Rational Silence and False Polarization: How Viewpoint Organizations and Recommender Systems Distort the Expression of Public Opinion

ATRISHA SARKAR*, Western University, Canada
GILLIAN K. HADFIELD, Johns Hopkins University, United States

Social media platforms are one of the most important domains in which artificial intelligence (AI) has already transformed the nature of economic and social interaction. AI enables the massive scale and highly personalized nature of online information sharing that we now take for granted. Extensive attention has been devoted to the polarization that social media platforms appear to facilitate. However, a key implication of the transformation we are experiencing due to these AI-powered platforms has received much less attention: how platforms impact what observers of online discourse come to believe about community views. These observers include policymakers and legislators, who look to social media to gauge the prospects for policy and legislative change, as well as developers of AI models trained on large-scale internet data, whose outputs may similarly reflect a distorted view of public opinion. In this paper, we present a nested game-theoretic model to show how observed online opinion is produced by the interaction of the decisions made by users about whether and with what rhetorical intensity to share their opinions on a platform, the efforts of viewpoint organizations (such as traditional media and advocacy organizations) that seek to encourage or discourage opinion-sharing online, and the operation of AI-powered recommender systems controlled by social media platforms. We show that signals from ideological viewpoint organizations encourage an increase in rhetorical intensity, leading to the *rational silence* of moderate users. This, in turn, creates a polarized impression of where average opinions lie. We also show that this observed polarization can also be amplified by recommender systems that, pursuant to a platform's incentive to maximize engagement, encourage the formation of viewpoint communities online that end up seeing a skewed sample of opinion. Unlike existing models, these well-known online phenomena are not here attributed to distortion in the formation of opinions nor to the seeking out of like-minded others, but rather to the interaction of the incentives of users, viewpoint organizations, and platforms implementing recommender systems. In addition to showing how these interactions can play out in simulations, we also identify practical strategies platforms can implement, such as reducing exposure to signals from ideological viewpoint organizations and a tailored approach to content moderation.

JAIR Track: AI and Society Track

JAIR Associate Editor: Toby Walsh

JAIR Reference Format:
Atrisha Sarkar and Gillian K. Hadfield. 2026. Rational Silence and False Polarization: How Viewpoint Organizations and Recommender Systems Distort the Expression of Public Opinion. *Journal of Artificial Intelligence Research* 85, Article 33 (March 2026), 29 pages. DOI: 10.1613/jair.1.20965

1 Introduction

Artificial intelligence (AI) is transforming the way communities learn about the views of their citizens. Without AI, we could not sustain the huge volume of viewpoint exchange on social media. AI-based recommender systems allow platforms to autonomously curate what will be served to an individual from a massive dynamic inventory

*Corresponding Author.

Authors' Contact Information: Atrisha Sarkar, ORCID: 0000-0003-3276-1683, atrisha.sarkar@uwo.ca, Western University, London, Ontario, Canada; Gillian K. Hadfield, ORCID: 0000-0003-4124-8187, ghadfield@jhu.edu, Johns Hopkins University, Baltimore, Maryland, United States.

DOI: 10.1613/jair.1.20965

of content, using fine-grained information about each user to curate a micro-targeted pinhole view of the world that can differ dramatically from person to person. This new landscape has transformed the public sphere. Gone is the era of one-to-many broadcasts of viewpoints by a limited number of broadcasters and publishers, replaced by a complex and dynamic media ecosystem. This transformation, mediated by recommender systems and online platforms, has important implications for our political economy. To the extent that policymakers, legislators, and courts look to social media to gauge public opinion on contentious issues, they may well be misled about where true public opinion lies. They may well come to believe that views are more polarized and extreme than they really are, and respond accordingly. This distortion in beliefs about public opinion then has the potential to distort policy and collective choice. Moreover, these dynamics also have implications for AI model training. Since large-scale internet data is commonly used to train AI models, any divergence between online and true public opinion could result in model behaviour that is similarly unrepresentative. Thus, we need more robust theoretical frameworks to understand why and how opinions expressed in the digital public sphere may differ from representative public opinion.

The phenomena of the digital public sphere have attracted a great deal of attention from researchers across various disciplines, including economics, computer science, psychology, sociology, and political science (Nyhan et al. 2023). Among these, polarization has become one of the most discussed topics in both academic and popular discourse (Arora et al. 2022; Lorenz-Spreen et al. 2023). Social media platforms that gained early prominence were applications that algorithmically connected users to friends, family, and strangers, and formed virtual communities (Aichner et al. 2021). For the most part, the analysis of polarization was based on the formation of echo chambers through these communities, patterns of online interactions that reinforce users' beliefs and opinions by isolating them from opposing points of view (Garimella et al. 2018; Golub and Jackson 2012). However, simplistic attempts to mitigate this problem by exposing people to opposing viewpoints have been shown to make the problem even worse by increasing polarization (Bail et al. 2018). Alternate models have shown that polarization, especially the kind in which people form contradictory beliefs when presented with the same factual information, can be a rational phenomenon based on differences in subjective past experiences (Haghtalab et al. 2021; Singer et al. 2019). And a large-scale experiment on the Facebook platform demonstrated that increasing the share of a user's feed that comes from like-minded people and publishers did not significantly affect any measures of polarization (Nyhan et al. 2023). Overall, we now view polarization on online social platforms as a far more complex phenomenon than initially hypothesized.

The shift toward a more nuanced understanding of polarization has taken several forms. First, there is an increasing focus away from issue-based polarization to affective polarization; the former refers to a shift over time in individual opinions to more extreme views, whereas the latter refers to deep-seated partisan animosity and increased hateful rhetoric toward opposing viewpoints (Iyengar et al. 2019). There is a rise in affective polarization not only in the political sphere but also in more general social issue discourse, and not only in the United States but worldwide (Boxell et al. 2022; Yarchi et al. 2021). Second, related to partisanship, there is growing attention on the role of key political entities on online platforms, such as partisan news media organizations and political influencers, who can broadcast perspectives to the public, thus exacerbating affective polarization. For example, between 2016 and 2020, influencers in the United States became both more political and more polarized. Similar trends have been observed in other countries, such as India (Dash et al. 2022). Given the well-documented impact of partisan media in traditional news on electoral outcomes (DellaVigna and Kaplan 2007), the influence of ideological media organizations on online platforms is an increasing concern. Third, there has been a shift from viewing polarization as solely an individual phenomenon to recognizing it as a population-level dynamic. While the former focuses on the factors that might cause an individual to shift their position to more extreme views, the latter examines the mechanisms of polarization operating at the population level. This broader perspective helps uncover phenomena such as false polarization or the perception gap, where differences in who expresses opinions versus who remains silent create the impression that the population is more polarized than it actually is. In fact,

the more partisan views are, the greater the misperception about the opinions of the out-group (Levendusky and Malhotra 2016; Yudkin et al. 2019). Reducing this misperception has shown promise in reducing affective polarization (Lees and Cikara 2020). Social recommendation systems are at the heart of these issues and play a key role in matching users with content to increase platform participation and engagement. As our understanding of these complex social processes evolves, we need improved formal frameworks to model social dynamics, from individual to population-level effects, and to develop strategies for better online platform design.

In this paper, we present a formal framework that models the interplay between these dynamics at the levels of individual users on platforms, various media organizations, and social recommender systems. From the perspective of individual users, we develop a game-theoretic model in which users strategically decide on the rhetorical intensity with which they will express their opinions. This decision is based not just on their personal opinion on a particular issue but also on their beliefs about the opinions and the rhetorical intensity they expect from out-group and in-group members in the community. We specifically model a user's costs and benefits of other's rhetoric: intense rhetoric from an in-group member boosts the utility a user expects from expressing their own opinion while intense rhetoric from an out-group member dampens their own utility. Using this model, we show that, in equilibrium, people with more extreme opinions are more likely to use intense rhetoric, while those with moderate views are more likely to remain silent, a phenomenon we refer to as *rational silence*. One implication of this model is that even when individuals' opinions do not change over time, the population can appear more polarized due to the suppressive effect increasing rhetorical intensity has on moderate opinion holders, leading to false polarization. At the organizational level, we integrate the individual opinion expression model into a model of *viewpoint stewarding* – a process by which media organizations deliberately shape long-term beliefs about out-group and in-group opinions within a community. We refer to media organizations that engage in this process as viewpoint organizations. We consider two types of viewpoint organizations: participatory organizations, which aim to maximize the percentage of users who express their opinions on a platform, and ideological organizations, which aim to shift public opinion to their preferred position and distort the perceived average opinion. We elucidate the optimal strategies for each type of organization and show that ideological organizations achieve their goals by making a community believe that the out-group is more extreme than it actually is. Finally, from the perspective of a platform's social recommender system, which seeks to maximize individual engagement, the platform adapts to users by selecting content from either ideological or participatory organizations. We show that this dynamic results in the stratification of the population into distinct organizational communities with different characteristics. Participatory communities have a greater representation of moderate opinion holders with moderate second-order beliefs about out-group and in-group opinions. In contrast, ideological communities are characterized by more extreme opinions and second-order beliefs, while the silent population tends to have moderate own opinions but extreme second-order beliefs about the opinions of others.

The rest of the paper is organized as follows. In Sec. 3, we begin with the model of the game that captures individual opinion expression and rational silence. In Sec. 4, we develop a model of viewpoint steering involving a single organization within a population. In this section, we derive the optimal signaling policies for the two types of viewpoint organizations, participatory and ideological. In Sec. 6, we extend this framework to an environment with multiple viewpoint organizations, a population with heterogeneous beliefs, and a setting where the recommender system of an online platform mediates the organizational stewarding process. We demonstrate the formation of organizational communities and discuss the opinion and belief characteristics of each community. Finally, we conclude by presenting policy implications for mitigating the population polarization identified in our model.

2 Related Work

Polarization as a phenomenon of politics of the 21st century has been one of the most widely discussed topics in the academic literature in the fields of economics (Boxell et al. 2017; Levy and Razin 2019), psychology (Jung et al. 2019), political science (Hare and Poole 2014), computer science (Lim and Bentley 2022), communication studies (Kubin and Sikorski 2021), and law (Fagan 2017). The literature is vast, and in this section, we present literature that is closest to the questions addressed in this paper, namely, models of individual-level and affective polarization, models of self-censorship and silence, and the role of media and recommendation algorithms in online social systems. For a more general coverage of social media and polarization, we refer to existing systematic reviews of the literature (Arora et al. 2022; Bramson et al. 2017; Kubin and Sikorski 2021; Tucker et al. 2018; Van Bavel et al. 2021).

Models of Individual-Level Polarization. There is a well-established body of literature focused on empirical studies identifying polarization on social networks (Bakshy et al. 2015; Cinelli et al. 2021; Tucker et al. 2018). However, understanding the mechanisms behind the complex social phenomenon of polarization of the public sphere through computational models is essential to explore interventions to mitigate the problem. DeGroote (DeGroot 1974) provided one of the earliest models of how the opinion of others influences an individual's opinion, and Golub and Jackson (Golub and Jackson 2012) applied a similar model to demonstrate how homophilic networks (a preference for associating with individuals who share similar opinions or beliefs) lead to belief and opinion polarization. However, the main focus of opinion dynamics models (see (Xia et al. 2011) and (Peralta et al. 2022) for an overview) is on how a network and interaction structure affect opinion change over time. In our model, the opinion of an individual stays fixed, and only the observation of the publicly expressed opinion at the population level changes over time. Polarization resulting from opinion dynamics over a network is mediated through the formation of filter bubbles (Pariser 2011) and echo chambers, and consequently, concerns about the rise of polarization in broader public discourse have led to empirical approaches to their detection in online social systems (Alatawi et al. 2021; Bakshy et al. 2015; Nguyen et al. 2014), and proposed models to break that effect (Helberger et al. 2018; Li et al. 2023). However, studies such as those of Bail et al. (Bail et al. 2018) also show that presenting people with information that they perceive to contradict their deeply held beliefs can result in further polarization.

Given the mixed evidence on interventions to reduce polarization based on filter bubbles, recent literature has followed two different paths of analysis of polarization. One looks at better mechanistic models to capture individual-level polarization, specifically, the question of why presenting people with the same information can lead to two separate conclusions. Within this branch of models, some are based on a cognitive process of motivated reasoning (Jost et al. 2022), and others are based on a rational choice-based model (Dorst 2023). The explanation of polarization through rational choice-based models predates social media; for example, Sunstein (Sunstein 1999) identifies social comparison (desire for favorable perception by one's ingroup) and subjective persuasion as two mechanisms that can lead to polarization. Recent literature often relies on differences across individuals based on their past experiences, and examples of specific modelling approaches include an agent-based approach (Singer et al. 2019), Bayesian approach (Jern et al. 2014), and learning theoretic approach (Haghtalab et al. 2021). In our model, although there is a similarity in the modelling paradigm of rational choice and its connection to polarization, the process that leads to polarization in our model is fundamentally different. Whereas the above literature provides a rational explanation of polarization through individuals coming to different conclusions, changing opinions, or rejecting factual information, our model identifies an individual's strategic decisions whether and how intensely to express their opinion as a key factor contributing to polarization.

Models of Affective Polarization. A second branch of the literature moves the focus from opinion polarization at the individual level to affective polarization (Iyengar et al. 2019) where the focus is on the extend of out-group

animosity. The dominant narrative around the rise of affective polarization appeals to social identity theory (Iyengar et al. 2019). The theory suggests that individuals increasingly perceive their primary identity along partisan lines, and the increase in affective polarization is a consequence of a cluster of cognitive processes rooted in-group identity-based social psychology. A smaller line of work uses rational choice-based models of affective polarization seek to provide a theoretical explanation for *why* such a phenomenon occurs. For example, based on national election data from four countries, Algara and Zur (Algara and Zur 2023) show that a Downsian model (Downs 1957) of strategic behaviour of voters based on ideological positions explains affective polarization more than partisan identification. As another example, Yaouanq (Le Yaouanq 2018) models ideological disagreement as arising from rational choice about how to interpret ambiguous evidence in the context of motivated beliefs based on preferred policy outcomes.

Models of Self-Censorship. The *spiral of silence* (Noelle-Neumann 1974) is a widely discussed theory in communication studies in which minority opinion holders' fear of isolation drives them into self-reinforcing silence, leading to an absence of minority views from the population. Since the original spiral of silence model predates social media, this framework has seen growing interest in the context of online opinion expression, specifically on the question of whether the model can explain self-censorship in social media. As an answer, a meta-analysis of sixty-six studies by Matthes et al. (Matthes et al. 2018) shows that a spiral-of-silence-like effect linking perceptions about opinions and the willingness to express own opinions can be established empirically in online social systems, too. A closely related work on the computational modelling of dynamics leading to a spiral of silence is by Gaisbauer et al. (Gaisbauer et al. 2020). Similar to our work, Gaisbauer et al. build a microfoundational account of opinion expression based on a game-theoretic model with incentives to stay silent or express opinion conditioned on the expression of other agents. However, there are a few key differences in the modelling constructs between our model and the Gaisbauer et al. model. First, the Gaisbauer et al. model uses a network structure and intensity of connection between agents on that network as the main factor that determines agents' perception of others' opinions. In comparison, viewpoint organizations play that role in our model, which helps us analyze a media organization's optimal strategies and how such organizations can distort beliefs. The second difference is that in Gaisbauer et al. (Gaisbauer et al. 2020), opinion expression is discrete, that is, individuals either choose to express or stay silent, whereas, in our model, we capture this process through a choice of rhetorical intensity on a continuous spectrum. This enables us to establish a relation between an ideological opinion and the rhetorical intensity, both of which can lie on a spectrum. Our approach also connects the spiral of silence phenomenon directly to affective polarization.

Media and Online Platforms. Media and its role in polarization has been well studied in the economics and political science literature (Prior 2013). The implications of the interaction between users' strategic information sharing and social media platform incentives have been developed in the context of misinformation sharing and fake news (Acemoglu et al. 2023; Hsu et al. 2020; Papanastasiou 2020). Although there are some similarities to our work, particularly in the use of game-theoretic models of interaction, such as Acemoglu's (Acemoglu et al. 2023) model, which frames user content sharing as a game of strategic complements, there are important qualitative differences. Unlike the focus on information reliability and misinformation, our model examines the interplay between private opinions, rhetorical choices, and their impact on publicly observed opinions.

Social Recommender Systems and Polarization. Approaches that explicitly model the algorithmic effect of social recommender systems often focus on the network topology, mainly how these systems help form homophilic networks by recommending connections between individuals with similar opinions. Musco et al. (Musco et al. 2018) propose a novel metric for recommender systems aimed at balancing the reduction of polarization (by connecting users with dissimilar opinions) without increasing the risk of disengagement due to disagreement. While homophilic community formation is one analysis, another set of models looks at how these connections

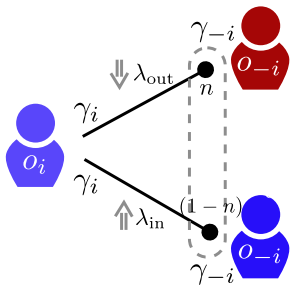

Fig. 1. Each player with a private opinion o_i is matched with another player drawn from an opinion distribution f. With probability n they are matched with someone from their in-group and probability $(1-n)$ from out-group. Each player strategically chooses their rhetorical intensity γ_i, γ_{-i}. If the player is matched with the in-group, utility is boosted by a factor $\lambda_{\text{in}} > 1$ and, conversely, utility is dampened by a factor $\lambda_{\text{out}} < 1$ when matched with an out-group member.

influence opinion dynamics over time. For example, Morales and Cointet (Ramaciotti Morales and Cointet 2021) have examined the interplay between recommendation systems and models of opinion dynamics. They combine network-based recommender systems with a DeGroot opinion dynamics model (DeGroot 1974). Using simulation data based on Twitter interactions among French Members of Parliament, Morales and Cointet find that different recommendation algorithms impact polarization differently. Specifically, they show that the Alternating Least Squares (Hu et al. 2008) and Bayesian Personalized Ranking (Rendle et al. 2012) algorithms increase polarization, while the Logistic Matrix Factorization algorithm (Johnson et al. 2014) reduces it. Adding to this line of approaches, Santos et al. (Santos et al. 2021) demonstrate that structural similarity—recommendations based on common neighbors in a network—can also exacerbate polarization. While these existing models provide valuable insights into how recommender systems contribute to polarization through opinion similarity and structural dynamics, they do not account for the strategic behaviours of users and the fact that non-engagement of users with the recommender systems itself can have a negative externality of polarization. Our work advances the latter line of research by incorporating the strategic behaviour of users that models silence and expression, as well as the nested incentives of both platforms and media organizations.

More recently, researchers have revisited the question of polarization in the context of decentralized social media platforms such as Mastodon, Gab, and Bluesky. La Cava et al. conduct a network analysis of Mastodon and identify four main "poles" or instances within the platform. Their findings suggest that moderation efforts primarily target instances dominated by bots or those explicitly designed to host controversial content (La Cava et al. 2024). Unlike the opinion-based polarization commonly observed on large centralized platforms, polarization on Mastodon appears to be shaped more by the governance rules of the individual instances. Left-leaning and right-leaning platforms, Bluesky and Gab respectively, also exhibit distinct characteristics. Acampa et al. find a higher proportion of conflict-driven and emotionally charged expression in Gab compared to the centralized platform of Facebook (Acampa et al. 2023). In contrast, Quelle and Bovet observe that while Bluesky generally exhibits lower levels of polarization on broad topics, significant division emerges around specific issues such as the Israel–Palestine conflict (Quelle and Bovet 2025).

3 Opinion Expression Game and Rational Silence

In this section, we present the opinion expression game that models users' expression of opinion on a given issue. We assume a context in which there is a matter of public debate in front of the community, such as whether to adopt a legislator's proposed policy or how a supreme court should resolve a case it is scheduled to decide. Individuals hold a spectrum of opinions on this issue but can be grouped based on a binary partition: approval

(the policy should be adopted, the court should hold in favor of the plaintiff) and disapproval (the policy should be rejected, the court should hold in favor of the defendant). We model individuals' choices about whether to express their opinion on the matter or remain silent as an iterated game in which individuals in each iteration are randomly matched with someone else from the community. This is a stylized representation of online interactions where users engage with a social media platform periodically and primarily interact with other users who are on the platform at (roughly) the same time. In any iteration, an individual may find themselves matched with someone in their 'in-group'who shares their binary position (approval or disapproval) on the question or 'out-group', who holds the opposite position. First, we present the main constructs of the game, followed by an equilibrium analysis that determines individual behaviour.

3.1 Opinions

Each individual, indexed as i, holds private opinions $o_i \in [0, 1]$ on a focal issue. Using game-theoretic terminology, the opinion of the player i is their *type*, which is private information. The opinions are drawn from an arbitrary opinion distribution $f_{(o_A, o_D)}$, where o_A, o_D are the mean approval and disapproval opinions on which the distribution is parameterized. The values $o_i < 0.5$ indicate *disapproval* on the focal issue and $o_i \geq 0.5$ indicates *approval*. Opinions lying on the continuum between 0 and 1 represent the diversity of opinions that go beyond a simple yes/no dichotomy. At the two extremes, an opinion of $o_i = 1$ or $o_i = 0$ denotes complete support for or against the matter under debate, respectively: support for all elements of a proposed policy, for example, or opposition to a finding for the plaintiff on any possible grounds in a court case. We then conceptualize an opinion between 0 and 1 as an indication that on average an individual may support (or not support) a particular outcome, but were a narrower question posed, their position might change. For example, suppose the matter under debate is whether animal testing of products should be prohibited. Someone might support a proposed total ban even though they would prefer that testing be allowed in a narrow set of cases for the development of life-saving drugs. The opinion o_i can be thought of as capturing the strength of approval or disapproval with respect to a particular contentious question. Note that unlike most work in the literature on social media, we assume people's opinions on a matter are fixed and not subject to influence or change.

Table 1. Parameters that define the interaction between two players within a group.

Parameters	Parameter type	Description
$\hat{n}, o_A, o_D$	Descriptive belief	Estimated proportion of the population that holds approval opinions, the estimate of the mean approval and disapproval opinion, respectively.
γ_{-i}	Strategic choices	Rhetorical intensity of the non-focal player
γ_i		Rhetorical intensity of the focal player
o_i	Private information	Opinion of the focal player
α	Exogenous parameters	Cost of using rhetoric to express opinions, including possible penalties imposed by platform's moderation policy
λ_{in}		Utility boosting factor from agreeing with in-group
λ_{out}		Utility reduction factor from disagreement with out-group

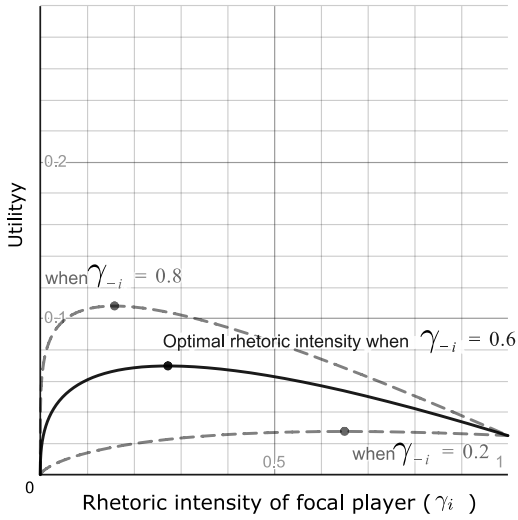

Fig. 2. The net utility of expressing an opinion (after deducting the cost) as a function of rhetorical intensity for the focal player. The utility-maximizing rhetorical intensity of the focal player decreases with an increase in other players' rhetorical intensity.

3.2 Rhetorical Intensity

Whereas opinion represents an individual private position on an issue, *rhetorical intensity* refers to *how* one chooses to express it. Continuing on the previous example, an individual who has absolute support for a total ban on animal testing ($o_i = 1$) and who chooses to express this opinion can choose to express themselves mildly, with a simple statement of their view. Or they could choose to express their opinion in a highly partisan way with verbal abuse of those holding an opposing view. We model this choice as a continuous variable $\gamma_i \in [0, 1]$. Values close to 0 represent remaining silent, low values represent a mild expression of one's opinion, and values close to 1 represent a threat of violence or other forms of extreme behaviour.

3.3 Cost and Utilities

The opinion expression game is played between an individual, indexed as i, and a randomly selected member of the community, indexed as $-i$. For purposes of our simulations we construct a utility function that captures the following ideas. First, a person experiences utility directly from expressing their opinion and this scales with the strength of their agreement with the proposal under debate. So, a person who is one-hundred percent in agreement with the question ($o = 1$) gets more utility from expression than someone who disagrees with some of the fine-grained elements of the matter under debate but is still overall supportive. (Using our earlier example, someone who supports a total ban on all animal testing gets more utility from expressing that view than someone who would prefer a policy that allowed exceptions for the development of life-saving drugs but will still choose to support rather than oppose the proposed total ban.) We also assume that a person's utility from expression is increasing in the level of rhetoric they use. Next, we take into account that the utility from expressing one's opinion depends also on whether one is matched with someone who is in overall agreement on a particular question (a member of one's in-group) or someone with whom one disagrees (a member of one's out-group.) Specifically, utility from opinion expression is increased (decreased) when one exchanges views with a member of the in-group (out-group). Finally, we assume that there is a positive externality associated with the rhetoric used by members of one's in-group. This affects an individual's own rhetoric level because we also assume that it is

costly to use rhetoric, and increasingly so as the intensity of rhetoric increases. This implies that a user can enjoy the same level of utility from opinion expression at lower levels of own rhetorical intensity when interacting with someone who engages in more intense rhetoric to express the same (similar) viewpoint. We capture these ideas in the following utility function shown here for a focal individual with an approval opinion:

$$
\begin{aligned}
u_i(\gamma_i, \gamma_{-i}; o_i) = \quad & \underbrace{\hat{n} \cdot o_i \cdot \lambda_{\text{in.}} \cdot \gamma_i^{(1-\gamma_{-i})}}_{\substack{\text{Utility from expression when matched} \\ \text{with in-group member}}} \\
& \underbrace{+(1-\hat{n}) \cdot o_i \cdot \lambda_{\text{out}}^{\gamma_{-i}} \cdot \gamma_i}_{\substack{\text{Utility from expression when matched} \\ \text{with out-group member}}} \\
& - \underbrace{\alpha \cdot \gamma_i}_{\text{Cost of expressing opinion}}
\end{aligned}
\tag{1}
$$

and for individuals with opinions of disapproval as:

$$
\begin{aligned}
u_i(\gamma_i, \gamma_{-i}; o_i) = & \underbrace{(1-\hat{n}) \cdot (1-o_i) \cdot \lambda_{\text{in.}} \cdot \gamma_i^{(1-\gamma_{-i})}}_{\substack{\text{Utility from expression when matched} \\ \text{with in-group member}}} \\
& \underbrace{+\hat{n} \cdot (1-o_i) \cdot \lambda_{\text{out}}^{\gamma_{-i}} \cdot \gamma_i}_{\substack{\text{Utility from expression when matched} \\ \text{with out-group member}}} \\
& - \underbrace{\alpha \cdot \gamma_i}_{\text{Cost of expressing opinion}}
\end{aligned}
\tag{2}
$$

where $\hat{n}$ is the proportion of the population with approval opinions. $\lambda_{in} > 1$ and $\lambda_{out} < 1$ are boosting and reducing constants that represent the increase and decrease of utility from being matched with the in-group or out-group, respectively. γ_i and γ_{-i} are the rhetorical intensity of expression of opinion for the focal and non-focal player, respectively. Note that we model out-group rhetorical intensity as a factor that increases the extent to which the utility enjoyed from own-opinion expression is dampened by being matched with someone with whom one disagrees. α is a cost factor that captures any costs incurred by rhetorical intensity. This could include personal psychological or time costs as well as costs imposed by a social media platform such as a flag, demoting a post in a newsfeed, or limiting a user's access to the platform. We thus also allow α to be interpreted as a pathway for the platform for imposing cost on the user through their moderation policy, although α will still generally be positive in the absence of platform moderation.

3.4 Optimal Rhetoric

The utility function given in Equations 1 and 2 is concave in γ for a focal individual and this means that we can solve for an optimal rhetorical intensity for this individual. Figure 2 shows this relationship. Rhetorical intensity of individual i is shown on the x axis and the y axis represents total utility. The optimal rhetorical intensity also depends on the rhetorical intensity of the matched player's expression of opinion (γ_{-i}), shown in the figure as the three separate utility curves. The optimal rhetorical intensity of the focal player i decreases with an increase in the other player $-i$'s rhetorical intensity. When matched with a member of the in-group, this effect arises because the other's rhetoric is a substitute for own rhetoric. When matched with a member of the out-group, this

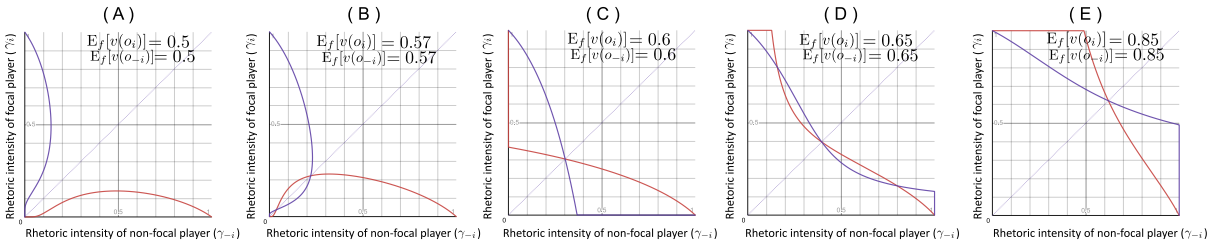

Fig. 3. Best response curves of the optimal rhetorical intensity for the focal (shown in red) and non-focal (shown in blue) player as a function of their opinion drawn from an opinion distribution of equal support for approval and disapproval. For the purpose of illustration, the constants $\alpha, \lambda_{\text{in}}, \lambda_{\text{out}}$ are set to 0.7, 2, and 0.5, respectively. The intersection of the two curves represents the Bayesian Nash equilibrium rhetorical intensity in the *ex-ante* form. The plots are shown for different opinions (types) of the focal and non-focal player, and we see that as the opinion moves to the extreme, the symmetric equilibrium rhetorical intensity becomes higher.

effect is because extreme rhetoric from the other side increases the extent to which utility from own expression is dampened.

3.5 Equilibrium

We can now calculate the equilibrium rhetorical intensity that determines the outcome of the game. In order to facilitate our analysis, we focus on the symmetric Bayesian Nash equilibrium in continuous strategies. Within Bayesian Nash equilibrium, there are two possible forms of analysis; in the *ex-ante* form, the analysis is based on the expected utility (over types) for each player, and in the *ex-interim* form, each player is aware of their own type, and therefore, the players create their own view of the game in which their utility depends on their type and the utility of the other player is an expected utility over all possible types. In this section, we focus the analysis on the former form, and we rely on the latter form and the connection between the two for simulation in Sec. 5.

Note that the opinion lies in the interval $[0, 0.5)$ for the disapproval group and in the interval $[0.5, 1]$ for the approval group. In order to make the value of expressing the opinion symmetric between the groups, we introduce the variable $v(o_i)$ such that $v(o_i) = o_i$ when $o_i \geqslant 0.5$ and $(1 - o_i)$, otherwise. Next, we consider this value of the opinion ($v(o_i)$) of each individual as their *type* as in standard game-theoretic terminology. One can treat the type here as capturing the extremity of the opinion holder; moderate opinion holders will have a lower type, and extreme opinion holders will have a higher type. Finally, the formal form of the *ex-ante* utility of a representative individual under equal support of approval and disapproval group size (i.e, $F(0.5) = 0.5$ for opinion distribution f) is given by the following equation (Lemma 1, *c.f* Appendix):

$$\begin{aligned}\mathrm{E}_f[u_i(\gamma_i, \gamma_{-i})] =& \hat{n} \cdot \mathrm{E}_f[v(o)] \cdot \lambda_{\text{in.}} \cdot \gamma_i^{(1-\gamma_{-i})} \\ &+ (1 - \hat{n}) \cdot \mathrm{E}_f[v(o)] \cdot \lambda_{\text{out}}^{\gamma_{-i}} \cdot \gamma_i - \alpha \cdot \gamma_i\end{aligned}$$

The best response of a representative player i to another representative player $-i$ can be found by taking the partial derivative of the above utility function with respect to γ_i. Similarly, the best response of a player $-i$ to a player i can be found by taking the partial derivative with respect to γ_{-i}. This gives us the best response curves for i and -i:

$$BR_i(\gamma_{-i}; o_i) = \min\left(1, \max\left(0, \left(\frac{\hat{n} \cdot \mathrm{E}_f[v(o_i)] \cdot \lambda_{\text{in}} \cdot (1 - \gamma_{-i})}{\alpha - (1 - \hat{n})(\mathrm{E}_f[v(o_i)]\lambda_{\text{out}}^{\gamma_{-i}})}\right)^{\frac{1}{\gamma_{-i}}}\right)\right) \tag{3}$$

$$BR_{-i}(\gamma_i; o_{-i}) = \min\left(1, \max\left(0, \left(\frac{\hat{n} \cdot \mathrm{E}_f[v(o_{-i})] \cdot \lambda_{\text{in}} \cdot (1-\gamma_i)}{\alpha - (1-\hat{n})(\mathrm{E}_f[v(o_{-i})]\lambda_{\text{out}}^{\gamma_i})}\right)^{\frac{1}{\gamma_i}}\right)\right) \tag{4}$$

The Bayesian Nash equilibrium of an imperfect information game assigns strategies for every possible *type* of a player. This means that in our case, there is an equilibrium rhetorical intensity as a function of the opinion value $v(o)$. We construct this function by solving the set of Equations 3 and 4 based on the symmetric equilibrium condition $BR_i(\gamma^*_{-i}; o_i) = BR_{-i}(\gamma^*_i; o_{-i})$ where $(\gamma^*_i, \gamma^*_{-i})$ represents the pair of equilibrium rhetoric intensities. Other than the parameters $\hat{n}$ and α, which are treated as constants for the purposes of solving for the equilibrium, the solutions to these equations are a function of the expected opinion value $v(o)$.

Fig. 3 shows the plot of the two best response curves (the focal player is shown in red, the best response curve of the non-focal player is shown in blue), the equilibria (intersection of the two curves), and how the equilibria change with increasing opinion value. Consider first optimal rhetorical intensity for the most moderate opinions ($\mathrm{E}_f[o_i] = 0.5$) (Panel A). For individuals with such opinions, if they anticipate engaging with someone who does not express an opinion ($\gamma_{-i} = 0$) their best response is to also remain silent (γ_i). This is attributable to the low rewards for opinion expression for moderates. When the cost of expression (mediated by α) is sufficiently high, these rewards do not warrant expression when there is no boost from being matched with someone who shares their views and whose rhetoric confers additional utility on the focal player. That boost occurs as the rhetorical intensity increases for the non-focal player, but it is counterbalanced by the possibility that the non-focal player is from the out-group and instead of a boost there is a dampening of the returns to opinion expression. We see that these combined effects initially lead the focal player to incur the cost of mild rhetoric but as the rhetorical intensity of the partner increases, the dampening effect comes to dominate and optimal rhetoric for the moderate focal player drops again, ultimately again inducing silence ($\gamma_i = 0$) as the partner's rhetoric reaches an extreme ($\gamma_{-i} = 1$). As opinions become more extreme, however ($\mathrm{E}_f[v(o_i)] = 0.6$, Panel C), the optimal response of a representative player to someone who is expected to remain silent is to use significant rhetoric to express themselves ($\gamma_i \approx 0.4$): there is no risk of bearing the cost of rhetoric from an out-group member. But as the rhetoric from the matched player increases, that cost reduces the expected return to the focal player's rhetoric and for a fixed cost of rhetoric, the optimum decreases. With very high rewards to opinion expression (Panel D), the focal player engages in maximal rhetoric until facing moderately intense rhetoric from the partner, at which point their own rhetoric begins to moderate.

Focusing on the symmetric equilibria, i.e equilibria in which the optimal rhetorical intensity of the focal player and the non-focal player are the same ($\gamma^*_i = \gamma^*_{-i}$), we see that with increasing opinion values o_i, the equilibrium rhetorical intensity becomes higher. This is because at more extreme opinion values, the higher utility generated by more extreme opinions offsets the higher cost incurred from increased rhetorical intensity for both players. The following theorem captures these results.

Theorem 1. *Let $v(o_i) = o_i$ if $o_i \geqslant 0.5$ and $v(o_i) = 1 - o_i$, if $o_i < 0.5$ and $o_i \sim f$ is drawn from any arbitrary prior opinion distribution with p.d.f f and c.d.f F, and $\hat{v}_o = \mathrm{E}_f[v(o_i)]$. Then, under the condition of equal support of approval and disapproval, i.e., $F(0.5) = 0.5$, there exists an ex-ante symmetric Bayesian Nash equilibrium $\gamma^*_i = 0, \forall i$, if and only if $\hat{v}_o < \frac{\alpha}{1-\hat{n}\cdot(1-\lambda_{in})}$*

Proof in Appendix A.. □

This theorem states that individuals with opinions that generate utility below a threshold, will choose to stay silent ($\gamma = 0$). This threshold increases as the cost of expressing opinions (α) increases, individuals' beliefs about the fraction of the population that shares their opinion (their in-group) rises, and the boost individuals get from sharing opinions with their in-group (λ_{in}) shrinks. Conversely, those with more extreme views, above this threshold, will express their views. This leads to a predictable pattern of online opinion sharing being biased to

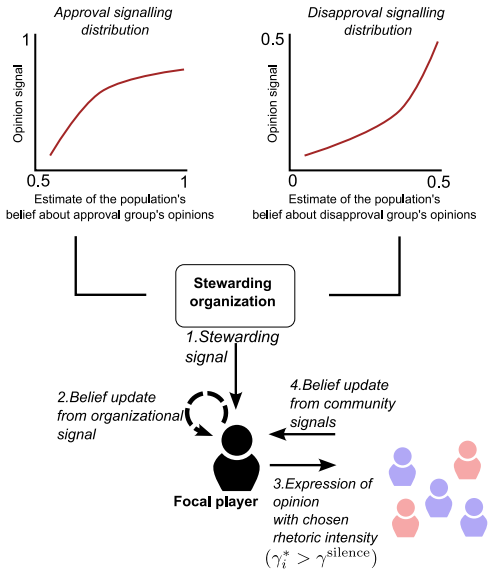

Fig. 4. Schematic representation of the organizational stewarding process. (1) An organization randomly samples a signal that conveys information about approval or disapproval group's opinions from the organization's signaling scheme. (2) The focal agent updates their descriptive belief based on the signal (3)). The focal agent estimates the equilibrium rhetorical intensity and expresses opinions. (4)) A community of agents also expresses opinions based on organizational signals. The focal agent and others updates descriptive beliefs based on the expressed opinion in the community.

those with more extreme views and is a theoretical demonstration of false polarization, with expressed views that are not representative of the true distribution of opinions in the population. We emphasize that this result holds even in the absence of platform moderation.

4 Organizational Stewarding

The equilibrium analysis presented in the previous section is still based on a static game, meaning that we establish a one-shot interaction in which the expressed rhetorical intensity is a function of the opinions players expect others to hold and the rhetorical intensity with which those players are expected to express their views. In this section, we present a higher-level dynamic that models the belief stewarding process, whereby individuals choose actions based on the static opinion expression game, and *viewpoint organizations*, which is a special category of users within the platform, can influence the beliefs in the game by conveying strategically relevant information, that is, information about the in-group and out-group's opinions. We show that based on their own long-run objectives, viewpoint organizations can use their coordination capacity to sway publicly expressed opinions by shaping who expresses their opinion and who stays silent.

In-group and out-group signaling. The nature of the viewpoint organizations we consider in our model is those that convey information about the population's opinions. Traditional media, such as newsprint and talk radio, as well as new media organizations, such as political influencers and independent online media, fall into this category. We model the actions of these organizations as a set of signals, one conveying information about the population that approves a particular position, and another about the population that disapproves of that position. As a practical matter, these signals are generated by actions such as the choice made by an organization

about who to interview for a particular story or whose opinion to share on social media, or what terms to use to describe a debate. These choices produce slant and bias (Mullainathan and Shleifer 2005) in the content of the reporting that conveys information about each of the positions of the group.

Signaling policy and bounded confidence constraints. By adding their own slant to a story that conveys information about a population's opinion, media organizations can generate biased signals that differ from the true opinion of the two groups (Rodrigo-Ginés et al. 2023). We refer to this slant and bias as the organization's signaling policy and can be represented as the distribution $\pi(o_I|\hat{o})$, where o_I is the organization's signal and $\hat{o}$ is the organization's estimate of the true approval or disapproval opinion. Even though an organization can bias its signaling policy, the generated signals cannot be arbitrary; rather, those signals need to convey some meaningful information about the population. Based on the Bounded Confidence model of Hegselmann and Krause (Rainer and Krause 2002), we model this requirement with the help of a constraining distribution C that connects the generated signal and the estimated true belief. The constraining distribution C models the random variable of the absolute difference between the signal and the beliefs. We choose C as a uniform distribution in the domain $\hat{o} \pm \frac{\tau}{2}$ and 0 outside of this range. Although we choose a specific distribution for our analysis, without loss of generality, one can choose an alternate distribution that better reflects the relation between the organization's signaling constraint and the true beliefs. The only requirement in the organizational stewarding process is that the signal should convey *some* information and not just be pure noise.

4.1 Organizational Stewarding Sequence

The organizational stewarding process involves repeated cycles of signal generation by the organization acting as sender and belief update by community members acting as receivers. Fig. 4 shows the sequence of steps involved in one cycle, and we describe the sequence in more detail below.

(1) *Organizational signal generation (stewarding signal):* The organization holds prior belief about the mean opinion of the approval and disapproval group. It generates a signal about the approval and disapproval group's opinion from two separately chosen distributions. Each of these distributions is conditioned on the organization's estimate of the groups' true mean opinions. We elaborate upon the choice of the signal generating distribution from the organization's perspective in the next section (Sec. 4.2)
(2) *Belief update (from organizational signal):* When a player receives the signal, they interpret the signal as one about their *in-group* or *out-group* based on their own opinion o_i, and update their descriptive belief about the corresponding group based on Bayes rule as follows:

$$f_{\text{posterior}}(\hat{o}_{t+1}|o_{I,t}) = \frac{C(o_{I,t};\hat{o}_t,\tau)\cdot f_{\text{prior}}(\hat{o}_t)}{\int C(o_{I,t};\hat{o}_t,\tau)\cdot f_{\text{prior}}(\hat{o}_t)\,d\hat{o}_t} \tag{5}$$

where $C(o_{I,t};\hat{o}_t,\tau)$ is the likelihood of the organization generating the signal based on the constraining distribution C and $f_{\text{prior}}(\hat{o}_t)$ is the prior belief about the corresponding group at time-step t. $f_{\text{posterior}}(\hat{o}_{t+1}|o_{I,t})$ is the posterior belief about the group after the receiver updates their belief based on the signal. The limits of the integral in the marginal are [0.5,1] or [0,0.5] depending on whether the signal is being generated for the approval or disapproval group, respectively.
(3) *Opinion expression:* Based on the posterior beliefs about the opinions, players estimate the equilibrium rhetorical intensity of the group and then best respond correspondingly. We rely on the equivalence between *ex-interim* and *ex-ante* Bayesian Nash equilibrium to simulate this process (Fujiwara-Greve 2015). Players first estimate the symmetric *ex-ante* equilibrium based on the belief about the mean opinions of the groups. Let $\gamma^*_{ex\text{-}ante}$ denote that value. Subsequently, each player best responds following Eqn. 3 as $BR_i(\gamma^*_{ex\text{-}ante};o_i)$. This response,

which simulates the individual, is the *ex-interim* response since own opinion (type) is known to the player. Next, individuals with rhetorical intensity less than a threshold γ^{silence} stay silent, whereas others express their opinions.

(4) *Belief update (from community signal):* Based on the mean disapproval and approval of the expressed opinions, everyone in the population (both those who expressed and who stayed silent) update their beliefs using Bayes rule. The update is similar to Eqn. 5, but this time, these signals come from the community expressing their opinions.

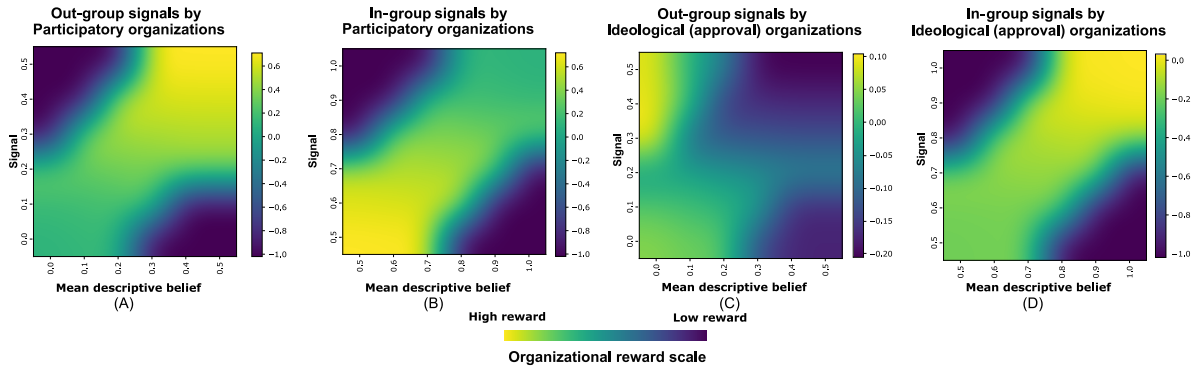

Fig. 5. Reward heatmap that shows the optimal signaling policy for different viewpoint organizations and for out-group and in-group signals. Brighter regions (yellow) represent higher reward. The y axis shows the organizations' estimate of the true out-group/in-group belief and the x axis shows the corresponding signals to generate. We see that the optimal signaling distribution follows different patterns for both organizations: ideological organizations get higher reward for more extreme signals about the groups' opinions whereas participatory organizations get higher reward from moderate signals.

4.2 Participatory and Ideological Organizations

In this section, we focus on step 1 of the organizational stewarding process, that is, the optimal signal generation on the part of the organizations. We analyze how different organizational incentives shape the signaling structure, and the effects of that on the beliefs and subsequent effect on the expression of opinion of the population. We consider two types of organizations: i) *Participatory organizations:* organizations whose objective is to maximize the proportion of the population that express their opinion, and ii) *ideological organizations:* organizations whose objective is to move mean opinion towards one extreme, that is, 0 or 1. We show that each of these incentives results in a different signaling strategy by the two types of organizations.

To compute the optimal signaling strategy, the organization needs to solve an optimization problem. The optimization problem can be formulated as a Markov Decision Process (MDP) with the organization's estimate of the mean descriptive belief of the approval and disapproval opinions as the state and the generated signal as the action; the optimal policy generates the mapping from one to the other. Note that the state transition, i.e. the change in the descriptive belief of the population from one time step to the next in the organizational stewarding process, depends only on the belief at the previous time step. A time step here refers to one cycle of the organizational stewarding process. Therefore, the Markovian nature of this transition makes it apt for the optimization problem from the perspective of the organization to be formulated as an MDP. For each organization, there are two separate MDPs to be solved; one that generates the optimal signaling policy for information about

the approval group and one for the disapproval group. In both MDPs, the usual constructs of *State, Action, Transition, Rewards* is as follows:

- *State:* The descriptive belief about the population's approval or disapproval opinion, for each of the two MDPs, respectively. For simplicity, we can denote this as $\hat{o}_t$, but it takes values between [0,0.5) for disapproval and [0.5,1] for approval.
- *Action:* The signals to generate. Similar to the state variable, this action, $o_{I,t}$, lies between [0,0.5) for disapproval and [0.5,1] for approval group signaling.
- *Transition:* The change in the belief, which is based on the two sequential Bayesian updates in step 2 to step 4 applied in sequential manner as described in Sec. 4.1.
- *Rewards:* We formulate two reward functions, one for participatory organizations (R_{pa}) and another for the ideological organizations (R_{id}).

$$R_{\text{pa}}(\hat{o}_t, o_{I,t}) = 2(\frac{|N_E^{t+1}|}{N} - 0.5)$$
$$R_{\text{id}}(\hat{o}_t, o_{I,t}, \hat{o}_{t+1}) = 4\bar{o}^E_{\geqslant 0.5} - 3$$

where $\frac{|N_E^{t+1}|}{N}$, is the proportion of agents who express their opinion at time $t+1$. For the ideological objective, $\bar{o}^E_{\geqslant 0.5}$ is the mean opinion of the population that expresses approval; a value closer to 1 means higher the rewards for the ideological organization. The reward structure is also constructed in this way to make both of them bounded in the interval [-1,1].

4.2.1 Optimal Signaling Policy. Based on the reward structure of the two organizations, we solve the MDP using the Value Iteration algorithm (Sutton and Barto 2018) after discretizing the state space. We choose this algorithm since the simplicity is sufficient for demonstration, however, for a more complex application in the real-world, such as when a media organization needs to determine the policy for their choice of content across different issues, a more sophisticated approach may be necessary. In Fig. 5, we plot the reward heatmap for the entire signaling space for both sets of signaling distributions (approval and disapproval). The y axis shows the descriptive beliefs about the disapproval group's opinions in the left panels and the belief about the approval group's opinions in the right panels. The x axis shows the signal values for the participatory organization in panels A and B, and the ideological organization in panels C and D. For the ideological organization, we show the solution for an organization that aims to move expressed opinion toward approval. The brighter regions in the heatmap indicate higher rewards. We can see from the plot that the organizational signal has to align with the descriptive beliefs (brighter section near the diagonal). This is not surprising since this follows straightforwardly from the bounded confidence constraints. Due to those constraints on the generated signals, signals further away from the descriptive beliefs fetch lower rewards for the organization since the receivers ignore those signals.

Given this constraint, however, we can see the different strategies that ideological and participatory organizations pursue. Participatory organizations want to moderate beliefs about both in-group and out-group members: when players expect others to hold moderate beliefs, they also expect lower rhetoric. This encourages more to share their views, promoting the organization's participation goal. But ideological organizations want to move the mean of expressed opinions toward their preferred extreme (which is approval, in the example shown in Fig. 5). This requires inducing moderates in the approval group to remain silent. This is achieved by causing approval group members to believe that members of both groups hold more extreme views than they do and hence that they will engage in more extreme rhetoric; moderate approval group members do not get a high enough reward from expressing their opinions, in the face of opposing rhetoric and a substitution effect when their in-group engages in rhetoric, to warrant saying anything. The optimal signal when the ideological organization believes that the out-group's true belief is around 0.4, for example, is around 0.2 and the optimal signal when

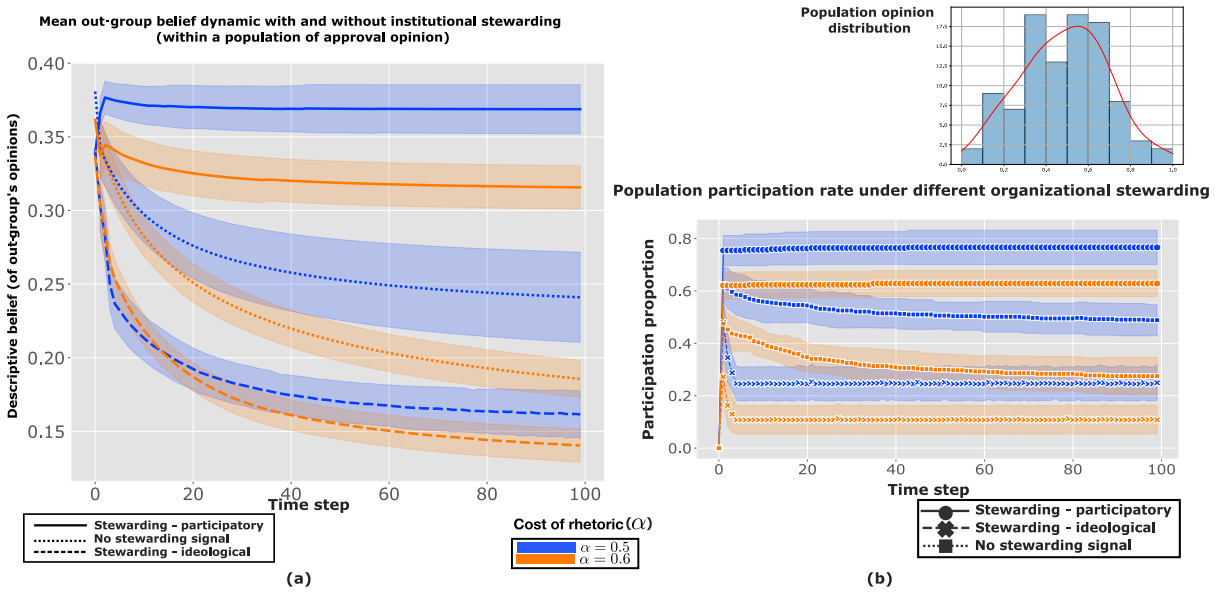

Fig. 6. Effect of organizational stewarding (participatory and ideological) on (a) descriptive beliefs about out-group's opinions; true out-group opinion (mean 0.4) is sampled from the opinion distribution shown in the inset (b) rate of participation within a population. Shown here for the population of approval opinion holders. Stewarding under participatory and ideological organization is shown in continuous and dashed lines, respectively. Belief dynamic under no organizational stewarding is shown with dotted lines.

the in-group is believed to be at 0.7 is close to 0.9. It is also interesting to note that when the descriptive belief about the out-group is at the most extreme (close to zero), then the optimal signal for the ideological organization is to moderate that effect to prevent even extreme opinion holders from staying silent. We see this from the bright reward spot on the upper right corner in panel C. Participatory organizations, on the other hand, want to make it seem that both the approval and disapproval groups are more moderate than they are. A participatory organization that believes the out-group is at 0.4 will send signals close to 0.5; if it believes the in-group is at 0.8 it will send signals as close as possible to 0.6. That strategy ensures that the equilibrium rhetorical intensity of the group is lower and, therefore, it is optimal for even moderate opinion holders to express their opinions with a higher rhetorical intensity.

Another important aspect to note here is that ideological organizations generally get higher rewards from signaling out-group opinions than in-group ones. (The brightest areas on the out-group heatmap correspond to higher values than the values for the brightest areas on in-group heatmap.) Meanwhile, for participatory organizations, there isn't a significant difference between out-group opinion and in-group opinion signaling. We can connect this observation to the empirical finding that in an ideological setting, content about political opponents is much more likely to be shared on social media (Rathje et al. 2021).

5 Simulation of Organizational Stewarding

Next, we analyze the long-run dynamics of the effects of organizational stewarding by ideological and participatory organizations using computational simulation. Each of the steps of the organization stewarding process described in Sec. 4.1 is simulated as follows:

• *Organizational signal generation:* We solve the optimization problem for ideological and participatory organizations as described in the previous section using the Value Iteration algorithm (Sutton and Barto 2018) which generates an optimal signaling policy for the two organizations.
• *Belief update (from organization signal):* In order to calculate the posterior beliefs in Eqn. 5, we use Monte Carlo estimation for Bayesian posteriors (Robert et al. 1999).
• *Opinion expression:* This step requires estimating the equilibrium rhetorical intensity from the perspective of each member of the population based on their own private type, that is, their opinion. As mentioned earlier, we use the equivalence between *ex-ante* and *ex-interim* Bayesian Nash equilibrium to simulate this process. Simply stated, the equivalence says that *ex-interim* response averaged over types is the same as the *ex-ante* response of the equilibrium (Lemma 6.1 in (Fujiwara-Greve 2015)). For simulation, as a first step, this involves estimating the *ex-ante* Bayesian Nash equilibrium of the opinion expression game played by two representative players with mean population opinion based on the prior beliefs about the opinion distribution. We calculate this by estimating the intersection of the Best Response curves of Eqns 3 and 4. Since finding a closed-form solution for the curves is not feasible due to the transcendental nature of the equations, we use numerical root finding (Bisection) to approximate the intersection of the curves. In the second step, once this value is estimated, we simulate the individual rhetorical intensity response to the *ex-ante* equilibrium by substituting it in Eqn. 3 as the non-focal player's rhetorical intensity.
• *Belief update (from community signals):* We model the beliefs about the approval and disapproval group in the population using a Beta distribution with parameters $a = 5, b = 3$ and $a = 3, b = 5$, respectively. Since the Bayesian update of a Beta distribution has a closed form solution, the updated distribution of the belief of the approval group's opinion is calculated as $Beta(a + o^E_{\geqslant 0.5}, b + (1 - o^E_{\geqslant 0.5}))$ where $o^E_{\geqslant 0.5}$ is the mean expressed opinion of the approval.

We run the simulation with a population of $N = 100$ agents with opinions drawn from a bimodal Gaussian distribution with parameters $\mu_1 = 0.4, \mu_2 = 0.6, \sigma_{\{1,2\}} = 0.2$, mixture coefficient 0.5. The rhetorical intensity threshold at which an individual stays silent (γ^{silence}) is set at 0.3.

We run the simulations under homogeneous beliefs, which means that within a group (approval or disapproval), all agents share the same beliefs about the population's mean approval and disapproval opinions. We run 10 batches of simulations for 100 timesteps. We also run three sets of simulations, with each set corresponding to the dynamic under the participatory, ideological, and *no organizational* stewarding. For the *no organizational* stewarding run, we include only steps 3 and 4 of the organizational stewarding process, eliminating the organizational signals completely and updating the beliefs solely based on the community interactions.

Fig. 6 shows the plot of the mean and standard deviation of the following set of attributes of the population calculated across the batches of run: i) descriptive belief about the disapproval group's opinion from the perspective of the approval group in Fig. 6a; ii) participation rate (measured as the proportion of the population who express their opinion, either of approval or disapproval) in Fig. 6b. The opinion distribution from which the simulation was run is shown in the inset of the Figure. We also repeat the simulation for different values of $\alpha = 0.5, 0.6$, which models different degrees of leniency in platform moderation policy. The dynamic of the descriptive beliefs and participation rate under participatory and ideological organization is shown with a smooth and dashed line, respectively. The dynamic without any kind of organization stewarding is shown in a dotted line. Although we show the plots for the beliefs of the approval group in Fig. 6 a, the beliefs of the disapproval group have identical patterns, but with the range of opinion values flipped for the in-group and out-group beliefs.

Based on the simulations, we find that participation stabilizes at a high level when it is stewarded by a participatory organization. Participation rates are lower in the absence of any stewarding and much lower when stewarded by an ideological organization. The cost of rhetoric (α) also influences participation: a lower cost leads

to higher overall participation. Moreover, lowering the cost of rhetoric has a greater impact on participation in populations without organizational stewarding than in those with any form of stewarding.

With respect to the beliefs about the out-group opinions shown in Fig. 6a, we see that stewarding under ideological organizations also produces the most distorting effect on the beliefs, where the population, over time, believes that the out-group is more extreme than it really is. This effect is primarily driven by the differences in participation between the moderate and extreme opinion holders. Meanwhile, under participatory organization stewarding, the distortion of beliefs is minimal.

We also note a surprising and unexpected effect of the cost of rhetoric, which can be increased by a moderation policy. Because such a policy imposes on everyone the same cost per unit on rhetoric, the distortion of belief becomes worse. This is because increased rhetorical costs affect moderates more than extreme opinion holders, resulting in lower participation by moderate opinion holders and greater distortion of beliefs. This is an important insight into the role of platform moderation. An effort to encourage moderates to participate by restricting the use of intense rhetoric on a platform in a coarse grained way might backfire: Moderation may further discourage moderate participation and exacerbate the distortion of second-order beliefs about the out-group.

6 Platforms and Organizational Communities

While the analysis of a single organization within a population with homogeneous beliefs offers valuable insights, the reality of online platforms is much more complex. On platforms like Reddit, Twitter, and YouTube, digital communities are formed around a central theme, such as subreddits on Reddit, influential accounts on Twitter (X), and channels on YouTube. In many of these digital communities, members share, discuss, and enforce their perceptions of right and wrong, often along a partisan axis (Conover et al. 2011; Waller and Anderson 2021). This speaks to the presence of a diverse population with heterogeneous beliefs and multiple viewpoint organizations, each stewarding the beliefs of the community that falls within its scope. In this section, using the organizational stewarding process presented earlier, we develop a model of organizational community formation for both participatory and ideological organizations within a platform. Furthermore, we analyze the characteristics of the beliefs of each organizational community and show how ideological organizational communities give rise to the phenomenon of false polarization, i.e., individuals in ideological communities hold more extreme opinions and, at the same time, perceive the out-group to be more extreme than in reality (Levendusky and Malhotra 2016).

6.1 Organizational Communities

For most online platforms, it is common to use a recommendation algorithm that connects users with communities or accounts that match their interests and engagement patterns. Although there are several differences between platforms in terms of how these processes are implemented in practice, we use a minimal set of key entities common to most platforms: organizations, users, and the platform.

The role of organizations and users is modeled as presented earlier in the organizational stewarding model. In reality, these organizations are often channels, subreddits, accounts etc. that generate information signals about approval and disapproval groups on normative issues. In this section, we use three categories of organizations in our simulation of the population under multiple organizations: one participatory organization, one ideological organization that has the objective of moving the expressed opinion towards approval, and one ideological organization that has the objective of moving the expressed opinion towards disapproval. As before, users within the population are modeled as receivers of organizational signals and they express opinion within the digital space provided by the platform.

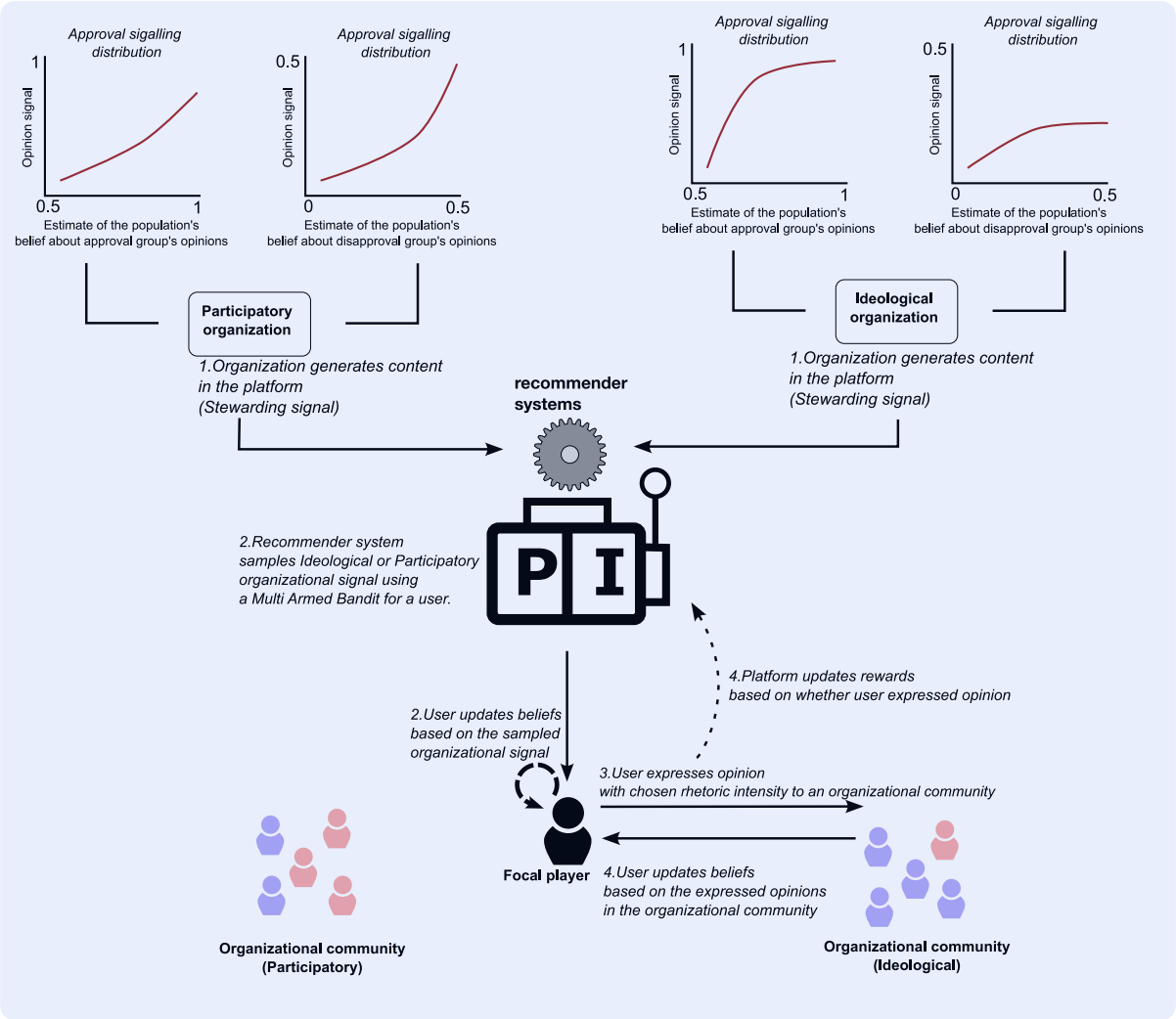

Fig. 7. Extended schematic representation of the organizational stewarding process in Fig. 4 with the addition of participatory and ideological organizations, recommender systems, and organizational communities.

6.2 Modelling the Platform

The platform is an intermediary between the organizations and the user, and it is responsible for the following set of functions in our model:

i. *Content recommendation* – A recommender system that recommends an organization and the organizational signal to a user based on their past interactions with the organization. In practice, a platform often uses a combination of various objectives, including engagement and diversity, to fine-tune the recommender system for a given user (Stray et al. 2022), however, in our model, the goal of the recommender system is to recommend an organization to a user to maximize the probability that the user expresses their opinion on the platform. In

terms of technology, Bandit algorithms have become one of the mainstays of recommender systems (Parapar and Radlinski 2021; Silva et al. 2022). With that motivation, we use a vanilla UCB algorithm (Auer et al. 2002) to simulate the platform's recommender engine, which recommends a participatory or ideological organizational signal to a user.
ii. *Digital space* – The platform maintains a digital space for each organization where users react to organizational signals by expressing their opinions and interacting with others in the community. One can think of this digital space as the comment and user interaction sections available in most platforms. We refer to this digital space for an organization as the organizational community.
iii. *Moderation* – The platform may choose to moderate the content within its digital space. In our case, this can be modeled as the platform increasing the per unit cost of rhetoric, α.

With the main components of the models in place, we present the sequence of steps involved in organizational community formation. The main idea behind the process is that the user's interaction with the content generated by a particular organization is visible and registered with the recommender system when the user expresses their opinion based on the organizational signal. The recommender system, in turn, learns to recommend the particular organizations based on the user's past participation and the predicted attributes of the user, in this case, the beliefs about the in-group and out-group. The objective of the platform is to maximize user participation within the platform. This is done by selecting the right organization for the user and ensuring that the signals it recommends from the selected organization induce the user to participate by expressing their opinion. Since participatory organizations generate signals that allow moderate opinion holders to participate in their community, and ideological organizational signals suppress moderate opinion participation, we see a higher representation of moderate opinion holders in participatory organizational communities. We elaborate these steps in more detail below.
Initial setup: Various participatory and ideological organizations in the platform generate signals based on their optimal signaling strategy. On a given issue, due to the diversity of the content and organizations in the platform, we assume that there is content available for the recommender system to present to a user from a wide spectrum of opinions for both approval and disapproval. For every user, the platform maintains a history of past interaction with two types of organization: the participatory organization and the ideological organization whose goal is to move the opinion in the direction that aligns with the user's opinion. We outline the modelling steps in more detail below.
At each time step t:

(1) For each user i, the platform's recommender system samples a participatory or an ideological organizational signal by pulling one of the two arms of the bandit based on the UCB values corresponding to each arm.
(2) The user receives the signal from the organization sampled by the recommender system. After a particular organization is chosen by the system, there is still the question of the type of the information signal: whether the signal is about the approval or the disapproval group's opinion since each organization maintains two separate signaling distributions. We model this choice as a stochastic policy where each type of information (approval or disapproval group's opinion) is selected in proportion to the maximum reward the organization can achieve from either of the two distributions. We model this choice based on the assumption that the overall proportion of information content about the two groups' opinions generated by the organization will be proportional to the rewards that each type of information fetches the organization. Once the user receives the signal from the organization, they update their belief based on the signal received.
(3) The user makes the choice to either *express* their opinion or *stay silent*. This decision is determined by the equilibrium rhetorical intensity discussed in step *opinion expression* in Sec. 4.1. Specifically, the user expresses if their best-response rhetorical intensity is higher than a threshold $\gamma^{\text{silence}} = 0.3$.

(4) Expression of the user's opinion is recorded in the digital space provided by the platform for the corresponding organizations. Each user who expresses their opinion in the digital space is added to the organizational community maintained by the platform.
(5) Each user who was added to the organizational community observes the opinion of other users within the community and updates their corresponding descriptive beliefs. For the participatory institutional communities, these community signals consist of both approval and disapproval opinions since both types of users interact in the participatory institutions. Whereas, for the ideological communities, since only one of the groups interacts in their corresponding community, the community signal is only either approval or disapproval opinions.
(6) The recommender system updates the rewards for the corresponding arm to +1 if the user expressed based on the sampled signal or 0 if they did not express.

Table 2. Summary of opinion and belief about out-group and in-group for participatory and ideological communities along with the characteristics of agents that do not participate in any community. The true population mean opinion is 0.4 for disapproval and 0.6 for approval.

Community Category	**Opinion Group**	**Opinion (Mean ± SD)**	**Belief about out-group (Mean ± SD)**	**Belief about in-group (Mean ± SD)**
Participatory	≥ 0.5	0.666 ± 0.114	0.378 ± 0.066	0.5 ± 0.001
Ideological	≥ 0.5	0.703 ± 0.104	0.180 ± 0.093	0.707 ± 0.002
Silent	≥ 0.5	0.516 ± 0.013	0.161 ± 0.134	0.907 ± 0.11

Table 3. Results of Cohen's d estimate for effect size for comparison of opinion, out-group belief, and in-group belief for each organizational community. $*$: small to medium, $**$: medium to large effect size, $***$: large to very large effect size.

Comparison	**Opinion Group**	**Cohen's d (Opinion)**	**Cohen's d (Out Belief)**	**Cohen's d (In Belief)**
Ideological vs Participatory	≥ 0.5	0.338^{*}	-2.488^{***}	100.459^{***}
Ideological vs Silent	≥ 0.5	1.852^{***}	0.201^{*}	-6.732^{***}
Participatory vs Silent	≥ 0.5	1.345^{**}	3.014^{***}	-15.548^{***}

6.3 Organizational Community Characteristics

We can analyze the characteristics of the organizational communities that form based on the combination of the organization signaling, user expression of opinion, and recommender systems described above. Specifically, we answer the following two questions about the characteristics of the community:

- Opinion differences: *How do the opinions differ among the members in each organizational community?*
- Second-order beliefs about out-group opinions: *How do the beliefs about the out-group differ among the members in each organizational community?*

• Second-order beliefs about in-group opinions: *How do the beliefs about the in-group differ among the members in each organizational community?*

Simulation setup: Similar to Sec. 5, to answer the above questions, we run simulations with the multiple institutions setup and heterogeneous beliefs, that is, users hold different beliefs about their in-group and out-group sub-population. The population size, opinion distribution are same as in simulation in Sec. 5), i.e, population of $N = 100$ for $T = 100$ time steps with opinions drawn from a bimodal Gaussian distribution with parameters $\mu_1 = 0.4, \mu_2 = 0.6, \sigma_{\{1,2\}} = 0.2$, mixture coefficient 0.5. We model the heterogeneous beliefs about the approval and disapproval group in the population using two randomly sampled Beta distribution for each user (corresponding to the belief about in-group and out-group) such that the mean value of the parameters a and b for the beliefs of approval and disapproval for the whole population are $a = 5, b = 3$ and $a = 3, b = 5$, respectively. The threshold at which an individual stays silent is 0.3. We answer the questions about the characteristics of the community in the second half of the simulation run in order for the recommender system to stabilize the learning for each user.

Table 2 shows the difference in opinion distribution for the participatory and ideological communities. There is a statistically significant difference in the distribution of opinions between the two communities, with the mean opinion of the ideological community more extreme than that of participatory communities. This difference results directly from higher likelihood of moderate opinion holders staying silent from signals of ideological organizations, and consequently, the recommender systems matching them to participatory organizations more often than extreme opinion holders. Next, we look at the differences in out-group beliefs of the two communities. We select the approval group for analysis, although the characteristics are the same for both groups. We see a significant difference between the two communities with respect to out-group and in-group beliefs; beliefs about both in-group and out-group are distorted towards the extreme for ideological communities. However, for participatory communities, the belief about in-group is more moderate than the true opinion. This relationship results from a combination of the optimal signaling schemes constructed by organizations and the recommender system that stratifies users to different communities. As extreme opinion holders participate more in ideological communities, their out-group belief also drift towards one extreme, a dynamic we observed in Sec. 5).

Finally, our analysis also sheds light on the characteristics of the population that does not participate in either of the communities. This population consists of users with moderate opinions and more extreme beliefs about others' opinions. This is because more moderate opinion holders with more extreme beliefs about the population are likelier to stay silent. Although the recommender system samples different organizations, it fails to bring those users to participate in either organization's community.

7 Discussion: Social Media Fragmentation and Decentralized Platforms

The social media landscape of the 2010s was dominated by centralized platforms such as Facebook, Twitter, and YouTube, operated by large corporations. However, in recent years, this centralized model has replaced a more fragmented ecosystem, populated by smaller and sometimes decentralized platforms such as Gab, Mastodon and Bluesky, each with its own governance structure (Jeong et al. 2024; Jhaver et al. 2023). In this section, we discuss how the insights from our model can be interpreted in the context of this fragmentation. Although the model presented in this paper offers a stylized representation of a centralized platform modeled through a single recommender system, the framework can also serve as an exploratory lens to analyze the behaviour of the population in the emerging fragmented landscape.

In 2020, in response to Twitter's content moderation policies and the perceived censorship of opinion expression, a significant user exodus occurred from Twitter to Gab, resulting in approximately three million daily users on the alternative platform. The users who made this transition likely held more extreme views and our model suggests they would have engaged in higher rhetorical intensity. Subsequent empirical studies confirm that Gab exhibits a higher degree of conflict-oriented and emotionally charged content (Acampa et al. 2023).

This trend of fragmentation has continued. Following Elon Musk's acquisition of Twitter and the subsequent changes in moderation policy which lowered restrictions on extreme expressions of opinions, many left- and center-left users migrated to alternative platforms, particularly the decentralized platform Bluesky. Due to its federated structure, Bluesky functions less as a single platform and more as a constellation of smaller communities. Unlike centralized platforms driven by organizational incentives, users can custom-design their feed oriented around non-political topics such as nature, music, and science. Popular accounts that manage these topic-specific feeds operate without strong incentives to distort content signals, making them analogous to participatory organizations. With respect to media consumption patterns on Bluesky, users show a high engagement with traditional outlets such as The Guardian and The New York Times (Quelle and Bovet 2025). As predicted by our model, such algorithmic environments are associated with reduced polarization, which is confirmed by empirical studies (Quelle and Bovet 2025), with the notable exception of the Israel–Palestine conflict, which remains a highly polarizing topic within the platform needing further analysis.

Although our model, focusing solely on a single centralized platform, can offer some insights, it cannot formally account for population dynamics in a fragmented social media environment. For example, it cannot distinguish whether low participation in platforms like Mastodon arises from algorithm-induced in-group–out-group dynamics, or simply from the availability of more appealing alternative platforms (Jeong et al. 2024). Future work can extend the model to incorporate multiple recommender systems, enabling the study of strategic user behaviour and polarization dynamics across fragmented platforms.

8 Conclusion

In this paper, we present a model that shows that user differences in the utility derived from the expression of opinions can cause polarization at the population and community levels, both in the opinions expressed and beliefs about others. The observed phenomenon arises from the differences in individual incentives for opinion expression, organizational incentives, and the recommender system of online platforms. As we have shown, polarization can arise not from a distortion in actual opinions but rather from rational choices about when and how to express opinions. Rhetorical intensity is shown to play a key role: rhetoric generates payoffs for individuals with strong views but also inhibits expression by those with moderate views. If a platform attempts to moderate intense rhetoric, however, this may only exacerbate the problem: moderates will be less willing to incur the costs of rhetoric than those with extreme views.

Importantly, we show these effects without assuming any change in the actual opinions held by individuals, only in the pattern of expression of opinions. This result highlights a risk from AI that has not been adequately appreciated: reliance on the expression of opinions on AI-powered social media platforms can distort perceptions in public debates and policy-making about the true distribution of views. That in turn can distort actual policy-making. Additionally, distortions in the expression of opinion also distort the data on which AI models are trained, amplifying the effect of a shift to AI-mediated interaction. Moreover, correction for the distorting effect is difficult: there are many reasons that users do not participate in particular social media discussions and so the signal from silence is difficult to extract.

The role we have identified for platform AI-based recommender systems opens up the possibility of designing mitigation strategies that do not require changing underlying opinions. We identify at least two such strategies for online platforms :

A content-based approach to platform moderation: In our model, we observe that if a platform implements a lenient moderation policy, more extreme opinion holders employ harsher rhetoric, which discourages participation by moderates. Conversely, a stricter moderation policy, making rhetorical intensity more costly, may make moderates even more likely to stay silent because they respond more to the cost change on the margin. This creates a difficult trade-off for platforms, as any shift in policy risks either amplifying extremes or discouraging

moderates. One solution to this problem could be a content-based approach to moderation, which grants greater leniency on the use of rhetoric by moderate opinion holders as compared to extreme opinion holders. Such a policy could help strike a balance between fair participation and reduced rhetorical intensity.
Prioritizing signals from participatory organizations: In conjunction with a tailored moderation policy, platforms could also tailor their recommender strategy to prioritize signals from participatory organizations over ideological organizations for extreme opinion holders. Since extreme opinion holders still express their opinion even under signals from participatory organizations, a higher proportion of participatory organizational signals can bring their beliefs about out-group opinions more in line with less polarized participatory organizational communities.

Both of these strategies require, however, that platforms engage in a form of content-based regulation of rhetoric. People who oppose animal-testing in any context would be more restricted in their platform speech than those who believe animal-testing is acceptable in some contexts but not others. That may itself be unacceptable for political communities. The only alternative would seem to be aggressive regulation of rhetoric across the board. In a sense, this is what political communities have historically sought to achieve, through strong rules of civility in political forums (such as legislatures) and strong editorial norms in publications such as widely-distributed newspapers. The challenge in the era of AI-based opinion platforms is to establish such norms in a highly decentralized setting. Our results emphasize that the benefits to developing such norms is important not merely to support individual expression rights but also to ensure the integrity of public and policy-makers perceptions of true public opinion on contentious policy matters and the integrity of the data on which our large language models are trained.

Acknowledgments

We thank the following people for their feedback on this work: Graham Noblit, Valerie Platsko, Kathryn E. Spier, Peter Marbach, Ashton Anderson. We thank Schwartz Reisman Institute for Technology and Society for providing support for this research.

References

S. Acampa, N. Crescentini, and G. M. Padricelli. 2023. "Between alternative and traditional social platforms: the case of GAB in exploring the narratives on the pandemic and vaccines". *Frontiers in Sociology*, 8, 1143263.

D. Acemoglu, A. Ozdaglar, and J. Siderius. 2023. "A model of online misinformation". *Review of Economic Studies*, rdad111.

T. Aichner, M. Grünfelder, O. Maurer, and D. Jegeni. 2021. "Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019". *Cyberpsychology, Behavior, and Social Networking*, 24, 4, 215–222.

F. Alatawi, L. Cheng, A. Tahir, M. Karami, B. Jiang, T. Black, and H. Liu. 2021. "A survey on echo chambers on social media: Description, detection and mitigation". *arXiv preprint arXiv:2112.05084*.

C. Algara and R. Zur. 2023. "The Downsian roots of affective polarization". *Electoral Studies*, 82, 102581.

S. D. Arora, G. P. Singh, A. Chakraborty, and M. Maity. 2022. "Polarization and social media: A systematic review and research agenda". *Technological Forecasting and Social Change*, 183, 121942.

P. Auer, N. Cesa-Bianchi, and P. Fischer. 2002. "Finite-time analysis of the multiarmed bandit problem". *Machine Learning*, 47, 235–256.

C. A. Bail et al.. 2018. "Exposure to opposing views on social media can increase political polarization". *Proceedings of the National Academy of Sciences*, 115, 37, 9216–9221.

E. Bakshy, S. Messing, and L. A. Adamic. 2015. "Exposure to ideologically diverse news and opinion on Facebook". *Science*, 348, 6239, 1130–1132.

L. Boxell, M. Gentzkow, and J. M. Shapiro. 2022. "Cross-country trends in affective polarization". *Review of Economics and Statistics*, 1–60.

L. Boxell, M. Gentzkow, and J. M. Shapiro. 2017. *Is the internet causing political polarization? Evidence from demographics*. Tech. rep. National Bureau of Economic Research.

A. Bramson, P. Grim, D. J. Singer, W. J. Berger, G. Sack, S. Fisher, C. Flocken, and B. Holman. 2017. "Understanding polarization: Meanings, measures, and model evaluation". *Philosophy of Science*, 84, 1, 115–159.

M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini. 2021. "The echo chamber effect on social media". *Proceedings of the National Academy of Sciences*, 118, 9, e2023301118.

M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. 2011. "Political polarization on twitter". In: *Proceedings of the International AAAI Conference on Web and Social Media* 1. Vol. 5, 89–96.
S. Dash, D. Mishra, G. Shekhawat, and J. Pal. 2022. "Divided we rule: Influencer polarization on Twitter during political crises in India". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 16, 135–146.
M. H. DeGroot. 1974. "Reaching a consensus". *Journal of the American Statistical Association*, 69, 345, 118–121.
S. DellaVigna and E. Kaplan. 2007. "The Fox News effect: Media bias and voting". *The Quarterly Journal of Economics*, 122, 3, 1187–1234.
K. Dorst. 2023. "Rational polarization". *Philosophical Review*, 132, 3, 355–458.
A. Downs. 1957. "An economic theory of political action in a democracy". *Journal of Political Economy*, 65, 2, 135–150.
F. Fagan. 2017. "Systemic social media regulation". *Duke L. & Tech. Rev.*, 16, 393.
T. Fujiwara-Greve. 2015. "Bayesian nash equilibrium". *Non-cooperative game theory*, 133–151.
F. Gaisbauer, E. Olbrich, and S. Banisch. 2020. "Dynamics of opinion expression". *Physical Review E*, 102, 4, 042303.
K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. 2018. "Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship". In: *Proceedings of the 2018 World Wide Web Conference*, 913–922.
B. Golub and M. O. Jackson. 2012. "How homophily affects the speed of learning and best-response dynamics". *The Quarterly Journal of Economics*, 127, 3, 1287–1338.
N. Haghtalab, M. O. Jackson, and A. D. Procaccia. 2021. "Belief polarization in a complex world: A learning theory perspective". *Proceedings of the National Academy of Sciences*, 118, 19, e2010144118.
C. Hare and K. T. Poole. 2014. "The polarization of contemporary American politics". *Polity*, 46, 3, 411–429.
N. Helberger, K. Karppinen, and L. D'acunto. 2018. "Exposure diversity as a design principle for recommender systems". *Information, communication & society*, 21, 2, 191–207.
C.-C. Hsu, A. Ajorlou, and A. Jadbabaie. 2020. "News sharing, persuasion, and spread of misinformation on social networks". *Persuasion, and Spread of Misinformation on Social Networks (July 1, 2020)*.
Y. Hu, Y. Koren, and C. Volinsky. 2008. "Collaborative filtering for implicit feedback datasets". In: *2008 Eighth IEEE International Conference on Data Mining*. Ieee, 263–272.
S. Iyengar, Y. Lelkes, M. Levendusky, N. Malhotra, and S. J. Westwood. 2019. "The origins and consequences of affective polarization in the United States". *Annual Review of Political Science*, 22, 129–146.
U. Jeong, P. Sheth, A. Tahir, F. Alatawi, H. R. Bernard, and H. Liu. 2024. "Exploring platform migration patterns between Twitter and Mastodon: A user behavior study". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 18, 738–750.
A. Jern, K.-M. K. Chang, and C. Kemp. 2014. "Belief polarization is not always irrational." *Psychological Review*, 121, 2, 206.
S. Jhaver, S. Frey, and A. X. Zhang. 2023. "Decentralizing platform power: A design space of multi-level governance in online social platforms". *Social Media+ Society*, 9, 4, 20563051231207857.
C. C. Johnson et al.. 2014. "Logistic matrix factorization for implicit feedback data". *Advances in Neural Information Processing Systems*, 27, 78, 1–9.
J. T. Jost, D. S. Baldassarri, and J. N. Druckman. 2022. "Cognitive–motivational mechanisms of political polarization in social-communicative contexts". *Nature Reviews Psychology*, 1, 10, 560–576.
J. Jung, P. Grim, D. J. Singer, A. Bramson, W. J. Berger, B. Holman, and K. Kovaka. 2019. "A multidisciplinary understanding of polarization." *American Psychologist*, 74, 3, 301.
E. Kubin and C. von Sikorski. 2021. "The role of (social) media in political polarization: a systematic review". *Annals of the International Communication Association*, 45, 3, 188–206.
L. La Cava, D. Mandaglio, and A. Tagarelli. 2024. "Polarization in decentralized online social networks". In: *Proceedings of the 16th ACM Web Science Conference*, 48–52.
Y. Le Yaouanq. 2018. *A model of ideological thinking*. Tech. rep. Discussion Paper.
J. Lees and M. Cikara. 2020. "Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts". *Nature Human Behaviour*, 4, 3, 279–286.
M. S. Levendusky and N. Malhotra. 2016. "(Mis) perceptions of partisan polarization in the American public". *Public Opinion Quarterly*, 80, S1, 378–391.
G. Levy and R. Razin. 2019. "Echo chambers and their effects on economic and political outcomes". *Annual Review of Economics*, 11, 303–328.
Z. Li, Y. Dong, C. Gao, Y. Zhao, D. Li, J. Hao, K. Zhang, Y. Li, and Z. Wang. 2023. "Breaking filter bubble: A reinforcement learning framework of controllable recommender system". In: *Proceedings of the ACM Web Conference 2023*, 4041–4049.
S. L. Lim and P. J. Bentley. 2022. "Opinion amplification causes extreme polarization in social networks". *Scientific Reports*, 12, 1, 18131.
P. Lorenz-Spreen, L. Oswald, S. Lewandowsky, and R. Hertwig. 2023. "A systematic review of worldwide causal and correlational evidence on digital media and democracy". *Nature Human Behaviour*, 7, 1, 74–101.
J. Matthes, J. Knoll, and C. von Sikorski. 2018. "The "spiral of silence" revisited: A meta-analysis on the relationship between perceptions of opinion support and political opinion expression". *Communication Research*, 45, 1, 3–33.
S. Mullainathan and A. Shleifer. 2005. "The market for news". *American Economic Review*, 95, 4, 1031–1053.

C. Musco, C. Musco, and C. E. Tsourakakis. 2018. "Minimizing polarization and disagreement in social networks". In: *Proceedings of the 2018 world wide web conference*, 369–378.
T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, and J. A. Konstan. 2014. "Exploring the filter bubble: the effect of using recommender systems on content diversity". In: *Proceedings of the 23rd international conference on World wide web*, 677–686.
E. Noelle-Neumann. 1974. "The spiral of silence a theory of public opinion". *Journal of Communication*, 24, 2, 43–51.
B. Nyhan et al.. 2023. "Like-minded sources on Facebook are prevalent but not polarizing". *Nature*, 620, 7972, 137–144.
Y. Papanastasiou. 2020. "Fake news propagation and detection: A sequential model". *Management Science*, 66, 5, 1826–1846.
J. Parapar and F. Radlinski. 2021. "Diverse user preference elicitation with multi-armed bandits". In: *Proceedings of the 14th ACM international conference on web search and data mining*, 130–138.
E. Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.
A. F. Peralta, J. Kertész, and G. Iñiguez. 2022. "Opinion dynamics in social networks: From models to data". *arXiv preprint arXiv:2201.01322*.
M. Prior. 2013. "Media and political polarization". *Annual Review of Political Science*, 16, 1, 101–127.
D. Quelle and A. Bovet. 2025. "Bluesky: Network topology, polarization, and algorithmic curation". *PLOS one*, 20, 2, e0318034.
H. Rainer and U. Krause. 2002. "Opinion dynamics and bounded confidence: models, analysis and simulation".
P. Ramaciotti Morales and J.-P. Cointet. 2021. "Auditing the effect of social network recommendations on polarization in geometrical ideological spaces". In: *Proceedings of the 15th ACM Conference on Recommender Systems*, 627–632.
S. Rathje, J. J. Van Bavel, and S. Van Der Linden. 2021. "Out-group animosity drives engagement on social media". *Proceedings of the National Academy of Sciences*, 118, 26, e2024292118.
S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. 2012. "BPR: Bayesian personalized ranking from implicit feedback". *arXiv preprint arXiv:1205.2618*.
C. P. Robert, G. Casella, and G. Casella. 1999. *Monte Carlo statistical methods*. Vol. 2. Springer.
F.-J. Rodrigo-Ginés, J. Carrillo-de-Albornoz, and L. Plaza. 2023. "A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it". *Expert Systems with Applications*, 121641.
F. P. Santos, Y. Lelkes, and S. A. Levin. 2021. "Link recommendation algorithms and dynamics of polarization in online social networks". *Proceedings of the National Academy of Sciences*, 118, 50, e2102141118.
N. Silva, H. Werneck, T. Silva, A. C. Pereira, and L. Rocha. 2022. "Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions". *Expert Systems with Applications*, 197, 116669.
D. J. Singer, A. Bramson, P. Grim, B. Holman, J. Jung, K. Kovaka, A. Ranginani, and W. J. Berger. 2019. "Rational social and political polarization". *Philosophical Studies*, 176, 2243–2267.
J. Stray et al.. 2022. "Building human values into recommender systems: An interdisciplinary synthesis". *ACM Transactions on Recommender Systems*.
C. R. Sunstein. 1999. "The law of group polarization". *University of Chicago Law School, John M. Olin Law & Economics Working Paper*, 91.
R. S. Sutton and A. G. Barto. 2018. *Reinforcement learning: An introduction*. MIT Press.
J. A. Tucker, A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan. 2018. "Social media, political polarization, and political disinformation: A review of the scientific literature". *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*.
J. J. Van Bavel, S. Rathje, E. Harris, C. Robertson, and A. Sternisko. 2021. "How social media shapes polarization". *Trends in Cognitive Sciences*, 25, 11, 913–916.
I. Waller and A. Anderson. 2021. "Quantifying social organization and political polarization in online platforms". *Nature*, 600, 7888, 264–268.
H. Xia, H. Wang, and Z. Xuan. 2011. "Opinion dynamics: A multidisciplinary review and perspective on future research". *International Journal of Knowledge and Systems Science (IJKSS)*, 2, 4, 72–91.
M. Yarchi, C. Baden, and N. Kligler-Vilenchik. 2021. "Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media". *Political Communication*, 38, 1-2, 98–139.
D. Yudkin, S. Hawkins, and T. Dixon. 2019. "The perception gap: How false impressions are pulling Americans apart". *PsyArXiv*. doi:doi:10.31234/osf.io/r3h5q.

A Proof of Theorem 1

Theorem statement: Let $v(o_i) = o_i$ if $o_i \geqslant 0.5$ and $v(o_i) = 1 - o_i$, if $o_i < 0.5$ and $o_i \sim f$ is drawn from any arbitrary prior opinion distribution with p.d.f f and c.d.f F, and $\hat{v}_o = \mathrm{E}_f[v(o_i)]$. Then, under the condition of equal support of approval and disapproval, i.e., $F(0.5) = 0.5$, there exists an ex-ante symmetric Bayesian Nash equilibrium $\gamma_i^* = 0, \forall i$, if an only if $\hat{v}_o < \frac{\alpha}{1-\hat{n}\cdot(1-\lambda_{in})}$

Proof. *Sketch:* The proof follows three steps: (i) we calculate the expected utility in ex-ante form for the players (Lemma 1), (ii) we calculate the best response functions based on the ex-ante utilities, and (iii) we show that the symmetric intersection of the best response function at $\gamma = 0$ exists only under the condition stated in the theorem.

Lemma 1. *The ex-ante utility of a representative player in a population with opinion distribution $o \sim f$ is*

$$\mathrm{E}_f[u_i(\gamma_i, \gamma_{-i})] = \hat{n} \cdot \mathrm{E}_f[v(o)] \cdot \lambda_{in.} \cdot \gamma_i^{(1-\gamma_{-i})} + (1-\hat{n}) \cdot \mathrm{E}_f[v(o)] \cdot \lambda_{out}^{\gamma_{-i}} \cdot \gamma_i - \alpha \cdot \gamma_i$$

The ex-ante utility is based on the expected utility of a random player in the game with opinion ($o \sim f$) drawn from an opinion distribution f. This utility can be calculated by taking expectation based on Eqn 1 and Eqn 2 as

$$\begin{aligned}
\mathrm{E}_f[u_i(\gamma_i, \gamma_{-i}; o_i)] &= \mathrm{E}_f[u_i(\gamma_i, \gamma_{-i}; o_i | o_i < 0.5)] + \mathrm{E}_f[u_i(\gamma_i, \gamma_{-i}; o_i | o_i \geqslant 0.5)] && (6)\\
&= \int_0^{0.5} [(1-\hat{n}) \cdot (1-o) \cdot \lambda_{\text{in.}} \cdot \gamma_i^{(1-\gamma_{-i})} && (7)\\
&\quad + \hat{n} \cdot (1-o) \cdot \lambda_{\text{out}}^{\gamma_{-i}} \cdot \gamma_i - \alpha \cdot \gamma_i] \cdot f(o)\, do && (8)\\
&\quad + \int_{0.5}^{1} [\hat{n} \cdot o \cdot \lambda_{\text{in.}} \cdot \gamma_i^{(1-\gamma_{-i})} && (9)\\
&\quad + (1-\hat{n}) \cdot o \cdot \lambda_{\text{out}}^{\gamma_{-i}} \cdot \gamma_i - \alpha \cdot \gamma_i] \cdot f(o)\, do && (10)
\end{aligned}$$

Replacing $v(o) = o$ if $o \geqslant 0.5$ and $v(o) = 1 - o$, if $o < 0.5$ in the above equation, the expected utility becomes

$$\begin{aligned}
\mathrm{E}_f[u_i(\gamma_i, \gamma_{-i}; o)] &= \int_0^{0.5} [(1-\hat{n}) \cdot v(o) \cdot \lambda_{\text{in.}} \cdot \gamma_i^{(1-\gamma_{-i})} + \hat{n} \cdot v(o) \cdot \lambda_{\text{out}}^{\gamma_{-i}} \cdot \gamma_i - \alpha \cdot \gamma_i] \cdot f(o)\, do\\
&\quad + \int_{0.5}^{1} [\hat{n} \cdot v(o) \cdot \lambda_{\text{in.}} \cdot \gamma_i^{(1-\gamma_{-i})} + (1-\hat{n}) \cdot v(o) \cdot \lambda_{\text{out}}^{\gamma_{-i}} \cdot \gamma_i - \alpha \cdot \gamma_i] \cdot f(o)\, do\\
&= ((1-\hat{n}) \cdot \lambda_{\text{in.}} \cdot \gamma_i^{(1-\gamma_{-i})}) \int_0^{0.5} v(o) \cdot f(o)\, do + (\hat{n} \cdot \lambda_{\text{in.}} \cdot \gamma_i^{(1-\gamma_{-i})}) \int_{0.5}^{1} v(o) \cdot f(o)\, do\\
&\quad + (\hat{n} \cdot \lambda_{\text{out}}^{\gamma_{-i}} \cdot \gamma_i) \int_0^{0.5} v(o) \cdot f(o)\, do + ((1-\hat{n}) \cdot \lambda_{\text{out}}^{\gamma_{-i}} \cdot \gamma_i) \int_{0.5}^{1} v(o) \cdot f(o)\, do\\
&\quad - \alpha \cdot \gamma_i \int_0^{1} f(o)\, do
\end{aligned}$$

$$\begin{aligned}
&= (\hat{n} \cdot \lambda_{\text{in.}} \cdot \gamma_i^{(1-\gamma_{-i})}) [\int_0^{0.5} v(o) \cdot f(o)\, do + \int_{0.5}^{1} v(o) \cdot f(o)\, do]\\
&\quad + ((1-\hat{n}) \cdot \lambda_{\text{out}}^{\gamma_{-i}} \cdot \gamma_i) [\int_0^{0.5} v(o) \cdot f(o)\, do + \int_{0.5}^{1} v(o) \cdot f(o)\, do]\\
&\quad - \alpha \cdot \gamma_i
\end{aligned}$$

Substituting $\mathrm{E}_f[v(o)] = \int_0^{0.5} v(o) \cdot f(o)\, do + \int_{0.5}^{1} v(o) \cdot f(o)\, do$ in the above equation and rearranging the terms

$$\mathrm{E}_f[u_i(\gamma_i, \gamma_{-i}; o)] = \hat{n} \cdot \mathrm{E}_f[v(o)] \cdot \lambda_{\text{in.}} \cdot \gamma_i^{(1-\gamma_{-i})} + (1-\hat{n}) \cdot \mathrm{E}_f[v(o)] \cdot \lambda_{\text{out}}^{\gamma_{-i}} \cdot \gamma_i - \alpha \cdot \gamma_i$$

Note that the above equation of the utility in ex-ante form has the same structure as the utility in Eqn. 1. Taking the partial derivative of the above function with respect to the strategic variable γ_i and γ_{-i} and solving the two set of equations:

$$\frac{\partial}{\partial \gamma_i} \mathrm{E}_f[u_i(\gamma_i, \gamma_{-i}; o)] = 0 \text{ and } \frac{\partial}{\partial \gamma_{-i}} \mathrm{E}_f[u_i(\gamma_i, \gamma_{-i}; o)] = 0$$

yields the two best response functions in ex-ante form as follows.

$$BR_i(\gamma_{-i}; o) = \min\left(1, \max\left(0, \left(\frac{\hat{n} \cdot \mathrm{E}_f[v(o)] \cdot \lambda_{\text{in}} \cdot (1-\gamma_{-i})}{\alpha - (1-\hat{n})(\mathrm{E}_f[v(o)]\lambda_{\text{out}}^{\gamma_{-i}})}\right)^{\frac{1}{\gamma_{-i}}}\right)\right) \tag{11}$$

$$BR_{-i}(\gamma_i; o_{-i}) = \min\left(1, \max\left(0, \left(\frac{\hat{n} \cdot \mathrm{E}_f[v(o)] \cdot \lambda_{\text{in}} \cdot (1-\gamma_i)}{\alpha - (1-\hat{n})(\mathrm{E}_f[v(o)]\lambda_{\text{out}}^{\gamma_i})}\right)^{\frac{1}{\gamma_i}}\right)\right) \tag{12}$$

For an ex-ante symmetric Bayesian Nash equilibrium $\gamma_i^* = 0, \forall i$, to exist, the following condition needs to hold true. $BR_i(\gamma_{-i} = 0; o_i) = 0$ and $BR_{-i}(\gamma_i = 0; o_{-i}) = 0$.

Since the closed-form best response in Eqn. 11 contains the term $\frac{1}{\gamma_{-i}}$, it is not defined at $\gamma_{-i} = 0$. Therefore, we evaluate the ex-ante utility directly at $\gamma_{-i} = 0$ to determine when $BR_i(0; o_i) = 0$ holds.

When $\gamma_{-i} = 0$, we have $\gamma_i^{(1-\gamma_{-i})} = \gamma_i$ and $\lambda_{\text{out}}^{\gamma_{-i}} = 1$.

Substituting into Lemma 1, the ex-ante utility becomes

$$\begin{aligned}\mathrm{E}_f[u_i(\gamma_i, 0)] &= \hat{n} \cdot \mathrm{E}_f[v(o)] \cdot \lambda_{\text{in}} \cdot \gamma_i + (1-\hat{n}) \cdot \mathrm{E}_f[v(o)] \cdot \gamma_i - \alpha \cdot \gamma_i \\ &= \gamma_i\Big(\mathrm{E}_f[v(o)]\big((1-\hat{n}) + \hat{n}\lambda_{\text{in}}\big) - \alpha\Big).\end{aligned}$$

Hence, $BR_i(\gamma_{-i} = 0; o_i) = 0$ if and only if the coefficient on γ_i is non-positive. That is, the following condition has to be true:

$$\mathrm{E}_f[v(o)]\big((1-\hat{n}) + \hat{n}\lambda_{\text{in}}\big) \leq \alpha. \tag{13}$$

Rearranging the terms yields

$$\mathrm{E}_f[v(o)] < \frac{\alpha}{1 - \hat{n} \cdot (1-\lambda_{in})} \tag{14}$$

The same argument applies symmetrically to $BR_{-i}(\gamma_i = 0; o_{-i}) = 0$.

□

Necessity. If there exists an ex-ante symmetric Bayesian Nash equilibrium $\gamma_i^* = 0, \forall i$, then 0 must be a best response to 0. Therefore the above condition must hold. Hence,

$$\mathrm{E}_f[v(o)] < \frac{\alpha}{1 - \hat{n} \cdot (1-\lambda_{in})}$$

is a necessary condition for the existence of the equilibrium at $\gamma^* = 0$.

Sufficiency. Conversely, if the condition in Eqn 14 holds, then the coefficient on γ_i in $\mathrm{E}_f[u_i(\gamma_i, 0)]$ is strictly negative. Therefore, $BR_i(\gamma_{-i} = 0; o_i) = 0$ and symmetrically $BR_{-i}(\gamma_i = 0; o_{-i}) = 0$. These two conditions imply that $\gamma^* = 0$ is an ex-ante symmetric Bayesian Nash equilibrium.

Since the condition is both necessary and sufficient, the theorem follows.

B Research Methods

C Online Resources

Code for running the simulations included in the paper can be found under https://github.com/atrisha/normative_stewarding. The simulation was run on Intel Core i9 processor with 32GB RAM.

Received 06 November 2025; accepted 23 January 2026