sigma). \end{aligned} \tag{B.7}$$

Similarly, the payoffs conditional on $\xi = \xi_L$ are

$$\begin{aligned} U_{1,i}^{\xi_L}(\Sigma) = b_i+&[\rho + (1-\rho)\sigma_L^j][\mu_i\tau_i\mathcal{R}^{ii}_{(1,\xi_H),0}(\Sigma) + \nu_i\mathcal{R}^{ji}_{(0,\xi_L),1}(\Sigma)] \\ &+ (1-\rho)(1-\sigma_L^j)[\mu_i\tau_i\mathcal{R}^{ii}_{(1,\xi_H),1}(\Sigma) + \nu_i\tau_j\mathcal{R}^{ji}_{(1,\xi_H),1}(\Sigma)]. \end{aligned} \tag{B.8}$$

and

$$\begin{aligned} U_{0,i}^{\xi_L}(\Sigma) =&[\rho + (1-\rho)\sigma_L^j][\mu_i\mathcal{R}^{ii}_{(0,\xi_L),0}(\Sigma) + \nu_i\mathcal{R}^{ji}_{(0,\xi_L),0}](\Sigma) \\ &+ (1-\rho)(1-\sigma_L^j)[\mu_i\mathcal{R}^{ii}_{(0,\xi_L),1}(\Sigma) + \nu_i\tau_j\mathcal{R}^{ji}_{(1,\xi_H),0}(\Sigma) + \nu_i(1-\tau_j)\mathcal{R}^{ji}_{(1,\xi_L),0}(\Sigma)]. \end{aligned} \tag{B.9}$$

We define $U_{1,j}^{\xi_H}(\Sigma), U_{0,j}^{\xi_H}(\Sigma), U_{1,j}^{\xi_L}(\Sigma)$ and $U_{0,j}^{\xi_L}(\Sigma)$ similarly. The profile of strategies Σ is an equilibrium if and only if the following system holds:

$$\begin{cases} \sigma_H^i \in \underset{\sigma}{\operatorname{argmax}}\, \sigma U_{1,i}^{\xi_H}(\Sigma) + (1-\sigma)U_{0,i}^{\xi_H}(\Sigma) \\ \sigma_L^i \in \underset{\sigma}{\operatorname{argmax}}\, \sigma U_{0,i}^{\xi_L}(\Sigma) + (1-\sigma)U_{1,i}^{\xi_L}(\Sigma) \\ \sigma_H^j \in \underset{\sigma}{\operatorname{argmax}}\, \sigma U_{1,j}^{\xi_H}(\Sigma) + (1-\sigma)U_{0,j}^{\xi_H}(\Sigma) \\ \sigma_L^j \in \underset{\sigma}{\operatorname{argmax}}\, \sigma U_{0,j}^{\xi_L}(\Sigma) + (1-\sigma)U_{1,j}^{\xi_L}(\Sigma). \end{cases}$$

B.3 Proof of Lemma 1

We proceed by contradiction and assume that $\sigma_H^i < 1$, which implies that $U_{1,i}^{\xi_H}(\Sigma) \leq U_{0,i}^{\xi_H}(\Sigma)$. We distinguish two cases as a function of the value of σ_L^i.

First case: $\sigma_L^i = 1$. We distinguish again two cases as a function of the value of σ_H^j. If $\sigma_H^j = 1$, then $U_{1,i}^{\xi_H}(\Sigma) = b_i + \mu_i\mathcal{R}^{ii}_{(1,\xi_H),1}(\Sigma) + \nu_i\mathcal{R}^{ji}_{(1,\xi_H),1}(\Sigma)$ and $U_{0,i}^{\xi_H}(\Sigma) = \mu_i\tau_i\mathcal{R}^{ii}_{(0,\xi_L),1}(\Sigma) + \nu_i\mathcal{R}^{ji}_{(1,\xi_H),0}(\Sigma)$. But substituting $\sigma_L^i = 1$ and $\sigma_H^j = 1$ into these expressions shows that $\mathcal{R}^{ii}_{(1,\xi_H),1}(\Sigma) \geq \rho, \mathcal{R}^{ji}_{(1,\xi_H),1}(\Sigma) \geq \rho, \mathcal{R}^{ii}_{(0,\xi_L),1}(\Sigma) \leq \rho, \mathcal{R}^{ji}_{(1,\xi_H),0}(\Sigma) \leq \rho$. Together with $b_i > 0$ this implies $U_{1,i}^{\xi_H}(\Sigma) > U_{0,i}^{H}(\Sigma)$, which contradicts $\sigma_H^i = 1$.

Suppose now that $\sigma_H^j < 1$. Then all posterior beliefs can be obtained by Bayes' rule. In addition, substituting $\sigma_L^i = 1$ we have

$$\begin{cases} \mathcal{R}^{ii}_{(1,\xi_H),1}(\Sigma) \geq \rho, \mathcal{R}^{ii}_{(1,\xi_H),0}(\Sigma) \geq \rho, \mathcal{R}^{ji}_{(1,\xi_H),1}(\Sigma) \geq \rho, \mathcal{R}^{ji}_{(0,\xi_L),1}(\Sigma) \geq \rho, \\ \mathcal{R}^{ii}_{(0,\xi_L),1}(\Sigma) \leq \rho, \mathcal{R}^{ii}_{(0,\xi_L),0}(\Sigma) \leq \rho, \mathcal{R}^{ji}_{(1,\xi_H),0}(\Sigma) \leq \rho, \mathcal{R}^{ji}_{(0,\xi_L),0}(\Sigma) \leq \rho. \end{cases}$$

Hence, by Equations (B.6) and (B.7) we have $U_{1,i}^{\xi_H}(\Sigma) > U_{0,i}^{\xi_H}(\Sigma)$, contradicting $\sigma_H^i < 1$.

Second case: $\sigma_L^i < 1$. All beliefs can be obtained by Bayes' rule. Consider first the case where $\sigma_H^j = \sigma_L^j = 1$.

Then $U_{1,i}^{\xi_H}(\Sigma) = b_i + \mu_i \mathcal{R}^{ii}_{(1,\xi_H),1}(\Sigma) + \nu_i \mathcal{R}^{ji}_{(1,\xi_H),1}(\Sigma) > 0 = U_{0,i}^{\xi_H}(\Sigma)$, which is again a contradiction.

Suppose then that $\min(\sigma_L^j, \sigma_H^j) < 1$. The conditions $\sigma_L^i < 1$ and $\sigma_H^i < 1$ imply respectively that $U_{1,i}^{\xi_L}(\Sigma) \geq U_{0,i}^{\xi_L}(\Sigma)$ and $U_{1,i}^{\xi_H}(\Sigma) \leq U_{0,i}^{\xi_H}(\Sigma)$. Hence $U_{1,i}^{\xi_H}(\Sigma) - U_{0,i}^{\xi_H}(\Sigma) \leq U_{1,i}^{\xi_L}(\Sigma) - U_{0,i}^{\xi_L}(\Sigma)$, which we rewrite using (B.6), (B.7), (B.8), and (B.9):

$$
\begin{aligned}
&[\rho + (1-\rho)\sigma_H^j][\mu_i \mathcal{R}^{ii}_{(1,\xi_H),1}(\Sigma) - \mu_i\tau_i \mathcal{R}^{ii}_{(0,\xi_L),1}(\Sigma) + \nu_i \mathcal{R}^{ji}_{(1,\xi_H),1}(\Sigma) - \nu_i \mathcal{R}^{ji}_{(1,\xi_H),0}(\Sigma)] \\
&+ (1-\rho)(1-\sigma_H^j)[\mu_i \mathcal{R}^{ii}_{(1,\xi_H),0}(\Sigma) - \mu_i\tau_i \mathcal{R}^{ii}_{(0,\xi_L),0}(\Sigma) \\
&+ \nu_i\tau_j \mathcal{R}^{ji}_{(0,\xi_L),1}(\Sigma) + \nu_i(1-\tau_j) \mathcal{R}^{ji}_{(0,\xi_H),1}(\Sigma) - \nu_i\tau_j \mathcal{R}^{ji}_{(0,\xi_L),0}(\Sigma)] \\
\leq &[\rho + (1-\rho)\sigma_L^j][\mu_i\tau_i \mathcal{R}^{ii}_{(1,\xi_H),0}(\Sigma) - \mu_i \mathcal{R}^{ii}_{(0,\xi_L),0}(\Sigma) + \nu_i \mathcal{R}^{ji}_{(0,\xi_L),1}(\Sigma) - \nu_i \mathcal{R}^{ji}_{(0,\xi_L),0}(\Sigma)] \\
&+ (1-\rho)(1-\sigma_L^j)[\mu_i\tau_i \mathcal{R}^{ii}_{(1,\xi_H),1}(\Sigma) - \mu_i \mathcal{R}^{ii}_{(0,\xi_L),1}(\Sigma) \\
&+ \nu_i\tau_j \mathcal{R}^{ji}_{(1,\xi_H),1}(\Sigma) - \nu_i\tau_j \mathcal{R}^{ji}_{(1,\xi_H),0}(\Sigma) - \nu_i(1-\tau_j) \mathcal{R}^{ji}_{(1,\xi_L),0}(\Sigma)].
\end{aligned}
\tag{B.10}
$$

Given that $\min(\sigma_H^j, \sigma_L^j) < 1$ we have $\mathcal{R}^{ii}_{(0,\xi_L),1}(\Sigma) > 0$ or $\mathcal{R}^{ii}_{(1,\xi_H),0}(\Sigma) > 0$ (or both). In the former case, the left-hand side of (B.10) is decreasing in τ_i (ignoring the dependence of the $\mathcal{R}^{ii}$s in τ_i), while in the latter case the right-hand side of (B.10) is increasing in τ_i. Similarly, the left-hand side of (B.10) is nonincreasing in τ_j because $\mathcal{R}^{ji}_{(0,\xi_L),1}(\Sigma) \leq \mathcal{R}^{ji}_{(0,\xi_H),1}(\Sigma) \leq \mathcal{R}^{ji}_{(0,\xi_H),1}(\Sigma) + \mathcal{R}^{ji}_{(0,\xi_L),0}(\Sigma)$, and the right-hand side of (B.10) is nondecreasing in τ_j because $\mathcal{R}^{ji}_{(1,\xi_H),0}(\Sigma) \leq \mathcal{R}^{ji}_{(1,\xi_L),10}(\Sigma) \leq \mathcal{R}^{ji}_{(1,\xi_L),0}(\Sigma) + \mathcal{R}^{ji}_{(1,\xi_H),1}(\Sigma)$.

The inequality in (B.10) is therefore true in a strict sense when one sets $\tau_i = \tau_j = 1$, which yields (after rearranging):

$$
\begin{aligned}
&[\rho + (1-\rho)(\sigma_H^j + \sigma_L^j - 1)] \times \\
&\quad [\mathcal{R}^{ii}_{(1,\xi_H),1}(\Sigma) - \mathcal{R}^{ii}_{(1,\xi_H),0}(\Sigma) + \mathcal{R}^{ii}_{(0,\xi_L),0}(\Sigma) - \mathcal{R}^{ii}_{(0,\xi_L),1}(\Sigma) \\
&\qquad + \mathcal{R}^{ji}_{(1,\xi_H),1}(\Sigma) - \mathcal{R}^{ji}_{(0,\xi_L),1}(\Sigma) + \mathcal{R}^{ji}_{(0,\xi_L),0}(\Sigma) - \mathcal{R}^{ji}_{(1,\xi_H),0}(\Sigma)] < 0.
\end{aligned}
\tag{B.11}
$$

But notice that, since the equilibrium is monotone, $\rho + (1-\rho)(\sigma_H^j + \sigma_L^j - 1) \geq 0$, and inequalities (B.4) and (B.5) imply that the second factor on the left-hand side of (B.11) is also nonnegative. This is a contradiction.

B.4 Analysis of co-partisan interactions

B.4.1 Proof of Lemma 2

By Lemma 1, any monotone equilibrium satisfies $\sigma_H^i = \sigma_H^j = 1$, and is therefore characterized by two values (σ^i, σ^j) (we drop the subscript L for convenience). Substituting the conditions $\sigma_H^i = \sigma_H^j = 1$,

the equilibrium conditions can be written as follows:

$$\begin{cases} \sigma^i = 1 \text{ and } b_i + (1-\rho)(1-\sigma^j)[\mu_i\tau_i\rho + \nu_i\tau_j\rho] \\ \qquad\qquad +[\rho + (1-\rho)\sigma^j][\mu_i\tau_i\mathcal{R}^{ii}_{(1,\xi_H),0}(\sigma^i=1,\sigma^j) + \nu_i\mathcal{R}^{ji}_{(0,\xi_L),1}(\sigma^i=1,\sigma^j)] \leq (\mu_i+\nu_i)\rho, \\ \sigma^i \in (0,1) \text{ and } b_i + (1-\rho)(1-\sigma^j)\Big[\dfrac{\mu_i\tau_i\lambda\rho}{\lambda + (1-\lambda)(1-\rho)^2\tau_i(1-\chi_i)(1-\sigma^i)(1-\sigma^j)} \\ \qquad\qquad + \dfrac{\nu_i\tau_j\lambda\rho}{\lambda + (1-\lambda)(1-\rho)^2\tau_j(1-\chi_j)(1-\sigma^i)(1-\sigma^j)}\Big] = (\mu_i+\nu_i)\dfrac{\rho}{\rho+(1-\rho)\sigma^i}, \\ \sigma^i = 0 \text{ and } b_i + (1-\rho)(1-\sigma^j)\Big[\dfrac{\mu_i\tau_i\lambda\rho}{\lambda + (1-\lambda)(1-\rho)^2\tau_i(1-\chi_i)(1-\sigma^j)} \\ \qquad\qquad + \dfrac{\nu_i\tau_j\lambda\rho}{\lambda + (1-\lambda)(1-\rho)^2\tau_j(1-\chi_j)(1-\sigma^j)}\Big] \geq (\mu_i+\nu_i), \end{cases}$$

where $\mathcal{R}^{ii}_{(1,\xi_H),0}(\sigma^i=1,\sigma^j)$ and $\mathcal{R}^{ji}_{(0,\xi_L),1}(\sigma^i=1,\sigma^j)$ are out-of-equilibrium beliefs.
Let us define the function

$$\Delta U^i(\sigma^i,\sigma^j) \coloneqq b_i + (1-\rho)(1-\sigma^j)\Big[\frac{\mu_i\tau_i\lambda\rho}{\lambda + (1-\lambda)(1-\rho)^2\tau_i(1-\chi_i)(1-\sigma^i)(1-\sigma^j)} + \frac{\nu_i\tau_j\lambda\rho}{\lambda + (1-\lambda)(1-\rho)^2\tau_j(1-\chi_j)(1-\sigma^i)(1-\sigma^j)}\Big] - (\mu_i+\nu_i)\frac{\rho}{\rho+(1-\rho)\sigma^i}.$$

This function captures the net expected utility gain from denial relative to realism given the equilibrium (σ^i,σ^j) (ignoring out-of-equilibrium reputation payoffs if $\sigma^i=1$). Fix σ^j. The function ΔU^i is continuous and increasing in σ^i. As a result, we have three possibilities:

- if $\Delta U^i(\sigma^i=0,\sigma^j) \geq 0$, then $\sigma^i=0$ is the unique best response;
- if $\Delta U^i(\sigma^i=1,\sigma^j) \leq 0$, then $\sigma^i=1$ is the unique best-response, supported by any out-of-equilibrium beliefs such that $\Delta U_i(\sigma=1,\sigma^j) + [\rho+(1-\rho)\sigma^j][\mu_i\tau_i\mathcal{R}^{ii}_{(1,\xi_H),0}(\sigma^i=1,\sigma^j) + \nu_i\mathcal{R}^{ji}_{(0,\xi_L),1}(\sigma^i=1,\sigma^j)] \leq 0$ (e.g. $\mathcal{R}^{ii}_{(1,\xi_H),0}(\sigma^i=1,\sigma^j) = \mathcal{R}^{ji}_{(0,\xi_L),1}(\sigma^i=1,\sigma^j) = 0$);
- if $\Delta U^i(\sigma^i=0,\sigma^j) < 0 < \Delta U^i(\sigma^i=1,\sigma^j)$, then there exists a unique $\sigma^i \in (0,1)$ with $\Delta U^i(\sigma^i=0,\sigma^j)=0$, which is therefore the unique best-response.

This establishes existence and uniqueness of the best response which, in addition, is continuous in all parameter values. The comparative statics then follows from the variations of the function $\Delta U^i(\sigma,\sigma^j)$ with the corresponding parameter values. This proves Lemma 2.

B.4.2 Proof of Proposition 2

B.4.2.1 Existence of an equilibrium

The previous section 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.

B.4.2.2 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.

Consider two values $\sigma^j < \tilde{\sigma}^j$ and the respective best responses σ^i and $\tilde{\sigma}^i$ by player i. The monotonicity of the best-response functions implies that $\tilde{\sigma}^i \geq \sigma^i$, and even $\tilde{\sigma}^i > \sigma^i$ since $\tilde{\sigma}^i = \sigma^i$ would imply $\sigma^j = \tilde{\sigma}^j$ by uniqueness of j's best response.

The condition $\tilde{\sigma}^i > \sigma^i$ implies $\tilde{\sigma}^i > 0$. Hence, $\Delta U^i(\tilde{\sigma}^i, \tilde{\sigma}^j) \leq 0$. Similarly, $\tilde{\sigma}^i > \sigma^i$ implies $\sigma^i < 1$ and thus $\Delta U^i(\sigma^i, \sigma^j) \geq 0$. Therefore $\Delta U^i(\sigma^i, \sigma^j) \geq \Delta U^i(\tilde{\sigma}^i, \tilde{\sigma}^j)$, which we rearrange to obtain

$$
\begin{aligned}
&(\mu_i + \nu_i)\Big[\frac{\rho}{\rho + (1-\rho)\sigma^i} - \frac{\rho}{\rho + (1-\rho)\tilde{\sigma}^i}\Big] \\
\leq &\frac{\mu_i \tau_i \lambda \rho (1-\rho)(1-\sigma^j)}{\lambda + (1-\lambda)(1-\rho)^2 \tau_i (1-\chi_i)(1-\sigma^i)(1-\sigma^j)} - \frac{\mu_i \tau_i \lambda \rho (1-\rho)(1-\tilde{\sigma}^j)}{\lambda + (1-\lambda)(1-\rho)^2 \tau_i (1-\chi_i)(1-\tilde{\sigma}^i)(1-\tilde{\sigma}^j)} \\
&+ \frac{\nu_i \tau_j \lambda \rho (1-\rho)(1-\sigma^j)}{\lambda + (1-\lambda)(1-\rho)^2 \tau_j (1-\chi_j)(1-\sigma^i)(1-\sigma^j)} - \frac{\nu_i \tau_j \lambda \rho (1-\rho)(1-\tilde{\sigma}^j)}{\lambda + (1-\lambda)(1-\rho)^2 \tau_j (1-\chi_j)(1-\tilde{\sigma}^i)(1-\tilde{\sigma}^j)} \\
\leq &\frac{\mu_i \tau_i \lambda \rho (1-\rho)(1-\sigma^j)}{\lambda + (1-\lambda)(1-\rho)^2 \tau_i (1-\chi_i)(1-\tilde{\sigma}^i)(1-\tilde{\sigma}^j)} - \frac{\mu_i \tau_i \lambda \rho (1-\rho)(1-\tilde{\sigma}^j)}{\lambda + (1-\lambda)(1-\rho)^2 \tau_i (1-\chi_i)(1-\tilde{\sigma}^i)(1-\tilde{\sigma}^j)} \\
&+ \frac{\nu_i \tau_j \lambda \rho (1-\rho)(1-\sigma^j)}{\lambda + (1-\lambda)(1-\rho)^2 \tau_j (1-\chi_j)(1-\tilde{\sigma}^i)(1-\tilde{\sigma}^j)} - \frac{\nu_i \tau_j \lambda \rho (1-\rho)(1-\tilde{\sigma}^j)}{\lambda + (1-\lambda)(1-\rho)^2 \tau_j (1-\chi_j)(1-\tilde{\sigma}^i)(1-\tilde{\sigma}^j)} \\
= &\Big[\frac{\mu_i \tau_i \lambda \rho (1-\rho)}{\lambda + (1-\lambda)(1-\rho)^2 \tau_i (1-\chi_i)(1-\tilde{\sigma}^i)(1-\tilde{\sigma}^j)} + \frac{\nu_i \tau_j \lambda \rho (1-\rho)}{\lambda + (1-\lambda)(1-\rho)^2 \tau_j (1-\chi_j)(1-\tilde{\sigma}^i)(1-\tilde{\sigma}^j)}\Big] \\
&\times [\tilde{\sigma}^j - \sigma^j].
\end{aligned}
$$

where the first inequality expresses $\Delta U^i(\sigma^i, \sigma^j) \geq \Delta U^i(\tilde{\sigma}^i, \tilde{\sigma}^j)$, and the second inequality comes from the fact that $\tilde{\sigma}^j > \sigma^j, \tilde{\sigma}^i > \sigma^i$.

Re-arranging the left-hand side we get

$$
\begin{aligned}
\frac{(\mu_i + \nu_i)\rho(\tilde{\sigma}^i - \sigma^i)}{[\rho + (1-\rho)\sigma^i]\rho + (1-\rho)\tilde{\sigma}^i]} \leq &\frac{\mu_i \tau_i \lambda \rho (\tilde{\sigma}^j - \sigma^j)}{\lambda + (1-\lambda)(1-\rho)^2 \tau_i (1-\chi_i)(1-\tilde{\sigma}^i)(1-\tilde{\sigma}^j)} \\
&+ \frac{\nu_i \tau_j \lambda \rho (\tilde{\sigma}^j - \sigma^j)}{\lambda + (1-\lambda)(1-\rho)^2 \tau_j (1-\chi_j)(1-\tilde{\sigma}^i)(1-\tilde{\sigma}^j)},
\end{aligned}
$$

which we rewrite

$$
\begin{aligned}
\tilde{\sigma}^i - \sigma^i \leq &\frac{\mu_i}{\mu_i + \nu_i} \underbrace{\frac{\tau_i \lambda [\rho + (1-\rho)\sigma^i][\rho + (1-\rho)\tilde{\sigma}^i]}{\lambda + (1-\lambda)(1-\rho)^2 \tau_i (1-\chi_i)(1-\tilde{\sigma}^i)(1-\tilde{\sigma}^j)}}_{<1} [\tilde{\sigma}^j - \sigma^j] \\
&+ \frac{\nu_i}{\mu_i + \nu_i} \underbrace{\frac{\tau_j \lambda [\rho + (1-\rho)\sigma^i][\rho + (1-\rho)\tilde{\sigma}^i]}{\lambda + (1-\lambda)(1-\rho)^2 \tau_j (1-\chi_j)(1-\tilde{\sigma}^i)(1-\tilde{\sigma}^j)}}_{<1} [\tilde{\sigma}^j - \sigma^j] < \tilde{\sigma}^j - \sigma^j.
\end{aligned}
$$

Thus, $\tilde{\sigma}^i - \sigma^i < \tilde{\sigma}^j - \sigma^j$, which proves that i's best-response function has slope strictly smaller than one. This concludes the proof.

B.4.2.3 Comparative statics

The comparative statics results 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.

Let $BR^i(\sigma^j, b_i)$ be i's best-response to σ^j conditional on b_i, and conversely $BR^j(\sigma^i)$ be j's best-response to σ^i, which 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 $b_i < b_i'$, and let $[\sigma^i(b_i), \sigma^j(b_i)]$ be the equilibrium when the parameter is b_i. We have

$$\forall \sigma^j \leq \sigma^j(b_i), BR^i(\sigma^j, b_i') \leq BR^i(\sigma^j, b_i) \leq BR^i[\sigma^j(b_i), b_i] = \sigma^i(b_i),$$

and

$$\forall \sigma^i \leq \sigma^i(b_i), BR^j(\sigma^i) \leq BR^j[\sigma^i(b_i)] = \sigma^j(b_i).$$

Therefore the best-response correspondence $(\sigma^i, \sigma^j) \mapsto [BR^i(\sigma^j, b_i'), BR^j(\sigma^i)]$ maps the compact set $[(0,0), (\sigma^i(b_i), \sigma^j(b_i))]$ 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 b_i' must therefore satisfy $\sigma^i(b_i') \leq \sigma^i(b_i)$ and $\sigma^j(b_i') \leq \sigma^j(b_i)$.

B.4.3 Non-monotone equilibrium

In this section we exhibit a non-monotone equilibrium where $\sigma_H^i = \sigma_L^i = \sigma_H^j = \sigma_L^j = 0$ and $\rho < 1/2$. Both players deny both signals. They do so because they expect each other to be very apt at rationalizing and almost perfectly naive, and hence convinced by their rationalizations. Upon seeing the realized value of ξ, each player understands that the other player is more likely not to play the action prescribed by the signal (as $\rho < 1/2$), and that their expected reputation vis-à-vis the other player will be better if they imitate them.

Formally, let $\chi_i \approx 1, \chi_j \approx 1$, $\tau_i \approx 1, \tau_j \approx 1$, $\mu_i \approx 0$ and $\mu_j \approx 0$. Fix $\sigma_H^j = \sigma_L^j = 0$ and $\sigma_L^i = 0$. Plugging all these conditions into (B.6) and (B.7) shows that $\sigma_H^i = 0$ is i's best-response following $\xi = \xi_H$ if and only if (in approximation) $b_i \leq \nu_i(1-2\rho)$. Similarly, assuming $\sigma_H^i = 0$ and using (B.8) and (B.9), $\sigma_L^i = 0$ is i's best-response following $\xi = \xi_L$ if and only if $b_i \geq \nu_i(2\rho - 1)$, which is implied by $\rho < 1/2$.

The analysis for j is symmetric: hence, if $b_i \leq \nu_i(1-2\rho)$ and $b_j \leq \nu_j(1-2\rho)$, then the strategies $(\sigma_H^i = \sigma_L^i = \sigma_H^j = \sigma_L^j = 0)$ constitute an equilibrium.

B.5 Analysis of counter-partisan interactions

B.5.1 Best-response

We start by analyzing the best-response of individual i to some $\sigma^j \in [0,1]$. We can invoke Lemma 1 again and infer that any monotone equilibrium satisfies $\sigma_H^i = \sigma_L^j = 1$, and is therefore characterized by two values $(\sigma^i = \sigma_L^i, \sigma^j = \sigma_H^j)$ (we drop the subscripts L and H for convenience).

The best-response behavior of i to j's strategy is summarized in Lemma 6.

Lemma 6 (Best-response among counter-partisans)**.** *Given some σ^j, there exists a unique best-response $\sigma^i \in [0,1]$ for i. This best-response level of compliance to inconvenient signals is nondecreasing in σ^j (strategic complementarity), in μ_i and in ν_i (image concerns), and nonincreasing in b_i (personal motive), τ_i and τ_j (rationalization efficiency).*

Proof. Let us start with the case where $\sigma^j = 1$. Then i's best-response condition can be rewritten

$$\begin{cases} \sigma^i = 1 \text{ and } b_i + \tau_i\mu_i\mathcal{R}^{ii}_{(1,\xi_H),0}(\sigma^i = 1, \sigma^j = 1) + \mathcal{R}^{ji}_{(0,\xi_L),1}(\sigma^i = 1, \sigma^j = 1) \leq (\mu_i + \nu_i)\rho, \\ \sigma^i \in (0,1) \text{ and } b_i = (\mu_i + \nu_i)\dfrac{\rho}{\rho + (1-\rho)\sigma^i}, \\ \sigma^i = 0 \text{ and } b_i \geq (\mu_i + \nu_i). \end{cases}$$

where $\mathcal{R}^{ii}_{(1,\xi_H),0}(\sigma^i = 1, \sigma^j = 1)$ and $\mathcal{R}^{ji}_{(0,\xi_L),1}(\sigma^i = 1, \sigma^j = 1)$ are out-of-equilibrium beliefs. It is clear that this system admits a unique solution σ^i that is nonincreasing in b_i, nondecreasing in σ^j, μ_i and ν_i, and independent of other parameters.

Suppose now that $\sigma^j < 1$. The best-response condition for i becomes

$$\begin{cases} \sigma^i = 1 \text{ and } b_i + (\mu_i\tau_i + \nu_i\tau_j)\rho \leq (\mu_i + \nu_i)\rho \\ \sigma^i \in (0,1) \text{ and } b_i + \dfrac{\mu_i\tau_i\lambda\rho(1-\sigma^j)}{\lambda(1-\sigma^j) + (1-\lambda)\tau_i(1-\chi_i)(1-\sigma^i)} \\ \qquad\qquad + \dfrac{\nu_i\tau_j\lambda\rho(1-\chi_j)(1-\sigma^j)}{\lambda\tau_j(1-\chi_j)(1-\sigma^j) + (1-\lambda)(1-\sigma^i)} = (\mu_i + \nu_i)\dfrac{\rho}{\rho + (1-\rho)\sigma^i}, \\ \sigma^i = 0 \text{ and } b_i + \dfrac{\mu_i\tau_i\lambda\rho(1-\sigma^j)}{\lambda(1-\sigma^j) + (1-\lambda)\tau_i(1-\chi_i)} \\ \qquad\qquad + \dfrac{\nu_i\tau_j\lambda\rho(1-\chi_j)(1-\sigma^j)}{\lambda\tau_j(1-\chi_j)(1-\sigma^j) + (1-\lambda)} \geq \mu_i + \nu_i. \end{cases}$$

This system also admits a unique solution whose comparative statics is given by the variations of the function

$$b_i + \frac{\mu_i\lambda\rho\tau_i(1-\sigma^j)}{\lambda(1-\sigma^j) + (1-\lambda)\tau_i(1-\chi_i)(1-\sigma^i)} + \frac{\nu_i\lambda\rho\tau_j(1-\chi_j)(1-\sigma^j)}{\lambda\tau_j(1-\chi_j)(1-\sigma^j) + (1-\lambda)(1-\sigma^i)} - (\mu_i + \nu_i)\frac{\rho}{\rho + (1-\rho)\sigma^i}.$$

Moreover, comparing the two systems, if σ^i is the best-response to $\sigma^j = 1$, then it is clearly weakly larger than the best-response to any $\sigma^j < 0$. This concludes the proof of Lemma 6. □

B.5.1.1 Proof of Proposition 3

B.5.1.2 Existence and uniqueness

Existence of an equilibrium is again given by Brouwer's fixed point theorem. We prove uniqueness by contradiction, assuming that (σ^i, σ^j) and $(\tilde{\sigma}^i, \tilde{\sigma}^j)$ are both equilibria, and (without loss) that $\sigma^i < \tilde{\sigma}^i$, $\sigma^j < \tilde{\sigma}^j$. The central argument in the proof is that player i must receive a larger payoff from denying $\xi = \xi_L$ in the equilibrium (σ^i, σ^j) than in the equilibrium $(\tilde{\sigma}^i, \tilde{\sigma}^j)$; similarly, player j must receive a larger payoff from denying $\xi = \xi_H$ in the former equilibrium as well. We show that both statements are inconsistent with each other.

First case : suppose that $(\tilde{\sigma}^i, \tilde{\sigma}^j) \neq (1,1)$. Then all posterior beliefs are formed by Bayes' rule.

The fact that $\sigma^i < 1$ is a best-response to $\sigma^j < 1$ implies that

$$b_i + \frac{\mu_i\tau_i\lambda\rho(1-\sigma^j)}{\lambda(1-\sigma^j)+(1-\lambda)\tau_i(1-\chi_i)(1-\sigma^i)} + \frac{\nu_i\tau_j\lambda\rho(1-\chi_j)(1-\sigma^j)}{\lambda\tau_j(1-\chi_j)(1-\sigma^j)+(1-\lambda)(1-\sigma^i)} \geq (\mu_i+\nu_i)\frac{\rho}{\rho+(1-\rho)\sigma^i}, \tag{B.12}$$

whereas the fact that $\tilde{\sigma}^i > 0$ implies that

$$b_i + \frac{\mu_i\tau_i\lambda\rho(1-\tilde{\sigma}^j)}{\lambda(1-\tilde{\sigma}^j)+(1-\lambda)\tau_i(1-\chi_i)(1-\tilde{\sigma}^i)} + \frac{\nu_i\tau_j\lambda\rho(1-\chi_j)(1-\tilde{\sigma}^j)}{\lambda\tau_j(1-\chi_j)(1-\tilde{\sigma}^j)+(1-\lambda)(1-\tilde{\sigma}^i)} \leq (\mu_i+\nu_i)\frac{\rho}{\rho+(1-\rho)\tilde{\sigma}^i}, \tag{B.13}$$

where all these expressions are well-defined since $(\tilde{\sigma}^i, \tilde{\sigma}^j) \neq (1,1)$.

Since $\sigma^i < \tilde{\sigma}^i$, the right-hand side of (B.12) is strictly larger than the right-hand side of (B.13). Comparing the left-hand sides, we then get

$$\begin{aligned} &\frac{\mu_i\tau_i\lambda\rho(1-\sigma^j)}{\lambda(1-\sigma^j)+(1-\lambda)\tau_i(1-\chi_i)(1-\sigma^i)} + \frac{\nu_i\tau_j\lambda\rho(1-\chi_j)(1-\sigma^j)}{\lambda\tau_j(1-\chi_j)(1-\sigma^j)+(1-\lambda)(1-\sigma^i)} \\ >&\frac{\mu_i\tau_i\lambda\rho(1-\tilde{\sigma}^j)}{\lambda(1-\tilde{\sigma}^j)+(1-\lambda)\tau_i(1-\chi_i)(1-\tilde{\sigma}^i)} + \frac{\nu_i\tau_j\lambda\rho(1-\chi_j)(1-\tilde{\sigma}^j)}{\lambda\tau_j(1-\chi_j)(1-\tilde{\sigma}^j)+(1-\lambda)(1-\tilde{\sigma}^i)}. \end{aligned} \tag{B.14}$$

And the symmetric reasoning for player j (expressing the conditions $\sigma^j < 1$ and $\tilde{\sigma}^j > 0$) yields similarly

$$\begin{aligned} &\frac{\mu_j\tau_j(1-\lambda)\rho(1-\sigma^i)}{\lambda\tau_j(1-\chi_j)(1-\sigma^j)+(1-\lambda)(1-\sigma^i)} + \frac{\nu_j\tau_i(1-\lambda)\rho(1-\chi_i)(1-\sigma^i)}{\lambda(1-\sigma^j)+(1-\lambda)\tau_i(1-\chi_i)(1-\sigma^i)} \\ >&\frac{\mu_j\tau_j(1-\lambda)\rho(1-\tilde{\sigma}^i)}{\lambda\tau_j(1-\chi_j)(1-\tilde{\sigma}^j)+(1-\lambda)(1-\tilde{\sigma}^i)} + \frac{\nu_j\tau_i(1-\lambda)\rho(1-\chi_i)(1-\tilde{\sigma}^i)}{\lambda(1-\tilde{\sigma}^j)+(1-\lambda)\tau_i(1-\chi_i)(1-\tilde{\sigma}^i)}. \end{aligned} \tag{B.15}$$

We will show that Conditions (B.14) and (B.15) are inconsistent, by distinguishing three cases:

1. If $\tilde{\sigma}^j = 1$, then condition (B.15) can be rewritten

$$\begin{aligned} &\frac{\mu_j\tau_j(1-\lambda)\rho(1-\sigma^i)}{\lambda\tau_j(1-\chi_j)(1-\sigma^j)+(1-\lambda)(1-\sigma^i)} + \frac{\nu_j\tau_i(1-\lambda)\rho(1-\chi_i)(1-\sigma^i)}{\lambda(1-\sigma^j)+(1-\lambda)\tau_i(1-\chi_i)(1-\sigma^i)} \\ >&[\mu_j\tau_j+\nu_j]\rho, \end{aligned}$$

   which is a contradiction since the multipliers of $\mu_j\tau_j$ and ν_j on the left-hand side are smaller than ρ.

2. If $\tilde{\sigma}^i = 1$, we find similarly a contradiction in inequality (B.14).

3. If $\tilde{\sigma}^i < 1$ and $\tilde{\sigma}^j < 1$, condition (B.14) can be rewritten

$$\frac{\mu_i\tau_i\lambda\rho}{\lambda+(1-\lambda)\tau_i(1-\chi_i)\dfrac{1-\sigma^i}{1-\sigma^j}}+\frac{\tau_j\nu_i\lambda\rho(1-\chi_j)}{\lambda\tau_j(1-\chi_j)+(1-\lambda)\dfrac{1-\sigma^i}{1-\sigma^j}}>$$
$$\frac{\mu_i\tau_i\lambda\rho}{\lambda+(1-\lambda)\tau_i(1-\chi_i)\dfrac{1-\tilde{\sigma}^i}{1-\tilde{\sigma}^j}}+\frac{\nu_i\tau_j\lambda\rho(1-\chi_j)}{\lambda\tau_j(1-\chi_j)+(1-\lambda)\dfrac{1-\tilde{\sigma}^i}{1-\tilde{\sigma}^j}},$$

from which we infer $(1-\sigma^i)/(1-\sigma^j) < (1-\tilde{\sigma}^i)/(1-\tilde{\sigma}^j)$. But going through the same steps for Condition (B.15) shows similarly that $(1-\sigma^j)/(1-\sigma^i) < (1-\tilde{\sigma}^j)/(1-\tilde{\sigma}^i)$. This is a contradiction.

Second case: suppose that $(\tilde{\sigma}^i, \tilde{\sigma}^j) = (1,1)$. We will show that Refinements 2 and 3 rule out this possibility. The fact that $\tilde{\sigma}^i = 1$ is the best-response to $\tilde{\sigma}^j = 1$ implies that the out-of-equilibrium beliefs $\mathcal{R}^{ii}_{(1,\xi_H),0}$ and $\mathcal{R}^{ji}_{(0,\xi_L),1}$ are such that

$$b_i + \mu_i\tau_i\mathcal{R}^{ii}_{(1,\xi_H),0} + \nu_i\mathcal{R}^{ji}_{(0,\xi_L),1} \leq (\mu_i+\nu_i)\rho. \tag{B.16}$$

In addition, the condition $\sigma^i < 1$ implies inequality (B.12) again. Combining (B.12) and (B.16) yields

$$\begin{aligned}\frac{(\mu_i+\nu_i)\rho(1-\rho)(1-\sigma^i)}{\rho+(1-\rho)\sigma^i} \leq{}& \mu_i\tau_i\Big[\frac{\lambda\rho(1-\sigma^j)}{\lambda(1-\sigma^j)+(1-\lambda)\tau_i(1-\chi_i)(1-\sigma^i)}-\mathcal{R}^{ii}_{(1,\xi_H),0}\Big]\\ &+\nu_i\Big[\frac{\lambda\rho\tau_j(1-\chi_j)(1-\sigma^j)}{\lambda\tau_j(1-\chi_j)(1-\sigma^j)+(1-\lambda)(1-\sigma^i)}-\mathcal{R}^{ji}_{(0,\xi_L),1}\Big].\end{aligned} \tag{B.17}$$

The symmetric reasoning for player j similarly yields

$$\begin{aligned}\frac{(\mu_j+\nu_j)\rho(1-\rho)(1-\sigma^j)}{\rho+(1-\rho)\sigma^j} \leq{}& \mu_j\tau_j\Big[\frac{(1-\lambda)\rho(1-\sigma^i)}{\lambda\tau_j(1-\chi_j)(1-\sigma^j)+(1-\lambda)(1-\sigma^i)}-\mathcal{R}^{jj}_{(0,\xi_L),1}\Big]\\ &+\nu_j\Big[\frac{(1-\lambda)\rho\tau_i(1-\chi_i)(1-\sigma^j)}{\lambda(1-\sigma^j)+(1-\lambda)\tau_i(1-\chi_i)(1-\sigma^i)}-\mathcal{R}^{ij}_{(1,\xi_H),0}\Big],\end{aligned} \tag{B.18}$$

where we define $\mathcal{R}^{jj}_{(0,\xi_L),1}$ as j's belief about their own type conditional on $[(a_j=0,\hat{\xi}_j=\xi_L),a_i=1]$, and $\mathcal{R}^{ij}_{(1,\xi_H),0}$ as i's belief about j's type conditional on $[(a_i=1,\hat{\xi}_i=\xi_H),a_j=0]$.

We can then use Refinement (2), which writes $\mathcal{R}^{ii}_{(1,\xi_H),0}+\mathcal{R}^{ij}_{(1,\xi_H),0} = \mathcal{R}^{jj}_{(0,\xi_L),1}+\mathcal{R}^{ji}_{(0,\xi_L),1} = \rho$, to rewrite inequality (B.18) into (after some algebra)

$$\begin{aligned}\frac{(\mu_j+\nu_j)\rho(1-\rho)(1-\sigma^j)}{\rho+(1-\rho)\sigma^j} \leq{}& \mu_j\tau_j\Big[\mathcal{R}^{ji}_{(0,\xi_L),1}-\frac{\lambda\rho\tau_j(1-\chi_j)(1-\sigma^i)}{\lambda\tau_j(1-\chi_j)(1-\sigma^j)+(1-\lambda)(1-\sigma^i)}\Big]\\ &+\nu_j\Big[\mathcal{R}^{ii}_{(1,\xi_H),0}-\frac{\lambda\rho(1-\sigma^j)}{\lambda(1-\sigma^j)+(1-\lambda)\tau_i(1-\chi_i)(1-\sigma^i)}\Big].\end{aligned} \tag{B.19}$$

But note that, given Refinement (3),

$$\begin{aligned}\mathcal{R}^{ji}_{(0,\xi_L),1} &\geq \frac{\lambda\rho\tau_j(1-\chi_j)(1-\sigma^j)}{\lambda\tau_j(1-\chi_j)(1-\sigma^j)+(1-\lambda)(1-\sigma^i)}\\ \Leftrightarrow \mathcal{R}^{ii}_{(1,\xi_H),0} &\geq \frac{\lambda\rho(1-\sigma^j)}{\lambda(1-\sigma^j)+(1-\lambda)\tau_i(1-\chi_i)(1-\sigma^i)}.\end{aligned}$$

This shows that one of inequalities (B.17) and (B.19) must have a nonpositive right-hand side, which is a contradiction since the left-hand sides are positive. This establishes equilibrium uniqueness.

B.5.1.3 Equilibrium multiplicity

In this section we exhibit multiple equilibria in a case where off-path beliefs violate Refinements (2) and (3): suppose that $\mathcal{R}^{ii}_{(1,\xi_H),0} = \mathcal{R}^{ji}_{(0,\xi_L),1} = \mathcal{R}^{ij}_{(1,\xi_H),0} = \mathcal{R}^{jj}_{(0,\xi_L),1} = 0$, that is, reputational punishments after a deviation are maximal, as every player infers that neither i nor j are reasonable. Then under the conditions

$$\frac{\lambda\rho\tau_j(1-\chi_j)}{\lambda\tau_j(1-\chi_j)+1-\lambda} > 1-\rho \text{ and } \frac{(1-\lambda)\rho\tau_i(1-\chi_i)}{\lambda+(1-\lambda)\tau_i(1-\chi_i)} > 1-\rho,$$

there exists parameter values (e.g., with $\rho \approx 1$) that satisfy the following two systems:

$$\begin{cases} b_i + \dfrac{\mu_i\tau_i\lambda\rho}{\lambda+(1-\lambda)\tau_i(1-\chi_i)} + \dfrac{\nu_i\lambda\rho\tau_j(1-\chi_j)}{\lambda\tau_j(1-\chi_j)+1-\lambda} \geq \mu_i+\nu_i, \\ \dfrac{\mu_j\tau_j(1-\lambda)\rho}{\lambda\tau_j(1-\chi_j)+(1-\lambda)} + \dfrac{\nu_j(1-\lambda)\rho\tau_i(1-\chi_i)}{\lambda+(1-\lambda)\tau_i(1-\chi_i)} \geq b_j+\mu_j+\nu_j, \end{cases} \tag{B.20}$$

and

$$\begin{cases} b_i \leq (\mu_i+\nu_i)\rho, \\ 0 \leq b_j + (\mu_j+\nu_j)\rho. \end{cases} \tag{B.21}$$

System (B.20) implies that $(\sigma^i = \sigma^j = 0)$ is an equilibrium, while system (B.21) guarantees that $(\sigma^i = \sigma^j = 1)$ is an equilibrium.

B.5.1.4 Comparative statics

The comparative statics results rely on similar arguments as in the co-partisan case. The only ambiguous case is the variation in χ_j, as i's best-response level of realism is nondecreasing in χ_j but j' best-response is nonincreasing in it.

B.6 Proof of Proposition 4

Co-partisan interaction $(b_i > 0, b_j > 0)$ We write $\mathbb{E}[p_i(\xi = \xi_H)]$ for player i's average posterior belief on $\xi = \xi_H$ (and similarly for all other variables).

To compute this expression, we first note that $\mathcal{I} = \{(a_i = 1, \hat{\xi}_i = \xi_H), a_j = 1\}$ is the only dataset for i under which $p_i[\xi = \xi_H \mid \mathcal{I}] > 0$. Indeed, each of the conditions $a_i = 0, a_j = 0$ or $\hat{\xi} = \xi_L$ reveals

that $\xi = \xi_L$ with certainty. Hence, we have

$$\begin{aligned}\mathbb{E}[p_i(\xi = \xi_H)] &= p[a_i = 1, \hat{\xi}_i = \xi_H, a_j = 1] \times p_i[\xi = \xi_H \mid a_i = 1, \hat{\xi}_i = \xi_H, a_j = 1] \\ &= \frac{[\lambda + (1-\lambda)(1-\rho)^2\tau_i(1-\sigma^i)(1-\sigma^j)]\lambda}{\lambda + (1-\lambda)(1-\rho)^2\tau_i(1-\chi_i)(1-\sigma^i)(1-\sigma^j)} \\ &= \lambda + \beta^{\text{co-p}},\end{aligned}$$

where we define

$$\beta^{\text{co-p}} := \frac{\lambda(1-\lambda)(1-\rho)^2\tau_i\chi_i(1-\sigma^i)(1-\sigma^j)}{\lambda + (1-\lambda)(1-\rho)^2\tau_i(1-\chi_i)(1-\sigma^i)(1-\sigma^j)} > 0$$

as the (average) belief bias in the co-partisan case.

Similar computations show that

$$\mathbb{E}[p_i(\theta(i) = R)] = \mathbb{E}[p_i(\theta(j) = R)] = \rho[1 + \beta^{\text{co-p}}] > \rho.$$

Counter-partisan interaction ($b_i > 0, b_j < 0$) In the case of counter-partisans, $p_i(\xi = \xi_H \mid \mathcal{I})$ takes positive value if and only if $\mathcal{I} \supset \{a_j = 1\}$ (in which case $p_i(\xi = \xi_H \mid \mathcal{I}) = 1$) or $\mathcal{I} = [(a_i = 1, \hat{\xi}_i = \xi_H), a_j = 0]$. We thus have

$$\begin{aligned}\mathbb{E}[p_i(\xi = \xi_H)] =& p[a_i = 1, \hat{\xi}_i = \xi_H, a_j = 0] \times p_i[\xi = \xi_H \mid a_i = 1, \hat{\xi}_i = \xi_H, a_j = 0] + p[a_j = 1] \\ =& \frac{[\lambda(1-\rho)(1-\sigma^j) + (1-\lambda)(1-\rho)\tau_i(1-\sigma^i)]\lambda(1-\sigma)^j}{\lambda(1-\sigma^j) + +(1-\lambda)\tau_i(1-\chi_i)(1-\sigma^i)} \\ &+ \lambda[\rho + (1-\rho)\sigma^j] \\ =& \lambda + \beta^{\text{counter-p}},\end{aligned}$$

where we now define

$$\beta^{\text{counter-p}} := \frac{\lambda(1-\lambda)(1-\rho)\tau_i\chi_i(1-\sigma^i)(1-\sigma^j)}{\lambda(1-\sigma^j) + (1-\lambda)\tau_i(1-\chi_i)(1-\sigma^i)} > 0.$$

Similar computations yield

$$\mathbb{E}[p_i(\theta(i) = R)] = \rho[1 + \beta^{\text{counter-p}}] \text{ and } \mathbb{E}[p_i(\theta(i) = R)] = \rho[1 - \beta^{\text{counter-p}}].$$

B.7 Proof of Lemma 3

As explained in the main text, we can restrict attention to the scenario where $a_i = a_{\underline{i}} = 1, a_{\underline{k}} = 0$.

Co-partisan communication (from $\underline{i}$ to i) In that case, $\underline{i}$ faces uncertainty about i's narrative $\hat{\xi}_i$. If $\hat{\xi}_i = \xi_L$, then i has inferred that $\theta(\underline{i}) = S$ with certainty, and hence communication of $\hat{\xi}_{\underline{i}}$ is inconsequential for $\underline{i}$'s reputation. If $\hat{\xi}_i = \xi_H$, then i's interim opinion of $\underline{i}$ equals

$$p_i[\theta(\underline{i}) = R \mid (a_i = 1, \hat{\xi}_i = \xi_H), a_{\underline{i}} = 1, a_{\underline{k}} = 0] = \frac{\lambda\rho)}{\lambda + (1-\lambda)(1-\rho)\tau_i(1-\chi_i)(1-\sigma^i)}.$$

If $\underline{i}$ were to share $\hat{\xi}_{\underline{i}} = \xi_H$, then i's beliefs would move to

$$\begin{aligned} &p_i[\theta(\underline{i}) = R \mid (a_i = 1, \hat{\xi}_i = \xi_H), (a_{\underline{i}} = 1, \hat{\xi}_{\underline{i}} = \xi_H), a_{\underline{k}} = 0] \\ =&\frac{\lambda\rho}{\lambda + (1-\lambda)(1-\rho)\tau_i(1-\chi_i)\tau_{\underline{i}}(1-\chi_i^{\underline{i}})(1-\sigma^i)}, \end{aligned}$$

which is strictly larger, while if $\underline{i}$ were to share $\hat{\xi}_{\underline{i}} = \xi_L$ then i would learn with certainty that $\underline{i}$ is a sophist. The standard unraveling argument then implies that all narratives are shared.

In addition, i's belief about ξ after communication of a congruent narrative $\hat{\xi}_{\underline{i}} = \xi_H$ equals

$$\begin{aligned} &p_i[\xi = \xi_H \mid (a_i = 1, \hat{\xi}_i = \xi_H), (a_{\underline{i}} = 1, \hat{\xi}_{\underline{i}} = \xi_H), a_{\underline{k}} = 0] \\ &= \frac{\lambda}{\lambda + (1-\lambda)(1-\rho)\tau_i(1-\chi_i)\tau_{\underline{i}}(1-\chi_i^{\underline{i}})(1-\sigma^i)}, \end{aligned}$$

which is strictly larger than both the interim belief and the belief conditional on $\hat{\xi}_{\underline{i}} = \xi_L$ (which equals 0). The variation of i's beliefs about $\theta(i)$ and $\theta(\underline{k})$ can be obtained similarly by applications of Bayes' rule.

Counter-partisan communication (from $\underline{k}$ to i) If $\hat{\xi}_i = \xi_L$, then i has inferred that $\xi = \xi_L$ and thus that $\underline{k}$ has a congruent narrative $\hat{\xi}_{\underline{k}} = \xi_L$ with certainty, and communication is inconsequential. If $\hat{\xi}_i = \xi_H$, then i's interim opinion of $\underline{k}$ equals

$$p_i[\theta(\underline{k}) = R \mid (a_i = 1, \hat{\xi}_i = \xi_H), a_{\underline{i}} = 1, a_{\underline{k}} = 0] = \frac{(1-\lambda)(1-\rho)\tau_i(1-\chi_i)(1-\sigma^i)\rho}{\lambda + (1-\lambda)(1-\rho)\tau_i(1-\chi_i)(1-\sigma^i)}.$$

If $\underline{k}$ shares a congruent narrative $\hat{\xi}_{\underline{k}} = \xi_L$, then i's beliefs move to

$$\begin{aligned} &p_i[\theta(\underline{k}) = R \mid (a_i = 1, \hat{\xi}_i = \xi_H), a_{\underline{i}} = 1, (a_{\underline{k}} = 0, \hat{\xi}_{\underline{k}} = \xi_L)] \\ =&\frac{(1-\lambda)(1-\rho)\tau_i(1-\chi_i)(1-\sigma^i)\rho}{\lambda\tau_{\underline{k}}(1-\chi_i^{\underline{k}}) + (1-\lambda)(1-\rho)\tau_i(1-\chi_i)(1-\sigma^i)}, \end{aligned}$$

which is strictly larger, while if $\underline{k}$ shares $\hat{\xi}_{\underline{k}} = \xi_H$ then i learns with certainty that $\underline{k}$ is a sophist. The rest of the proof relies on arguments similar to the co-partisan case.

B.8 Proof of Proposition 5

Fix σ^i, and consider first the case without communication, where i observes both speakers' actions but not their narratives.

Using similar steps as in the proof of Proposition 4, we find i's expected posterior beliefs:

$$\begin{cases} \mathbb{E}[p_i(\xi = \xi_H)] = \lambda + \beta^{\varnothing}, \\ \mathbb{E}[p_i(\theta(i) = R)] = \mathbb{E}[p_i(\theta(\underline{i}) = R)] = \rho(1 + \beta^{\varnothing}), \\ \mathbb{E}[p_i(\theta(\underline{k}) = R)] = \rho(1 - \beta^{\varnothing}) \end{cases}$$

where

$$\beta^{\varnothing} := \frac{\lambda(1-\lambda)(1-\rho)^2\tau_i\chi_i(1-\sigma^i)}{\lambda + (1-\lambda)(1-\rho)\tau_i(1-\chi_i)(1-\sigma^i)} \tag{B.22}$$

is the belief bias without communication.

We obtain similar expressions for the case where i observes $\underline{i}$'s narrative ex post, replacing the belief bias $\beta^{\varnothing}$ by

$$\beta^{\text{echo chamber}} := \frac{\lambda(1-\lambda)(1-\rho)^2\tau_i\tau_{\underline{i}}[1-(1-\chi_i)(1-\chi_i^{\underline{i}})](1-\sigma^i)}{\lambda+(1-\lambda)(1-\rho)\tau_i\tau_{\underline{i}}(1-\chi_i)(1-\chi_i^{\underline{i}})(1-\sigma^i)}. \tag{B.23}$$

The comparison of i's posterior belief bias with and without communication boils down to the comparison of expressions (B.22) and (B.23). Some algebra shows that $\beta^{\text{echo chamber}} > \beta^{\varnothing}$ if and only if $\iota(\tau_{\underline{i}}, \chi_i^{\underline{i}}) > 0$, where we define

$$\iota(\tau_{\underline{i}}, \chi_i^{\underline{i}}) := \lambda[\tau_{\underline{i}}(\chi_i + \chi_i^{\underline{i}} - \chi_i\chi_i^{\underline{i}}) - \chi_i] + (1-\lambda)(1-\rho)\tau_i\tau_{\underline{i}}(1-\chi_i)\chi_i^{\underline{i}}(1-\sigma^i).$$

If i observes $\underline{k}$'s narrative $\hat{\xi}_{\underline{k}}$ ex post, then we obtain the following expression for the belief bias:

$$\beta^{\text{agora}} := \frac{\lambda(1-\lambda)(1-\rho)^2\tau_i\tau_{\underline{k}}(\chi_i - \chi_i^{\underline{k}})(1-\sigma^i)}{\lambda\tau_{\underline{k}}(1-\chi_i^{\underline{k}}) + (1-\lambda)(1-\rho)\tau_i(1-\chi_i)(1-\sigma^i)}, \tag{B.24}$$

which is strictly smaller than the expression in (B.22). This concludes the proof.

B.9 Proof of Corollary 1

It suffices to compare the expressions in (B.23) and (B.24), which determine i's average self-esteem. Some algebra shows that $\beta^{\text{echo chamber}} > \beta^{\text{agora}}$ if and only if (we write $\underline{\tau}$ for the common value of $\tau_{\underline{i}}$ and $\tau_{\underline{k}}$)

$$\begin{aligned}&\lambda\big[\underline{\tau}(1-\chi_i^{\underline{k}})[1-(1-\chi_i)(1-\chi_i^{\underline{i}})] - (\chi_i - \chi_i^{\underline{k}})\big] > \\ &(1-\lambda)(1-\rho)\tau_i(1-\chi_i)(1-\sigma^i)[\chi_i - \chi_i^{\underline{k}} - 1 + (1-\chi_i)(1-\chi_i^{\underline{i}})].\end{aligned}$$

The right-hand side of this inequality is negative. Thus, it is sufficient that the left-hand side be nonnegative, which amounts to

$$\underline{\tau} \geq \tau^* := \frac{\chi_i - \chi_i^{\underline{k}}}{(1-\chi_i^{\underline{k}})[1-(1-\chi_i)(1-\chi_i^{\underline{i}})]}.$$

B.10 Proof of Proposition 6

We start with the case of counter-partisans. As the equilibrium analysis of Section 3 reveals, the expression that governs i's best-response is the net expected utility from denial. Note that since i's behavior is not public, we only need to compare i's expected self-image from self-deception with and without communication. Player i's expected self-image following denial equals

$$\frac{\mu_i\tau_i\lambda\rho(1-\rho)}{\lambda+(1-\lambda)(1-\rho)\tau_i(1-\chi_i)(1-\sigma^i)} \tag{B.25}$$

without communication,

$$\frac{\mu_i \tau_i \tau_{\underline{i}} \lambda \rho (1-\rho)}{\lambda + (1-\lambda)(1-\rho)\tau_i \tau_{\underline{i}} (1-\chi_i)(1-\chi_i^{\underline{i}})(1-\sigma^i)} \tag{B.26}$$

with narrative transmission from $\underline{i}$ to i, and

$$\frac{\mu_i \tau_i \tau_{\underline{k}} (1-\chi_i^{\underline{k}}) \lambda \rho (1-\rho)}{\lambda \tau_k (1-\chi_i^{\underline{k}}) + (1-\lambda)(1-\rho)\tau_i (1-\chi_i)(1-\sigma^i)} \tag{B.27}$$

with narrative transmission from $\underline{k}$ to i. It is easy to check that the expression in (B.27) is always strictly smaller than the one in (B.25), while the one in (B.25) is strictly larger than the one in (B.25) if and only if $\upsilon(\tau_{\underline{i}}, \chi_i^{\underline{i}}) > 0$, where $\upsilon(\tau_{\underline{i}}, \chi_i^{\underline{i}}) = \lambda(\tau_{\underline{i}} - 1) + (1-\lambda)(1-\rho)\tau_i \tau_{\underline{i}} (1-\chi_i)\chi_i^{\underline{i}}(1-\sigma^i)$.

Therefore, if $\upsilon(\tau_{\underline{i}}, \chi_i^{\underline{i}}) > 0$, the net benefit from denial at the habitual strategy σ^i, which is nonnegative without communication (since $\sigma^i < 1$), is positive with communication from $\underline{i}$ to i, which implies that i's best-response in that case is (weakly) lower than σ^i. If $\upsilon(\tau_{\underline{i}}, \chi_i^{\underline{i}}) < 0$ the comparative statics goes in the opposite direction. For communication from $\underline{k}$ to i, the net benefit from self-deception is smaller at every σ^i than without communication, and hence i's best-response features more realism. This concludes the proof.