# When do Misinformation Policies (not) Work?

**Abstract**

We use a simple model to analyze several policies currently proposed in the public sphere to reduce the effects of misinformation. We show that the efficacy of these policies crucially depends on the strategic sophistication and reasoning abilities in the population. We focus on the following policies: *censorship*, where news can be moderated by governments or social media platforms; *content diversification*, where agents are given news representing different viewpoints or are shown news that is counter to their prevailing beliefs; *accuracy nudging*, where agents are encouraged to think more critically about news they receive; and *performance targets*, where social media outlets try to regulate the amount of misinformation on their platforms. We show that policies that work well for naive agents can perform poorly or completely backfire for Bayesian agents and vice versa. These insights highlight the importance of sophistication as a factor that regulators should consider when deploying policies to fight misinformation.

## 1 Introduction

The challenges that misinformation presents to learning and decision making touch on all aspects of our lives. The slow uptake of the COVID-19 vaccine in many parts of the world provides a recent and important example. Policymakers and researchers have attributed a substantial portion of this vaccine hesitancy to misinformation,[1] and policies to combat the misinformation problem are currently being proposed and debated in the public sphere. These policies range from gentle nudges that encourage people to carefully consider the veracity of the news they receive to more interventionist strategies like outright censorship of content.

In this paper, we argue that an important factor currently missing from this policy debate is the strategic sophistication of the population. There is recent work documenting the role that this sophistication plays in how agents respond to misinformation, e.g., Pennycook and Rand (2018, 2021). Yet, this has not been an element that regulators have incorporated into their discourse. We examine several policies through the lens of agent sophistication. These policies are currently proposed or deployed by regulators, platforms, and researchers across multiple

[1] See for example Loomba et al. (2021) and the press release from the U.S. Surgeon General (2021).

disciplines. For all these policies, strategic and cognitive sophistication emerges as an important factor that can make the same policy successful in one context but unsuccessful or even harmful in another.

We model sophistication using the classical models of the social learning literature: (sophisticated) Bayesian learning and (naive) DeGroot learning.[2] In these models, agents try to uncover an underlying state of the world, e.g., whether a vaccine is safe or not, and do so by learning from news they receive as well as by exchanging opinions with each other. The news that agents receive contains a mixture of organic news (which is correlated with the correct state but not completely accurate) and misinformation (which is orthogonal to the state). Agents do not know if a piece of news is organic or misinformation, but as documented empirically in van der Linden et al. (2020), they know that misinformation exists but do not agree on whether it comes mostly from the left or right. Mostagir and Siderius (2021) show that in this setting, Bayesian agents can be *less* immune to misinformation and more prone to mislearning than DeGroot agents . Other recent work has studied how different learning mechanisms can contribute to polarization and spread of misinformation. For example, in Haghtalab et al. (2021), agents who start with the same information but use different learning mechanisms can reach conflicting conclusions. In a similar vein, Jackson et al. (2019) demonstrate that limiting the number of times a message can be shared in a social network can make DeGroot agents more robust to information distortion compared to their Bayesian counterparts. We show that these differences between learning mechanisms extend to how agents respond to different policies. The policies we consider are those currently discussed in the public sphere to curb misinformation.

Our paper is organized as follows. In Section 2, we discuss an overview of our results, including the intuitions for our findings, and the ties to existing empirical literature on misinformation policies. In Section 3, we supply our formal model of reasoning and networked learning. In Section 4, we provide the four main results about the efficacy of our four different misinformation policies. In Section 5, we conclude. All proofs and simulations are provided in the Appendix.

## 2 Overview of Results

**Censorship.** Platforms sometimes remove content that they believe is false and/or harmful. For example, in the early stages of the pandemic Twitter stated that it would begin removing Coronavirus-related posts that deny expert recommendations or promote fake treatments and

[2] For more context, see the survey of Golub and Sadler (2017). Additionally, for even more details, see the works of Gale and Kariv (2003), Mueller-Frank (2013) and Mossel et al. (2014) for Bayesian learning, and Degroot (1974) and Golub and Jackson (2010) for DeGroot learning.

prevention techniques.[3] Theorem 1 shows that such policies do not work for a Bayesian population, who start *rationally* spinning narratives about *what else* the platform might be censoring, and use that rationalization to hold on to their existing beliefs and not learn. On the other hand, DeGroot populations can be influenced and have their minds changed through censorship, with the caveat that the platform is reasonably sure that what they are censoring is highly likely to be false information.

**Content Provision.** Balanced provision of content is another contender for a policy that might dilute the effects of misinformation. In this policy, agents are shown a diverse set of news that might be counter-attitudinal to their beliefs. The idea is that social media, through microtargeting, keeps users in "filter bubbles," where they mostly see content that agrees with their beliefs. Exposure to different viewpoints can move agents out of these bubbles and improve their learning outcomes. This policy echoes the Fairness Doctrine, which required news providers to present both sides of a controversial issue. The policy was eliminated in 1987, in a decision that has been cited as one of the reasons for the rise of partisan talk radio. Theorem 2 shows that the negative consequences that resulted from biased coverage cannot be reversed by simply reinstating an equal-coverage policy. While the policy helps DeGroot agents learn the true state, its effects on Bayesian agents is more complex: when the amount of misinformation in the system is manageable, showing diverse content can help learning. However, once the amount of misinformation crosses a certain threshold, diverse content does not help.

**Accuracy Nudging.** Mild interventions that only nudge people towards certain actions have become ubiquitous in fields like healthcare and financial planning. In the context of misinformation, accuracy nudging encourages and reminds agents to carefully consider the veracity of the news they receive. Experiments have shown that these nudges can increase awareness and help people spot false news and eventually learn better, although replications suggest that the size of the effect might not be as large as initially believed. Our model offers several interesting insights when it comes to how nudging interacts with strategic sophistication, and offers a possible explanation for these experimental observations. Bayesian agents are aware of the presence of misinformation and take that into account when learning from news, and so nudging leaves them unaffected. DeGroot agents however can benefit from nudging, but interestingly, it is also possible that encouraging them to think more carefully about news can lead to worse learning outcomes. When nudging is beneficial, DeGroot agents are endowed with "moderate sophistication": they are not as gullible as traditional DeGroots, in the sense that the nudge induces

[3] https://blog.twitter.com/en_us/topics/company/2020/covid-19#misleadinginformation

them to consider the accuracy of the news they receive, but they are also not as sophisticated as Bayesians, in the sense that they aggregate the information they receive from their friends without worrying about how these friends arrived at their beliefs. It turns out that this moderate sophistication can outperform both the naive DeGroots and the sophisticated Bayesians.

On the other hand, consider a scenario where a lot of the misinformation actually argues for the correct state, i.e., the information is false but the conclusion that one would draw from it is correct (e.g., the true state is that a vaccine is safe and effective, but the misinformation pushes this conclusion on the agents through outlandish claims about all the good things that will happen if a person is vaccinated). In this case, accuracy nudging might actually lead to DeGroots failing to learn the correct state *when they would have learned it correctly without the nudge*. The reason is that as agents start questioning the veracity of information they receive (e.g., they think there are a lot of false claims about the vaccine's effectiveness) they incorrectly discard correct information as misinformation. This can tip the scale towards the information arguing for the incorrect state and eventually makes the agents believe that this state is indeed the correct one. This suggests that nudging should be used carefully, as it may not be as innocuous as it appears.

**Performance Targets.** Removal of all misinformation from a platform is a desirable but unrealistic goal. We consider regulations that are either internal (to a platform) or external (e.g., government policy that platforms must obey) that decree a threshold for an "acceptable" level of misinformation floating on a platform. These performance targets have been discussed in a recent Facebook whitepaper (Facebook, 2020). The amount of misinformation on a platform leads to a certain *mislearning rate* – a probability with which agents fail to learn. Theorem 4 shows a somewhat surprising finding: for the same mislearning rate, Bayesian societies should be more regulated than DeGroot societies, i.e., the platform should strive to remove more misinformation when the population is strategic compared to when it is naive.

## 3 A Model of Reasoning with Misinformation

In this section, we present a model of reasoning based on networked social interactions, building on the work of Mostagir and Siderius (2021).

**Basic Setup**. There is a true state $\theta \in \{L, R\}$ which advocates for either a left-wing or right-wing proposal. Agents learn about $\theta$ both from content (via "news articles") and from others in the population (via "social learning"). We consider a discrete-time model with a finite (but long) learning horizon $T$, and each agent $i$ updates her beliefs $\pi_{i,t}$ about $\theta$ over time $t = 0, 1, 2, \ldots, T$

(as described below).

At $t = 0$, every agent $i$ is born with a heterogenous prior belief $\pi_{i,0}$ (about $\theta = R$) which is drawn from a (continuous) distribution $H$. We assume $H$ is symmetric about belief $1/2$ and has positive density $h$ almost everywhere. The symmetry assumption guarantees that for every agent with prior belief leaning toward $L$, there is a corresponding agent (with the same tenacity) holding a prior belief learning toward $R$.[4] At $t = 1$, content is generated (as described below) and no more content is sent into the society.[5] At $t \geq 2$, agents employ social learning to form beliefs about whether the true state is $\theta = L$ or $\theta = R$.

**Content Generation**. Content is generated at $t = 1$, with some of this content being *organic* and some of this content containing *misinformation*. Each agent $i$ gets a message $m_i \in \{L, R\}$ and the agent cannot tell whether this content is organic or contains misinformation. If the content is organic, it is generated as $\mathbb{P}[m_i = \theta] = p > 1/2$ (i.e., organic content is more likely to argue for the truth than not). If the content contains misinformation, then it is orthogonal to the state $\theta$.

The proportion of misinformation arguing for state $R$ is given by $r$, randomly generated according to distribution $F$ (but independent of the truth $\theta$), with positive density $f$ almost everywhere. The proportion of misinformation is $q < 1/2$ and the proportion of organic news is $1 - q > 1/2$ (i.e., most content is organic and does not contain misinformation). As empirically demonstrated in van der Linden et al. (2020), we assume agents agree on the amount of misinformation in the system but do not necessarily agree on whether this misinformation leans more left or right (per Figure 1).

**Social Network**. We assume that all agents are arranged in an undirected social network $\mathbf{G}$. A link $i \leftrightarrow j$ denotes that agent $i$ and agent $j$ observe (or talk to) each other. We let $\mathcal{N}_i$ denote the neighborhood of agent $i$ (i.e., the set of agents $j$ with $i \leftrightarrow j$). The adjacency matrix $\mathbf{A}$ of $\mathbf{G}$ is a binary matrix with $[\mathbf{A}]_{ii} = 1$ and $[\mathbf{A}]_{ij} = 1$ if and only if $i \leftrightarrow j$. We assume that the network is connected and contains no overly-influential agents (see Appendix C.1 for details).

**Agents**. Following previous literature, we assume agents in the society are one of two sophistication types, Bayesian or DeGroot (naive). Bayesian agents are fully strategic agents and update their beliefs in a way fully consistent with Bayes' rule and common knowledge of strategic interactions, whereas DeGroot agents employ only heuristic-based, "rule-of-thumb" belief updating. We describe the exact updating next:

[4] This allows us to isolate the effects of misinformation policies without concerns about the entire population (as a whole) being initially biased toward or away from the truth.

[5] In Appendix C.2 we show that our results apply identically if content is injected persistently over time, but complicates the timing of the updating process. Hence, for simplicity, we assume that content arrives only at $t = 1$.

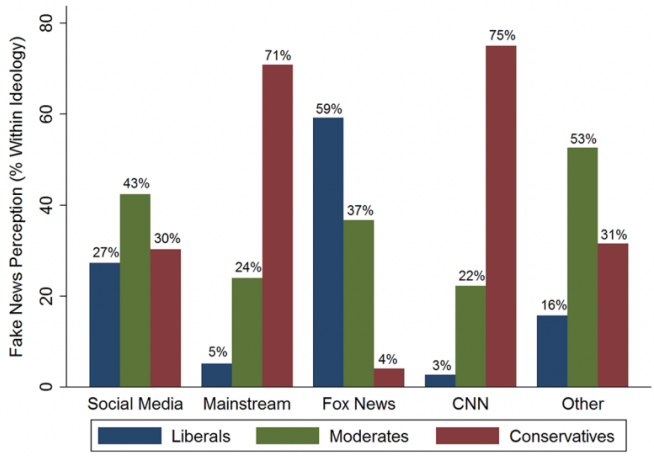



Figure 1. Perception of misinformation by ideological belief (courtesy of van der Linden et al. (2020)).

(i) *Bayesian Society*: At $t = 1$, each Bayesian agent forms a posterior update about the state, $\pi_{i,1}$, given the article with message $m_i$ and knowing content may contain misinformation:

$$\pi_{i,1}(m_i = R) = \mathbb{E}[\mathbf{1}_{\theta=R}|m_i = R] = \int_0^1 \frac{(p(1-q)+qr)\pi_{i,0}}{p(1-q)\pi_{i,0} + (1-p)(1-q)(1-\pi_{i,0}) + qr} f(r)\ dr$$

$$\pi_{i,1}(m_i = L) = \mathbb{E}[\mathbf{1}_{\theta=R}|m_i = L] = \int_0^1 \frac{((1-p)(1-q)+q(1-r))\pi_{i,0}}{(1-p)(1-q)\pi_{i,0} + p(1-q)(1-\pi_{i,0}) + q(1-r)} f(r)\ dr$$

At all times $t \geq 2$, agents form Bayesian posterior estimates about the state, $\pi_{i,t}$, given their article with message $m_i$ and the beliefs of agents in their social neighborhood $\{\pi_{j,t}\}_{j\in\mathcal{N}_i\,;\,t\geq 0}$, again, fully aware that there may be misinformation in the system. This is akin to the updating process in Acemoglu et al. (2016), where agents are uncertain about the underlying message distribution.

(ii) *DeGroot Society*: DeGroot agents are boundedly rational agents who use a learning heuristic to learn $\theta$. At $t = 1$, each DeGroot agent updates her belief of the state using Bayes' rule taking the news at *face value* (i.e., assuming there is no misinformation in the system). This is similar to how these agents update their beliefs in Jadbabaie et al. (2012):

$$\pi_{i,1}(m_i = R) = \mathbb{E}[\mathbf{1}_{\theta=R}|m_i = R,\ q = 0] = \frac{p\pi_{i,0}}{p\pi_{i,0} + (1-p)(1-\pi_{i,0})} \tag{1}$$

$$\pi_{i,1}(m_i = L) = \mathbb{E}[\mathbf{1}_{\theta=R}|m_i = L,\ q = 0] = \frac{(1-p)\pi_{i,0}}{(1-p)\pi_{i,0} + p(1-\pi_{i,0})} \tag{2}$$

Based on these observations, each DeGroot takes an average of all the time $t-1$ beliefs of

the agents in her neighborhood to form her time $t \geq 2$ beliefs, i.e., $\pi_{i,t} = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \pi_{j,t-1}$.

**Learning**. At $t = T$, each agent has formed her belief from both the message she has received but also the beliefs shared by other agents. At $t = T$, agents take a binary terminal action $a_i \in \{L, R\}$ that minimizes her quadratic loss $\mathbb{E}[(a_i - \mathbf{1}_{\theta=R})^2]$ given her belief, $\pi_{i,T}$. We follow the standard definition of learning (e.g., Acemoglu et al. (2011)) and say that society *learns* if all agents take the correct action ($a_i = \theta$); otherwise, society *mislearns*.[6]

**Regulator**. There is a regulator who has an array of policy tools at their disposal (described below). The regulator is born ex-ante unbiased with a belief of $\Pi = 1/2$. Before the messages are generated, the regulator conducts research to learn about the true state of the world $\theta$. We assume this process recovers a signal $s = \theta$ with probability $1 - \varepsilon$ and $s \neq \theta$ with probability $0 < \varepsilon < 1/2$. When $\varepsilon$ is large, the regulator has little confidence in its ability to uncover the truth. When $\varepsilon$ is small, the regulator is more confident that it can get a tighter estimate of the truth relative to what information appears to the agents. The value of $\varepsilon$ is common knowledge. After the research step, the belief of the regulator is $\Pi' \in (0, 1/2) \cup (1/2, 1)$.

# 4 Learning with Misinformation Policies

In this section, we discuss four regulatory policies currently proposed to curb misinformation. The objective of the regulator is to maximize the likelihood of learning. We say a policy is *effective* if it always (weakly) improves the likelihood of learning, *ambiguous* if it can sometimes strictly improve and sometimes strictly harm the likelihood of learning, and *backfires* if it always (weakly) harms the likelihood of learning. We analyze the impact of these policies for both Bayesian and DeGroot societies.

## 4.1 Policy I: Censorship

Censorship is one of the oldest policies for controlling access to information. The practice is controversial because it typically involves a unilateral decision (e.g., by a platform or a government) to remove certain information and make it inaccessible. The use of the policy in modern times is more complicated as platforms try to regulate content by balancing freedom of expression with reducing harmful speech, e.g., the aforementioned example of Twitter removing COVID-19 misinformation from the platform.

[6] Appendix B.5.1 in Mostagir and Siderius (2021) considers alternative learning definitions (e.g., expected proportion of agents who mislearn) and shows that the main insights do not qualitatively change.

The interesting aspect of the role of censorship in our model is that the outcome of the policy depends on the targeted audience. A benevolent censor (who is interested in agents learning the correct state) can show the same information to two different groups and have one group correctly learn the state of the world while leading the other farther from the truth. Underlying these diverging outcomes is how agents with different sophistication levels perceive censorship. No censoring technology is completely accurate: even a benevolent platform might still remove content that is correct and/or leave content that is false. Bayesian agents use this observation to *rationally* reason about what else might be censored, and dismiss information that disagrees with their existing positions. This is reminiscent of the empirical documentation showing that there is disagreement across ideologies about what content is censored on social media Vogels et al. (2020). We show that censorship always hurts the likelihood of learning for Bayesian agents but weakly improves it for DeGroots.

*Analysis.* There is a regulator who can control the flow of information by censoring a fraction of the misinformation that advocates for either $\theta = L$ or $\theta = R$ (or both). First, the regulator makes a binary decision to either **Not Censor** or **Censor**. The decision to **Not Censor** permits all information, regardless of whether it contains misinformation. The decision to **Censor** allows the regulator to censor $0 < \delta < 1$ fraction of the misinformation that appears (where $\delta$ is a given technology parameter). Agents can observe whether the regulator has chosen to **Not Censor** or **Censor**. This assumption is natural given, for example, observed tags that certain content was deleted due to policy violations on Twitter.[7]

However, if the regulator chooses **Censor**, it selects a value $0 \leq \rho \leq 1$ of the $\delta$ misinformation arguing for $\theta = R$ to censor (and censors $1 - \rho$ of the $\delta$ misinformation arguing for $\theta = L$). That is, a policy that chooses $\rho = 0$ removes only the $L$ misinformation (up to the total amount of $L$ misinformation) but a policy that chooses $\rho = 1$ removes only $R$ misinformation (up to the total amount of $R$ misinformation). The type of misinformation removed (i.e., the regulator's choice of $\rho$) cannot be observed by the users.

We assume the regulator's objective is to minimize the likelihood of society mislearning the truth. The following characterizes the optimal strategy of the regulator:

**Theorem 1.** *The regulator should implement* **Censor** *for the DeGroot society but* **Not Censor** *for the Bayesian society.*

Theorem 1 states that censorship is always beneficial to a DeGroot society but will necessarily backfire for a Bayesian one, so the optimal censorship policy critically depends on the so-

[7] See, for example, https://help.twitter.com/en/rules-and-policies/notices-on-twitter.

phistication of the agents. With DeGroot agents, who take information at face value, censoring misinformation advocating for (likely) the incorrect state is more likely to nudge beliefs toward this state, and is better off being removed. Thus, censoring can be useful for learning. However, we remark that the regulator does not necessarily employ a bang-bang censorship policy ($\rho = 0$ or $\rho = 1$) for DeGroots, as seen in the simulations presented in Appendix B.1.

On the other hand, any censorship policy allows (Bayesian) agents to rationalize their prior beliefs more strongly. Because $\varepsilon > 0$, it is always possible the regulator is erroneous in his assessment of the truth. When the censored content is unobservable, more extreme narratives are possible: right-wing believers can spin a narrative the regulator is removing right-wing content, whereas left-wing believers can spin a narrative the regulator is removing left-wing content. This forces the hand of the regulator, who has no choice but to allow all content to exist (i.e., **Not Censor**) in Bayesian communities.

### 4.2 Policy II: Provision of Diverse Content

In 1949, the FCC introduced a policy known as the "fairness doctrine" which required news providers to present both sides of a controversial issue (see Simmons (1976)). This policy was eliminated in 1987, giving rise to news outlets that were heavily one-sided and paving the way for partisan talk radio (Clogston (2016)). While one-sided news does not necessarily contain misinformation, it provides an avenue for presenting content that is skewed toward one perspective, with misinformation (when it exists) also likely arising from this perspective. For example, a liberal media outlet that is not required to present diverse content is more likely to present misinformation in favor of $\theta = L$. Hence, we assume a policy that requires "equal coverage" of both sides of a topical issue, and model this as content being less likely to present strongly misleading information toward one perspective.

A modern-day equivalent of the fairness doctrine is the idea of requiring social media platforms to provide more diverse news feeds. As seen in Levy (2021), platforms typically try to increase user engagement by (algorithmically) recommending stories that match users' profiles, and this can result in "filter bubbles" that limit the scope of counter-attitudinal content that users see. This has been linked to the propagation of misinformation and its influence on outcomes such as the 2016 presidential race (see Allcott and Gentzkow (2017)). A proposed solution is to regulate these algorithms in order to provide more diverse news, for example by requiring content be shown "uniformly at random" from one's social network, and not selectively filtered (e.g., Sunstein (2018); Cen and Shah (2020)). This is similar to an equal-coverage policy, where users of the social media platform ideally learn from a variety of sources with different perspec-

tives. As we discuss, the efficacy of this policy is highly dependent on the population type and amount of misinformation in the system.

*Analysis.* We frame the provision of diverse (or counter-attitudinal) content as reducing the probability that misinformation is heavily slanted toward one ideology or another. Formally, we assume the policy introduces a new misinformation distribution $\tilde{F}(r) = \gamma\delta_{\text{Dirac}}(r-1/2)+\gamma F(r)$, where $\delta_{\text{Dirac}}(r-1/2)$ is a Dirac delta function centered at $r = 1/2$ and $\gamma$ is a (fixed) scalar parameter specifying the intensity of a policy to implement content diversity. See.

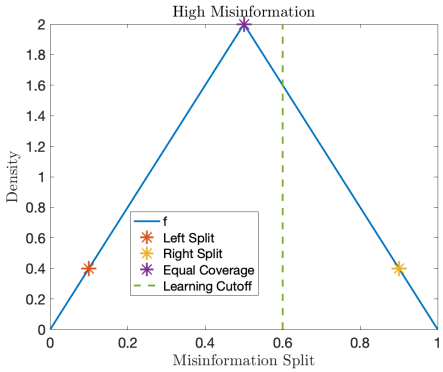
Moreover, we let $q^*$ denote the threshold at which misinformation is large enough that both Bayesian and DeGroot societies mislearn with positive probability, which is incidentally the same across both society types, given our assumptions on $H$ and $F$. Our main result establishes:

**Theorem 2.** *Implementing the equal-coverage policy always improves learning for DeGroot agents. For Bayesian agents, there exists $q^{**} > q^*$ such that for $q \in (q^*, q^{**})$, implementing equal-coverage improves learning, but for $q > q^{**}$, the equal-coverage policy backfires and reduces the probability of learning.*

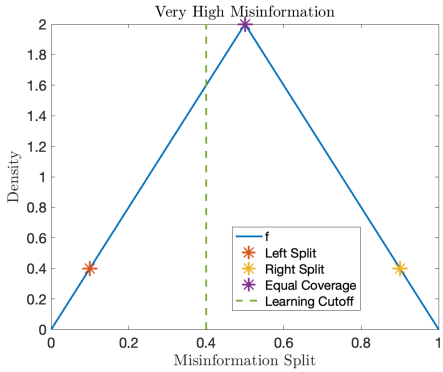
Theorem 2 states that the equal-coverage policy works well for DeGroot agents, but that a "Pandora's box" effect exists for the Bayesians: the policy works well as long as misinformation is not extensive, but once the amount of misinformation crosses a certain threshold, the policy becomes ineffective. Essentially, this result states that the effects of eliminating the fairness doctrine and equal coverage (partisan news and consistent disagreement) cannot simply be reversed by requiring that news is covered in an equal and fair manner. Recent empirical evidence provides support for this backfire potential: by encouraging Twitter users to follow informational sources that oppose their ideology (e.g., encouraging conservatives to follow liberals and vice versa) can actually lead to more disagreement (e.g., make the conservatives more conservative) and hurt learning Bail et al. (2018). Similar outcomes were observed when personalized Twitter feeds were replaced by more diverse ones during the Argentine presidential debates Di Tella et al. (2021). The gap between $q^*$ and $q^{**}$ is explored numerically in Appendix B.2.

The intuition behind this result is as follows. When misinformation is evenly split between left and right ideologies, and because DeGroot agents take their messages at face value, the organic messages are more likely to dominate and allow the truth to be revealed. This means the equal-coverage policy unambiguously improves learning outcomes for this group of agents.

However, this effect is not necessarily true with a Bayesian society, as in Figures 2a and 2b. To see this, suppose the true state is $L$. With low amounts of misinformation (as in Figure 2a), the amount of right-leaning misinformation needed to spin multiple narratives (given by the verti-




(a) When misinformation is above $q^*$ but below $q^{**}$




(b) Misinformation exceeds the threshold $q^{**}$

Figure 2. Figures 2a and 2b compare the equal-coverage policy when misinformation is high (but manageable) as opposed to when it is unmanageable.

cal dashed line) is low enough that making misinformation more balanced improves outcomes. However, when misinformation is exorbitantly high (as in Figure 2b), the right-leaning misinformation threshold is much lower. In this way, moving the extreme cases of misinformation toward the center actually permits *more* narrative spinning. This occurs because both ideological perspectives are more easily able to (rationally) dismiss the coverage from the other side as misinformation.

### 4.3 Policy III: Accuracy Nudging

Nudging has emerged as one of the less-interventionist choices that policymakers have at their disposal (Thaler and Sunstein, 2009), the idea being that a gentle pointer towards desired behavior can be enough to influence outcomes in meaningful ways. The policy has been recently studied in the context of misinformation in field experiments such as Pennycook et al. (2021, 2020). Following this line of work, we consider a policy that shifts the attention of agents toward the accuracy of the messages they receive, with the goal of making agents more aware of the possible presence of misinformation.

*Analysis.* Recall that Bayesian agents have a fully specified learning model, in that they are *aware* of the presence of misinformation. Reminding these agents that the information they receive is possibly inaccurate does not affect their posterior beliefs or how they process information. On the other hand, recall that DeGroot agents form their beliefs by doing a Bayesian updating on the news they receive and a simple weighted averaging over the beliefs of their neighbors. The nudge/reminder provides a more accurate Bayesian posterior of their belief based on content alone, i.e., it alerts them to the possibility that news contains inaccurate information. However,

note it does not endow the DeGroots with fully "Bayesian" abilities in processing beliefs from others (i.e., ignore redundant information or factor in other agents' ideological priors). Thus, in some ways nudged DeGroots have intermediate sophistication between plain DeGroots and plain Bayesians.

Formally, at $t = 1$, a DeGroot agent $i$ who receives message $m_i = R$, forms beliefs about $\theta = R$ according to:

$$\pi_{i,1} = \int_0^1 \frac{(p(1-q)+qr)\pi_{i,0}}{p(1-q)\pi_{i,0} + (1-p)(1-q)(1-\pi_{i,0}) + qr} f(r)\ dr$$

When there is no misinformation, this reduces to the standard DeGroot update based on the observed content (but with no nudging):

$$\pi_{i,1} = \frac{p\pi_{i,0}}{p\pi_{i,0} + (1-p)(1-\pi_{i,0})}$$

However, in the more general case with misinformation (i.e., $q > 0$), this policy nudges the posterior closer to the truth (in expectation) for all prior beliefs $\pi_{i,0}$. The case of $m_i = L$ follows a similar update (see Appendix B.3).

In our next result, we characterize the impact of an accuracy nudging policy:

**Theorem 3.** *Accuracy nudging does not affect Bayesian learning and when $F$ is symmetric, does not affect DeGroot learning. When $F$ is asymmetric, accuracy nudging has an ambiguous effect on learning outcomes.*

As Theorem 3 shows, accuracy nudging can be an effective policy for DeGroot societies, but has similar backfire potential. Most prominently, accuracy nudging can be too subtle of a policy to be effective. When misinformation is not likely to be skewed toward the incorrect state (e.g., $F$ is symmetric), then the net effect from nudging is nothing – nudging makes DeGroot agents more tethered to their ideological priors, like Bayesian agents, but does not facilitate learning. However, if the regulator knows that most misinformation is probably opposing $\theta$ (e.g., misinformation arguing COVID vaccines are unsafe), then nudging *can* be effective (see the proof of Theorem 3 in Appendix A). Simulations in Appendix B.3 show how learning outcomes change as a result of the nudge. This is summarized in Table 1.

| No Effect | Helps | Hurts |
|---|---|---|
| 88.4% | 7.8% | 3.8% |

Table 1. Outcome of Accuracy Nudging Simulation (Appendix B.3).

Table 1 shows that, generally, the policy is too gentle (i.e., nudging usually has no effect), but

the policy typically helps learning more often than it hurts learning. This corroborates replication studies on accuracy nudging that question the statistical significance of this policy in reducing the influence of misinformation (Roozenbeek et al., 2021).

More surprisingly, an accuracy nudge can backfire because it makes DeGroot agents more skeptical about *all* the information they receive. While the extra diligence in vetting information leads them to identify misinformation more often, it can also make them discard accurate information as misinformation. This implies that, depending on how much misinformation there is in the system, encouraging agents to think more carefully about the information they receive can lead to an overall decline in learning. This mechanism is reminiscent of the "distrust mindset" identified in previous work such as Ognyanova et al. (2020); Kwon and Barone (2020); Park et al. (2020).

Another interesting aspect of nudging is that a nudged DeGroot society can perform better than both (unnudged) DeGroots and Bayesians. Because nudging improves the awareness of DeGroots, it pushes them towards more Bayesian-like behavior, but importantly, they do not become full-on Bayesian. This intermediate reasoning ability allows them to outperform both extremes and can be, in fact, first-best for learning the true state.

*Remark* — We assume accuracy nudging is effective for all DeGroots. An alternative formulation is that accuracy nudging only works with some probability. For example, it is only effective for $\zeta \in (0,1)$ proportion of DeGroot agents but is dismissed by the other $1-\zeta$ proportion. We note that the statement of Theorem 3 remains unchanged under this alternative formulation, but for lower values of $\zeta$, the "no effect" outcome from nudging (seen in Table 1) will become even more pronounced.

### 4.4 Policy IV: Performance Targets

We consider a policy where the regulator imposes a *performance target* – a target that requires misinformation on the platform to be below a certain level.[8] A natural, but unrealistic, regulation is to require social media platforms to set the misinformation target at 0%. As Candogan and Drakopoulos (2020) identify, this is likely to decrease engagement on the platform . It also provides a plethora of perverse incentives; for example, it can shift the attention of the platform toward eradicating misinformation at the cost of neglecting other unmeasured/unregulated obli-

[8] In particular, Facebook (2020) proposes the following possible regulatory action:

> "Governments could also consider requiring companies to hit specific performance targets, such as decreasing the prevalence of content that violates a site's hate speech policies."

gations, or it can lead to a narrower definition of what constitutes misinformation and make reporting it harder. Thus, while decreasing misinformation is beneficial, setting a performance target too low can have undesired effects. In this section, we consider the problem of setting the optimal performance target for the platform.

*Analysis.* The motivation for adopting a misinformation performance target is to decrease the likelihood that agents will be influenced and learn incorrectly from misleading content that appears. Unsurprisingly, mislearning is monotonically increasing in misinformation (as proven in Mostagir and Siderius (2021)). Thus, if we assume the regulator wants to hit a *mislearning* target $\lambda > 0$,[9] there is a unique corresponding misinformation target $q(\lambda)$ it needs to achieve. Setting the misinformation target too high under-regulates: mislearning is high relative to the moral hazard cost. However, setting the misinformation target too low over-regulates: it leads to very little mislearning at the neglect of other obligations.

For the same $\lambda$, this misinformation threshold $q(\lambda)$ may depend on the society type. We denote by $q_B(\lambda)$ and $q_D(\lambda)$ the thresholds needed for the Bayesian and DeGroot societies, respectively, to obtain mislearning rate $\lambda$.

**Theorem 4.** *The regulator's optimal policy sets* $q_B(\lambda) < q_D(\lambda)$ *for all* $\lambda > 0$.

There are two key takeaways from Theorem 4. First, it contradicts the classical intuition that more sophisticated societies are more resistant to misinformation. For example, experimental evidence that shows that a sophisticated population requires a performance target of $q(\lambda)$ implies that a less sophisticated population requires even *less* regulation to obtain the same mislearning rate. Second, it implies that if the regulator is ignorant about the sophistication of the agents on the platform, setting a performance target can backfire. If a DeGroot society is mistaken as a Bayesian one, it will be over-regulated whereas if a Bayesian society is mistaken as a DeGroot one, it will be under-regulated. Thus, controlling misinformation through performance targets depends critically on the sophistication type of the population. Simulations (in Appendix B.4) show that the gap between $q_B$ and $q_D$ can be quite significant, especially when the regulator sets a mislearning target $\lambda$ in the intermediate range between 0 and 1.

## 5 Conclusion

This paper argues that agent sophistication is an important factor to consider when drawing up policies to stop misinformation. Agents learn in different ways and as a result their responses to

[9] As stated in the white paper from Facebook (2020), a performance target should "hold companies accountable for the ends they have achieved rather than ensuring the validity of how they got there."

| Policy | Bayesian Population | DeGroot Population |
|---|---|---|
| Censorship | Backfire | Effective |
| Diverse Content | Ambiguous | Effective |
| Accuracy Nudging | No Effect | Ambiguous |
| Performance Targets | Ambiguous | Ambiguous |

Table 2. Summary of Policy Efficacy. (**Key**: Backfire = Guaranteed to make learning worse; Effective = Guaranteed to make learning better; Ambiguous = Sometimes improves and sometimes worsens learning; No Effect = Guaranteed to have no effect on learning.)

regulations might also be different or even unexpected. Not accounting for sophistication might lead to ineffective policies or to outcomes that are worse than if the policy was not put in place to begin with.

The policies we study are directly informed by the current conversation in the public sphere between policymakers, researchers, and platforms. In evaluating these policies, we have considered the two main models in the social learning literature, Bayesian and DeGroot learning. Table 2 summarizes our findings: some policies, like censorship or provision of diverse content, are clearly sensitive to the population type. They unambiguously help the DeGroot agents but could be ineffective or outright harmful if used in a Bayesian society. Accuracy nudging and setting performance targets have more subtle effects, where for the same population the policy might help or hurt depending on the parameters of the environment.

An interesting point that comes out of our focus on sophistication type is the following. Section 4.3 shows that it is possible to boost the sophistication of DeGroots through nudging in a way that makes them more sophisticated than DeGroots but less sophisticated than Bayesians, and that this moderate sophistication can outperform both pure types. A natural question is whether there is a policy that can do the same for Bayesians, i.e., makes them behave in a less sophisticated fashion in order to obtain better learning outcomes. Another question is to think how these results would extend to a *mixed* society that combines both types of agents, or that has even more heterogeneity when it comes to strategic sophistication.

Finally, while we studied the effects of distinct policies, it is possible that a combination of these (and other) policies can be used in conjunction with one another to deliver outcomes that improve on any of the policies used in isolation. This is similar to the idea of "policy cocktails" in Jackson (2021), where a mixture of policies is proposed to address inequality. Thinking about combination of policies is a natural next step in the fight against misinformation and constitutes another area of promising future work.

## References


Acemoglu, Daron, Victor Chernozhukov, and Muhamet Yildiz (2016), "Fragility of asymptotic agreement under bayesian learning." *Theoretical Economics*, 11, 187–225.

Acemoglu, Daron, Munther A Dahleh, Ilan Lobel, and Asuman Ozdaglar (2011), "Bayesian learning in social networks." *The Review of Economic Studies*, 78, 1201–1236.

Allcott, Hunt and Matthew Gentzkow (2017), "Social media and fake news in the 2016 election." *Journal of economic perspectives*, 31, 211–36.

Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky (2018), "Exposure to opposing views on social media can increase political polarization." *Proceedings of the National Academy of Sciences*, 115, 9216–9221. Publisher: National Academy of Sciences Section: Social Sciences.

Candogan, Ozan and Kimon Drakopoulos (2020), "Optimal Signaling of Content Accuracy: Engagement vs. Misinformation." *Operations Research*, 68, 497–515. Publisher: INFORMS.

Cen, Sarah H and Devavrat Shah (2020), "Regulating algorithmic filtering on social media." *arXiv preprint arXiv:2006.09647*.

Clogston, Juanita "Frankie" (2016), "The Repeal of the Fairness Doctrine and the Irony of Talk Radio: A Story of Political Entrepreneurship, Risk, and Cover." *Journal of Policy History*, 28, 375–396. Publisher: Cambridge University Press.

Degroot, Morris H. (1974), "Reaching a Consensus." *Journal of the American Statistical Association*, 69, 118–121.

DeMarzo, Peter M., Dimitri Vayanos, and Jeffrey Zwiebel (2003), "Persuasion Bias, Social Influence, and Unidimensional Opinions*." *The Quarterly Journal of Economics*, 118, 909–968.

Di Tella, Rafael, Ramiro H. Gálvez, and Ernesto Schargrodsky (2021), "Does Social Media cause Polarization? Evidence from access to Twitter Echo Chambers during the 2019 Argentine Presidential Debate." Working Paper 29458, National Bureau of Economic Research. Series: Working Paper Series.

Facebook (2020), "Charting a Way Forward on Online Content Regulation."

Gale, Douglas and Shachar Kariv (2003), "Bayesian learning in social networks." *Games and Economic Behavior*, 45, 329–346.

Golub, Benjamin and Matthew O Jackson (2010), "Naive learning in social networks and the wisdom of crowds." *American Economic Journal: Microeconomics*, 2, 112–49.

Golub, Benjamin and Evan Sadler (2017), "Learning in Social Networks." SSRN Scholarly Paper ID 2919146, Social Science Research Network, Rochester, NY.

Haghtalab, Nika, Matthew O Jackson, and Ariel D Procaccia (2021), "Belief polarization in a complex world: A learning theory perspective." *Proceedings of the National Academy of Sciences*, 118.

Jackson, Matthew (2021), "Policy cocktails: Attacking the roots of persistent inequality." *Policy*.

Jackson, Matthew O., Suraj Malladi, and David McAdams (2019), "Learning through the Grapevine: The Impact of Noise and the Breadth and Depth of Social Networks." SSRN Scholarly Paper ID 3269543, Social Science Research Network, Rochester, NY.

Jadbabaie, Ali, Pooya Molavi, Alvaro Sandroni, and Alireza Tahbaz-Salehi (2012), "Non-bayesian social learning." *Games and Economic Behavior*, 76, 210–225.

Kwon, Mina and Michael J. Barone (2020), "A World of Mistrust: Fake News, Mistrust Mind-Sets, and Product Evaluations." *Journal of the Association for Consumer Research*, 5, 206–219. Publisher: The University of Chicago Press.

Levy, Ro'ee (2021), "Social media, news consumption, and polarization: Evidence from a field experiment." *American economic review*, 111, 831–70.

Loomba, Sahil, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson (2021), "Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa." *Nature human behaviour*, 1–12.

Mossel, Elchanan, Allan Sly, and Omer Tamuz (2014), "Asymptotic learning on Bayesian social networks." *Probability Theory and Related Fields*, 158, 127–157.

Mostagir, Mohamed and James Siderius (2021), "Learning in a post-truth world." *Management Science*.

Mueller-Frank, Manuel (2013), "A general framework for rational learning in social networks." *Theoretical Economics*, 8, 1–40.

Mueller-Frank, Manuel (2014), "Does one bayesian make a difference?" *Journal of Economic Theory*, 154, 423–452.

Ognyanova, Katherine, David Lazer, Ronald E. Robertson, and Christo Wilson (2020), "Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power." *Harvard Kennedy School Misinformation Review*.

Park, Sora, Caroline Fisher, Terry Flew, and Uwe Dulleck (2020), "Global Mistrust in News: The Impact of Social Media on Trust." *International Journal on Media Management*, 22, 83–96. Publisher: Routledge _eprint: https://doi.org/10.1080/14241277.2020.1799794.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand (2021), "Shifting attention to accuracy can reduce misinformation online." *Nature*, 592, 590–595.

Pennycook, Gordon, Jonathon McPhetres, Yunhao Zhang, Jackson G. Lu, and David G. Rand (2020), "Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention." *Psychological Science*, 31, 770–780. Publisher: SAGE Publications Inc.

Pennycook, Gordon and David G Rand (2018), "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning." *Cognition*.

Pennycook, Gordon and David G Rand (2021), "The psychology of fake news." *Trends in cognitive sciences*.

Roozenbeek, Jon, Alexandra L. J. Freeman, and Sander van der Linden (2021), "How Accurate Are Accuracy-Nudge Interventions? A Preregistered Direct Replication of Pennycook et al. (2020)." *Psychological Science*, 32, 1169–1178. Publisher: SAGE Publications Inc.

Simmons, Steven J. (1976), "Fairness Doctrine: The Early History." *Federal Communications Bar Journal*, 29, 207.

Sunstein, Cass R. (2018), *#Republic: Divided Democracy in the Age of Social Media*, ned - new edition edition. Princeton University Press.

Thaler, Richard H. and Cass R. Sunstein (2009), *Nudge: Improving Decisions About Health, Wealth, and Happiness*, revised & expanded edition edition. Penguin Books, New York.

U.S. Surgeon General (2021), "Surgeon general issues advisory during covid-19 vaccination push warning american public about threat of health misinformation."

van der Linden, Sander, Costas Panagopoulos, and Jon Roozenbeek (2020), "You are fake news: political bias in perceptions of fake news." *Media, Culture & Society*, 42, 460–470.

Vogels, Emily a, Rew Perrin, MONICA, and ERSON (2020), "Most Americans Think Social Media Sites Censor Political Viewpoints."

# Appendix

## A Proofs

*Proof of Theorem 1.* For a Bayesian society, we show (i) the regulator always has a (weak) preference for **Not Censor** over **Censor**, and (ii) has a strict preference for **Not Censor** over **Censor** when choosing a strategy that depends on the research signal $s$. This implies that **Not Censor** (weakly) dominates **Censor** for Bayesian societies. For a DeGroot society, we show (i) the regulator can always obtain the same likelihood of learning as **Not Censor** by choosing **Censor** and $\rho = 1/2$, and (ii) can do strictly better than **Not Censor** when $\varepsilon$ is sufficiently small. This implies that **Censor** (weakly) dominates **Not Censor** for DeGroot societies.

Part 1: We establish that if the regulator chooses **Censor** and some distribution $\sigma$ over $\rho$ (which is unobservable to the society),[10] mislearning occurs with (weakly) higher probability that when **Not Censor** is chosen. Note that if $\sigma$ is a pure strategy independent of the signal $s$ that the regulator receives, then the original signal distribution can be backed out perfectly by every Bayesian agent. To see this, note that if the distribution of $R$ messages is $\kappa'$ after censorship, then the original distribution of $R$ messages ($\kappa$) is given by:

$$\begin{aligned} \kappa' &= \frac{\kappa - \delta\rho}{1-\delta} \\ \implies \kappa &= \kappa'(1-\delta) + \delta\rho \end{aligned}$$

which implies that learning is unaffected by the censorship policy.

If $\sigma$ is a mixed strategy, but still independent of $s$, then there is a set of possible original message distributions $\{\kappa_1, \kappa_2, \ldots, \kappa_k\}$, given the observed $\kappa'$, one for each of the $k$ choices of $\rho$ that have support in $\sigma$. Let each of these have likelihood $\{\zeta_1, \zeta_2, \ldots, \zeta_k\}$ in the mixed strategy $\rho$ of the regulator. Then the posterior belief of a Bayesian agent at $t = 2$ is given by:

$$\pi_{i,2} = \sum_{\ell=1}^{k} \zeta_\ell \frac{\eta_\ell^R \pi_{i,0}}{\eta_\ell^R \pi_{i,0} + \eta_\ell^L (1 - \pi_{i,0})} \tag{3}$$

where $\eta_\ell^R$ and $\eta_\ell^L$ are the likelihood of the $R$ narrative and $L$ narrative, respectively, given (unobserved, but inferred) original signal distribution $\kappa_\ell$ before policy $\ell$. Because $H$ and $F$ have full support, the probability of learning is given by:

$$\mathbb{P}[\theta = L \,|\, s] \sum_{\ell=1}^{k} \zeta_\ell \mathbb{P}\left[\eta_\ell^R = 0 \,|\, s\right] + \mathbb{P}[\theta = R \,|\, s] \sum_{\ell=1}^{k} \zeta_\ell \mathbb{P}\left[\eta_\ell^L = 0 \,|\, s\right]$$

This is a direct consequence of the fact that if $\eta_\ell^R > 0$ but $\theta = L$, then there exists an agent $i$ with sufficiently large $\pi_{i,0} < 1$ (via the full support assumption on $H$) such that all agents $j$ with $\pi_{j,0} > \pi_{i,0}$ mislearn, so society mislearns, via Equation (3). A symmetric argument applies for $\eta_\ell^L > 0$ but $\theta = R$.

For $\sigma$ to be a best response for the regulator, it must be the case that:

$$\mathbb{P}[\theta = L \,|\, s]\mathbb{P}\left[\eta_\ell^R = 0 \,|\, s\right] + \mathbb{P}[\theta = R \,|\, s]\mathbb{P}\left[\eta_\ell^L = 0 \,|\, s\right] = \mathbb{P}[\theta = L \,|\, s]\mathbb{P}\left[\eta_{\ell'}^R = 0 \,|\, s\right] + \mathbb{P}[\theta = R \,|\, s]\mathbb{P}\left[\eta_{\ell'}^L = 0 \,|\, s\right]$$

[10] For simplicity, we will assume the mixed strategy here is a discrete probability distribution over $\rho$, but the proof technique generalizes to continuous distributions of $\sigma$, albeit slightly more involved.

for all $\ell, \ell'$ (otherwise, the regulator has a profitable deviation). Thus, the rate of mislearning is the same as the rate of mislearning if the regulator instead played one of $\ell \in \{1, \ldots, k\}$ with probability 1, which as we established before does not improve or hurt learning.

We next consider a strategy $\sigma$ for the regulator that depends on $s$; in particular, the regulator plays $\sigma_L$ when $s = L$ and plays $\sigma_R$ when $s = R$. We assume the former has support over $k_L$ policies $\{\rho_1^L, \rho_2^L, \ldots, \rho_{k_L}^L\}$ whereas the latter has support over $k_R$ policies $\{\rho_1^R, \rho_2^R, \ldots, \rho_{k_R}^R\}$. For a given policy action $\rho$, there are two cases:

(i) $\underline{s = L}$: Then the distribution of $R$ messages is given by $\kappa_L' = \frac{\kappa_1' - \delta\rho_L^\ell}{1-\delta}$.

(ii) $\underline{s = R}$: Then the distribution of $R$ messages is given by $\kappa_L' = \frac{\kappa_2' - \delta\rho_R^\ell}{1-\delta}$.

Thus for a given realization $\kappa'$ distribution of messages, we can back out each of the true distributions $\kappa_L, \kappa_R$:

$$\kappa_L = \kappa_L'(1-\delta) + \delta\rho_\ell^L$$
$$\kappa_R = \kappa_R'(1-\delta) + \delta\rho_\ell^R$$

Let $\eta_{L,\ell}^R$ and $\eta_{L,\ell}^L$ denote the likelihood of the $R$ narrative and $L$ narrative (respectively) conditional on the regulator's signal being $s = L$ and $\eta_{L,\ell}^R$ and $\eta_{L,\ell}^L$ denote the likelihood of the $R$ narrative and $L$ narrative (respectively) conditional on the regulator's signal being $s = R$. By the full support assumption on $H$ and $F$ and assuming $\varepsilon > 0$, for any censorship policy $(\rho_\ell^L, \rho_\ell^R)$, the probability of learning is similarly given by:

$$\mathbb{P}[\theta = L \,|\, s] \sum_{\ell=1}^{k} \zeta_\ell^s \mathbb{P}\left[\eta_{L,\ell}^R = 0 \cap \eta_{R,\ell}^R = 0 \,|\, s\right] + \mathbb{P}[\theta = R \,|\, s] \sum_{\ell=1}^{k} \zeta_\ell^s \mathbb{P}\left[\eta_{L,\ell}^L = 0 \cap \eta_{R,\ell}^L = 0 \,|\, s\right]$$

which is strictly less than the learning probability under a strategy where $\sigma$ is independent of $s$:[11]

$$\mathbb{P}[\theta = L \,|\, s] \sum_{\ell=1}^{k} \zeta_\ell \mathbb{P}\left[\eta_\ell^L = 1 \,|\, s\right] + \mathbb{P}[\theta = R \,|\, s] \sum_{\ell=1}^{k} \zeta_\ell \mathbb{P}\left[\eta_\ell^R = 1 \,|\, s\right]$$

which yields the same learning probability as **Not Censor**, as mentioned before. This means **Censor** with a $\sigma$ strategy that depends on $s$ is never a best response for the regulator.

Part 2: Unlike the Bayesian society, recall that the learning dynamics of a DeGroot society do not react to the censorship strategy (due to a lack of strategic sophistication). Note that when $\theta = L$, the fraction of $R$ messages ($\kappa$) is given by

$$\kappa = \frac{(1-p)(1-q) + qr - \delta\rho}{1-\delta}$$

and mislearning occurs with probability

$$\mathbb{P}\left[r \geq \frac{\delta(2\rho - 1) + 2p(1-q) + 2q - 1}{2q}\right] = \mathbb{P}\left[r \geq \frac{1 - 2(1-q)(1-p)}{2q}\right]$$

[11] This follows from the fact that the event $\{\eta_{L,\ell}^R = 0 \cap \eta_{R,\ell}^R = 0\}$ is a subset of the event $\{\eta_\ell^R = 0\}$ for any censorship policy $\ell$.

when $\rho = 1/2$. By Lemma 1 in Mostagir and Siderius (2021), this is exactly the probability of mislearning for DeGroot agents in the baseline model (i.e., with **Not Censor**) when $\theta = L$.

Similarly, when $\theta = R$, the fraction of $R$ messages is given by

$$\kappa = \frac{p(1-q) + qr - \delta\rho}{1-\delta}$$

and mislearning occurs with probability

$$\mathbb{P}\left[r \leq \frac{1 + \delta(2\rho - 1) - 2p(1-q)}{2q}\right] = \mathbb{P}\left[r \leq \frac{1 - 2(1-q)p}{2q}\right]$$

when $\rho = 1/2$. Again, by Lemma 1 in Mostagir and Siderius (2021), this is the probability of mislearning when the regulator plays **Not Censor** when $\theta = R$.

Finally, note that as $\varepsilon \to 0$, the regulator can (strictly) improve DeGroot learning outcomes by electing to take action **Censor** and $\rho \to 1$ (resp. $\rho \to 0$) when $\theta = L$ (resp. $\theta = R$) instead of **Not Censor**. Because $\varepsilon \to 0$, the probability that $s = \theta$ converges to 1, and the optimal regulation is to maximize the likelihood of more messages arguing for $s$. (This is a consequence of Lemma 1 in Mostagir and Siderius (2021).) Note the distribution of $R$ messages when $\theta = L$ is given by $\frac{qr+(1-q)(1-p)-\delta\rho}{1-\delta}$ and when $\theta = R$ is given by $\frac{qr+(1-q)p-\delta\rho}{1-\delta}$. A policy of $\rho = 1$ maximizes the likelihood that there are more than half $R$ messages and $\rho = 0$ maximizes the likelihood that there are more than half $L$ messages (which strictly outperforms **Not Censor**). Both of these strictly improve on $\rho = 1/2$, and by continuity, $\rho = 0$ and $\rho = 1$ strictly improve on $\rho = 1/2$ as $\varepsilon \to 0$. Thus, **Censor** is optimal for the regulator. □

*Proof of Theorem 2.* As in the proof of Theorem 1, we first prove the result for the Bayesian society and then prove it for the DeGroot society. Because the regulator's policy affects only $F$, which is considered common knowledge for Bayesians and does not affect learning dynamics for DeGroots, we can apply Lemmas 1 and 2 from Mostagir and Siderius (2021) identically under this new misinformation distribution.

Part 1: Note that $\tilde{F}$ still has full support for all $\gamma$. Using Lemma 2 in Mostagir and Siderius (2021), the Bayesian society mislearns if and only $r \geq \frac{(2p-1)(1-q)}{q}$ when $\theta = L$ and mislearns if and only if $r \leq 1 - \frac{(2p-1)(1-q)}{q}$ when $\theta = R$. Let $r^*$ denote the cutoff where mislearning happens, i.e., $r_L^* = \frac{(2p-1)(1-q)}{q}$ for $\theta = L$ and $r_R^* = 1 - \frac{(2p-1)(1-q)}{q}$ for $\theta = R$. Observe that when $q < \frac{2(2p-1)}{4p-1}$ both $r_L^*$ and $r_R^*$ are greater than $1/2$, but when $q > \frac{2(1-2p)}{1-4p}$ both $r_L^*$ and $r_R^*$ are less than $1/2$. Moreover, note that when $q > \frac{2p-1}{2p}$, then $r_L^*, r_R^* \in (0,1)$ so mislearning occurs with positive probability.

Let us take $q^{**} = \frac{2(2p-1)}{4p-1}$ and $q^* = \frac{2p-1}{2p}$, where it is clear that $q^{**} > q^*$. When $q \in (q^*, q^{**})$, it is clear there is mislearning with positive probability under both the original $F$ and $\tilde{F}$. We focus on the case of $\theta = L$, noting that the analysis for $\theta = R$ follows the exact same technique. Note,

however, for values of $q \in (q^*, q^{**})$:

$$\begin{aligned}
\mathbb{P}_{\tilde{F}}[r \geq r_L^*] &= \mathbb{P}_{\tilde{F}}[r \geq r_L^* | r > 1/2]\mathbb{P}_{\tilde{F}}[r > 1/2] + \mathbb{P}_{\tilde{F}}[r \geq r_L^* | r \leq 1/2]\mathbb{P}_{\tilde{F}}[r \leq 1/2] \\
&= \mathbb{P}_F[r \geq r_L^* | r > 1/2](\mathbb{P}_F[r > 1/2] - \gamma) + \mathbb{P}_F[r \geq r_L^* | r \leq 1/2](\mathbb{P}_F[r \leq 1/2] + \gamma) \\
&= \mathbb{P}_F[r \geq r_L^* | r > 1/2](\mathbb{P}_F[r > 1/2] - \gamma) \\
&\leq \mathbb{P}_F[r \geq r_L^* | r > 1/2]\mathbb{P}_F[r > 1/2] \\
&= \mathbb{P}_F[r \geq r_L^* | r > 1/2]\mathbb{P}_F[r > 1/2] + \mathbb{P}_F[r \geq r_L^* | r \leq 1/2]\mathbb{P}_F[r \leq 1/2] \\
&= \mathbb{P}_F[r \geq r_L^*]
\end{aligned}$$

so mislearning occurs more often under $F$ than under $\tilde{F}$. On the other hand, when $q > q^{**}$:

$$\begin{aligned}
\mathbb{P}_{\tilde{F}}[r \geq r_L^*] &= \mathbb{P}_{\tilde{F}}[r \geq r_L^* | r \geq 1/2]\mathbb{P}_{\tilde{F}}[r \geq 1/2] + \mathbb{P}_{\tilde{F}}[r \geq r_L^* | r < 1/2]\mathbb{P}_{\tilde{F}}[r < 1/2] \\
&= \mathbb{P}_F[r \geq r_L^* | r \geq 1/2](\mathbb{P}_F[r \geq 1/2] + \gamma) + \mathbb{P}_F[r \geq r_L^* | r < 1/2](\mathbb{P}_F[r < 1/2] - \gamma) \\
&= \mathbb{P}_F[r \geq 1/2] + \gamma + \mathbb{P}_F[r \geq r_L^* | r < 1/2](\mathbb{P}_F[r < 1/2] - \gamma) \\
&= \mathbb{P}_F[r \geq 1/2] + \mathbb{P}_F[r \geq r_L^* | r < 1/2]\mathbb{P}_F[r < 1/2] + \gamma(1 - \mathbb{P}_F[r \geq r_L^* | r < 1/2]) \\
&\geq \mathbb{P}_F[r \geq 1/2] + \mathbb{P}_F[r \geq r_L^* | r < 1/2]\mathbb{P}_F[r < 1/2] \\
&= \mathbb{P}_F[r \geq r_L^* | r \geq 1/2]\mathbb{P}_F[r \geq 1/2] + \mathbb{P}_F[r \geq r_L^* | r < 1/2]\mathbb{P}_F[r < 1/2] \\
&= \mathbb{P}_F[r \geq r_L^*]
\end{aligned}$$

so mislearning occurs more often under $\tilde{F}$ than $F$.

Part 2: Note that by Lemma 1 of Mostagir and Siderius (2021), the DeGroot society mislearns if $r \geq \frac{1-2(1-q)(1-p)}{2q} \equiv r_L^*$ when $\theta = L$ and $r \leq \frac{1-2(1-q)p}{2q} \equiv r_R^*$ when $\theta = R$. It is easy to verify that when $p > 1/2$ and $q < 1$, both $r_L^* > 1/2$ and $r_R^* < 1/2$. Therefore, when $\theta = L$, for all values of $q$:

$$\begin{aligned}
\mathbb{P}_{\tilde{F}}[r \geq r_L^*] &= \mathbb{P}_{\tilde{F}}[r \geq r_L^* | r > 1/2]\mathbb{P}_{\tilde{F}}[r > 1/2] + \mathbb{P}_{\tilde{F}}[r \geq r_L^* | r \leq 1/2]\mathbb{P}_{\tilde{F}}[r \leq 1/2] \\
&= \mathbb{P}_F[r \geq r_L^* | r > 1/2](\mathbb{P}_F[r > 1/2] - \gamma) + \mathbb{P}_F[r \geq r_L^* | r \leq 1/2](\mathbb{P}_F[r \leq 1/2] + \gamma) \\
&= \mathbb{P}_F[r \geq r_L^* | r > 1/2](\mathbb{P}_F[r > 1/2] - \gamma) \\
&\leq \mathbb{P}_F[r \geq r_L^* | r > 1/2]\mathbb{P}_F[r > 1/2] \\
&= \mathbb{P}_F[r \geq r_L^* | r > 1/2]\mathbb{P}_F[r > 1/2] + \mathbb{P}_F[r \geq r_L^* | r \leq 1/2]\mathbb{P}_F[r \leq 1/2] \\
&= \mathbb{P}_F[r \geq r_L^*]
\end{aligned}$$

so mislearning occurs more often under $F$ than $\tilde{F}$. A similar approach shows there is more mislearning under $F$ than $\tilde{F}$ when $\theta = R$ using the cutoff $r_R^*$. □

*Proof of Theorem 3.* By definition, nudging does not affect the Bayesian society, so learning remains unaffected. The rest of the proof (DeGroot learning) involves three parts: (i) showing that learning is unaffected when $F$ is symmetric, (ii) an example showing that learning can improve when $F$ is asymmetric, and (iii) an example showing that learning can become worse when $F$ is asymmetric.

Let $\pi_{i,1}^M$ be the posterior belief of a DeGroot agent $i$ who takes into account possible misinformation, e.g., $\pi_{i,1}^M = \int_0^1 \frac{(p(1-q)+qr)\pi_{i,0}}{p(1-q)\pi_{i,0}+(1-p)(1-q)(1-\pi_{i,0})+qr} f(r)\, dr$ when $m_i = R$, whereas let $\pi_{i,1}^N$ be the posterior belief of a DeGroot agent $i$ who is naive and does not, e.g., $\pi_{i,1}^N = \frac{p\pi_{i,0}}{p\pi_{i,0}+(1-p)(1-\pi_{i,0})}$ when $m_i = R$. (M is to represent misinformation-cognizant and N is to represent naive.)

Part (i): Suppose $F$ is symmetric. First, we show $\pi_{i,1}^N - \pi_{i,1}^M$ when the message is $m_i = R$ for belief $\pi_{i,0} < 1/2$ is the same as $\pi_{i,1}^M - \pi_{i,1}^N$ when the message is $m_i = L$ for belief $1 - \pi_{i,0} > 1/2$. Consider:

$$\begin{aligned}
\pi_{i,1}^N - \pi_{i,1}^M \,|\, m_i = R &= \int_0^1 \left[ \frac{p\pi_{i,0}}{p\pi_{i,0} + (1-p)(1-\pi_{i,0})} - \frac{(p(1-q)+qr)\pi_{i,0}}{p(1-q)\pi_{i,0} + (1-p)(1-q)(1-\pi_{i,0}) + qr} \right] f(r)\, dr \\
&= \int_0^1 \left[ \frac{p\pi_{i,0}}{p\pi_{i,0} + (1-p)(1-\pi_{i,0})} - \frac{(p(1-q)+q(1-r))\pi_{i,0}}{p(1-q)\pi_{i,0} + (1-p)(1-q)(1-\pi_{i,0}) + q(1-r)} \right] f(r)\, dr \\
&= \int_0^1 \frac{(2p-1)q(1-r)(1-\pi_{i,0})\pi_{i,0}}{(1-p+\pi_{i,0}(2p-1))(1-\pi_{i,0}+q(\pi_{i,0}-r) - p(1-q)(1-2\pi_{i,0}))} f(r)\, dr \\
&= \int_0^1 \left[ \frac{((1-p)(1-q)+q(1-r))(1-\pi_{i,0})}{(1-p)(1-q)(1-\pi_{i,0}) + p(1-q)\pi_{i,0} + q(1-r)} - \frac{(1-p)(1-\pi_{i,0})}{(1-p)(1-\pi_{i,0}) + p\pi_{i,0}} \right] f(r)\, dr \\
&= \pi_{j,1}^M - \pi_{j,1}^N \,|\, m_j = L
\end{aligned}$$

for an agent $j$ with $\pi_{j,0} = 1 - \pi_{i,0}$.

Next we consider the mass of agent $i$ with prior $\pi_{i,0}$ (equal to $h(\pi_{i,0})$) and the mass of agent $j$ with prior $\pi_{j,0} = 1 - \pi_{i,0}$ (equal to $h(\pi_{j,0})$). Notice that $h(\pi_{i,0}) = h(\pi_{j,0})$ because of symmetry in $H$. Note that both the set of agents with the same belief as $i$ (call these $i$-class agents) and the set of agents with the same belief as $j$ (call these $j$-class agents) receive the same distribution of $L$ and $R$ messages as $N \to \infty$. Let us consider the average belief of $i$-class and $j$-class agents, and further partition these into $(i,L),(i,R),(j,L),(j,R)$ classes depending on which message the agent received. Notice the average belief is given by:

$$\begin{aligned}
&\frac{1}{2h(\pi_{i,0})} \int_{k \in \cup(i,L),(i,R),(j,L),(j,R)} \pi_{k,1}^M - \pi_{k,1}^N \\
=&\frac{1}{2h(\pi_{i,0})} \left[ \int_{k \in (i,R)} \pi_{k,1}^M - \pi_{k,1}^N + \int_{k \in (j,L)} \pi_{k,1}^M - \pi_{k,1}^N + \int_{k \in (i,L)} \pi_{k,1}^M - \pi_{k,1}^N + \int_{k \in (j,R)} \pi_{k,1}^M - \pi_{k,1}^N \right]
\end{aligned}$$

When there are more $L$ messages than $R$ messages, the first expression can be offset entirely by a fraction of the second, and the fourth expression can be offset entirely by a fraction of the first. Thus, to measure $\frac{1}{2h(\pi_{i,0})} \int_{k \in \cup(i,L),(i,R),(j,L),(j,R)} \pi_{k,1}^M$, it is enough to consider the leftover fraction of agents in $(i,L)$ and $(j,L)$ who update but do not offset the updates of the agents in $(j,R)$ and $(i,R)$. By symmetry of $H$, we know that $\frac{1}{2h(\pi_{i,0})} \int_{k \in \cup(i,L),(i,R),(j,L),(j,R)} \pi_{k,0} = 1/2$, and because agents $k$ in $(i,L)$ and $(j,L)$ update such that $\pi_{k,1} < \pi_{k,0}$, it is clear that $\frac{1}{2h(\pi_{i,0})} \int_{k \in \cup(i,L),(i,R),(j,L),(j,R)} \pi_{k,1}^M < 1/2$. Integrating over all agents in the population (for all priors) and leveraging the symmetry of $H$ shows that the consensus is $\pi_{k,2} < 1/2$ for all agents $k$. Identical reasoning shows $\pi_{k,2} > 1/2$ if there are more $R$ messages than $L$ messages.

Finally, we note that if and only if $r \geq (1-2(1-q)(1-p))/(2q)$ there are more $R$ messages when $\theta = L$ and if and only if $r \leq (1-2(1-q)p)/(2q)$, there are more $L$ messages when $\theta = R$. Exactly as in Lemma 1 of Mostagir and Siderius (2021), this shows that the DeGroot society mislearns if and only if $r \geq (1-2(1-q)(1-p))/(2q)$ when $\theta = L$ and $r \leq (1-2(1-q)p)/(2q)$ when $\theta = R$. This proves that accuracy nudging has no effect on learning in DeGroot societies when $F$ is symmetric.

Part (ii): We construct an example with an asymmetric $F$ where learning improves from accuracy nudging. Suppose the state is $\theta = L$, total misinformation is $q = 1/5$, signal strength is $p = 3/5$, and misinformation distribution $F$ is as follows: with probability $\alpha$, $r$ is $r_1 = 1/2$, and with

probability $1-\alpha$, $r$ is $r_2 = 1$.[12] Moreover the population consists of half left-wing agents (with $\pi_{i,0} = 0.4$) and half right-wing agents (with $\pi_{i,0} = 0.6$).[13]

When $\theta = R$, the fraction of $R$ messages is $p(1-q) + qr_1 = 0.58$ with probability $\alpha$ and $p(1-q) + qr_2 = 0.68$ with probability $1-\alpha$, whereas when $\theta = L$, the fraction of $R$ messages is $(1-p)(1-q) + qr_1 = 0.42$ with probability $\alpha$ and $(1-p)(1-q) + qr_2 = 0.52$ with probability $1-\alpha$. Without accuracy nudging, the DeGroot society learns with probability $\alpha$ and mislearns with probability $1-\alpha$ (this follows from symmetry of $H$ and Lemma 1 from Mostagir and Siderius (2021)).

With accuracy nudging, beliefs in the population at $t = 1$ conditional on an $R$ message is given by:

$$\pi_{i,1} = \alpha \frac{(p(1-q)+qr_1)\pi_{i,0}}{p(1-q)\pi_{i,0} + (1-p)(1-q)(1-\pi_{i,0}) + qr_1} + (1-\alpha)\frac{(p(1-q)+qr_2)\pi_{i,0}}{p(1-q)\pi_{i,0} + (1-p)(1-q)(1-\pi_{i,0}) + qr_2}$$

For $\pi_{i,0} = .4$, we have $\pi_{i,1} \approx .466 + .014\alpha$, whereas when $\pi_{i,0} = .6$, we have $\pi_{i,1} \approx .662 + .012\alpha$. The average belief is given approximately by $.564 + .012\alpha$. Conversely, beliefs in the population at $t = 1$ conditional on an $L$ message is given by:

$$\begin{aligned}\pi_{i,1} =& \alpha \frac{((1-p)(1-q)+q(1-r_1))\pi_{i,0}}{(1-p)(1-q)\pi_{i,0} + p(1-q)(1-\pi_{i,0}) + q(1-r_1)} \\ &+ (1-\alpha)\frac{((1-p)(1-q)+q(1-r_2))\pi_{i,0}}{(1-p)(1-q)\pi_{i,0} + p(1-q)(1-\pi_{i,0}) + q(1-r_2)}\end{aligned}$$

For $\pi_{i,0} = .4$, we have $\pi_{i,1} \approx .308 + 0.018\alpha$, whereas when $\pi_{i,0} = .6$, we have $\pi_{i,1} \approx .500 + .020\alpha$. The average belief is given approximately by $.404 + .019\alpha$.

When $r = r_1$, terminal beliefs are given by:

$$\pi_{i,2} \approx ((1-p)(1-q) + qr_1)(.564 + .012\alpha) + (p(1-q) + q(1-r_1))(.404 + .019\alpha) = .471 + .017\alpha$$

and when $r = r_2$, terminal beliefs are given by:

$$\pi_{i,2} \approx ((1-p)(1-q) + qr_1)(.564 + .012\alpha) + (p(1-q) + q(1-r_1))(.404 + .019\alpha) = .487 + .016\alpha$$

Thus, provided that $\alpha \leq 0.81$, the DeGroot society that is accuracy nudged learns with probability 1 which is strictly greater than the $\alpha$ probability of learning without the nudge. (When $\alpha$ is too large, there is too little likelihood that $r = r_2$ and nudged DeGroots are more skeptical of $L$ content, even though most of the misinformation is in fact arguing for $R$, so the nudge is less effective.)

Part (iii): We consider the same parameters as from Part 2, with the exception of a different asymmetric $F$ given by $r = r_1 = 0$ with probability $\alpha$ and $r = r_2 = 0.8$ with probability $1-\alpha$. When $\theta = L$, the fraction of $R$ messages is given by $(1-p)(1-q) + qr_1 = 0.32$ with probability $\alpha$ and $(1-p)(1-q) + qr_2 = 0.48$ with probability $1-\alpha$. When the true state is $L$, the DeGroot society without an accuracy nudge learns $\theta$ with probability 1.

Using the same technique as before, when $r = r_1$, terminal beliefs are given by:

$$\pi_{i,2} \approx ((1-p)(1-q) + qr_1)(.569 + .028\alpha) + (p(1-q) + q(1-r_1))(.413 + .023\alpha) = .463 + .025\alpha$$

[12] Note that while this is a discrete distribution, the same conclusions hold for a continuous distribution that approaches $\alpha\delta(r - 1/4) + (1-\alpha)\delta(r - 4/5)$.

[13] See Footnote 12.

and when $r = r_2$, terminal beliefs are given by:

$$\pi_{i,2} \approx ((1-p)(1-q) + qr_1)(.569 + .028\alpha) + (p(1-q) + q(1-r_1))(.413 + .023\alpha) = .487 + .025\alpha$$

Notice that when $\alpha \geq 0.52$, when $r = r_2$, the accuracy nudged DeGroot agents mislearn. Thus, when $0.52 < \alpha < 1$, the accuracy nudged DeGroot agents learn with probability $\alpha < 1$, which is strictly worse than the baseline of almost sure learning with no accuracy nudge. □

*Proof of Theorem 4.* Let us fix the state $\theta = L$ for concreteness (an identical analysis applies for $\theta = R$). Given that $H$ and $F$ have full support and $H$ is symmetric, we know by Lemma 1 (resp. Lemma 2) in Mostagir and Siderius (2021) that mislearning occurs in a DeGroot (resp. Bayesian) population if and only if $r \geq \frac{1-2(1-q)(1-p)}{2q} \equiv r_D^*$ (resp. $r \geq \frac{(2p-1)(1-q)}{q} \equiv r_B^*$). Note that this implies that both types of societies mislearn with positive probability if and only if $q \geq \frac{2p-1}{2p} \equiv q^*$. Thus, if the mislearning target is $\lambda > 0$, it must be the case that $q > q^*$.

From the proof of Theorem 1(ii) in Mostagir and Siderius (2021), we know the DeGroot society mislearns with strictly less probability than the Bayesian society when $q > q^*$. Moreover, it is clear from the cutoffs $r_D^*, r_B^*$ (in both populations) that the probability of mislearning is increasing in $q$. Thus, the Bayesian society requires a lower misinformation threshold $q_B$ than $q_D$ for all $\lambda > 0$. □

# B Simulation Details

## B.1 Censorship

For simulations, we assume the regulator gets signal $s = L$ but is not ex-ante biased toward one belief of $\theta$ or another, starting with $\pi_r = 1/2$. Under this assumption, the regulator has posterior belief about $\theta$ after receiving signal $s$ given by:

$$\mathbb{P}[\theta = R \,|\, s = L] = \frac{\mathbb{P}[s = L \,|\, \theta = R]\pi_r}{\mathbb{P}[s = L \,|\, \theta = R]\pi_r + \mathbb{P}[s = L \,|\, \theta = L](1-\pi_r)} = \varepsilon$$

By removing $\rho\delta$ (resp. $(1-\rho)\delta$) content containing misinformation that argues for state $R$ (resp. state $L$), the proportion of content arguing for state $R$ is given by:

$$\kappa_R = \begin{cases} \frac{p(1-q)+qr-\rho\delta}{1-\delta} \text{, if } \theta = R \\ \frac{(1-p)(1-q)+qr-\rho\delta}{1-\delta} \text{, if } \theta = L \end{cases}$$

the likelihood of DeGroot mislearning (by Lemma 1 in Mostagir and Siderius (2021)) is given by:

$$\begin{aligned} & (1-\varepsilon) \cdot \mathbb{P}\left[\frac{(1-p)(1-q)+qr-\rho\delta}{1-\delta} \geq 1/2\right] + \varepsilon \cdot \mathbb{P}\left[\frac{p(1-q)+qr-\rho\delta}{1-\delta} \leq 1/2\right] \\ = \; & (1-\varepsilon) \cdot \mathbb{P}\left[r \geq \frac{\delta(2\rho-1)+2p(1-q)+2q-1}{2q}\right] + \varepsilon \cdot \mathbb{P}\left[r \leq \frac{1+\delta(2\rho-1)-2p(1-q)}{2q}\right] \end{aligned}$$

In the simulation that follows, we assume that $F$ follows a triangular distribution $a = 0, b = 1, c = 1/2$ (i.e., there is roughly misinformation equally from both sides). Moreover, we let $p = 3/5$ and $q = 2/5$.

Figure 4 shows the optimal choice of $\rho$ for the regulator who always elects **Censor** for DeGroot agents, as per Theorem 1. A few observations emerge. First, when $\varepsilon$ is low (so the research

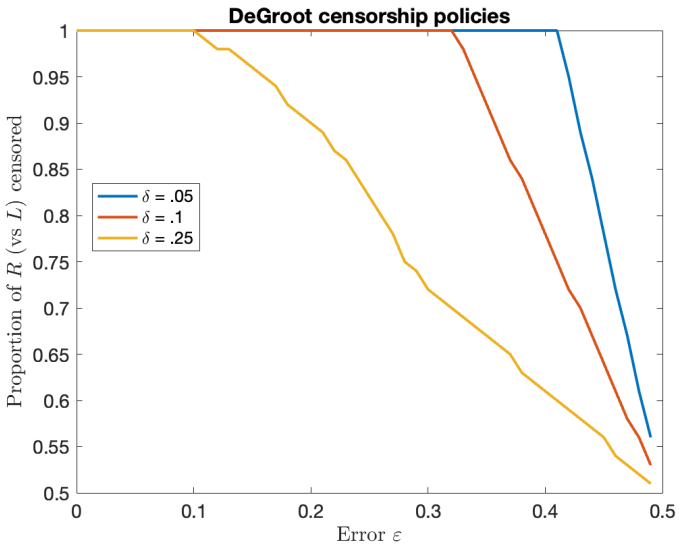



Figure 3. Optimal censorship split $\rho$ for the regulator.

signal $s$ is very indicative of $\theta$), the optimal censorship policy is always to remove all misinformation arguing against the signal $s$. Second, as $\varepsilon$ starts increasing (and the research signal becomes noisier), the regulator still removes misinformation but becomes less tethered to signal $s$ in this removal, creating a more even split. As the research signal becomes very imprecise, the regulator mostly focuses resources on removing misinformation from both sides equally ($\rho \approx 0.5$). Lastly, we note that as the misinformation detection technology $\delta$ improves, the regulator has "great power and great responsibility." In other words, the regulator should not over-exert her belief of $\theta$ in the censorship policy, but provide more even censorship across ideological-charged misinformation. In other words, a regulator

## B.2 Provision of Diverse Content

Recall that DeGroot learning always benefits from more diverse content (per Theorem 2), so here we focus on Bayesian learning. In Figure 4, we simulate $q^*$ (the level of misinformation at which Bayesian mislearning occurs with positive probability) and $q^{**}$ (the level of misinformation at which Bayesian mislearning is exacerbated by more diverse content) for different values of $p$. The former is given by the solid blue line and the latter by the dashed red line, with the gap between them denoting the range for $q$ where the diverse content policy *strictly* improves learning outcomes for the Bayesian society. (Below the blue solid line learning occurs with probability 1 regardless of the policy, above the orange dashed line learning is made worse by the diverse content policy, and in-between learning is improved by the diverse content policy.)

With highly informative content (i.e., $p > 3/4$), note that $q^{**} > 1/2$, so the diverse content policy is also effective in Bayesian societies (by assumption that $q < 1/2$). However, with organic content that is not strongly correlated with $\theta$, there is backfire potential when providing more diverse content. Because the gap between $q^*$ and $q^{**}$ is increasing as the informativeness of content ($p$) decreases, topics with more controversy are more likely to have a small "gap" for $q$ where the diverse content policy might actually be effective.

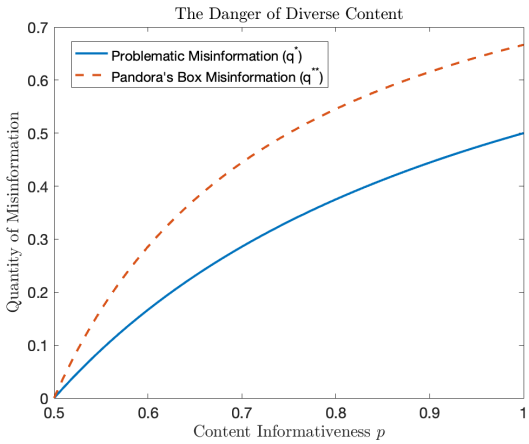



Figure 4. $q^*$ and $q^{**}$ from Theorem 2.

### B.3 Accuracy Nudging

We run a simulation with $K = 10^6$ trials to see the effectiveness (and backfire potential) of accuracy nudging. We fix $\theta = L$ for concreteness. In this simulation, we assume $p \sim U[1/2, 1]$, $\alpha \sim U[0, 1]$, $q \sim U[0, 1/2]$, $r_1 \sim [1/10, 1/2]$, and $r_2 \sim [3/5, 1]$. Moreover, we let half the population begin with a left-leaning belief $\pi_L \sim [0, 1/2]$ and half the population begin with the opposite right-leaning belief $\pi_R \equiv 1 - \pi_L$.

Without nudging, DeGroot agents will mislearn whenever $(1-p)(1-q) + qr > 1/2$, and with the two values for $r \in \{r_1, r_2\}$ (with $r_1 < r_2$), mislearning occurs with probability:

$$\begin{cases} 0, & \text{if } (1-p)(1-q) + qr_2 < 1/2 \\ 1/2, & \text{if } (1-p)(1-q) + qr_1 < 1/2 < (1-p)(1-q) + qr_2 \\ 1, & \text{if } (1-p)(1-q) + qr_1 > 1/2 \end{cases}$$

With nudging, we assume $\alpha \sim U[0, 1]$, which determines the likelihood of drawing $r_1$ (probability $\alpha$) or $r_2$ (or probability $1 - \alpha$). Beliefs at time $t = 1$, conditional on message $R$, are given by:

$$\{\pi'_L \,|\, R\} = \alpha \frac{(p(1-q) + qr_1)\pi_L}{p(1-q)\pi_L + (1-p)(1-q)(1-\pi_L) + qr_1} + (1-\alpha)\frac{(p(1-q) + qr_2)\pi_L}{p(1-q)\pi_L + (1-p)(1-q)(1-\pi_L) + qr_2}$$

$$\{\pi'_R \,|\, R\} = \alpha \frac{(p(1-q) + qr_1)\pi_R}{p(1-q)\pi_R + (1-p)(1-q)(1-\pi_R) + qr_1} + (1-\alpha)\frac{(p(1-q) + qr_2)\pi_R}{p(1-q)\pi_R + (1-p)(1-q)(1-\pi_R) + qr_2}$$

Let $\{\pi \,|\, R\} = (\{\pi'_L \,|\, R\} + \{\pi'_R \,|\, R\})/2$. Beliefs at time $t = 1$, conditional on message $L$, are given by:

$$\{\pi'_L \,|\, L\} = \alpha \frac{((1-p)(1-q) + qr_1)\pi_L}{(1-p)(1-q)\pi_L + p(1-q)(1-\pi_L) + qr_1} + (1-\alpha)\frac{((1-p)(1-q) + qr_2)\pi_L}{(1-p)(1-q)\pi_L + p(1-q)(1-\pi_L) + qr_2}$$

$$\{\pi'_R \,|\, L\} = \alpha \frac{((1-p)(1-q) + qr_1)\pi_R}{(1-p)(1-q)\pi_R + p(1-q)(1-\pi_R) + qr_1} + (1-\alpha)\frac{((1-p)(1-q) + qr_2)\pi_R}{(1-p)(1-q)\pi_R + p(1-q)(1-\pi_R) + qr_2}$$

Once again, let $\{\pi \mid L\} = (\{\pi'_L \mid L\} + \{\pi'_R \mid L\})/2$. Finally, mislearning occurs with probability:

$$\begin{cases} 0, & \text{if } ((1-p)(1-q)+qr_2)\{\pi \mid R\} + (p(1-q)+q(1-r_2))\{\pi \mid L\} < 1/2 \\ 1/2, & \text{otherwise} \\ 1, & \text{if } ((1-p)(1-q)+qr_1)\{\pi \mid R\} + (p(1-q)+q(1-r_1))\{\pi \mid L\} > 1/2 \end{cases}$$

The results are presented in Table 1 of the main body.

### B.4 Performance Targets

We suppose that $F$ is uniformly distributed on $[0,1]$. Then one can explicitly solve for the performance targets in each population:

$$q_D = \frac{2p-1}{2(p-\lambda)} \qquad q_B = \frac{2p-1}{2p-\lambda}$$

Letting $p = 3/5$, we obtain the performance targets for both the DeGroot and Bayesian societies as a function of the mislearning target $\lambda$ in Figure 5. As argued in Theorem 4, perhaps counterintuitively, the DeGroot target should always be more lenient than the Bayesian target. Most interestingly, the gap in optimal targets between the sophistication types is first increasing and then decreasing. Thus, sophistication plays a critical role for regulators who set moderate goals for mislearning rates.

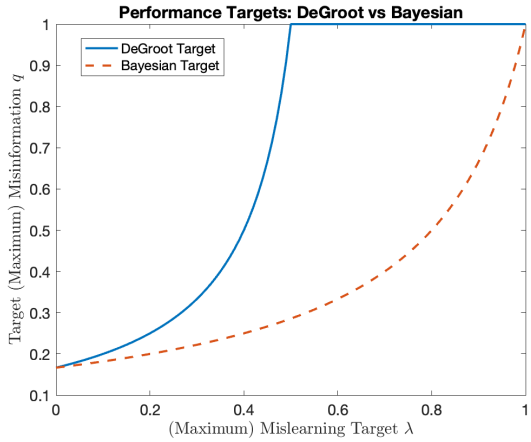



Figure 5. Performance targets across sophistication type.

## C Details of Networked Learning

We highlight the networked learning details present in Appendix B.4 of Mostagir and Siderius (2021).

### C.1 DeGroot and Bayesian Network Learning

In Section 3, we considered a model of learning where all agents observe the beliefs of all other agents. However, this is often an unrealistic assumption, and there is a wide array of literature

that considers the subtleties of learning when these observations are incomplete (see Golub and Sadler (2017) for a survey). The common approach to modeling this incompleteness is to assume there is a social network with pairwise connections that determines who can observe (or talk to) whom. In this context, the model in Section 3 assumes a *complete* social network, which simplifies the relevant dynamics to two periods.

In this section, we relax this assumption by considering arbitrary network architectures and the richer learning dynamics that occur over a longer time horizon. Under relatively mild conditions on the network structure, we show network learning leads to the same outcomes and insights found in the more parsimonious complete network setting, thereby rendering our assumption to be largely without loss of generality. We do this by building off of the previous literature on network learning in both Bayesian and DeGroot populations.

**Network Preliminaries**. Recall, we assume that all agents are arranged in an undirected social network $\mathbf{G}$. A link $i \leftrightarrow j$ denotes that agent $i$ and agent $j$ observe (or talk to) each other. We let $\mathcal{N}_i$ denote the neighborhood of agent $i$ (i.e., the set of agents $j$ with $i \leftrightarrow j$). The adjacency matrix $\mathbf{A}$ of $\mathbf{G}$ is a binary matrix with $[\mathbf{A}]_{ii} = 1$ and $[\mathbf{A}]_{ij} = 1$ if and only if $i \leftrightarrow j$. Let $d_i^{\mathbf{G}}$ be the degree of agent $i$ in $\mathbf{G}$. We say a network $\mathbf{G}$ is $k$-regular if all agents have degree $k$.

We consider a discrete time model (as before) but with a much longer learning horizon $T$, $t = 0, 1, 2, \ldots, T$. We let $\tilde{\pi}_{i,t}$ denote the belief of agent $i$ at time $t$ under network learning, whereas $\pi_{i,0}, \pi_{i,1}$, and $\pi_{i,2}$ denote the beliefs of agent $i$ at time $0, 1$, and $2$, respectively, in the baseline model (i.e., a complete network).

**Bayesian Population**. Network learning in settings with fully rational (i.e., Bayesian) agents has been studied in many contexts, most notably in Acemoglu et al. (2011) and Gale and Kariv (2003). As is common in many models of Bayesian network learning,[14] we assume that the network $\mathbf{G}$ and initial priors $\pi_{i,0}$ are common knowledge.[15] Bayesian agents observe the beliefs of all agents in their neighborhoods for all $t \geq 1$ (i.e., agent $i$ observes at time $t$ the beliefs from $t-1$, $\{\pi_{j,t-1}\}_{j \in \mathcal{N}_i}$). Our next result shows that terminal beliefs in network learning indeed converge to the terminal beliefs of the baseline model:

**Lemma 1.** *Suppose* $\mathbf{G}$ *is connected. Then as* $T \to \infty$, $\tilde{\pi}_{i,T} \to \pi_{i,2}$.

This claim follows directly from Mueller-Frank (2013). While agents do not hold a common prior about $\theta$, common knowledge of the heterogenous priors $\{\pi_{j,0}\}_{j=1}^N$ allows agents to recalibrate the (updated) beliefs they see to their own prior. It is clear that the private information at $t = 1$ (i.e., the messages) are drawn from a finite partition of the $\theta$ state space (conditional on misinformation split $r$). Thus, by Theorem 4 of Mueller-Frank (2013), all Bayesian agents uncover the private information (i.e., $t = 1$ messages) of all other agents (including non-neighbors) in the network as $T \to \infty$, as is the case at $t = 2$ in the baseline model.

**DeGroot Population**. Due to demanding assumptions about the reasoning abilities of Bayesian agents, "rule-of-thumb" learning has become a popular alternative model. The most common model is that of Degroot (1974), and later expanded upon in works such as Golub and Jackson (2010) and DeMarzo et al. (2003). In these models, agents are assumed to update their beliefs using the simple heuristic of taking linear combinations of their neighbors' beliefs. Formally,

[14] In addition to Acemoglu et al. (2011) and Gale and Kariv (2003), see Mueller-Frank (2014) and Mossel et al. (2014).

[15] An alternative assumption, which does not require strong common knowledge assumptions of non-neighbor priors or the network structure, is that the size of the smallest neighborhood grows unboundedly as $N \to \infty$.

agent $i$ forms belief $\pi_{i,t+1}$ at each time $t$ by computing:

$$\pi_{i,t+1} = \frac{1}{1+d_i^{\mathbf{G}}}\left(\pi_{i,t} + \sum_{j\in\mathcal{N}_i}\pi_{j,t}\right)$$

Our next result provides conditions under which DeGroot learning over the network $\mathbf{G}$ leads to the same terminal beliefs as in our baseline model:

**Lemma 2.** *Suppose* $\mathbf{G}$ *is a connected,* $k$*-regular network. Then as* $T\to\infty$, $\tilde{\pi}_{i,T}\to\pi_{i,2}$.

This claim follows directly from Golub and Jackson (2010). First, by Proposition 1 in Golub and Jackson (2010), observe that consensus is reached (as in the baseline model) because the normalized adjacency matrix $\mathbf{A}$ is irreducible and aperiodic, the former following from the connectedness assumption and the latter following from a positive diagonal on $\mathbf{A}$. Second, by Theorem 3 in Golub and Jackson (2010), the consensus belief of the agents as $T\to\infty$ is given by $\tilde{\pi}_{i,\infty} = \sum_{j=1}^N v_j^{\mathbf{G}}\pi_{j,1}$ for all agents $i$, where $v_j^{\mathbf{G}}$ is the (eigenvector) centrality of agent $j$ (according to the row-stochastic normalized adjacency matrix $\mathbf{A}$). Because $v_j^{\mathbf{G}} = d_j^{\mathbf{G}}/\sum_{\ell=1}^N d_\ell^{\mathbf{G}}$, we obtain by $k$-regularity that $\tilde{\pi}_{i,\infty} = \frac{1}{N}\sum_{j=1}^N \pi_{j,1} = \pi_{i,2}$.

Observe that Lemma 2 requires an additional condition not present in Lemma 1, which is that no agent is more "influential" than any other agent in the network $\mathbf{G}$, as measured by her degree. This is easily satisfied by many network topologies, including several classes of random networks such as Erdos-Renyi networks (where links between agents occur uniformly at random).

## C.2 Multiple Messages

Let us consider the complete network setting of Section 3 for simplicity, but note that the reduction from arbitrary network learning discussed previously still applies.

In a Bayesian society with $N\to\infty$, by the strong law of large numbers, the first round of messages reveals the true fraction of $R$ messages, $\rho_R$, and the true fraction of $L$ messages, $\rho_L$, almost surely. Obtaining additional messages in subsequent rounds does not alter the (almost surely) known values of $\rho_R$ or $\rho_L$, thus, learning is entirely unaffected by more incoming messages.

In a DeGroot society, after the first round of messages, agents converge to a consensus about $\theta$ which is a function of $\rho_R$ (and $\rho_L$) alone. When $H$ is symmetric, whether $\rho_R > 1/2$ or $\rho_L > 1/2$ determines if the consensus, call it $\pi_2$, lies more toward state $R$ (i.e., $\pi_2 > 1/2$) or state $L$ (i.e., $\pi_2 < 1/2$). By the martingale property of Bayesian updating, it is easy to see that $\mathbb{E}[\text{BU}(\pi_2)\,|\,\rho_R > 1/2\,;\,\pi_2 > 1/2] > 1/2$ and $\mathbb{E}[\text{BU}(\pi_2)\,|\,\rho_R < 1/2\,;\,\pi_2 < 1/2] < 1/2$ (where BU is the Bayesian update for DeGroot agents conditioning on the message, given by Equation (1) and Equation (2)). Therefore, one can show by induction that beliefs remain on the same side of belief $1/2$ as they are at $t = 2$ for all $t \geq T$, even with additional messages. Consequently, the likelihood of (mis)learning is unaffected by any further stream of messages.