Information stewards and normative persuasion in online social systems

Atrisha Sarkar
Schwartz Reisman Institute for Technology and Society
Vector Institute for Artificial Intelligence
University of Toronto
atrisha.sarkar@utoronto.ca

Gillian K. Hadfield
Schwartz Reisman Institute for Technology and Society
Vector Institute for Artificial Intelligence
University of Toronto
g.hadfield@utoronto.ca

ABSTRACT

In online social communities, influencers, moderators of communities, new media figures, etc., hold a privileged position on account of their visibility, and have been referred to as norm entrepreneurs because of their ability to broadcast their views about right and wrong. In recent years, there have been calls to consider regulation and reevaluate their roles and responsibilities towards the health of our social institutions. However, any act of regulation needs further understanding of their macro-level impact on society and expectation of an ideal behaviour. In this paper, we take a step in that direction by formalising the following question. What should be the ideal behaviour of such highly visible entities towards mitigating adverse societal phenomena such as opinion polarisation. We refer to such entities as *information stewards* and answer the above question in three steps. First, we develop a model of opinion exchange mediated by descriptive norms, i.e. expectations of others' opinions, based on a coordination game. Next, we demonstrate the conditions under which opinion can become polarised organically due to imperfect observation and the incentives to stay silent. Finally, we formalise the optimal behaviour of information stewards as an information design problem solved using sequential Bayesian persuasion. Our results show that in the absence of information stewarding, opinion polarisation can develop through a process similar to *spiral of silence* effect. Furthermore, we also show that whether an optimal information steward can mitigate societal polarisation depends on the proximity of their own opinion to the descriptive beliefs of community members, thereby highlighting the importance of the process through which the stewards are selected.

KEYWORDS

game theory, normative systems, computational social science

1 INTRODUCTION

Driven by the rise of social networks, online interactions have become an important aspect of social life [23, 26]. It is becoming increasingly clear that online social connections between people play an important role in the health of our political and social institutions. They have been an agent of positive change; acting as a catalyst in the fight for democratic rights and in giving voice to marginalised populations [11, 13]. However, the past few years have also shown ample evidence of the role of online communities in disrupting established institutions of public good, such as public health and democratic norms [16, 41]. In response, there have been calls for the oversight and regulation of online spaces [21, 31, 33]. However, when the drivers of the target of regulation, such as the spread of misinformation and "fake news", are consequences of partisanship [34], it can be an arduous task to achieve consensus on the form of regulation, as has been observed in US[1].

Among other considerations of privacy, data governance issues, and corporate power concentration, arguably some of the most widespread concerns around social media platforms have been the polarisation of public discourse [3, 19]. Part of the blame has been attributed to the inherent design of large social media platforms and the connection between the spread of information in homophilic groups and the rise of echo chambers [12, 38]. In response, computer scientists have developed several useful fairness oriented techniques vis-à-vis training data augmentation and controlling information flow with the goal of reducing group homophily [1, 35, 37]. However, other works have also pointed to the established correlation between increased political participation and polarisation with the difficulty of disentangling the causal chain, and harm mitigation can focus on ways to reduce extremism rather than addressing polarisation as the end goal [5].

An alternate perspective points to a much deeper problem. Various factors, including the scale of our social networks combined with algorithmic feedback, collectively have created an information ecosystem that is unprecedented in human society [4]. Social institutions and cultural norms evolve by latching on to human psychological elements that have evolved at a much slower biological time scale [20], and the growth and maintenance of inclusive norms of tolerance, democratic values, and the advancement of civil rights are not guaranteed to naturally flow from technology-mediated cultural change. Bak Coleman et al. [4] argue that there is a need for global stewardship of information to achieve cooperative behaviour in this new information ecosystem. Similarly, Fagan [15] highlights that influencers and public entities of note in social networks that take on the job of normative classification of behaviour, i.e., sharing opinions about right and wrong, act as *de-facto* information stewards describing them as *norm entrepreneurs.* Using their privileged position in the network, such entities coordinate behaviour by choosing which performances are up for judgements, as well as their own classification of approval and disapproval.

In this social system, the incentives at play for information stewards may not align with the broader social good that arises from the mitigation of polarisation and disinformation. From a legal standpoint, Fagan argues that within the tenets of Sec. 230 of the United States Code, it is prudent to consider the platform design, which includes the question of how stewards are selected for and their behaviour, as the focus of regulation rather than the content expressed by individual community users. Furthermore, as a part of

[1]https://rollcall.com/2021/12/07/partisan-bickering-could-doom-efforts-to-regulate-social-media-companies/

cost-benefit analysis of any such regulation, it is also necessary to take into account the self-regulation action taken by platforms, for example Reddit enforcing higher standards of behaviour of their community moderators compared to general users[2]. In that context, we present a computational approach to answer the following question. What should be the ideal behaviour of information stewards for mitigating polarisation in a population? This question is important for lawmakers and platform designers since both need insight into an ideal behaviour expected from information stewards as part of any legislative process and evaluating the effect of platform design changes, respectively. Specifically, we make the following contributions.

- *A normative microfoundation of opinion expression:* Analysing macro-level phenomena such as polarisation requires a model of individual behaviour. To that end, we present a normative model of opinion expression in which actions are mediated by descriptive norms, i.e., an expectation of the opinion of others' and a preference towards opinion conformance. We model agent's behaviour of choosing to stay silent or share their opinion as equilibrium actions of an underlying coordination game of opinion expression.
- *A mechanistic model for polarisation.* Similar to the *spiral of silence* effect [32], we demonstrate how a normative model under some mild conditions, namely people having different degrees of opinion and the option to stay silent on a focal issue, can result in self-reinforcing cycles of belief distortions leading to polarisation of opinions. Furthermore, we show that as a consequence of that effect, individuals with extreme opinions are over-represented in the set that share their opinions.
- *Optimal information design for information stewards.* Using a Bayesian persuasion framework [22], we cast the problem of optimal behaviour for information stewards as an information design problem. Using recent results from Gan et al. [18], we formulate the information design problem as a form of sequential Bayesian persuasion that can be modelled using a Markov Decision Process. Based on simulation and formal analysis, we demonstrate and quantify the reduction in polarisation and increase in community participation an optimal information stewardship can achieve.

2 THE MODEL

A normative microfoundation of individual behaviour relies on a few key elements: i) an individual's own incentives, ii) their belief about others' incentives, iii) beliefs about others' expectation of their own actions, and iv) the awareness of sanctions that can be levied based on their action. These constructive elements give rise to a wide category of norms that are typically present in a society (see Bicchieri [8] for a taxonomy of norms based on these constructs). Our model considers the minimal criteria of a *descriptive norm*, which is a behavioural rule that says that an individual prefers to conform if they believe that a sufficiently large subset of the population also conforms to the rule [8]. One can observe that behaviour driven by such descriptive norms is commonly prevalent in online communities, where the expression of opinion is often

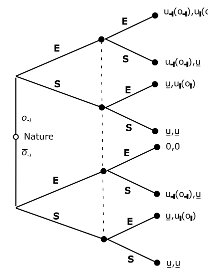


Figure 1: The coordination game structure of individual interactions of opinion exchange.

mediated by the belief about the approval and disapproval of their opinion.

- A population of N agents indexed as $i \in \{1, .., N\}$
- An opinion $o_i \sim f_{\theta^*}$ of an agent i drawn from the opinion distribution f_{θ^*}; $o_i < 0.5$ indicates *disapproval* on the focal issue and $o_i \geq 0.5$ indicates *approval.*
- A set of actions $A = \{E, S\}$ available to each agent, where the action E represents an agent expressing their opinion, and the action S represents the agent staying silent.
- A common descriptive belief θ, which is the parameter of the opinion distribution. Note, when $\theta \neq \theta^*$, it implies a distortion of belief with respect to the real distribution of opinion.
- Utility $u_i(a_i|o_i)$ of an action $a_i \in A$ to an agent i conditioned on the opinion of the agent.

Utilities. Agents can acquire utility based on the expression of their opinion and they can also secure a basic utility $\underline{u}$ when they remain silent. Regarding the semantic meaning of utility, one can consider that if the opinion values encode some notion of the strength of conviction, assertiveness, loudness, or any other affective attributes of an opinion, higher values in terms of its intensity will bring greater utility to the player, which can only be achieved by expressing an opinion. Therefore, we model the utility of sharing an opinion as a linearly increasing function of opinion value centered at 0.5.

$$u_i(E|o_i) = \begin{cases} o_i, & \text{if } o_i \geq 0.5 \\ 1 - o_i, & \text{otherwise} \end{cases} \tag{1}$$

Next, we present the interactions of agents based on the above model with the help of a coordination game.

2.1 Interaction model

Interactions between individuals, or players, in this model can be described in the form of a Bayesian game, as shown in Fig. 1. Prior to interaction, a player is not informed of the opinion of the other player they are about to interact; they may hold a prior belief about the other player's opinion, which is a function of the descriptive belief that the player holds about the members of the community, however, they do not have complete information about the opinion of the interacting agent. Interaction of this nature can be modelled as an incomplete information game with the opinion value of the other player, o_{-i}, acting as a *type* in standard game-theoretic terminology. From the perspective of a player, indexed as i and as

[2] https://www.redditinc.com/policies/moderator-code-of-conduct

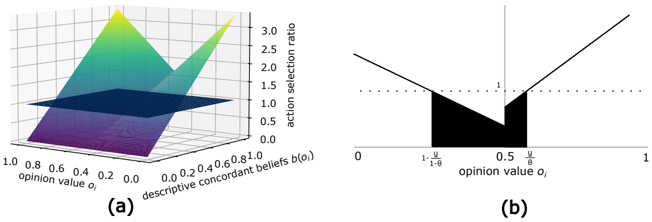


Figure 2: (a)The Bayesian Nash equilibrium action selection ratio at different levels of opinion value and descriptive concordant beliefs. (b) Cut of the the BNE action election

the second player in the game, nature draws the type (or opinion) of the other player, who in turn decides to stay silent (S) or express (E) their opinion. Based on the utilities of the actions, one can achieve the utility corresponding to the opinion value by expressing their opinion, while staying silent gets them a constant utility of $\underline{u}$. The game captures a preference of agents for conformation, and higher payoffs are achieved by both agents expressing their concordant opinions. Although in the general sense, the parameter θ may not have semantic meaning beyond just the parameters of the opinion distribution f_θ, for ease of analysis, we consider θ as the expectation of approval. This makes analysis easier since under a common prior, if b is the concordant belief for $o_i \geqslant 0.5$, i.e., for approval, then the corresponding concordant beliefs for opinions of disapproval, $o_i < 0.5$, is $1-b$. Specifically, we use the Bernoulli distribution as the common prior distribution and the concordant descriptive beliefs $b(o_i;\theta^0)$, that is, the belief of agent i that the interacting agent $-i$ shares the same opinion, can be calculated as θ^0 for $o_i \geqslant 0.5$ and $1-\theta^0$ otherwise. With θ^0 as the common prior in the Bayesian game, depending on the prior descriptive belief $b(o_i;\theta^0)$ and the opinion values of the two players, there are multiple Bayesian Nash equilibria (BNE) in the game.

Theorem 1. *The Bayesian Nash equilibrium, $\pi^*(o_i)$ of the normative opinion interaction game is*

$$\sigma^*(o_i) = \begin{cases} E, & \text{if } \frac{b(o_i;\theta^0)\cdot u_i(o_i)}{\underline{u}} > 1 \\ S, & \text{otherwise} \end{cases} \tag{2}$$

where $b(o_i;\theta^0)$ is the descriptive belief of a concordant opinion parameterized on the common prior distribution θ^0, $\underline{u}$ the utility for silence and $u_i(o_i)$ the utility of expressing opinion.

Proof. In Appendix. □

The equilibrium action of each agent is mediated by three factors, the opinion value o_i, the concordant descriptive belief $b(o_i;\theta^0)$, and the security utility $\underline{u}$. The probability of choosing to express one's opinion is positively correlated with the first two factors; a player is more likely to share their opinion if they have higher opinion value or higher descriptive beliefs. On the other hand, a higher security utility $\underline{u}$ carry an incentive to stay silent, and therefore the probability of staying silent increases. Fig.2a shows the BNE action selection ratio from Eqn. 2 at different opinion values ($x-$axis), as the well as different concordant descriptive beliefs $b(o_i;\theta^0)$ ($y-$axis) at a fixed value of $\underline{u} = 0.3$. The optimal action for any combination of opinions and concordant descriptive belief values for which the surface lies above $z = 1$ plane is to express their opinion. Notice that mid-range opinion values (near 0.5) need a higher concordant beliefs to express their opinion compared to opinion values in the extremes.

Under a common prior, there are two separate cuts of the surface, one for the concordant beliefs of $o_i \geqslant 0.5$ and another for $o_i < 0.5$. Fig. 2b shows the corresponding cuts of the BNE action selection surface for common prior $\theta^0 = 0.55$, which results in a cuts corresponding to approval opinions at $b(o_i;\theta^0) = 0.55$ and for opinions of disapproval at $b(o_i;\theta^0)$ or 0.45. Shaded regions are constructed from the BNE action of Eq. 2, and represent the opinion values for which the optimal action is to stay silent.

Lemma 1. *The maximum opinion value of minority opinion expression is $1-\frac{\underline{u}}{\theta^0}$ and the minimum opinion value of majority opinion expression is $\frac{\underline{u}}{\theta^0}$*

Proof. Solving for o_i based on the inequalities $\frac{\theta^0\cdot u_i(o_i)}{\underline{u}} - 1 > 0$ and $\frac{(1-\theta^0)\cdot u_i(o_i)}{\underline{u}} - 1 > 0$ for opinions for approval and disapproval, respectively. □

Corollary 1.1. *The minimum utility of majority opinion expression is greater than or equal to the minimum utility of minority expression.*

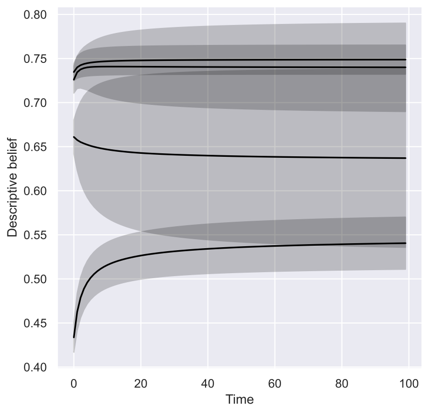
2.2 System dynamic and polarisation.

In this section, we present the evolution of the descriptive beliefs of the population and its corresponding impact on the population characteristics over time. We model the population as a set of myopic agents, meaning that the agents act by repeatedly playing the equilibrium action of the one-shot game described in the above sections[3]. We model agents' uncertainty over the common prior θ^0 as a Beta distribution with parameter α^0, β^0 where $\theta^0 = \frac{\alpha^0}{\alpha^0+\beta^0}$. At each subsequent time step, the sequence of interactions is as follows.

- *Action step:* At time step t, each agent selects the action of expression or silence based on Eqn. 2 and the descriptive belief at that time step, θ^t.
- *Update step:* Each agent observes the opinions expressed and updates their prior in accordance with only the expressed opinion values with the help of Bayes rule. The updated belief parameters are $\alpha^t + (c\cdot \bar{o_E}), \beta^t + (c\cdot(1-\bar{o_E}))$ where c is the update rate and $\bar{o_E}$ is the average opinion values of the expressed opinions. Agents update θ^{t+1} as the mean of the posterior distribution.

In an online community, it is imperceptible to identify why an individual stays silent; one may stay silent due to the incentives at play as modelled in the opinion exchange game or for a variety of other factors that result in the individual not being engaged in the conversation. The update step captures such imperfect observation by incorporating only the opinions that are expressed into the Bayesian update process. The update rate encodes the degree to which uncertainty about the approval rate is reduced with repeated

[3]This is in contrast to far-sighted agents who treat the interactions as a dynamic game and follow behaviour based on the equilibrium of the dynamic game.



(a)

Figure 3: Evolution of descriptive beliefs for different initial starting priors $\alpha^0, \beta^0 = \{(4,2),(3,1.3),(5,2),(2,3.3)\}$.

interactions. A higher value implies a faster rate of uncertainty reduction per time step. The new posterior parameters follow from the closed-form expression of the Bayesian update rule for Beta distributions.

Fig. 3a shows evolution trajectories of the descriptive belief of the population for $N = 100$ agents, 100 time steps, $\underline{u} = 0.2, c = 10$, and four different initial priors for $\alpha^0, \beta^0 = \{(4,2),(3,1.3),(5,2),(2,3.3)\}$. x-axis shows time steps and the y-axis shows the mean descriptive belief of the population. For each starting prior, we generate 100 uniform opinion distributions (f_{θ^*}) with the true rate of approval $\theta^* = 0.55$ and simulate the trajectories. The mean and standard deviation of the descriptive beliefs grouped by the sets of starting priors are shown in the figure.

In the evolution of the beliefs, there are two relevant aspects in the context of polarisation. First, depending on the value of security utility $\underline{u}$, repeated interactions can lead to more extreme opinions over time. This results from the fact that the individuals for whom silence is the optimal action are concentrated near moderate opinion values as seen in Fig. 2. The following lemma formalises the conditions under which the expected opinion value of the expressed opinions in the population will move to a more extreme value.

Lemma 2 (Polarization). *For opinion of approval, if $\underline{u} \geqslant \frac{\theta}{2}$, and $\theta^t < o_E$, then $\mathbb{E}_f[|o_A|]^{t+1} \geqslant \mathbb{E}_f[|o_A|]^t$. Conversely, for opinion of disapproval, if $\underline{u} \geqslant \frac{1-\theta}{2}$ and $\theta^t > o_E$ then $\mathbb{E}_f[|o_D|]^{t+1} \geqslant \mathbb{E}_f[|o_D|]^t$, where $o_A = \{o_i; o_i \geqslant 0.5\}$, $o_D = \{1-o_i; o_i < 0.5\}$, θ^t is the common expected belief of approval at time t, f is any arbitrary opinion distribution, and $\bar{o_E}$ is the average opinion values of the expressed opinions given by*

$$\bar{o_E} = \int_0^{\max\{0,1-\frac{\underline{u}}{1-\theta^t}\}} o \cdot f_{\theta^*}(o)do + \int_{\max\{0.5,\frac{\underline{u}}{\theta^t}\}}^{1} o \cdot f_{\theta^*}(o)do \qquad (3)$$

Second, extreme values of common descriptive beliefs result in the complete silence of minority opinion holders, and only majority opinions are expressed in the population. This phenomenon is similar to a *spiral of silence* phenomenon [32], however, it is interesting to note that prior to when the spiral is triggered, the residual minority opinions expressed will be the ones with extreme opinion values. One can see from the figure that when the population starts from a high prior $\theta^0 > 0.7$, the population never reaches the true approval rate, while starting from 0.66, only a small proportion of the belief trajectories reach the true opinion value. At the same value of opinion utility, agents with minority opinion values need higher concordant beliefs to express their opinion compared to majority opinion holders. Therefore, if the prior beliefs (θ^0) start from a high value, or alternatively, if the system trajectory reaches a high value, the optimal action of minority opinion holders is to stay silent, and subsequently the belief never reaches the level needed for the minority opinion holders to express their opinion regardless of the opinion value, therefore causing a spiral of silence. In our model, the threshold at which the spiral of silence in the direction of majority opinion is triggered is $\theta^t \geqslant 1 - \underline{u}$. If the mean true opinion is of approval (as in this case with $\theta^* = 0.55$), although the prior may be of disapproval, as long as $\theta \geqslant \bar{u}$, the beliefs will converge to the true mean following the standard Bayesian update process. This is seen in Fig.3a with the lower set of trajectories.

Lemma 3 (Spiral of silence). *If approval is a minority opinion and $\underline{u} \geqslant \theta^t$, then $\mathbb{E}_f[|o_A|]^{t'} = 0, \forall t' \geqslant t$. On the contrary, if disapproval is a minority opinion and $\underline{u} \geqslant 1-\theta^t$, then $\mathbb{E}_f[|o_D|]^{t'} = 0, \forall t' \geqslant t$*

3 PUBLIC PERSUASION

Having presented a dynamic that can organically lead to opinion polarisation, in this section, we analyse the role of information stewards in this social process. Specifically, we address two questions: How can we model the impact of a information steward who generates normative signals? What should be the optimal behaviour of such a steward with the goal of mitigating the polarisation observed in the population? We highlight the point that even in the presence of information stewards, the behaviour of community members are still mediated by purely descriptive beliefs; in other words, we do not model information stewards as entities with sanctioning powers over the community members. However, due to the privileged position of the information steward, their actions can act as public signals, which influence the descriptive beliefs of the community members, which in turn influence the community behaviour. In a formal sense, we cast the problem of optimal behaviour of information stewards as a problem of optimal signals generated in a Bayesian persuasion framework [22]. We first provide the necessary background on Bayesian persuasion, and then discuss the application of the framework to our problem.

3.1 Background

The basic framework of Bayesian persuasion is a game of information between a sender and a receiver, which in our case are the steward and the community members, respectively. The sender has the choice of designing a signalling distribution $\pi(s|\theta)$ over an arbitrary space of signals $s \in S$ conditioned on a state $\theta \in \Theta$.

The sender's action is a signal realisation $s \sim \pi(s|\theta)$ sampled from the signaling distribution. The structure of the game imposes a constraint on the sender to truthfully convey the signal realisation from the chosen publicly observed distribution π to the receiver. The utility of the receiver $u(a, \theta)$ depends on its action a and the state θ. The receiver holds a belief μ over the state and selects their optimal action that maximises their own utility based on their beliefs according to the following equation.

$$a_r^* = \arg\max_a \mathbb{E}_{\theta\sim\mu}[u_r(a, \theta)] \tag{4}$$

The action of the sender depends on the action taken by the receiver and the sender's utility is represented as $u_s(a_r^*, \theta)$. Both share a common prior over the state, that is, $\mu_0(\theta)$. Based on the signal s observed by the receiver, the receiver updates their belief μ based on Bayes rule, and chooses the optimal action based on the posterior beliefs. Note that the optimal action of the receiver depends on their posterior belief (Eq. 4), which *influences* the utility of the sender through the receiver's action a_r^*, and in turn is *influenced by* the signal $s \sim \pi(s|\theta)$ that the sender generates. Therefore, the main decision problem of the sender is to choose the optimal signalling distribution $\pi(s|\theta)$ such that they can maximise their own utility when the behaviour of the receiver is constrained by the possible set of posterior beliefs that can be induced by the choice of the signalling distribution.

4 STEWARD SIGNALS AS NORMATIVE PERSUASION

Let us now consider the case of a steward whose goal is to mitigate polarisation and the spiral of silence as discussed in Sec.2. Recall that in our model, information stewards act as the sender in a Bayesian persuasion framework. Although the space of signals S the stewards can generate can be arbitrary, as long as a signal does not convey any extra information, without loss of generality, one can consider the space of signal realisations to be the direct action recommendations that the steward wants the community to do, similar to *revelation principle* [18, 30]. In our case, the realisation of the signal consists of approval and disapproval recommendations, indexed as 1 and 0, $s = \{0, 1\}$, which are drawn from a chosen Bernoulli distribution $\pi(s|\theta) = Bern.(\theta_s)$, where the subscript s stands for steward. Therefore, the problem of an optimal steward is choosing the optimal parameter θ_s^* with the goal of mitigating polarisation.

There are two unresolved aspects of this information design problem; first, there is the commitment issue of whether the steward truthfully gives the sampled signal realisation from the signalling distribution. Once we interpret the signalling distribution as the publicly observed behaviour of the steward, it follows straightforwardly that their specific action recommendation of approval or disapproval (the signal s) is representative of the behaviour they have chosen for themselves (the distribution $Bern.(\theta_s)$). The second unresolved question is the connection between the state θ and the choice of the steward's distribution θ_s. In a typical Bayesian persuasion model, since $\pi(s|\theta)$ is a conditional distribution on the state, it conveys information about the true state of nature, which the sender uses to update the posterior belief about the true state. However, in our case, there are two different interpretations of the

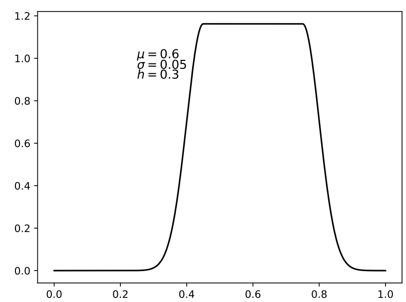


Figure 4: Constraining distribution connection the descriptive belief θ^t and signalling distribution θ_s^t

true state that can be considered. One can consider the state to represent the expectation of the entire opinion distribution $\mathbb{E}_{f_{\theta^*}}(o_i)$, or an alternative choice is to consider the state that represents the expected opinion of only the agents who express their opinion as encoded in the descriptive belief θ^t at each time step. Since the latter is the one that shapes the behaviour of the population and it imposes less restriction on the information that the steward can convey, we consider the second alternative.

Steward constraints One key aspect of information stewards is that they are members of the community who come to that privileged position through some underlying social choice mechanism, including informal ones such as through individuals "liking" and "subscribing" to online media content, and also formal ones where many online communities on Reddit and elsewhere vote on their moderators. Therefore, it is conceivable that the behaviour of the steward (as in their signalling distribution) would be in some degree of alignment with the behaviour an individual would expect of their community members, which is captured by their descriptive beliefs. Therefore, from a Bayesian perspective, this means that the steward's signals convey information about the underlying belief of the population. We model this relation between the steward's distribution parameter θ_s and the descriptive belief θ^t by imposing a constraining distribution C that connects the two parameters. One can interpret this constraint as the degree to which the behaviour of the steward is representative of the descriptive beliefs of the community. We choose a constraining distribution with the condition that the absolute difference between the descriptive beliefs and the parameters of the steward's signal distribution is bound by a threshold τ, $|\theta^t - \mathbb{E}_C[\theta_s^t]| \leqslant \tau$. Fig. 4 shows an example of such a distribution parameterized by a width parameter $h = \tau$, mean $\mu = \theta_s^t$, and standard deviation $\sigma = 0.05$. Details of the distribution are included in the appendix. 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 behaviour of the steward and descriptive beliefs specific to a community.

Putting it all together, the system dynamics of the model under signals by a information steward is as follows:

- A performance arrives for deliberation.
- The information steward chooses a publicly observed signalling distribution $\pi_s(\theta^t, \theta_s^t)$, where θ^t is the common descriptive belief about probability of approval and θ_s^t is the parameters of the distribution from a family of distributions chosen by the steward.
- The steward generates a signal $s \sim \pi_s(\theta^t, \theta_s^t)$ from the signaling distribution.
  - The community observes the signal sample and updates their descriptive belief according to Bayes formula.

$$\pi(\theta'^t|s) = \frac{\pi(s|\theta^t;\theta_s^t) \cdot \text{Beta}(\alpha^t, \beta^t)}{\int \pi(s|\theta^t;\theta_s^t) \cdot \text{Beta}(\alpha^t, \beta^t)\, d\theta^t} \tag{5}$$

  where θ'^t is the posterior descriptive belief after seeing the signal s.

The rest of the steps are identical to the *Action* and *Update* steps of Sec. 2.2.

- Each member of the community chooses their Bayesian Nash equilibrium action based on the posterior belief $\pi(\theta'^t|s)$ and Eqn. 2.
- Each member observes the opinions values expressed in the population and performs another round of Bayesian belief update as described in the *update* step of Sec. 2.2.

In eqn. 5, $\pi(s|\theta^t;\theta_s^t)$ represents the likelihood of observing the signal realization s in state θ^t. The likelihood function can be expressed as the product of the probability of the signal being generated based on the parameter θ_s^t and the constraint distribution C, both of which are available knowledge to the receivers of the signal. Following is the expanded Bayesian update expression from eqn. 5 incorporating the constraining distribution C.

$$\pi({\theta'}^t|s) = \frac{Bern(s;\theta_s^t) \cdot C(\theta^t, \theta_s^t) \cdot \text{Beta}(\alpha^t, \beta^t)}{\int Bern(s;\theta_s^t) \cdot C(\theta^t, \theta_s^t) \cdot \text{Beta}(\alpha^t, \beta^t)\, d\theta^t}$$

4.1 Optimal signals

The minimal criterion for an effective steward is that their signals should not induce greater polarisation in the population compared to the absence of the steward which was modelled in Sec. 2.2. Formally, this means that the following two conditions should hold for such a steward.

$$(\mathbb{E}_f[|o_A|]^{t+1} - \mathbb{E}_f[|o_A|]^t) - (\mathbb{E}_{f_s}[|o_A|]^{t+1} - \mathbb{E}_{f_s}[|o_A|]^t) \geqslant 0$$

$$(\mathbb{E}_f[|o_D|]^{t+1} - \mathbb{E}_f[|o_D|]^t) - (\mathbb{E}_{f_s}[|o_D|]^{t+1} - \mathbb{E}_{f_s}[|o_D|]^t) \geqslant 0$$

where o_A and o_D follows the same definition from Lemma 2, f^t, f^{t+1} are the distribution of expressed opinion values in two subsequent time steps following the system dynamic in Sec. 2.2 (absence of steward) and $f_s^t, f_s^{t+\Delta t}$ are the distribution following the system dynamic in Sec. 4 under the signals of the steward. The utility of the steward depends on how much their signals can reduce polarization, and for an *optimal* steward, their signals should not only keep the reduction in polarisation positive (as in the above equations) but also induce maximum reduction. The rest of the section focuses on the solution to the latter problem of optimal signals.

Both under the presence and absence of the steward, the system dynamic of Secs. 2.2 and 4 are Markovian in nature, since the next state (the descriptive belief θ^{t+1}) depends only on the current state of the descriptive belief θ^t and optionally the action of the steward. Therefore, the problem of optimal signal generation for the steward can be formulated as a solution to the following Markov Decision Process (MDP) with the usual constructs of *State, Action, Transition, Rewards* as follows:

- *State:* The descriptive belief of the population θ^t
- *Action:* The parameters of the signaling distribution θ_s^t
- *Transition:* The transition from θ^t to $\theta^{t+\Delta t}$ governed by a transition function

$$T(\theta^t, \theta_s^t, \theta^{t+\Delta t}) = \sum_{s\in\{0,1\}} Bern.(s;\theta_s^t) \cdot \pi(\theta'^t|s)\mathbb{E}_{f_{\theta'^t}}[o_E] \tag{6}$$

  where $o_E = \{o_i : \sigma^*(o_i) = E\}$ and the expectation $\mathbb{E}_{f_{\theta'^t}}[o_E]$ is calculated the same way as in Eqn. 3 but with the posterior parameter θ'^t.
- *Rewards:* Since an increase in participation leads to decrease in polarization (one can see this by looking at the opinion values corresponding to silence actions in Fig. 2b), we design the rewards based on the level of participation in the community. Maximum reward +1 is achieved when the signal induces full participation. $R(\theta^t, \theta_s^t) = 2(\frac{N_E}{N} - 0.5)$ where $N_E = \{i : \sigma(o_i)^* = E\}$, that is, the number of agents expressing their opinion.

The transition function in Eqn. 6 is constrained by multiple factors. First, the committed signaling distribution of the steward ($Bern.(s;\theta_s^t)$), the posterior beliefs of the agents ($\pi(\theta'^t|s)$, Eqn. 5), and the expected opinion of the population based on the equilibrium action of the agents ($\mathbb{E}_{f_{\theta'^t}}[o_E]$) with respect to the posterior descriptive beliefs. The formulation of the MDP is one of sequential Bayesian persuasion [18] and the policy of an optimal steward is one that maximises the following value function:

$$V(\theta^t) = \max_{\theta_s^t\in[0,1]}[R(\theta^t, \theta_s^t) + \sum_{t=1}^{\infty} T(\theta^t, \theta_s^t, \theta^{t+\Delta t})V(\theta^{t+\Delta t})]$$

We solve for the optimal signaling policy in the MDP using dynamic programming after discretising the state and the action space with uniform increments of 0.1 [39]. Fig. 5 shows the evolution of population descriptive beliefs for the same four starting priors as in Fig. 3a, i.e., $\alpha^0, \beta^0 = \{(4,2), (3,1.3), (5,2), (2,3.3)\}$ with two different constraining distributions ($\tau = 0.1$ in the bottom panel and τ=0.2 in the top panel). It is evident from the figure of $\tau = 0.2$ that for the priors that triggered a spiral of silence, effect ((3,1.3),(5,2)), the optimal signals by an information steward were able to bring the descriptive beliefs down to moderate levels, thereby preventing the spiral from taking hold. Similar results are also seen for the prior (4,2) in which the spiral of silence took hold for only a few trajectories (as a result of randomness in the initial distribution of opinions). Also note that for prior (2,3.3), in which the spiral was not observed, the addition of the steward signal is able to maintain the absence of the spiral. On the other hand, when the constraining distribution imposes a stricter restriction on the distribution of the steward signals ($\tau = 0.1$), in all cases, the steward cannot mitigate the development of a spiral of silence.

The above results demonstrate an important finding of our paper with respect to the role of stewards. In order to mitigate polarisation

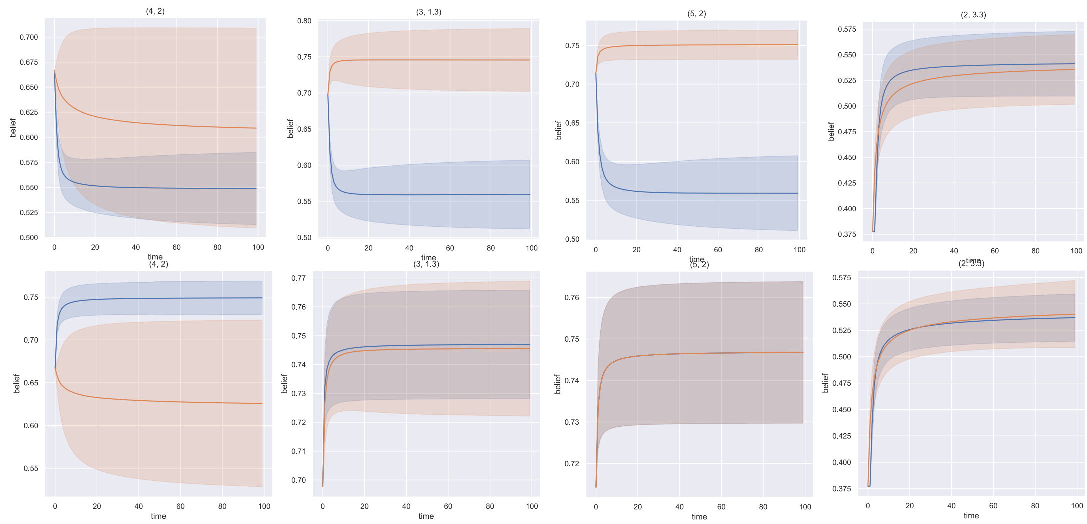


Figure 5: Evolution of descriptive belief in the presence of a steward signal with constraining distribution parameter $h = 0.2$ (top panel) and $h = 0.1$ (bottom panel)

and prevent the spiral of silence, it is not necessary for minority and moderate opinion holders to express their opinion when the optimal action in the opinion exchange game is to stay silent. Polarisation can also be mitigated when the following two conditions hold: i) the community has an underlying mechanism through which minority and moderate opinion holders can select information stewards. This would lead to a relaxation of the constraining distribution C by increasing the threshold value τ, ii) the selected information stewards generate signals that are optimal based on the model presented in Sec. 4.1. Returning to the main motivation of the paper, these results demonstrate the importance of information stewards, and whether the consideration is one of new regulation for lawmakers or design decisions for platform designers, it is necessary to reevaluate the role of information stewards in the community.

5 RELATED WORK

In this section we discuss the three contributions of the paper in relation to the relevant literature on computational models of normative systems and dynamics of the spiral of silence. Normative models are an important framework for studying cooperation in human groups, and the existing literature spans multiple disciplines (see Legros and Cislaghi [25] for a general review). Specific recurring research questions revolve around the role of norms in economic and social behaviour [14], the design of norm-driven intervention strategies towards positive behaviour change [9], and the role of norms in experimental and behavioural game theory [10]. From a theoretical point of view, social norms act as a resolution mechanism for social interactions that present a conflict between individual incentives and collective benefit, commonly seen in public goods problems and the management of the commons [2]. Within social norms literature, significant attention has been paid to problems in which individual and group welfare are in direct conflict *a. la.* prisoners' dilemma type social interactions. However, the role of social norms has also been established in a more broader category of social interactions beyond where the interest of collective good is in direct opposition with individual incentives. Examples of such interactions include a coordination problem and the type of conflict that materialises as between the high-risk choice with higher collective utility and the low-risk choice of lower collective utility, as exemplified in a stag hunt game. We derive our opinion exchange game from such type of interactions.

The dynamic of polarization in our model is due to the distortion of descriptive beliefs in the community. A related work on distorted beliefs is the recent work by Morsky and Akćay [28], which addresses the role of false beliefs in sustaining cooperation. Similar to our model, the paper models both micro-level individual behaviour as a coordination game mediated by conditional conformance and the population impact at the macro-level. However, there are two key differences from our work. First, Morsky and Akćay focus on the problem of norm maintenance and, therefore, use a public goods game as the interaction scenario. Whereas we model the interaction process relevant to our problem through a game of opinion exchange. However, there is commonality between the models with respect to the coordination structure and descriptive norms. Second, the aetiology of the belief distortion in Morsky and Akćay is exogenous to the model and is modelled as specific biases of over- and underestimation of the true level of cooperation in the population. In our case, the origin of belief distortion is endogenous and follows minimally from the agents following their optimal action in the opinion expression game and incomplete observations.

Another related phenomenon in social psychology literature in relation to belief distortion is false consensus effect (FCE), i.e., the overestimation of one's first-order belief about others' opinions due to egocentric bias [36]. FCE has been one of the most consistently and well-established phenomena supported by experimental evidence [29]. The connection of false consensus effect on social norms has also been studied empirically and theoretically [6, 7, 40]. The main difference between our model and FCE models is that, in FCE, the distortion in belief is correlated with the individual opinion values. In comparison, belief distortion in our model is independent of the individual opinion values, and descriptive belief is common across all opinion values; therefore, in that respect the resulting effect of polarisation is different from those that arise due to FCE [24].

Spiral of silence is a widely discussed theory in communication studies in which minority opinion holders' fear of isolation drives them into a self-reinforcing silence and thereby we observe the absence of minority views from the population. The phenomenon has seen growing interest in light of online opinion expression and a meta-analysis of sixty-six studies by Matthes et al. [27] show that the effect between opinion perception and expression that contributes to a spiral is strong in online social systems. A closely related work on the computational modelling of dynamics leading to a spiral of silence is by Gaisbauer et al. [17]. Similar to our work, Gaisbauer et al. build a microfoundation 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 few key differences in the modelling constructs between our model and the Gaisbauer et al. model. First, their model uses a network structure and intensity of connection between agents on that network to model the optimal action of agents to stay silent or express opinion. In comparison, ours is a normative model, where the optimal action is mediated by beliefs and opinion values without reference to a specific network structure. Additionally, we also develop a framework for mitigating the spiral using the Bayesian persuasion framework as a contribution along with the opinion expression model.

6 CONCLUSION

In this paper we make the case that information stewards (entities that hold a privileged position in online communities and can broadcast normative signals about right and wrong) have an important role to play in the health of social communities and in mitigating adverse phenomena, such as polarisation. We formulate the role of the information steward as entities generating normative signals, and members of the community use those signals to update their descriptive beliefs in a game of opinion exchange with other community members. Our work highlights that there are two countervailing forces on the steward; first, the fact that they are selected from within the community imposes a constraint of how much they can deviate from the descriptive belief of the population, and second, whether the signals generated by the stewards are optimal with respect to mitigating polarisation. Based on our results, we show that the steward's overall impact depends on how these two forces play out, and that with some relaxation on how far the steward can deviate from the descriptive belief combined with an optimal policy, they can mitigate polarisation. In our view, the computational models presented in the paper is valuable both for regulatory analysis and as a guidance for policy design for online social communities.

7 ACKNOWLEDGEMENTS

We thank Graham Noblit and Valerie Platsko for theit helpful feedback on this work.

REFERENCES

[1] Faisal Alatawi, Lu Cheng, Anique Tahir, Mansooreh Karami, Bohan Jiang, Tyler Black, and Huan Liu. 2021. A survey on echo chambers on social media: Description, detection and mitigation. *arXiv preprint arXiv:2112.05084* (2021).

[2] John M Anderies and Marco A Janssen. 2016. *Sustaining the commons*. Independent.

[3] Swapan Deep Arora, Guninder Pal Singh, Anirban Chakraborty, and Moutusy Maity. 2022. Polarization and social media: A systematic review and research agenda. *Technological Forecasting and Social Change* 183 (2022), 121942.

[4] Joseph B Bak-Coleman, Mark Alfano, Wolfram Barfuss, Carl T Bergstrom, Miguel A Centeno, Iain D Couzin, Jonathan F Donges, Mirta Galesic, Andrew S Gersick, Jennifer Jacquet, et al. 2021. Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences* 118, 27 (2021), e2025764118.

[5] Pablo Barberá. 2020. Social media, echo chambers, and political polarization. *Social media and democracy: The state of the field, prospects for reform* 34 (2020).

[6] Kathleen Patricia Bauman. 1997. *The role of perceived social norms on attitudes and behavior: An examination of the false consensus effect.* University of New Hampshire.

[7] Kathleen P Bauman and Glenn Geher. 2002. We think you agree: The detrimental impact of the false consensus effect on behavior. *Current Psychology* 21 (2002), 293–318.

[8] Cristina Bicchieri. 2005. *The grammar of society: The nature and dynamics of social norms.* Cambridge University Press.

[9] Cristina Bicchieri, Eugen Dimant, Michele Gelfand, and Silvia Sonderegger. 2022. Social norms and behavior change: The interdisciplinary research frontier.

[10] Colin F Camerer, Ernst Fehr, et al. 2004. Measuring social norms and preferences using experimental games: A guide for social scientists. *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies* 97 (2004), 55–95.

[11] Bronwyn Carlson and Ryan Frazer. 2020. "They got filters": Indigenous social media, the settler gaze, and a politics of hope. *Social Media+ Society* 6, 2 (2020), 2056305120925261.

[12] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118.

[13] Shelley L Craig, Andrew D Eaton, Lauren B McInroy, Vivian WY Leung, and Sreedevi Krishnan. 2021. Can social media participation enhance LGBTQ+ youth well-being? Development of the social media benefits scale. *Social Media+ Society* 7, 1 (2021), 2056305121988931.

[14] Jon Elster. 1989. Social norms and economic theory. *Journal of economic perspectives* 3, 4 (1989), 99–117.

[15] Frank Fagan. 2017. Systemic social media regulation. *Duke L. & Tech. Rev.* 16 (2017), 393.

[16] Elia Gabarron, Sunday Oluwafemi Oyeyemi, and Rolf Wynn. 2021. COVID-19-related misinformation on social media: a systematic review. *Bulletin of the World Health Organization* 99, 6 (2021), 455.

[17] Felix Gaisbauer, Eckehard Olbrich, and Sven Banisch. 2020. Dynamics of opinion expression. *Physical Review E* 102, 4 (2020), 042303.

[18] Jiarui Gan, Rupak Majumdar, Goran Radanovic, and Adish Singla. 2022. Bayesian persuasion in sequential decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5025–5033.

[19] Andrew M Guess and Benjamin A Lyons. 2020. Misinformation, disinformation, and online propaganda. *Social media and democracy: The state of the field, prospects for reform* 10 (2020).

[20] Joseph Henrich. 2015. The secret of our success. In *The Secret of Our Success.* princeton University press.

[21] Tim Hwang. 2020. Dealing with disinformation: evaluating the case for Amendment of Section 230 of the communications decency act. *Social Media and Democracy: The State of the Field and Prospects for Reform* (2020), 252–285.

[22] Emir Kamenica and Matthew Gentzkow. 2011. Bayesian persuasion. *American Economic Review* 101, 6 (2011), 2590–2615.

[23] Kawaljeet Kaur Kapoor, Kuttimani Tamilmani, Nripendra P Rana, Pushp Patil, Yogesh K Dwivedi, and Sridhar Nerur. 2018. Advances in social media research: Past, present and future. *Information Systems Frontiers* 20 (2018), 531–558.

[24] Tim Kenyon. 2014. False polarization: debiasing as applied social epistemology. *Synthese* 191 (2014), 2529–2547.

[25] Sophie Legros and Beniamino Cislaghi. 2020. Mapping the social-norms literature: An overview of reviews. *Perspectives on Psychological Science* 15, 1 (2020), 62–80.

[26] Eden Litt, Siyan Zhao, Robert Kraut, and Moira Burke. 2020. What are meaningful social interactions in today's media landscape? A cross-cultural survey. *Social Media+ Society* 6, 3 (2020), 2056305120942888.

[27] Jörg Matthes, Johannes Knoll, and Christian 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 (2018), 3–33.

[28] Bryce Morsky and Erol Akçay. 2021. False beliefs can bootstrap cooperative communities through social norms. *Evolutionary Human Sciences* 3 (2021), e36.

[29] Brian Mullen, Jennifer L Atkins, Debbie S Champion, Cecelia Edwards, Dana Hardy, John E Story, and Mary Vanderklok. 1985. The false consensus effect: A meta-analysis of 115 hypothesis tests. *Journal of Experimental Social Psychology* 21, 3 (1985), 262–283.
[30] Roger B Myerson. 1981. Optimal auction design. *Mathematics of operations research* 6, 1 (1981), 58–73.
[31] Philip M Napoli. 2019. *Social media and the public interest: Media regulation in the disinformation age.* Columbia University Press.
[32] Elisabeth Noelle-Neumann. 1974. The spiral of silence a theory of public opinion. *Journal of communication* 24, 2 (1974), 43–51.
[33] Jonathan A Obar and Steven S Wildman. 2015. Social media definition and the governance challenge-an introduction to the special issue. *Obar, JA and Wildman, S.(2015). Social media definition and the governance challenge: An introduction to the special issue. Telecommunications policy* 39, 9 (2015), 745–750.
[34] Mathias Osmundsen, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann, and Michael Bang Petersen. 2021. Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review* 115, 3 (2021), 999–1015.
[35] Bashir Rastegarpanah, Krishna P Gummadi, and Mark Crovella. 2019. Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. In *Proceedings of the twelfth ACM international conference on web search and data mining*. 231–239.
[36] Lee Ross, David Greene, and Pamela House. 1977. The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology* 13, 3 (1977), 279–301.
[37] Jakob Schoeffer, Alexander Ritchie, Keziah Naggita, Faidra Monachou, Jessie Finocchiaro, and Marc Juarez. 2022. Online Platforms and the Fair Exposure Problem Under Homophily. *arXiv preprint arXiv:2202.09727* (2022).
[38] Cass Sunstein. 2018. # Republic. In *# Republic*. Princeton university press.
[39] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction.* MIT press.
[40] Magdalena Wojcieszak and Vincent Price. 2009. What underlies the false consensus effect? How personal opinion and disagreement affect perception of public opinion. *International Journal of Public Opinion Research* 21, 1 (2009), 25–46.
[41] Thomas Zeitzoff. 2017. How social media is changing conflict. *Journal of Conflict Resolution* 61, 9 (2017), 1970–1991.

8 APPENDIX

Proof of Eqn. 2: Each player i knows their own opinion, therefore, their own type (o_i), however, they are unaware of the type of the other player (o_{-i}). Player i playing as the row player can face essentially two types of players. One who match their opinion (e.g., $o_i < 0.5$ and $o_{-i} < 0.5$) or another type who doesn't. We model these two *coarse* types in the game since these two sets are the ones that are relevant for the payoffs. Table 1 shows the payoff table from the normal form strategies of the game with the row player i's strategies E and S, and column player's strategies EE, SS, ES, EE, where the first letter denotes the strategy of the type with matched opinion and second letter denotes the strategy of the type who doesn't match the opinion. One can see that under the condition $b \cdot u_i(o_i) > \underline{u}$, E is a strictly dominant strategy.

Constraining distribution. The probability density function of the constraining distribution C is

$$C(x; \mu, h, \sigma) = \begin{cases} \frac{c}{\sqrt{2\pi}\sigma} \exp^{-\frac{1}{2\sigma^2}(x-\mu+h/2)^2}, & \text{if } x \leqslant \mu - (h/2) \\ \frac{c}{\sqrt{2\pi}\sigma}, & \text{if } \mu - h/2 < x \leqslant \mu + h/2 \\ \frac{c}{\sqrt{2\pi}\sigma} \exp^{-\frac{1}{2\sigma^2}(x-\mu-h/2)^2}, & \text{if } x > \mu + (h/2) \end{cases}$$

where $c = \frac{1}{1+h/\sqrt{2\pi}\sigma}$.

	SE	SS	ES	EE
E	$b \cdot u_i(o_i), b \cdot \underline{u}$	$u_i(o_i), \underline{u}$	$u_i(o_i), (1-b) \cdot \underline{u} + b \cdot u_{-i}(o_{-i})$	$b \cdot u_i(o_i), b \cdot u_{-i}(o_{-i})$
S	$\underline{u}$,$b \cdot \underline{u} + (1-b) \cdot u_{-i}(o_{-i})$	$\underline{u}, \underline{u}$	$\underline{u}, b \cdot u_{-i}(o_{-i}) + (1-b) \cdot \underline{u}$	$\underline{u}, u_{-i}(o_{-i})$

Table 1: Payoff table for the opinion expression game.