Perverse Ethical Concerns: Online Platforms and Offline Conflicts

Chang, Dongkyu and Vong, Allen

City University of Hong Kong, University of Macau

5 November 2021

Perverse Ethical Concerns: Online Platforms and Offline Conflicts[*]

Dongkyu Chang[†] Allen Vong[‡]

(Click here for the latest version.)

November 5, 2021

Abstract

We study a model where many citizens learn a hidden state individually on an online platform. The platform slants news and imperfectly filters misinformation, triggering conflicts about the state among the citizens. We show that a platform that faces an ethical concern to internalize conflict costs could perversely aggravate conflicts. This cautionary observation highlights that societal efforts to mitigate conflicts, such as investments in ethical algorithms, public awareness campaigns, and government policies, are effective if and only if their implementations are sufficiently aggressive.

JEL codes: C72, D83, L86.

Keywords: platforms, social media, polarization, conflicts.

1 Introduction

Online platforms are commonly viewed as facilitating the spread of misinformation and flaming offline conflicts such as hate crimes and political violence (see, e.g., Sunstein, 2001, 2018; Müller and Schwarz, 2018, 2020; Settle, 2018; Bursztyn, Egorov, Enikolopov and Petrova, 2019; Karell, 2021). The 2016 Pizzagate incident, the 2019 Christchurch

[*] We are grateful to Johannes Hörner, Larry Samuelson, and Marina Halac for their constant support. For helpful comments, we thank Yeon-Koo Che, Yi Chen, Syngjoo Choi, Chris Edmond, Ilwoo Hwang, Shota Ichihashi, Bart Lipman, Dmitry Shapiro, seminar participants in Korea University, SAET 2021, Seoul National University, Sungkyunkwan University, and University of Macau. Allen Vong acknowledges the Asia-Pacific Academy of Economics and Management at the University of Macau for financial support.

[†] City University of Hong Kong. Email: donchang@cityu.edu.hk.

[‡] University of Macau. Email: allenvongecon@gmail.com.

mosque shootings, and the 2021 Capitol riot are prominent recent examples. Societies respond with efforts that lead platforms to face ethical concerns to internalize the cost of the conflicts that their contents induce. These efforts include public awareness campaigns, such as the *Wall Street Journal*'s investigative reports and podcast series, as well as congressional hearings regarding platforms' behavior,[1] that cause platforms to worry about their reputations for mitigating conflicts. These efforts also include a growing research program on ethical algorithms that spans multiple disciplines, including computer science, engineering, sociology, and philosophy (see, e.g., Wu, 2017; Kearns and Roth, 2019; Cusumano, Gawer and Yoffie, 2021). Governments worldwide also adopt various policies to motivate platforms to act as if they face ethical concerns to mitigate conflicts (see, e.g., Napoli, 2019; Funke and Flamini, 2021).

This paper offers a cautionary observation concerning these societal efforts. We show that by internalizing the conflict costs, platforms could perversely aggravate conflicts: citizens who anticipate platforms' ethical concerns might become too confident of the personalized contents that they read on the platforms and, in turn, become more hostile against disagreeing opinions. Importantly, our results highlight that the societal efforts to mitigate conflicts are effective if and only if their implementations are sufficiently aggressive.

We deliver our results in a model where a large number of citizens learn a hidden state, for instance, the change in vaccine efficacy against a new COVID-19 variant, by using a platform. The platform is an information intermediary. It receives news reports about the state and then creates a noisy idiosyncratic signal for each citizen, based on two algorithms that the platform develops at a cost. One algorithm slants news contents for each citizen, and the other algorithm filters misinformation in the news reports. Each citizen's received signal summarizes the personalized contents that she reads on the platform. Upon receiving their signals, the citizens' beliefs about the state typically disagree. Their disagreements trigger costly conflicts.

We begin with a baseline version of the model where the platform is self-interested. The platform develops the algorithms to create contents that citizens enjoy reading for higher revenues. Citizens like signals that are informative about the state as well as signals that conform to their individual biases. Some citizens are "rational," who perform Bayesian inferences; others are "credulous," who are not Bayesians and plainly

[1] See, e.g., "Big Tech CEOs Face Lawmakers in House Hearing on Social Media's Role in Extremism, Misinformation," *The Washington Post*, April 9, 2021.

believe that the state is equal to their received signals. We next consider a version of the model where the platform faces an additional ethical concern to internalize the conflict costs when developing its algorithms. We contrast the equilibria in the two versions of the model and deliver our main result: the platform's ethical concern could perversely aggravate conflicts between any two rational citizens, but unambiguously mitigate conflicts between any two citizens involving at least one credulous citizen.

The intuition of the main result is as follows. Given its ethical concern, the platform filters misinformation more aggressively and adjusts its slants for the citizens so as to mitigate conflicts. On the one hand, such changes of the platform's algorithms improve the citizens' learning about the state. We call this phenomenon the learning effect. On the other hand, the rational citizens correctly anticipate such changes of the algorithms and are more confident about their own inferences. We call this phenomenon the confidence effect. If the confidence effect dominates the learning effect, which is the case when the platform's ethical concern is too weak to induce a sufficiently aggressive change of its algorithms, then conflicts between any two rational citizens escalate. In contrast, this perverse outcome never arises in the credulous citizens' inferences. Because the credulous citizens interpret signals at their face value, the confidence effect is absent when they acquire information from the platform. As a result, the platform's ethical concern unambiguously mitigates conflicts between any two citizens involving at least one credulous citizen.

Policy implications. In Section 6, we return to the baseline model where the platform is self-interested and analyze several popular government efforts that motivate platforms to act as if they face ethical concerns to mitigate conflicts. These efforts range from legislation against misinformation to proposing a version of the FCC fairness doctrine for online platforms. We find that when adopting these efforts, the governments are confronted with the same difficulty that platforms with ethical concerns face: a perverse outcome may arise unless the implementation of the government efforts is sufficiently aggressive.

While government efforts to mitigate conflicts typically target platforms that represent the supply side of online information, some governments also adopt efforts that target the demand side, such as media literacy campaigns that aim to educate these credulous citizens. We show that such campaigns could aggravate conflicts unless they are coupled with aggressive supply-side efforts that ensure sufficient filtering of misinformation: by improving the credulous citizens' ability to process information, these campaigns disrupt the platform's incentive to filter misinformation as it is more difficult for the platform to manipulate these citizens' beliefs.

More broadly, our result speaks to debates concerning the transparency of platform algorithms (see, e.g., MacCarthy, 2020). While a typical argument for transparency is to promote effective monitoring of platforms,[2] our result highlights alternatively that transparency allows platforms to correctly internalize their social responsibilities. We show that if the algorithms were observable to the citizens, then when the platform chooses its algorithms in the presence of its ethical concern, it anticipates that the rational citizens perform inferences based on its actual choice of algorithms. The platform thus correctly internalizes the conflict costs, and no perverse outcome arises.

Related literature. This paper speaks to a burgeoning research program on ethical algorithms, as noted at the outset, that covers topics beyond offline conflicts, such as privacy, addiction, and fairness issues. We contribute to this literature from a game-theoretic perspective by elucidating the strategic implications of platforms' ethical concerns. Limiting to the context of offline conflicts, our results offer a cautionary observation against the conventional wisdom that arguably underlies this burgeoning research program, namely that ethical concerns are unambiguously socially desirable.

Within economics, our work contributes to the literature of disagreements among Bayesian agents driven by heterogeneous prior beliefs (see, e.g., Dixit and Weibull, 2007; Andreoni and Mylovanov, 2012; Sethi and Yildiz, 2012; Baliga, Hanany and Klibanoff, 2013; Zanardo, 2017; Kartik, Lee and Suen, 2021) or by competition among information providers (see, e.g., Chen and Suen, 2021; Perego and Yuksel, 2021). Departing from the literature, our analysis zooms in on conflicts driven by citizens' heterogeneous interim beliefs that are induced by platforms' optimizing algorithms. To highlight this phenomenon, our setup considers citizens who share a common prior belief and learn individually on the platform; we later show that assuming heterogeneous prior beliefs do not alter our insights.[3] We also depart from the literature by considering a society that consists of both Bayesians and non-Bayesians, namely rational and credulous

[2] See, e.g., "Whistle-blower Unites Democrats and Republicans in Calling for Regulation of Facebook," *The New York Times*, October 5, 2021.

[3] Our baseline setup that citizens share a common prior belief and learn individually is reminiscent of models of common learning (e.g., Cripps, Ely, Mailath and Samuelson, 2008). While this literature focuses on asymptotic beliefs given an exogenous learning process, our analysis focuses on non-asymptotic beliefs given an endogenous learning process that results from the platform's optimization.

citizens, and examining disagreements within and between the two groups.[4] In our view, the analysis of credulous citizens, alongside their conflicts with rational citizens, is not only theoretically attractive but also important for policy implications.

Finally, our analysis contributes broadly to the literature of media economics (see, e.g., Prat and Strömberg, 2013; Anderson, Waldfogel and Strömberg, 2015). In particular, by introducing a model of credence information intermediation, our analysis sheds light on media biases to cater to the demand side of information (see, e.g., Suen, 2004; Mullainathan and Shleifer, 2005; Gentzkow and Shapiro, 2006)[5] and the role of social media in political conflicts (see, e.g., Zhuravskaya, Petrova and Enikolopov, 2020). A key feature of our model is that rational citizens cannot ascertain the veracity of their signals and thus assess the signals based only on their expectations of the platform's hidden algorithms. This credence nature of the platform's signals, in addition to driving our main result, yields notable positive implications regarding platforms' slanting and filtering incentives that accord well with empirical findings:

Slanting—our model predicts that platforms slant for both rational and credulous citizens in the direction of their individual biases. This prediction sheds light on evidence that social media users encounter more contents aligned with their ideology (see, e.g., Bakshy, Messing and Adamic, 2015) and that extreme contents tend to trend on platforms (see, e.g., Lang, Erickson and Jing-Schmidt, 2021).

Filtering misinformation—our model predicts that platforms spend costs to filter misinformation in order to cater to citizens' biases even when they face no ethical concerns. This prediction offers a reconciling perspective on platforms' significant investments in their filtering algorithms despite often being criticized for catering to citizens' biases at the expense of filtering misinformation.[6]

[4] Our modeling of credulous citizens as non-Bayesian agents who take signals at their face value follows from Kartik, Ottaviani and Squintani (2007); see also Little (2017).

[5] Gentzkow and Shapiro (2010) provide an empirical analysis that shows that news consumers have strong preference for like-minded news and news providers respond strongly to such preference. In general, media slanting might also be driven by the supply side of information that is biased. See Gentzkow, Shapiro and Stone (2015) for a survey concerning media bias.

[6] A recent example of these conflicting perspectives is the ongoing exchanges of "conversations" between *Facebook* and the *Wall Street Journal*. Before the *Wall Street Journal* launched the investigative reports and podcast series that are noted in the opening paragraphs to publicly investigate *Facebook*'s efforts in mitigating conflicts, it published an article claiming *Facebook*'s lack of effort in filtering misinformation to mitigate conflicts. *Facebook* responded by publicly outlining its investment efforts to mitigate conflicts and what the *Wall Street Journal* "got wrong." See "Facebook Executives Shut Down Efforts to Make the Site Less Divisive," *The Wall Street Journal*, May 26, 2020 and "Investments to Fight Polarization," *Facebook*, May 27, 2020.

2 Model

A unit mass of citizens, indexed by $i \in [0,1]$, wants to learn a hidden state $\theta \in \mathbb{R}$ from an online platform. They share a common prior belief that θ is normally distributed with mean normalized to 0 and precision $p > 0$. The platform is an information intermediary that receives noisy news reports about the state from external sources and then creates personalized contents for each citizen based on the reports.

The platform develops two algorithms to create the contents. One algorithm filters misinformation among the news reports that the citizens receive, and the other algorithm slants the reports individually for each citizen. Specifically, the platform chooses a filter $f \in \mathbb{R}_+$ and an (integrable) slant function $s : [0,1] \to \mathbb{R}$, where $s_i \equiv s(i)$ is the slant designated for citizen i. The algorithms are hidden from the citizens. The citizens take no actions.

Given algorithms (f, s), each citizen i receives a signal y_i, which summarizes the personalized contents that citizen i reads on the platform, given by

$$y_i = s_i + \theta + \varepsilon_i, \tag{1}$$

where ε_i represents misinformation in the contents and is normally distributed with mean 0 and precision $q + f$. The parameter $q > 0$ is exogenous and represents the default precision of the signal absent any filtering;[7] the noise ε_i is independent of the state θ and is independent across citizens. We interpret a higher filter f as a more aggressive filter. We also interpret the platform as receiving a large amount of both negative and positive news reports, and the platform chooses some negative (resp., positive) slant $s_i < 0$ (resp., $s_i > 0$) for citizen i by omitting an appropriate amount of positive (resp., negative) news reports.[8]

The signal y_i is private to citizen i. In practice, citizens might communicate the contents they read on platforms with their peers. Our insights carry over to settings where the citizens observe not only their own signals, but also a few other citizens' signals. What is crucial to our results is that the unit mass of citizens does not commonly observe the same signals, so that there is some posterior disagreement about the state among them. Alternatively, one can interpret the signal y_i as citizen

[7] The assumption that q is positive is immaterial for our results. It simply rules out a trivial equilibrium with zero filtering.

[8] See Mullainathan and Shleifer (2005, Section V) for a microfoundation of such slanting technology.

i's acquired information about the state after reading contents on the platform and communicating with peers. Finally, in reality, citizens might acquire individual private signals about the state before acquiring information from the platform. We analyze this setting in Appendix A.2 and show that our main insights carry over.

Upon receiving signal y_i, citizen i forms an estimate $\hat{\theta}_i(y_i)$ of the state. Citizens differ in their reading preferences and their ability to interpret the signals. Each citizen i's reading preference is characterized by a bias $b_i \in \mathbb{R}$, capturing the value that she would like the state to be equal to. Regarding ability, each citizen is either rational or credulous. Rational citizens are Bayesian. Each rational citizen i's estimate given her signal y_i and her expectation $(f^{*,i}, s^{*,i})$ of the platform's (hidden) algorithms is plainly her posterior mean of the state. In contrast, credulous citizens are not Bayesians. They believe that the state is equal to their received signals and do not form any expectation of the platform's algorithms.

We assume that each citizen i's bias and ability are commonly known. This assumption is stronger than necessary and plainly simplifies the exposition. As will be clear, for our results, it suffices to impose the following two assumptions. First, the mass of credulous citizens and their aggregate bias, namely $\int_C b_i \, \mathrm{d}i$ where $C \subseteq [0,1]$ is the set of credulous citizens, are commonly known. Second, each citizen's bias and reading ability are commonly known between the citizen and the platform. Such bilateral relationship between each citizen and the platform can be interpreted as a long-run phenomenon where both parties understand that the platform has collected a sufficient amount of information from the citizen concerning her reading preference and ability. Finally, without loss, we assume that for some $r \in [0,1]$, $C = (r,1]$ so that each citizen $i \in [0,r]$ is rational and each citizen $i \in (r,1]$ is credulous.

The platform's payoff is equal to its revenue minus its cost to develop the algorithms. By choosing a filter f, the platform incurs a quadratic cost $cf^2/2$, where $c > 0$ measures how costly it is for the platform to filter more aggressively; by choosing a slant s_i for citizen i, the platform incurs a quadratic cost $ks_i^2/2$, where $k > 0$ measures how costly it is for the platform to slant more aggressively. On the other hand, the platform derives a higher revenue by attracting more citizens' activities on the platform, and citizens are more active if they enjoy the contents more. Specifically, given signals $y := (y_i)_{i\in[0,1]}$ and the rational citizens' expectations of the platform's algorithms

$(f^*, s^*) := (f^{*,i}, s^{*,i})_{i \in [0,r]}$, the platform's realized revenue is

$$v(y; f^*, s^*) := \beta \int_0^1 -(\hat{\theta}_i(y_i) - b_i)^2 \, \mathrm{d}i + \tau \int_0^r -\mathbf{Var}_i(\theta | y_i) \, \mathrm{d}i, \tag{2}$$

where $\beta > 0$ and $\tau > 0$ are exogenous parameters. The parameter β measures how beneficial it is for the platform to cater to the citizens' biases, as captured by the quadratic loss of the citizens' estimate from their biases. The parameter τ measures how beneficial it is for the platform to improve the quality of the citizens' learning, as captured by each rational citizen's negative posterior variance. The revenue is independent of the quality of credulous citizens' learning about the state, as they are assumed to believe that their signals fully reveal the state.

To summarize, the platform's payoff given its algorithms (f, s) and rational citizens' expectations (f^*, s^*) is given by

$$\mathbf{E}[v(y; f^*, s^*)] - \frac{cf^2}{2} - \int_0^1 \frac{ks_i^2}{2} \, \mathrm{d}i, \tag{3}$$

where the expectation $\mathbf{E}$ is taken over the distribution of signal profiles y induced by its algorithms (f, s).

The solution concept that we use is Bayesian Nash equilibrium in pure strategies, henceforth equilibria. We focus on equilibria in pure strategies to facilitate tractable belief updating by the rational citizens; nonetheless, we allow the platform to contemplate deviations to arbitrary strategies. In any such equilibrium, the platform chooses its algorithms (f, s) to maximize its payoff (3) given the rational citizens' expectations (f^*, s^*), such that the expectations are correct. Thus, the rational citizens' equilibrium expectations of the algorithms must be identical. In the remainder of this paper, when we say that the rational citizens' expectation of the algorithms is (f^*, s^*), without loss, we refer to the event that they expect the same algorithms and we abuse notation to denote such algorithms by (f^*, s^*) to ease the exposition. Moreover, throughout, we write $\mathbf{E}^*[\cdot]$ as each rational citizen's expectation by expecting algorithms (f^*, s^*) and write $\mathbf{E}[\cdot]$ as the platform's expectation by choosing algorithms (f, s).

In this baseline version of the model, we say that the platform is self-interested as its objective (3) is to maximize profits. In the next section, we analyze the equilibria given a self-interested platform. Then, we turn to define and analyze equilibrium conflicts among the citizens.

3 Equilibrium

Proposition 1 below characterizes the essentially unique equilibrium of the baseline model. The equilibrium is essentially unique in the sense that it is unique except for the slants chosen for a subset of users with zero Lebesgue measure.

Proposition 1. *There exists an essentially unique equilibrium. In the equilibrium, the self-interested platform chooses* $(f, s) = (f^{\mathrm{S}}, s^{\mathrm{S}})$ *where:*

1. *The filter* f^{S} *is positive and is characterized by*

$$\beta\left(\frac{r}{(p+q+f^{\mathrm{S}})^2} + \frac{1-r}{(q+f^{\mathrm{S}})^2}\right) = cf^{\mathrm{S}}. \tag{4}$$

2. *For each rational citizen* i*, the slant* s_i^{S} *is characterized by*

$$s_i^{\mathrm{S}} = \frac{2\beta}{k}\left(\frac{q+f^{\mathrm{S}}}{p+q+f^{\mathrm{S}}}\right)b_i. \tag{5}$$

3. *For each credulous user* i*, the slant* s_i^{S} *is characterized by*

$$s_i^{\mathrm{S}} = \frac{2\beta}{k+2\beta}b_i. \tag{6}$$

In the equilibrium, the platform filters and slants *solely* to cater to the citizens' biases: the filter is strictly increasing in the benefit β to cater to the citizens' biases and vanishes as β vanishes; the slants are proportional to the citizens' biases. To see why this is the case, observe that given any rational citizens' expectation (f^*, s^*), the component of the platform's revenue (2) that corresponds to improving citizens' learning is independent of the platform's actual choice of algorithms (f, s):

$$\mathbf{E}\left[\tau\int_0^r -\mathbf{Var}^*\left[\theta|y_i\right]\mathrm{d}i\right] = \mathbf{E}\left[\tau\int_0^r \frac{-1}{p+q+f^*}\,\mathrm{d}i\right] = \frac{-\tau r}{p+q+f^*},$$

where the first equality follows from Bayesian updating. Thus, given the expectation (f^*, s^*), the platform's incentives to develop the algorithms rely solely on the other component of its revenue, namely the component that corresponds to bias catering:

$$\mathbf{E}\left[\beta\int_0^1 -(\hat{\theta}_i\left(y_i\right)-b_i)^2\,\mathrm{d}i\right]. \tag{7}$$

From the platform's perspective, when it chooses the algorithms, the citizens' estimates are random (because their signals are random). By filtering more aggressively, the platform reduces the dispersion of the citizens' estimates and better caters to their biases, improving (7). Equation (4) pins down the unique equilibrium filter f^{S} by equating the diminishing marginal benefit to filter on the left side and the increasing marginal cost to filter on the right side.

In addition, the platform improves (7) by inducing citizens' estimates that are closer to their biases. Such incentive for the credulous citizens is immediate as these citizens' estimates are plainly their received signals. Thus, slanting for the credulous citizens is independent of the prior precision p of the state and the default precision q of the platform's signal absent filtering. In contrast, slanting for the rational citizens is a "rat-race" phenomenon. Each rational citizen i discounts her received signal y_i according to her expectation of the slant s_i^* to form her estimate. By Bayes' rule, her estimate is

$$\mathbf{E}^*\left[\theta|y_i\right]=\frac{q+f^*}{p+q+f^*}\left(y_i-s_i^*\right)+\frac{p}{p+q+f^*}\mathbf{E}^*[\theta]=\frac{q+f^*}{p+q+f^*}\left(y_i-s_i^*\right). \tag{8}$$

Thus, by slanting more aggressively for the rational citizens than what they expect in the direction of their biases, the platform pulls their estimates closer to their biases and improves (7). Given the increasing marginal cost to slant more aggressively, equations (5) and (6) pin down the equilibrium slant s_i^{S} for each citizen i.[9]

Notably, the filter is determined independently of the slants in the equilibrium. This is because a change in the filter affects only the dispersion of the citizens' signals, but does not affect how the slants influence the citizens' inferences: the rational citizens correctly remove the slants and the credulous citizens do not remove the slants at all.

In contrast, the slant for each rational citizen depends on the filter in the equilibrium. When the rational citizens expect a higher filter, the platform's deviation to more aggressive slants in the direction of their biases would be more effective in pulling the

[9] This "rat-race" phenomenon departs from existing models of media bias, and is reminiscent of career concerns à la Holmström (1999).

rational citizens' estimates towards their biases.

Finally, the slant for each credulous citizen is independent of the filter in the equilibrium, as the credulous citizens take the signals at face value irrespective of the filter. As a result, the platform might slant more or less aggressively for a rational citizen than for a credulous citizen with the same bias, depending on the exogenous parameters. We provide two numerical instances to illustrate this in Appendix A.1.

More generally, the tractable characterization in Proposition 1 facilitates comparative statics results concerning how the equilibrium filter and slants depend on the exogenous parameters. We relegate the comparative statics results to Appendix A.1, as they are not central to our main insight of perverse ethical concerns in Section 5. We next introduce our measure of offline conflicts and then turn to a platform that faces an ethical concern to mitigate conflicts.

4 Offline Conflicts and Ethical Concern

Given the rational citizens' expectation (f^*, s^*) and the realized signals y, the citizens' inferences of the state upon receiving their signals typically disagree. We measure such disagreement between citizens i and j by the distance between their estimates $\hat{\theta}_i(y_i)$ and $\hat{\theta}_j(y_j)$. To provide a concrete context, consider, for example, that a government is contemplating a policy that affects the citizens' welfare and the best policy for their welfare is the one that matches the hidden state θ. After receiving their signals, each citizen believes that the optimal policy is precisely her own estimate, and they disagree about the optimal policy.[10]

The citizens' disagreements lead to conflicts that are costly for a regulator who is not a player in the game in Section 2 but observes the interactions between the platform and the citizens. Given the rational citizens' expectation (f^*, s^*) and realized signals y, the regulator's (realized) conflict cost is measured by

$$\kappa(y; f^*, s^*) := \frac{1}{2} \int_0^1 \int_0^1 \left(\hat{\theta}_j(y_j) - \hat{\theta}_i(y_i)\right)^2 \mathrm{d}j \, \mathrm{d}i,$$

[10] The measure of disagreement as the distance in the citizens' expected value of the state is familiar in the literature (see, e.g., Kartik et al., 2021), and is natural in applications where the citizens disagree regarding the optimal government policy. In general, such notion of disagreement is limiting when one is interested in comparing the citizens' posterior distributions. Zanardo (2017) examines the notion of disagreement between probability distributions axiomatically.

where the scalar $1/2$ accounts for double-counting of the citizens' disagreements in the double integral.

We now consider an alternative version of the model in which the platform faces an ethical concern to internalize the regulator's conflict cost. The platform's payoff by choosing (f, s) given the rational citizens' expectation (f^*, s^*) is

$$\mathbf{E}\left[v\left(y; f^*, s^*\right) - h \cdot \kappa(y; f^*, s^*)\right] - \frac{cf^2}{2} - \int_0^1 \frac{ks_i^2}{2} \,\mathrm{d}i, \tag{9}$$

where $h > 0$ measures the strength of the platform's ethical concern and, as in (2), the expectation $\mathbf{E}$ is taken over signal profiles y with respect to the platform's actual choice of algorithms (f, s). The model is otherwise identical to the baseline version in Section 2. Thus, contrary to the baseline version of the model, the platform in this alternative version receives a lower payoff if it induces a higher conflict cost.

Proposition 2 below characterizes the essentially unique equilibrium in this alternative version of the model. To ease the exposition, we let $B_r := \int_r^1 b_i \,\mathrm{d}i$ denote the aggregate bias of the credulous citizens and for each citizen i, we let

$$\hat{b}_i := b_i + \frac{2hB_r}{k + 2\beta + 2hr}$$

denote her bias adjusted by the aggregate bias B_r of the credulous citizens.

Proposition 2. *There exists an essentially unique equilibrium. In the equilibrium, the platform chooses $(f, s) = (f^{\mathrm{E}}, s^{\mathrm{E}})$ where:*

1. *The filter f^{E} strictly exceeds f^{S} and is characterized by*

$$\frac{(1-r)\,(\beta + h)}{\left(q + f^{\mathrm{E}}\right)^2} + \frac{r\,(\beta + h)}{\left(p + q + f^{\mathrm{E}}\right)^2} = cf^{\mathrm{E}}. \tag{10}$$

2. *For each rational citizen i, the slant s_i^{E} is characterized by*

$$s_i^{\mathrm{E}} = \frac{2\beta}{k}\left(\frac{q + f^{\mathrm{E}}}{p + q + f^{\mathrm{E}}}\right)\hat{b}_i. \tag{11}$$

3. *For each credulous citizen i, the slant s_i^{E} is characterized by*

$$s_i^{\mathrm{E}} = \frac{2\beta}{2\beta + k + 2h}\hat{b}_i. \tag{12}$$

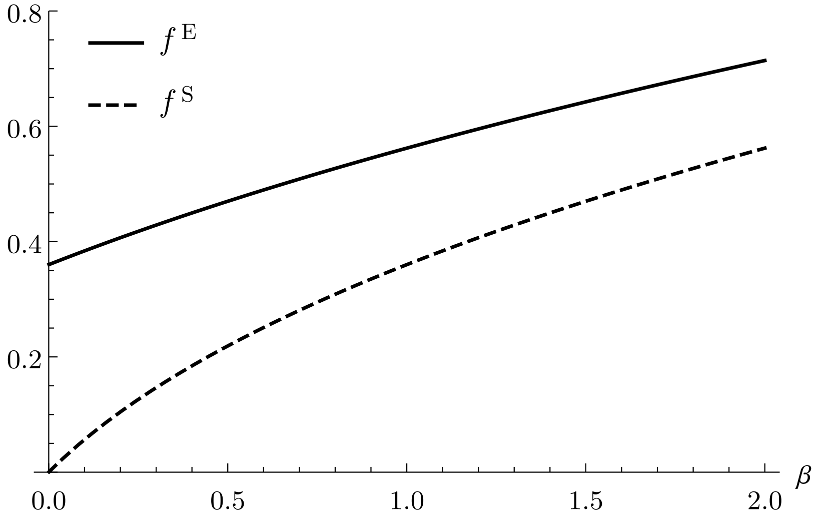
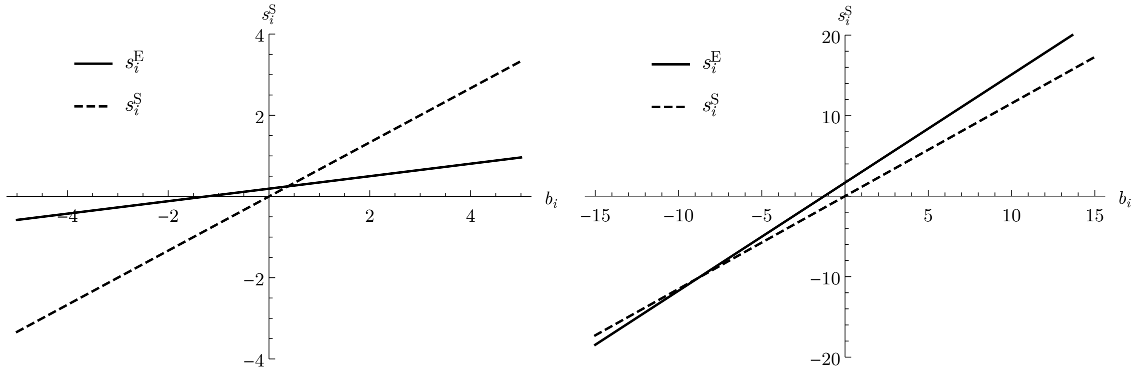


Figure 1: Filtering given $p = q = c = B_r = k = \beta = 1, r = \frac{1}{2}, h = 5$

As in the baseline model, the platform's actual choice of algorithms does not affect the rational citizens' inferences given their expectation of the algorithms and their realized signals. Unlike in the baseline model, the platform benefits by mitigating conflicts given its ethical concern. By filtering more aggressively, the platform reduces the dispersion of the citizens' estimates and mitigates conflicts. Thus, as depicted in Figure 1, the ethical concern causes the platform to filter more aggressively.

Notably, the ethical concern need not cause the platform to slant less aggressively. In the equilibrium, the slant for each citizen depends on the aggregate slant for the credulous citizens, which in turn depends on their aggregate bias B_r. On the one hand, slanting more aggressively for a credulous citizen towards her bias, or slanting more aggressively for a rational citizen than what she expects, unambiguously aggravates the citizen's disagreement with other rational citizens who correctly remove their slants in equilibrium. On the other hand, doing so could mitigate (resp., aggravate) her conflicts with other credulous citizens if the citizen is biased towards the same (resp., opposite) direction as the other credulous citizens are on aggregate.

To illustrate, given the platform's ethical concern, Figure 2a depicts its slanting for a credulous citizen and Figure 2b depicts the counterpart for a rational citizen. In both figures, the credulous citizens' aggregate bias is set to one. The platform's ethical concern leads to less (resp., more) aggressive slants for the credulous citizens with extreme (resp., small) biases. Since each credulous citizen takes her signals at face value, the platform mitigates the conflict that she faces by reducing the difference



(a) Credulous citizen i **(b)** Rational citizen i

Figure 2: Slanting given $p = q = c = B_r = k = \beta = 1, r = \frac{1}{2}, h = 5$

between her slant and the aggregate of the other credulous citizens' slants.

As for a positively biased rational citizen, deviating to slant more aggressively towards her bias mitigates her disagreement with the credulous citizens (of which the majority receive positive signals). The ethical concern thus drives the platform to slant more aggressively for such citizen. In contrast, for a negatively biased rational citizen, deviating to slant more aggressively towards her bias aggravates her conflicts with the credulous citizens. Thus, if such citizen's bias is not too negative, then the platform slants less aggressively because the benefit of mitigating her conflict with the credulous citizens dominates the benefit of catering to her bias. Otherwise, the platform slants more aggressively.

As in the case for the equilibrium absent any ethical concern, we relegate the comparative statics results to Appendix A.1. We next turn to examine the equilibrium structure of conflicts among the citizens.

5 Equilibrium Conflicts

In this section, we present our main result, elucidating how the platform's ethical concern affects the equilibrium conflict costs between and within the two groups of rational and credulous citizens. Specifically, in an equilibrium where the platform chooses algorithms (f, s), we analyze the structure of the following three objects. The first object is the equilibrium conflict cost among rational citizens:

$$K_R(f,s) := \mathbf{E}\left[\frac{1}{2}\int_0^r \int_0^r \left(\hat{\theta}_j(y_j) - \hat{\theta}_i(y_i)\right)^2 \mathrm{d}j\,\mathrm{d}i\right], \tag{13}$$

where the scalar 1/2, as in Section 4, accounts for double-counting of the citizens' disagreements in the double integral. Observe that in (13), we do not distinguish between the platform's actual choice of algorithms and the rational citizens' expectation of its algorithms, because the rational citizens' expectation is correct in equilibrium. The second object is the equilibrium conflict cost between any pair of rational and credulous citizens:

$$\begin{aligned} K_B(f,s) &:= \mathbf{E}\left[\frac{1}{2}\int_0^r \int_r^1 \left(\hat{\theta}_j(y_j) - \hat{\theta}_i(y_i)\right)^2 \mathrm{d}j\,\mathrm{d}i + \frac{1}{2}\int_r^1 \int_0^r \left(\hat{\theta}_j(y_j) - \hat{\theta}_i(y_i)\right)^2 \mathrm{d}j\,\mathrm{d}i\right] \\ &= \mathbf{E}\left[\int_0^r \int_r^1 \left(\hat{\theta}_j(y_j) - \hat{\theta}_i(y_i)\right)^2 \mathrm{d}j\,\mathrm{d}i\right]. \end{aligned} \tag{14}$$

The last object is the equilibrium conflict cost among the credulous citizens:

$$K_C(f,s) := \mathbf{E}\left[\frac{1}{2}\int_r^1 \int_r^1 \left(\hat{\theta}_j(y_j) - \hat{\theta}_i(y_i)\right)^2 \mathrm{d}j\,\mathrm{d}i\right]. \tag{15}$$

Proposition 3 below shows that the platform's ethical concern aggravates conflicts among rational citizens whenever the ethical concern is not strong enough, but it unambiguously mitigates conflicts involving credulous citizens.

Proposition 3. *The following holds.*

1. *There exists $\bar{h} \geq 0$ such that $K_R(f^{\mathrm{S}}, s^{\mathrm{S}}) < K_R(f^{\mathrm{E}}, s^{\mathrm{E}})$ if and only if $h < \bar{h}$.*
2. *$K_B(f^{\mathrm{S}}, s^{\mathrm{S}}) > K_B(f^{\mathrm{E}}, s^{\mathrm{E}})$.*
3. *$K_C(f^{\mathrm{S}}, s^{\mathrm{S}}) > K_C(f^{\mathrm{E}}, s^{\mathrm{E}})$.*

Thus, whenever the mass r of rational citizens is sufficiently large, the platform's ethical concern could perversely aggravate overall conflicts, as formalized in Corollary 1 below. Let

$$K(f,s) := K_R(f,s) + K_B(f,s) + K_C(f,s) \tag{16}$$

denote the (overall) conflict cost in an equilibrium where the algorithms are (f, s).

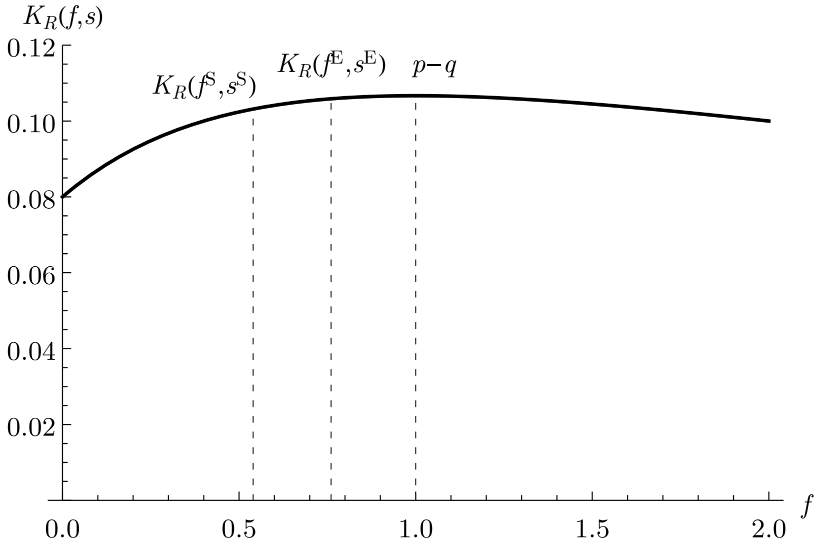


Figure 3: Rational conflicts given $\beta = c = k = h = 1, r = \frac{4}{5}, p = \frac{3}{2}, q = \frac{1}{2}$

Corollary 1. *There exists $\bar{r} \in (0,1)$ such that for every $r \in [\bar{r}, 1)$, there exists $h' \geq 0$ such that $K(f^{\mathrm{S}}, s^{\mathrm{S}}) < K(f^{\mathrm{E}}, s^{\mathrm{E}})$ whenever $h > h'$.*

Because the rational citizens correctly remove their slants in their equilibrium inferences, the conflict cost (13) among the rational citizens is constant in the equilibrium slant and is solely driven by misinformation. More aggressive equilibrium filtering driven by the ethical concern has two opposing effects for the rational citizens' inferences. On the one hand, there is a learning effect that mitigates conflicts: it improves the rational citizens' learning about the state. On the other hand, there is a confidence effect that aggravates conflicts: rational citizens correctly anticipate the more aggressive filtering and thus put more weights on their own signals in their inferences. Given the two opposing effects, the platform's ethical concern mitigates conflicts among the rational citizens if and only if the filter f^{E} given ethical concern is sufficiently large, or equivalently, if and only if the concern h is sufficiently large, so that the learning effect dominates the confidence effect.

On the other hand, the cost (14) due to conflicts between each pair of rational and credulous citizens is driven by both misinformation and the slants for the credulous citizens. The platform's ethical concern unambiguously reduces such cost. Here, when compared to conflicts among the rational citizens, the learning effect is stronger. The ethical concern not only leads to more aggressive filtering, but also causes the platform to adjust its slants to mitigate any conflict involving a credulous citizen who fails to remove her slant. Further, the confidence effect due to more aggressive filtering is absent in the credulous citizens' information acquisition, as they take the platform's signals at face value irrespective of whether the platform faces ethical concern. Likewise, the cost (15) driven by conflicts among the credulous citizens is also driven by the dispersion of misinformation in their signals as well as the distance between their slants, and the platform's ethical concern unambiguously reduces this cost.

In the remainder of this section, we analyze in more detail the structure of the threshold $\bar{h}$ in Proposition 3. The goal is to understand how strong the platform's ethical concern needs to be in order to preempt the perverse outcome among the rational citizens. Proposition 4 below shows that the threshold increases in p, and that the threshold is vacuous if and only if the prior precision p of the state is small.

Proposition 4. *There exists $\bar{p} > 0$ such that:*

1. *If $p > \bar{p}$, then $\bar{h} \equiv \bar{h}(p) > 0$ and is strictly increasing in p.*
2. *If $p \leq \bar{p}$, then $\bar{h} \equiv \bar{h}(p) = 0$.*

By (4) and (10) in Propositions 1 and 2, a higher p leads to smaller equilibrium filters f^{S} and f^{E}. When p is higher, the marginal improvement that the platform can ever contribute to the rational citizens' learning of the hidden state is small. As a result, the rational citizens put a smaller weight on their received signals relative to the prior mean when forming their estimates given a higher p, which makes it more costly for the platform to generate revenue by filtering misinformation in favor of the citizens' biases. The platform plays a limited role in the rational citizens' learning, and hence the perverse outcome hardly arises.

In turn, although the platform's ethical concern strengthens its filtering incentive, unless the concern is sufficiently strong, the resulting filter is still relatively small. As a result, among the two opposing effects that the ethical concern makes to the rational citizens' equilibrium conflicts, the confidence effect dominates the learning effect. Thus, the ethical concern aggravates the conflicts among the rational citizens. In contrast, when p is small, the learning effect is significant and always dominates the confidence effect, no matter how weak the ethical concern is.

6 Government Efforts

As noted in the introduction, governments worldwide adopt different efforts to provide incentives for platforms to act as if they face ethical concerns. In this section, we return to our baseline version of the model where the platform does not face any ethical concern and cast several popular government efforts in our model to analyze their effects on offline conflicts. Funke and Flamini (2021) provide a concise summary of the popular efforts adopted by different countries.

6.1 Supply-side Efforts

Propositions 5 and 6 below analyze several efforts that target the supply side of online information. We show that these efforts echo our insight concerning perverse ethical concerns in Section 5. Their implementation must be sufficiently aggressive in order to be effective.

Legislation against misinformation. We first consider legislation that holds the platform accountable for the misinformation that it displays to the citizens, ensuring a sufficient level of filtering by the platform. To capture such legislation, we consider a filtering floor $\underline{f} > f^{\mathrm{S}}$, where the filter f^{S} is the equilibrium filter characterized by (4) in the baseline model, such that the platform's filter must be at least $\underline{f}$. Given the floor, there is an essentially unique equilibrium $(f^{\mathrm{L}}, s^{\mathrm{L}})$ where the platform sets its filter to be precisely $f^{\mathrm{L}} = \underline{f}$, and the slant s^{L} is determined accordingly by (5) and (6) given the filter f^{L}.

Proposition 5. *The following holds.*

1. $K_B(f^{\mathrm{L}}, s^{\mathrm{L}}) < K_B(f^{\mathrm{S}}, s^{\mathrm{S}})$ *and* $K_C(f^{\mathrm{L}}, s^{\mathrm{L}}) < K_C(f^{\mathrm{S}}, s^{\mathrm{S}})$.
2. *If* $f^{\mathrm{S}} \geq p - q$, *then* $K_R(f^{\mathrm{L}}, s^{\mathrm{L}}) < K_R(f^{\mathrm{S}}, s^{\mathrm{S}})$.
3. *If* $f^{\mathrm{S}} < p - q$, *then there exists* $F > f^{\mathrm{S}}$ *such that* $K_R(f^{\mathrm{L}}, s^{\mathrm{L}}) < K_R(f^{\mathrm{S}}, s^{\mathrm{S}})$ *if and only if* $\underline{f} > F$.

Part 1 shows that the filtering floor unambiguously mitigates conflicts involving the credulous citizens. As discussed in Section 5, in equilibrium, such conflicts are driven by misinformation and the slants for the credulous citizens. The more aggressive filter due to the filtering floor reduces misinformation and does not affect the platform's incentives to slant for the credulous citizens, as is evident in (6), yielding the result. In contrast, part 2 and part 3 show that the more aggressive filter unambiguously mitigates conflicts among the rational citizens if and only if the filtering is sufficiently aggressive, in light of the confidence effect and the learning effect that oppose each other.

Proposition 5 also sheds light on policy discussions concerning the modification and elimination of platforms' immunity of Section 230 of the Communications Decency Act. Such immunity is commonly viewed as a "legal shield" that protects platforms from liability for third-party content that they host.[11] Plainly, modifying or eliminating the platforms' immunity introduces a cost that platforms incur due to their insufficient filtering of misinformation, and motivate platforms to filter misinformation more aggressively. Our results highlight that for such changes to the platforms' immunity to effectively mitigate conflicts, they must be implemented sufficiently aggressively.

Arrests and cyber task forces. We next consider arrests of misinformation spreaders and cyber task forces against misinformation campaigns. We cast such efforts in our model as an increase of the default precision absent filtering from an initial value q^{B} to some $q^{\mathrm{A}} > q^{\mathrm{B}}$, and denote the corresponding (essentially unique) equilibria as characterized in Proposition 1 by $(f^{\mathrm{B}}, s^{\mathrm{B}})$ and $(f^{\mathrm{A}}, s^{\mathrm{A}})$, respectively.

Proposition 6. *The following holds.*

1. $K_B(f^{\mathrm{A}}, s^{\mathrm{A}}) < K_B(f^{\mathrm{B}}, s^{\mathrm{B}})$ *and* $K_C(f^{\mathrm{A}}, s^{\mathrm{A}}) < K_C(f^{\mathrm{B}}, s^{\mathrm{B}})$.
2. *If* $f^{\mathrm{B}} \geq p - q^{\mathrm{B}}$, *then* $K_R(f^{\mathrm{A}}, s^{\mathrm{A}}) < K_R(f^{\mathrm{B}}, s^{\mathrm{B}})$.
3. *If* $f^{\mathrm{B}} < p - q^{\mathrm{B}}$, *then there exists* $Q > q^{\mathrm{B}}$ *such that* $K_R(f^{\mathrm{A}}, s^{\mathrm{A}}) < K_R(f^{\mathrm{B}}, s^{\mathrm{B}})$ *if and only if* $q^{\mathrm{A}} > Q$.

The intuition of Proposition 6 is analogous to that of Proposition 5, and so their statements share an analogous structure. Part 1 shows that these efforts unambiguously mitigate conflicts involving the credulous citizens; part 2 and part 3 show that these efforts unambiguously mitigate conflicts among the rational citizens if and only if they are sufficiently aggressive. While a higher default precision crowds out the platform's

[11] See, e.g., "Legal Shield for Social Media Is Targeted by Lawmakers," *The New York Times*, May 28, 2020.

filtering incentives in view of Proposition 1, the overall precision of the platform's signal, namely the sum of the default precision and the platform's filter, increases. The effect on the equilibrium conflicts given the efforts is thus identical that given a fixed default precision and a higher filter, which is the case in Proposition 5.

6.2 Demand-side Efforts: Media Literacy Campaigns

We next turn to media literacy campaigns that target the demand side of online information. Such campaigns aim to educate the credulous citizens to process online information "rationally." In this section, we argue that the effect of such campaigns on mitigating conflicts is ambiguous, because these campaigns disrupt the platform's incentives to filter misinformation despite improving the citizens' ability to process information.

Consider a successful media literacy campaign given which the credulous citizens become rational before the platform chooses its algorithms, and this event is common knowledge. To state our result, we introduce the following notations. For any two citizen i, j, we define

$$K_{ij}(f, s) := \mathbf{E}[(\hat{\theta}_i(y_i) - \hat{\theta}_j(y_j))^2] \tag{17}$$

as the equilibrium conflict cost between the two citizens given algorithms (f, s). As in (13)—(15), (17) does not distinguish between the actual algorithms chosen by the platform and the rational citizens' expectation of the algorithms, since their expectation is correct in equilibrium. Moreover, we denote by $(f^{\mathrm{M}}, s^{\mathrm{M}})$ the platform's (essentially unique) equilibrium algorithms upon a successful campaign, which are characterized by (4)—(6) and evaluated at $r = 1$.

Because the rational citizens correctly remove slants from their received signals but the credulous citizens do not, the conflict cost between any two rational citizens are smaller than that between any two citizens involving at least one credulous citizen absent the campaign. Thus, if the equilibrium filter remains fixed upon a campaign, then the campaign unambiguously mitigates conflicts. However, the campaign disrupts the platform's filtering incentive, as the platform finds it more difficult to influence the citizens' beliefs. Proposition 7 below makes this intuition precise.

Proposition 7. *Let $r \in (0, 1)$ denote the mass of rational citizens such that each* *citizen $i \in [0, r]$ is rational and each citizen $i \in (r, 1]$ is credulous before the campaign is implemented. The following holds.*

1. *For any citizens $i, j > r$, $K_{ij}(f^{\mathrm{M}}, s^{\mathrm{M}}) < K_{ij}(f^{\mathrm{S}}, s^{\mathrm{S}})$.*

2. *For any citizens $i \leq r$ and $j > r$, $K_{ij}(f^{\mathrm{M}}, s^{\mathrm{M}}) < K_{ij}(f^{\mathrm{S}}, s^{\mathrm{S}})$.*

3. *If $p - q > f^{\mathrm{M}} > 0$, then there exists $\bar{r} \in [0, 1)$ such that $K_{ij}(f^{\mathrm{M}}, s^{\mathrm{M}}) > K_{ij}(f^{\mathrm{S}}, s^{\mathrm{S}})$ for any citizens $i, j \leq r$ if and only if $r < \bar{r}$; otherwise, $K_{ij}(f^{\mathrm{M}}, s^{\mathrm{M}}) > K_{ij}(f^{\mathrm{S}}, s^{\mathrm{S}})$ for any citizens $i, j \leq r$.*

The first two parts of Proposition 7 show that the campaign unambiguously mitigates conflicts between any two citizens involving at least one citizen that was credulous before the campaign. This is because the credulous citizens learn to discount the signals when forming their state estimates upon the campaign. Part 3 of the proposition shows that the effect of the campaign on conflicts between citizens who were already rational before the campaign is ambiguous. A fall in the equilibrium filter upon the campaign reduces both the learning effect and the confidence effect. The net effect on their conflicts thus depends on the relative magnitudes of the fall.

6.3 Mixed Efforts

The above discussion points to an appeal of performing a mix of demand-side and supply-side efforts, which is indeed a common practice in many countries:

Corollary 2. *Given a filtering floor $\underline{f} \geq f^{\mathrm{S}}$, implementing a media literacy campaign unambiguously reduces aggregate conflict cost $K(f, s)$.*

Absent a campaign, the platform filters at the binding level $\underline{f}$; upon a campaign, the platform filters no less than $\underline{f}$ despite a disrupted incentive to filter, while all citizens discount their received signals.

6.4 Potential Regulation Efforts

We close this section by analyzing several potential regulation efforts that are commonly discussed in ongoing policy debates. We begin with platform transparency. Then, we turn to a version of the FCC fairness doctrine for online platforms.

Transparency. Suppose that the platform's algorithms are publicly observable. Then the platform anticipates that the rational citizens perform inferences based on its actual choice of algorithms (f, s). Thus, contrary to (9), the payoff of a platform with ethical concern is

$$\begin{aligned} &\mathbf{E}\left[v\left(y; f, s\right) - h \cdot \kappa(y; f, s)\right] - \frac{cf^2}{2} - \int_0^1 \frac{ks_i^2}{2}\,\mathrm{d}i \\ = &\mathbf{E}\left[v\left(y; f, s\right)\right] - \frac{cf^2}{2} - \int_0^1 \frac{ks_i^2}{2}\,\mathrm{d}i - h \cdot K(f, s) \end{aligned} \tag{18}$$

where the expectation is taken with respect to the platform's choice of algorithms (f, s). In view of (18), transparency allows the platform to correctly internalize its "social responsibility" and therefore, no perverse outcome arises:

Proposition 8. *Suppose that the platform's algorithms are observable to the citizens. Then given any equilibrium* $(\tilde{f}^{\mathrm{S}}, \tilde{s}^{\mathrm{S}})$ *in the model absent ethical concern and any equilibrium* $(\tilde{f}^{\mathrm{E}}, \tilde{s}^{\mathrm{E}})$ *in the model given ethical concern,* $K(\tilde{f}^{\mathrm{S}}, \tilde{s}^{\mathrm{S}}) \geq K(\tilde{f}^{\mathrm{E}}, \tilde{s}^{\mathrm{E}})$.

In practice, calls for transparency are primarily motivated by the conventional wisdom that transparency is essential to accountability measures for platforms and consumer protection (see, e.g., MacCarthy, 2020). Proposition 8 offers an alternative case for transparency by highlighting its role to complement platforms' ethical concerns.

Fairness doctrine. We next discuss a version of the FCC fairness doctrine for online media. The doctrine was originally applied to radio and television broadcasters, requiring that the broadcasters provide a fair and balanced presentation of information.[12] To cast the doctrine in our model, suppose that the platform is not allowed to slant. Let s^{F} denote a slant function such that for all citizens $i \in [0, 1]$, $s_i^{\mathrm{F}} = 0$. Since the platform's filter is determined independently of its slants in equilibrium, absent any ethical concern, the equilibrium filter remains as f^{S} given the doctrine. In addition, the platform's slants affect only the conflicts between any pair of citizens involving at least one credulous citizen in equilibrium. Thus, Proposition 9 below follows, showing that the doctrine unambiguously mitigates conflicts involving credulous citizens but does not affect conflicts among rational citizens.

[12] See "Lessons for Social Media from the Fairness Doctrine," *Columbia Journalism Review*, August 13, 2020.

Proposition 9. *It holds that* $K_R(f^{\mathrm{S}}, s^{\mathrm{F}}) = K_R(f^{\mathrm{S}}, s^{\mathrm{S}})$*,* $K_B(f^{\mathrm{S}}, s^{\mathrm{F}}) < K_B(f^{\mathrm{S}}, s^{\mathrm{S}})$ *and* $K_C(f^{\mathrm{S}}, s^{\mathrm{F}}) < K_C(f^{\mathrm{S}}, s^{\mathrm{S}})$*.*

It is worth mentioning that some media scholars (see, e.g., Napoli, 2019) urge for a version of the doctrine for online media, although the FCC eliminated the doctrine for broadcasters in 1987. The core justification of the elimination was that the doctrine was no longer necessary, as the growing number of media outlets available facilitates consumers' access to diverse information. Our analysis highlights that such justification is limiting in the context of online media. While consumers' access to diverse information is also a defining feature of online media, platforms possess the power to slant news and personalize the slanting for each citizen. These citizens include those who are credulous and hence lack the sophistication to process information. As Proposition 9 highlights, slanting for the credulous citizens flames the offline conflicts.

7 Conclusion

Public concerns that online platforms flame offline conflicts are paramount, so are societal efforts to address them. These efforts either directly introduce ethical concerns to internalize the conflict costs to the platforms or indirectly motivate the platforms to act as if they face ethical concerns. *Prima facie*, these efforts appear unambiguously effective for mitigating conflicts.

In this paper, we have offered a cautionary observation. We have elucidated the strategic implications of platforms' ethical concerns, highlighting their potential perverse consequences. In particular, we have shown that rational citizens who anticipate platforms' aggressive filtering of misinformation might become "too confident" of the individual online contents that they read and in turn become more hostile to others' disagreeing opinions. A critical message that our results put forward is that for the societal efforts to mitigate conflicts to be effective, their implementations must be sufficiently aggressive.

Appendices

A Omitted Details

A.1 Comparative Statics

In this section, we report results concerning comparative statics in the baseline version of the model where the platform does not face any ethical concern, elucidating how the equilibrium filter and the equilibrium slant function depend on the exogenous parameters. Corollary 3 below immediately follows from (4)—(6) in Proposition 1.

Corollary 3. *Consider the baseline version of the model where the platform does not have any ethical concern. In the equilibrium* $(f^{\mathrm{S}}, s^{\mathrm{S}})$*, the filter* f^{S} *is strictly increasing in* β *and is strictly decreasing in* (r, c, p, q)*. For each rational citizen* i*, the slant* s_i^{S} *is proportional to* b_i*; its magnitude is strictly increasing in* (β, q) *and is strictly decreasing in* (k, p, r)*. Finally, for each credulous citizen* i*, the slant* s_i^{S} *is proportional to* b_i*; its magnitude is strictly increasing in* β *and is strictly decreasing in* k*.*

To illustrate, Figure 4 depicts several numerical instances of the filter f^{S}. As noted in Section 3 in the main text, the filter f^{S} is strictly increasing in the benefit β to cater to the citizens' biases and vanishes as β vanishes. Further, as noted briefly in Section 4, prior consensus about the state crowds out filtering incentives, in the sense that the filter f^{S} is strictly decreasing in the prior state precision p. Next, there are diminishing returns to filtering. Given a higher default precision q of the platform's signals, the platform chooses a smaller filter f^{S}, since additional filtering by the platform has a smaller impact on the rational citizens' estimates (8). Finally, given a higher mass r of rational citizens, the platform filters less aggressively as filtering is less effective in affecting the rational citizens' estimates than the credulous citizens' estimates.

Figure 5 reports several numerical instances of the slants s_i^{S}. As depicted in Figure 5a, given a higher default precision q, the platform slants more aggressively for a rational citizen towards her bias. Given a higher q, the precision of the platform's signal $q + f^{\mathrm{S}}$ is higher, despite the filter f^{S} being smaller. By (8), the citizen puts more weight on the signal in her inference. In turn, the platform's slanting to cater to her bias is more effective. On the other hand, Figure 5b shows that given a higher precision p, the platform slants less aggressively for a rational citizen, as the citizen

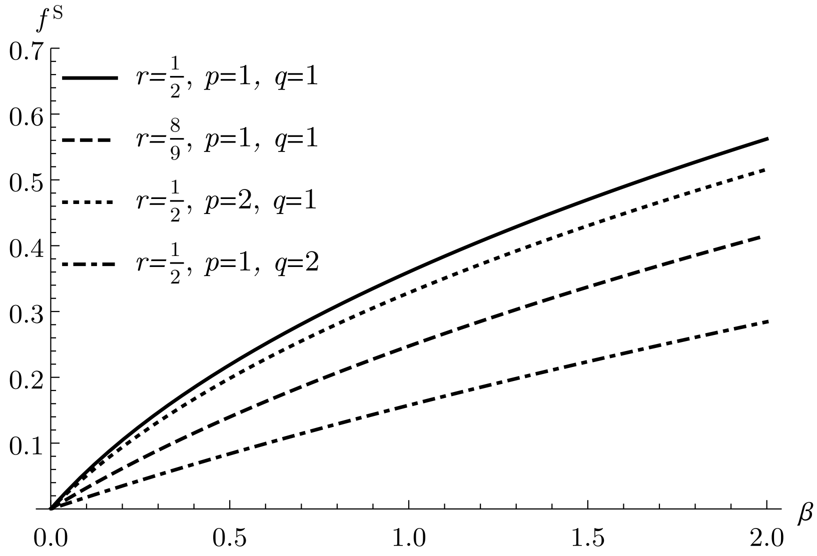
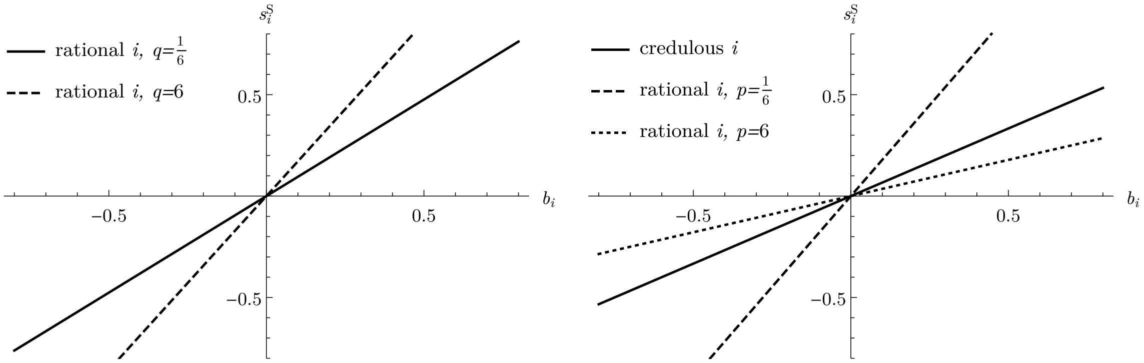


Figure 4: Equilibrium filter characterized by (4), given that $c = 1$



(a) $p = 1$

(b) $q = 1$

Figure 5: Equilibrium slanting given $c = k = 1$ and $r = \frac{1}{2}$

then puts less weight on the platform's signals in their inferences by (8). In turn, the platform's slanting to cater to her bias is less effective.

Notably, as briefly mentioned in Section 3, the platform might slant more or less aggressively for a rational citizen than for a credulous citizen with the same bias. When the precision p is relatively small (resp., large), a rational citizen expects a higher (resp., small) filter f^{S} and the platform's deviation to slant more aggressively than what she expects is more (resp., less) effective in pulling her estimates towards her bias. In contrast, the platform's incentive to cater to a credulous citizen's bias is independent of the precision p. When p is small (resp., large), the platform slants more (resp., less) aggressively for a rational citizen than for a credulous citizen with the same bias.

We next turn to the comparative statics in the alternative version of the model where the platform faces an ethical concern. Corollary 4 below immediately follows from (4)—(6) in Proposition 2.

Corollary 4. *Consider the version of the model where the platform faces ethical concern. In the equilibrium $(f^{\mathrm{E}}, s^{\mathrm{E}})$, the filter f^{E} is strictly increasing in (β, h) and is strictly decreasing in (r, c, p, q). For each rational user i, the slant s_i^{E} is proportional to the adjusted bias $\hat{b}_i$; its magnitude is strictly increasing in q and is strictly decreasing in p. The comparative statics of its magnitude with respect to (β, k, r, h) is ambiguous and depends on the value of B_r. For each credulous user i, the slant s_i^{E} is proportional to the adjusted bias $\hat{b}_i$. The comparative statics of its magnitude with respect to (β, k, r, h) is ambiguous and depends on the value of B_r.*

The comparative statics concerning the filter f^{E} is analogous to that concerning the filter f^{S} except that the filter f^{E} depends on the ethical concern h and is strictly increasing in h. Given a higher h, the platform derives a higher benefit by filtering more aggressively to reduce the dispersion of citizens' signals. Likewise, the comparative statics concerning the slant s_i^{E} for a rational citizen with respect to p and q is identical to those given a self-interested platform. Finally, the ambiguity of the comparative statics concerning the slant s^{E} with respect to the parameters (β, k, r, h) follows from our discussion in Section 4 that the platform with ethical concern might slant more or less aggressively than a self-interested platform does, depending on the the aggregate slant for the credulous citizens. The aggregate slant for the credulous citizens is determined by the parameters (β, k, r, h) and their aggregate bias B_r.

A.2 Extension: Heterogeneous Beliefs

Suppose that each citizen i can access additional a private signal x_i before they receive personalized signals y_i from the platform, where

$$x_i|\theta \sim_{i.i.d.} N\left(\theta, \frac{1}{z}\right) \quad z > 0.$$

For simplicity, suppose in this section that all citizens are rational (i.e., $r = 1$). Let *interim stage* refer to the moment after each citizen i receives x_i but does not receive

y_i yet. After observing x_i in the interim stage, each citizen i believes

$$\theta \sim N\left(\frac{z}{p+z}x_i, \frac{1}{p+z}\right), \tag{19}$$

Let $p^\dagger := p + z$ stand for the common precision of the citizens' beliefs in the interim stage. On the other hand, define $A^\dagger := \frac{z}{p+z}$ as the weight that each citizen i puts to x_i in (19). Then, the aggregate conflict cost in the interim stage is given by

$$\frac{1}{2}\int_0^1\int_0^1 \mathbf{E}\left[\left(\frac{z}{z+p}\right)^2 (x_i - x_j)^2\right] \mathrm{d}i\, \mathrm{d}j = \frac{z}{(z+p)^2} = \frac{A^\dagger}{p^\dagger}.$$

The proof of the next proposition is analogous to that of Proposition 1 and 2 and hence is omitted.

Proposition A.1. *Suppose that $r = 1$ and the citizens have heterogeneous beliefs* (19) *before they receive personalized signals from the platform.*

1. *Suppose that the platform is self-interested. In the essentially unique equilibrium, the platform chooses $(f, s) = (f^{\mathrm{S}}, s^{\mathrm{S}})$ such that*

$$\frac{\beta}{(p^\dagger + q + f^{\mathrm{S}})^2} = cf^{\mathrm{S}} \quad \textit{and} \quad s_i^{\mathrm{S}} = \frac{2\beta}{k}\left(\frac{q + f^{\mathrm{S}}}{p^\dagger + q + f^{\mathrm{S}}}\right) b_i \quad \forall i \in [0,1].$$

2. *Suppose that the platform is ethical. In the essentially unique equilibrium, the platform chooses $(f, s) = (f^{\mathrm{E}}, s^{\mathrm{E}})$ such that*

$$\frac{\beta + h}{(p^\dagger + q + f^{\mathrm{E}})^2} = cf^{\mathrm{E}} \quad \textit{and} \quad s_i^{\mathrm{E}} = \frac{2\beta}{k}\left(\frac{q + f^{\mathrm{E}}}{p^\dagger + q + f^{\mathrm{E}}}\right) b_i \quad \forall i \in [0,1].$$

3. *$f^{\mathrm{S}} < f^{\mathrm{E}}$, and both f^{S} and f^{E} strictly decrease in $p^\dagger$.*

In either case $(f, s) = (f^E, f^E)$ or $(f, s) = (f^S, s^S)$, each citizen i's estimate of the hidden state is given by

$$\hat{\theta}_i(x_i, y_i) := \mathbf{E}[\theta|x_i, y_i] = \frac{p^\dagger A^\dagger}{p^\dagger + q + f}x_i + \frac{q + f}{p^\dagger + q + f}(y_i - s_i), \tag{20}$$

hence, the equilibrium aggregate conflict cost is

$$\begin{aligned} K(f,s) &= \frac{1}{2}\int_0^1\int_0^1 \mathbf{E}\left[\left(\frac{p^\dagger}{p^\dagger+q+f}A^\dagger(x_i-x_j)+\frac{q+f}{p^\dagger+q+f}(\varepsilon_i-\varepsilon_j)\right)^2\right]\mathrm{d}i\,\mathrm{d}j \\ &= \frac{p^\dagger}{(p^\dagger+q+f)^2}A^\dagger+\frac{q+f}{(p^\dagger+q+f)^2}. \end{aligned}$$

In addition, $K(f,s)$ is independent of s, and

$$\frac{\partial K(f,s)}{\partial f}>0 \quad\iff\quad f<p^\dagger(1-2A^\dagger)-q \tag{21}$$

If $A^\dagger \geq 1/2$, $\partial K(f,s)/\partial f \leq 0$ at all $f>0$, hence an ethical platform generates *less* conflict cost. Recall that the perverse outcome in Corollary 1 is driven by the confidence effect: the citizens put more weight on the platform's signals as the filter increases from f^{S} to f^{E}, thereby yielding more dispersion in their estimates of θ. The citizens' private signals are not susceptible to such a perverse confidence effect because, as is evident in (20), the weight on the private signals decreases in the filter. A larger filter leads the citizens to put more weight on their private signals more than the platform's signals hence prevents the perverse outcome.

On the other hand, an ethical platform may perversely generate more conflict cost if $A^\dagger < 1/2$, provided that they already have precise information before observing the platform's signals. Formally, both f^{E} and f^{S} become smaller than the threshold $p^\dagger(1-2A^\dagger)-q>0$ in (21) when $A^\dagger<1/2$ and $p^\dagger$ is sufficiently large, so that the conflict cost $K(f^{\mathrm{E}},s^{\mathrm{E}})>K(f^{\mathrm{S}},s^{\mathrm{S}})$.

The next proposition summarizes the discussion so far.

Proposition A.2. *Suppose that $r=1$ and the citizens have heterogeneous beliefs (19) before they receive personalized signals from the platform.*

1. *$K(f^{\mathrm{E}},s^{\mathrm{E}})<K(f^{\mathrm{S}},s^{\mathrm{S}})$ whenever $A^\dagger \geq 1/2$.*

2. *If $A^\dagger<1/2$, there is $\bar{p}^\dagger>0$ such that $K(f^{\mathrm{E}},s^{\mathrm{E}})>K(f^{\mathrm{S}},s^{\mathrm{S}})$ whenever $p^\dagger>\bar{p}^\dagger$.*

B Proofs

B.1 Proof of Proposition 1

Throughout the appendix, let

$$A(f) := \frac{q+f}{p+q+f} \quad \forall f > 0.$$

denote the weight that the rational citizens put on their personalized signals in inference of the hidden state; see (8). By choosing (f, s) given citizens' expectation (f^*, s^*), the platform's expected revenue is

$$\mathbf{E}\left[v(y; f^*, s^*)\right] = \mathbf{E}\Big[-\int_0^r \beta\left(\mathbf{E}^*\left[\theta|y_i\right] - b_i\right)^2 + \tau \mathbf{Var}^*\left[\theta|y_i\right] \mathrm{d}i - \int_r^1 \beta\left(y_i - b\right)^2 \mathrm{d}i\Big]. \quad (22)$$

By direct calculations,

$$\mathbf{E}\left[(\mathbf{E}^*[\theta|y_i]-b_i)^2\right] = \frac{A(f^*)^2}{pA(f)} + \left(A(f^*)(s_i - s_i^*) - b_i\right)^2 \quad (23)$$

$$\mathbf{E}\left[(y_i - b_i)^2\right] = \frac{1}{pA(f)} + (s_i - b_i)^2 \quad (24)$$

$$\mathbf{Var}^*[\theta|y_i] = \frac{1}{p+q+f^*}. \quad (25)$$

Substituting (23)-(25) into the expected revenue (22), the platform's payoff given its algorithm (f, s) and rational citizens' expectations (f^*, s^*) is

$$\begin{aligned} &\mathbf{E}\Big[v(y; f^*, s^*)\Big] - \frac{c}{2}f^2 - \frac{k}{2}\int_0^1 s_i^2 \,\mathrm{d}i \\ &= -\beta\left[\frac{1 - r + rA(f^*)^2}{pA(f)} + \int_0^r \left(A(f^*)(s_i - s_i^*) - b_i\right)^2 \mathrm{d}i + \int_r^1 (s_i - b_i)^2 \,\mathrm{d}i\right] \quad (26) \\ &\quad - \tau\frac{r}{p+q+f^*} - \frac{c}{2}f^2 - \frac{k}{2}\int_0^1 s_i^2 \,\mathrm{d}i. \end{aligned}$$

The first order conditions with respect to f and s_i yield (4)-(6).

B.2 Proof of Proposition 2

By direct calculations, the expected conflict costs are

$$\begin{aligned}
\mathbf{E}\left[(y_j - y_i)^2\right] &= (s_i - s_j)^2 + \frac{2}{q+f} \\
\mathbf{E}\left[(\mathbf{E}^*[\theta|y_j] - \mathbf{E}^*[\theta|y_i])^2\right] &= A(f^*)^2\left[\frac{2}{q+f} + \left(s_i - s_j - s_i^* + s_j^*\right)^2\right]
\end{aligned}$$

and

$$\mathbf{E}\left[(y_j - \mathbf{E}^*[\theta|y_i])^2\right] = \left(s_j - A(f^*)\left(s_i - s_i^*\right)\right)^2 + \frac{1 + A(f^*)^2}{q+f} + \frac{p}{(p+q+f^*)^2}$$

where

$$A(f^*) = \frac{q+f^*}{p+q+f^*}$$

as defined in the proof of Proposition 1. Hence, the expected conflict cost that each rational user $i \in [0, r]$ suffers from is

$$\begin{aligned}
K_i(f, s; f^*, s^*) &= \frac{1}{2}\mathbf{E}\left[\int_0^r \left(\mathbf{E}^*\left[\theta|y_j\right] - \mathbf{E}^*\left[\theta|y_i\right]\right)^2 \mathrm{d}j + \int_r^1 \left(y_j - \mathbf{E}^*\left[\theta|y_i\right]\right)^2 \mathrm{d}j\right] \\
&= \frac{1}{2}\left[\int_0^r A(f^*)^2(s_i - s_j - s_i^* + s_j^*)^2\,\mathrm{d}j + \int_r^1 \left(s_j - A(f^*)\left(s_i - s_i^*\right)\right)^2 \mathrm{d}j\right] \\
&\quad + \frac{1}{2}\left[\frac{(1+r)A(f^*)^2 + 1 - r}{q+f} + \frac{p(1-r)}{(p+q+f^*)^2}\right]. \qquad (27)
\end{aligned}$$

Similarly, each credulous user $i \in (r, 1]$ suffers from the exepcted conflict cost

$$\begin{aligned}
K_i(f, s; f^*, s^*) &= \frac{1}{2}\mathbf{E}\left[\int_0^r \left(\mathbf{E}^*\left[\theta|y_j\right] - y_i\right)^2 \mathrm{d}j + \int_r^1 \left(y_j - y_i\right)^2 \mathrm{d}j\right] \\
&= \frac{1}{2}\left[\int_0^r (s_i - A(f^*)(s_j - s_j^*))^2\,\mathrm{d}j + \int_r^1 (s_i - s_j)^2\,\mathrm{d}j\right] \\
&\quad + \frac{1}{2}\left[\frac{rA(f^*)^2 + 2 - r}{q+f} + \frac{rp}{(p+q+f^*)^2}\right]. \qquad (28)
\end{aligned}$$

The ethical platform's expected payoff is

$$V^E(f, s; f^*, s^*) := \mathbf{E}\left[v(y; f^*, s^*)\right] - \frac{cf^2}{2} - \int_0^1 \frac{ks_i^2}{2}\,\mathrm{d}i - h\int_0^1 K_i\left(f, s; f^*, s^*\right)\mathrm{d}i,$$

where the exact form of $\mathbf{E}\left[v(y; f^*, s^*)\right]$ is given in (26) in the proof of Proposition 1.

The characterization of f^{E} in (10) directly follows the first-order condition with respect to f. To solve for the ethical platform's equilibrium slants, define for any smooth $\eta : [0,1] \to \mathbb{R}$, integrable $s : [0,1] \to \mathbb{R}$, and $\varepsilon \in \mathbb{R}$

$$Z_\eta(\varepsilon|s) := V^E(f^*, s + \varepsilon\eta; f^*, s^*).$$

A necessary condition for the optimality of s^* given citizens' expectation (f^*, s^*) is

$$Z_\eta(\varepsilon|s^*) = \frac{\mathrm{d}}{\mathrm{d}\varepsilon} V^E(f^*, s^* + \varepsilon\eta; f^*, s^*)\bigg|_{\varepsilon=0} = 0$$

for any smooth η. Substituting (26), (27), and (28) into this necessary condition and then rearranging and simplifying terms, we obtain

$$\int_0^r A(f^*)\left[h\int_r^1 s_j^*\,\mathrm{d}j + \beta b_i\right]\eta_i\,\mathrm{d}i + \int_r^1\left[h\int_r^1 s_j^*\,\mathrm{d}j - (\beta+h)s_i^* + \beta b_i\right]\eta_i\,\mathrm{d}i = \int_0^1 \frac{ks_i^*}{2}\eta_i\,\mathrm{d}i.$$

By the fundamental lemma of calculus of variations, the above equation holds for any smooth η if and only if

$$A(f^*)\left(\beta b_i + \int_r^1 hs_j^*\,\mathrm{d}j\right) = \frac{ks_i^*}{2}, \qquad \text{a.e. } i \in [0, r], \tag{29}$$

$$\beta b_i + \int_r^1 hs_j^*\,\mathrm{d}j = \frac{ks_i^*}{2} + s_i^*\left(\beta + h\right), \qquad \text{a.e. } i \in (r, 1]. \tag{30}$$

Rearranging and simplifying (30), we have, for almost every $i \in (r, 1]$,

$$s_i^* = \frac{2\beta b_i}{k + 2\beta + 2h} + \frac{2h}{k + 2\beta + 2h}\int_r^1 s_j^*\,\mathrm{d}j. \tag{31}$$

Integrating both sides of (31) with respect to i over $(r, 1]$ gives

$$\int_r^1 s_i^*\,\mathrm{d}i = \int_r^1 \frac{2\beta b_i}{k + 2\beta + 2h}\,\mathrm{d}i + \frac{2h\,(1-r)}{k + 2\beta + 2h}\left(\int_r^1 s_j^*\,\mathrm{d}j\right).$$

Rearranging gives

$$\int_r^1 s_j^* \,\mathrm{d}j = \frac{2\beta}{k+2\beta+2hr}\int_r^1 b_j \,\mathrm{d}j.$$

Substituting this into (29) and (31), we obtain (11) and (12) as desired. Finally, to check the sufficiency of the conditions (29) and (30), note that the second derivative $\partial^2 Z_\eta(\varepsilon|s)/\partial\varepsilon^2$ is negative for any $\varepsilon \in \mathbb{R}$ and any integrable $\eta : [0,1] \to \mathbb{R}$ given the quadratic structure of $V^E(f,s;f^*,s^*)$. Thus, any non negligible perturbation from s^{E} must yield a strictly smaller payoff for the platform. This completes the proof.

B.3 Proof of Proposition 3

Preliminary observations. Recall from Proposition 2 that f^{E} are continuous and increasing in h. In this proof, we will often denote f^{E} by $f^{\mathrm{E}}(h)$ to emphasize their dependence on h. Note that $f^{\mathrm{E}}(0) = f^{\mathrm{S}}$. Also, recall from the proof of Proposition 2 that

$$\mathbf{E}\left[(y_j - y_i)^2\right] = (s_i^* - s_j^*)^2 + \frac{2}{q+f^*} \tag{32}$$

$$\mathbf{E}\left[(\mathbf{E}^*[\theta|y_j] - \mathbf{E}^*[\theta|y_i])^2\right] = \frac{2(q+f^*)}{(p+q+f^*)^2} \tag{33}$$

$$\mathbf{E}\left[(y_j - \mathbf{E}^*[\theta|y_i])^2\right] = (s_j^*)^2 + \frac{1}{q+f^*} + \frac{1}{p+q+f^*} \tag{34}$$

where (f^*, s^*) denotes $(f^{\mathrm{S}}, s^{\mathrm{S}})$ or $(f^{\mathrm{E}}, s^{\mathrm{E}})$ depending on whether the platform is self-interested or ethical.

Proof of statement 1. Note that the conflict cost among rational citizens

$$K_R(f,s) = \mathbf{E}\left[\frac{1}{2}\int_0^r\int_0^r \left(\hat\theta_j(y_j) - \hat\theta_i(y_i)\right)^2 \mathrm{d}j\,\mathrm{d}i\right] = \frac{r^2(q+f)}{(p+q+f)^2}$$

is independent of s. Moreover, for any fixed p, q, and s, the mapping $f \mapsto K_R(f,s)$ is single-peaked at $f = p-q$.

First, consider the case $p - q \le f^{\mathrm{S}}$. In this case, $K_R(f^{\mathrm{E}}, s^{\mathrm{E}}) \ge K_R(f^{\mathrm{S}}, s^{\mathrm{S}})$ for any $h \ge 0$, hence the statement holds with $\bar h = 0$. Next, consider the case $f^{\mathrm{S}} < p - q$. Let

$h^* > 0$ denote the level of h such that $f^{\mathrm{S}} < f^{\mathrm{E}}(h^*) = p - q$. Also, define

$$\Delta K_R(h) := \frac{r^2(q + f^{\mathrm{E}}(h))}{(p + q + f^{\mathrm{E}}(h))^2} - \frac{r^2(q + f^{\mathrm{S}})}{(p + q + f^{\mathrm{S}})^2}$$

by the difference between $K_R(f^{\mathrm{E}}, s^{\mathrm{E}})$ and $K_R(f^{\mathrm{S}}, s^{\mathrm{S}})$. Then, $\Delta K_R(h)$ increases over $[0, h^*]$ and then decreases over (h^*, ∞). Also, $\Delta K_R(0) = 0$ and $\lim_{h\to\infty} \Delta K_R(h) < 0$ because $\lim_{h\to\infty} f^{\mathrm{E}}(h) = \infty$. Hence, there is $\bar{h} > 0$ such $\Delta K_R(h) \geq 0$ if and only if $h \in [0, \bar{h}]$.

Proof of statement 2. The conflict cost between any pair $(i, j) \in [0, r] \times (r, 1]$, as given in (33) in the proof of Proposition 3, decreases in f^*. Hence, it suffices to show

$$\int_r^1 \left(s_i^{\mathrm{E}}\right)^2 \mathrm{d}i \leq \int_r^1 \left(s_i^{\mathrm{S}}\right)^2 \mathrm{d}i.$$

From Propositions 1 and 2,

$$\int_r^1 (s_i^{\mathrm{S}})^2 \,\mathrm{d}i = \left(\frac{2\beta}{2\beta + k}\right)^2 \int_r^1 b_i^2 \,\mathrm{d}i$$

and

$$\begin{aligned}\int_r^1 (s_i^{\mathrm{E}})^2 \,\mathrm{d}i &= \left(\frac{2\beta}{2\beta + 2h + k}\right)^2 \left[\int_r^1 b_i^2 \,\mathrm{d}i + \Gamma(r) \left(\int_r^1 b_i \,\mathrm{d}i\right)^2\right] \\ &\leq \left(\frac{2\beta}{2\beta + 2h + k}\right)^2 \left[\int_r^1 b_i^2 \,\mathrm{d}i + \Gamma(0) \int_r^1 b_i^2 \,\mathrm{d}i\right],\end{aligned}$$

where the last inequality follows Jensen's inequality and

$$\Gamma(r) := (1 - r) \left(\frac{2h}{2\beta + k + 2hr}\right)^2 + \frac{4h}{2\beta + k + 2hr} \leq \left(\frac{2h}{2\beta + k}\right)^2 + \frac{4h}{2\beta + k} = \Gamma(0).$$

Hence,

$$\int_r^1 (s_i^{\mathrm{E}})^2 \,\mathrm{d}i - \int_r^1 (s_i^{\mathrm{S}})^2 \,\mathrm{d}i \leq \underbrace{\left[(1 + \Gamma(0)) \left(\frac{2\beta}{2\beta + 2h + k}\right)^2 - \left(\frac{2\beta}{2\beta + k}\right)^2\right]}_{=0} \int_r^1 b_i^2 \,\mathrm{d}i = 0.$$

Proof of statement 3. The expected conflict cost between any two credulous citizens, as given in (32), strictly decreases in f^*. Hence, it suffices to show

$$\int_r^1 \int_r^1 \left(s_i^{\mathrm{E}} - s_j^{\mathrm{E}}\right)^2 \mathrm{d}i\, \mathrm{d}j \leq \int_r^1 \int_r^1 \left(s_i^{\mathrm{S}} - s_j^{\mathrm{S}}\right)^2 \mathrm{d}i\, \mathrm{d}j.$$

Indeed, from Propositions 1 and 2,

$$\begin{aligned} &\int_r^1 \int_r^1 \left(s_i^{\mathrm{S}} - s_j^{\mathrm{S}}\right)^2 \mathrm{d}i\, \mathrm{d}j = \left(\frac{2\beta}{2\beta + k}\right)^2 \int_r^1 \int_r^1 (b_i - b_j)^2 \,\mathrm{d}i\, \mathrm{d}j \\ \leq &\int_r^1 \int_r^1 \left(s_i^{\mathrm{E}} - s_j^{\mathrm{E}}\right)^2 \mathrm{d}i\, \mathrm{d}j = \left(\frac{2\beta}{2\beta + 2h + k}\right)^2 \int_r^1 \int_r^1 (b_i - b_j)^2 \,\mathrm{d}i\, \mathrm{d}j. \end{aligned}$$

B.4 Proof of Proposition 4

Recall from Propositions 1 and 2 that (i) $f^{\mathrm{S}} < f^{\mathrm{E}}$, (ii) both f^{S} and f^{E} are continuous and decreasing in p, and (iii) f^{E} is increasing in h. In this proof, we will denote f^{S} and f^{E} by $f^{\mathrm{S}}(p)$ and $f^{\mathrm{E}}(h|p)$ respectively to emphasize their dependence on p and h. For any p and h, define

$$\Delta K_R(h|p) := \frac{r^2(q + f^{\mathrm{E}}(h|p))}{(p + q + f^{\mathrm{E}}(h|p))^2} - \frac{r^2(q + f^{\mathrm{S}}(p))}{(p + q + f^{\mathrm{S}}(p))^2}$$

by the difference between $K_R(f^{\mathrm{E}}, s^{\mathrm{E}})$ and $K_R(f^{\mathrm{S}}, s^{\mathrm{S}})$ (see the proof of the first statement of Proposition 3 for the basic properties of K_R). The monotonicity of $f^{\mathrm{S}}(p)$ with respect to p gaurantees

$$\exists \bar{p} > 0 \quad \text{such that} \quad f^{\mathrm{S}}(p) < p - q \quad \text{iff} \quad p > \bar{p}.$$

Here, the significance of the threshold $p - q$ lies in the fact that the mapping $f \mapsto K_R(f, s)$ is single-peaked at $f = p - q$ for any given p, q, and s. We first prove statement 2, and then turn to statement 1.

Proof of statement 2. First, consider the case such that $p \leq \bar{p}$, thereby

$$f^{\mathrm{E}}(h|p) \geq f^{\mathrm{S}}(p) \geq p - q \quad \forall h \geq 0.$$

$K_R(f, s)$ is decreasing at any $f > p - q$. Hence, $\Delta K_R(h|p) < 0$ at all $h \geq 0 = \bar{h}(p)$.

Proof of statement 1. Next, consider the case such that $p > \bar{p}$, thereby $f^{\mathrm{S}}(p) < p - q$. Again, $f^{\mathrm{S}}(p) < f^{\mathrm{E}}(h|p)$ for any $h > 0$, and K_R is single-peaked at $f = p - q$. Hence, $\bar{h}(p)$ is strictly positive in this case. To show that $\bar{h}(p)$ is strictly increasing in p, note that

$$\frac{q + f^{\mathrm{E}}(\bar{h}(p)|p)}{(p + q + f^{\mathrm{E}}(\bar{h}(p)|p))^2} = \frac{q + f^{\mathrm{S}}(p)}{(p + q + f^{\mathrm{S}}(p))^2} \tag{35}$$

by the definition of $\bar{h}(p)$ and the continuity of $f^{\mathrm{E}}(h|p)$. Differentiating both sides of (35) with respect to p, and then rearranging terms, we obtain

$$\frac{p - q - f^{\mathrm{E}}}{(p + q + f^{\mathrm{E}})^3} \frac{\partial f^{\mathrm{E}}}{\partial h} \frac{\partial \bar{h}}{\partial p} - \frac{p - q - f^{\mathrm{S}}}{(p + q + f^{\mathrm{S}})^3} \frac{\partial f^{\mathrm{S}}}{\partial p} = 2 \left[\frac{q + f^{\mathrm{E}}}{(p + q + f^{\mathrm{E}})^3} - \frac{q + f^{\mathrm{S}}}{(p + q + f^{\mathrm{S}})^3} \right].$$

The right hand side of the last equation is negative:

$$\begin{aligned} \frac{q + f^{\mathrm{E}}}{(p + q + f^{\mathrm{E}})^3} - \frac{q + f^{\mathrm{S}}}{(p + q + f^{\mathrm{S}})^3} &= \frac{q + f^{\mathrm{E}}}{(p + q + f^{\mathrm{E}})^2} \cdot \frac{1}{p + q + f^{\mathrm{E}}} - \frac{q + f^{\mathrm{S}}}{(p + q + f^{\mathrm{S}})^3} \\ &= \frac{q + f^{\mathrm{S}}}{(p + q + f^{\mathrm{S}})^2} \left[\frac{1}{p + q + f^{\mathrm{E}}} - \frac{1}{p + q + f^{\mathrm{S}}} \right] < 0. \end{aligned}$$

Hence,

$$\underbrace{\frac{p - q - f^{\mathrm{E}}}{(p + q + f^{\mathrm{E}})^3}}_{<0} \underbrace{\frac{\partial f^{\mathrm{E}}}{\partial h}}_{>0} \frac{\partial \bar{h}}{\partial p} < \underbrace{\frac{p - q - f^{\mathrm{S}}}{(p + q + f^{\mathrm{S}})^3}}_{>0} \underbrace{\frac{\partial f^{\mathrm{S}}}{\partial p}}_{<0} \quad \Longrightarrow \quad \frac{\partial \bar{h}}{\partial p} > 0.$$

B.5 Proof of Proposition 5

We first prove the first statement in Proposition 5. One can show that $s_j^{\mathrm{S}} = s_j^L$ for all credulous citizens $j \in (r, 1]$ directly from the platform's first order condition with respect to s_i. Also, recall from (32) and (34) in the proof of Proposition 2 that the conflict cost between a credulous citizen $j \in (r, 1]$ and any other citizen $i \in [0, 1]$ always decreases in f. Finally, $f^{\mathrm{L}} > f^{\mathrm{S}}$ as discussed in the main text. In conclusion, $K_B(f^{\mathrm{L}}, s^{\mathrm{L}}) < K_B(f^{\mathrm{S}}, s^{\mathrm{S}})$ and $K_C(f^{\mathrm{L}}, s^{\mathrm{L}}) < K_C(f^{\mathrm{S}}, s^{\mathrm{S}})$.

The second and third statements in Proposition 5 directly follow the fact that $K_R(f, s)$ is independent of s and the mapping $f \mapsto K_R(f, s)$ is single-peaked at $f = p - q$ (see the proof of the first statement of Proposition 3 for the basic properties of K_R).

B.6 Proof of Proposition 6

From (4) in Proposition 1,

$$\frac{(1-r)\,\beta}{(q^{\mathrm{B}}+f^{\mathrm{B}})^2}+\frac{r\beta}{(p+q^{\mathrm{B}}+f^{\mathrm{B}})^2}=cf^{\mathrm{B}}\quad\text{and}\quad\frac{(1-r)\,\beta}{(q^{\mathrm{A}}+f^{\mathrm{A}})^2}+\frac{r\beta}{(p+q^{\mathrm{A}}+f^{\mathrm{A}})^2}=cf^{\mathrm{A}}$$

hence, $f^{\mathrm{B}} > f^{\mathrm{A}}$ but $q + f^{\mathrm{B}} < q^{\mathrm{A}} + f^{\mathrm{B}}$. Also, by (6) in Proposition 1, $s_j^{\mathrm{B}} = s_j^{\mathrm{A}}$ for all credulous citizens $j \in (r, 1]$.

To prove the first statement in Proposition 6, recall from (32) and (34) in the proof of Proposition 2 that the conflict cost between a credulous citizen $j \in (r, 1]$ and any other citizen $i \in [0, 1]$ decreases in $z \equiv q + f$. Hence, $K_B(f^{\mathrm{A}}, s^{\mathrm{A}}) < K_B(f^{\mathrm{B}}, s^{\mathrm{B}})$ and $K_C(f^{\mathrm{A}}, s^{\mathrm{A}}) < K_C(f^{\mathrm{B}}, s^{\mathrm{B}})$.

The second and third statements in Proposition 6 directly follow the fact that $K_R(f, s)$ is independent of s and the mapping $z \equiv f + q \mapsto K_R(z - q, s)$ is single-peaked at $z \equiv q + f = p$ (see the proof of the first statement of Proposition 3 for the basic properties of K_R).

B.7 Proof of Corollary 2

Let $(f^{\mathrm{B}}, s^{\mathrm{B}})$ and $(f^{\mathrm{M}}, s^{\mathrm{M}})$ respectively denote the slants that the platform chooses before and after a media campaign, where a filter floor $\underline{f} \geq f^{\mathrm{S}}$ imposes in both cases. Let $(f^{\mathrm{S}}, s^{\mathrm{S}})$ denote the algorithm that the platform would choose without any regulations.

First, from (4) in Proposition 1, the platform's choice of f decreases in r if there were no filter floor. Hence, any filter floor $\underline{f} > f^{\mathrm{S}}$ binds both before and after a successful media campaign, and therefore, $f^{\mathrm{B}} = f^{\mathrm{M}} = \underline{f}$. Also, one can show that $s^{\mathrm{M}} = s^{\mathrm{B}}$ directly from the platform's first order condition with respect to s_i. Summing up, with a filter floor $\underline{f} > f^{\mathrm{S}}$, the platform chooses the same algorithm before and after a media literacy campaign, which we will denote by (f^*, s^*) in the remaining part of the proof.

Next, recall from (32)—(34) in the proof of Proposition 2 that, with the platform's algorithm (f^*, s^*) being fixed, the conflict cost between two rational citizens, $\mathbf{E}\left[(\mathbf{E}^*[\theta|y_j] - \mathbf{E}^*[\theta|y_i])^2\right]$, is strictly smaller than the conflict cost between any other possible pair of citizens, $\mathbf{E}\left[(y_j - y_i)^2\right]$ and $\mathbf{E}\left[(y_j - \mathbf{E}^*[\theta|y_i])^2\right]$. Hence, a media campaign only reduces the conflict cost between any pair of citizens, and therefore, the aggregate conflict cost decreases.

B.8 Proof of Proposition 7

In this proof, let $(f^{\mathrm{S}}|_{r=\hat{r}}, s^{\mathrm{S}}|_{r=\hat{r}})$ denote the equilibrium algorithms as characterized by (4), where the fraction of rational citizens r is evaluated at $\hat{r} \in [0,1]$. Let r_B denote the fraction of rational citizens *before* the media literacy campaign. By Proposition 1, $f^{\mathrm{S}}|_{r=\hat{r}}$ is strictly decreasing in $\hat{r}$. With these notations, $(f^{\mathrm{M}}, s^{\mathrm{M}}) = (f^{\mathrm{S}}|_{r=1}, s^{\mathrm{S}}|_{r=1})$ and $(f^{\mathrm{S}}, s^{\mathrm{S}}) = (f^{\mathrm{S}}|_{r=r_B}, s^{\mathrm{S}}|_{r=r_B})$, respectively refer to the algorithms that the platform would choose before and after the media literacy campaign.

To show the first two statements in the proposition, note that the first order condition (4) and the observation that $f^{\mathrm{M}} < f^{\mathrm{S}}$ together imply

$$\frac{\beta}{(q+f^{\mathrm{S}})^2} > \frac{(1-r)\,\beta}{\left(q+f^{\mathrm{S}}\right)^2} + \frac{r\beta}{\left(p+q+f^{\mathrm{S}}\right)^2} = cf^{\mathrm{S}} > cf^{\mathrm{M}} = \frac{\beta}{\left(p+q+f^{\mathrm{M}}\right)^2},$$

and thus

$$\begin{aligned}
\frac{q+f^{\mathrm{M}}}{(p+q+f^{\mathrm{M}})^2} &< \frac{1}{p+q+f^{\mathrm{M}}} < \frac{1}{q+f^{\mathrm{S}}}, \\
\frac{q+f^{\mathrm{M}}}{(p+q+f^{\mathrm{M}})^2} &< \frac{q+f^{\mathrm{S}}}{(p+q+f^{\mathrm{S}})(p+q+f^{\mathrm{M}})} < \frac{1}{p+q+f^{\mathrm{S}}},
\end{aligned}$$

where the first inequality in the second line holds as $(f+q)/(p+q+f)$ is strictly increasing in f. By combining these two inequalities and (32)–(34), the first two statements in the propositin follow.

To prove the third statement in the proposition, let $\kappa_R(\hat{f}) := 2(q+\hat{f})/(p+q+\hat{f})^2$ denote the equilibrium conflict cost between two rational citizens when the equilibrium filtering algorithm is given by $f = \hat{f}$. First, consider the case $p - q > f^M = f^{\mathrm{S}}|_{r=1} > 0$. Because $f^{\mathrm{S}}|_{r=\hat{r}}$ is strictly decreasing in $\hat{r}$ and $\kappa_R(\hat{f})$ is single-peaked at $\hat{f} = p - q$, there exists $\bar{r} \in [0,1)$ such that the following inequality holds for any $i, j \le r_B$ if and only if $r_B < \bar{r}$:

$$K_{ij}(f^{\mathrm{M}}, s^{\mathrm{M}}) = \kappa(f^{\mathrm{S}}|_{r=1}) > \kappa(f^{\mathrm{S}}|_{r=r_B}) = K_{ij}(f^{\mathrm{S}}, s^{\mathrm{S}}) \quad \forall i, j \le r_B. \tag{36}$$

Finally, if either $f^M \ge p - q > 0$ or $p - q < 0$ holds, $\kappa_R(f|_{r=r_B}) = \kappa_R(f^{\mathrm{S}})$ is strictly increasing in r_B, so that (36) always holds.

B.9 Proof of Proposition 8

Suppose that the algorithms are publicly observable. Then, by choosing algorithms (f, s), the self-interested platform's payoff is given by

$$\pi(f, s) := \mathbf{E}\left[v\left(y; f, s\right)\right] - \frac{cf^2}{2} - \int_0^1 \frac{ks_i^2}{2}\,\mathrm{d}i,$$

where the expectation is taken with respect to the platform's algorithms (f, s). On the other hand, by choosing algorithms (f, s), the platform's payoff given ethical concern is

$$\pi(f, s) - h \cdot K(f, s)$$

in view of (18). Then, given any equilibrium $(\tilde{f}^{\mathrm{S}}, \tilde{s}^{\mathrm{S}})$ absent ethical concern and any equilibrium $(\tilde{f}^{\mathrm{E}}, \tilde{s}^{\mathrm{E}})$ in the presence of ethical concern, it must hold that

$$\pi(\tilde{f}^{\mathrm{E}}, \tilde{s}^{\mathrm{E}}) \leq \pi(\tilde{f}^{\mathrm{S}}, \tilde{s}^{\mathrm{S}}).$$

To complete the proof, suppose, towards a contradiction, that $K(\tilde{f}^{\mathrm{S}}, \tilde{s}^{\mathrm{S}}) < K(\tilde{f}^{\mathrm{E}}, \tilde{s}^{\mathrm{E}})$. Then,

$$\pi(\tilde{f}^{\mathrm{E}}, \tilde{s}^{\mathrm{E}}) - h \cdot K(\tilde{f}^{\mathrm{E}}, \tilde{s}^{\mathrm{E}}) \leq \pi(\tilde{f}^{\mathrm{S}}, \tilde{s}^{\mathrm{S}}) - h \cdot K(\tilde{f}^{\mathrm{E}}, \tilde{s}^{\mathrm{E}}) < \pi(\tilde{f}^{\mathrm{S}}, \tilde{s}^{\mathrm{S}}) - h \cdot K(\tilde{f}^{\mathrm{S}}, \tilde{s}^{\mathrm{S}}).$$

This contradicts the assumption that $(\tilde{f}^{\mathrm{E}}, \tilde{s}^{\mathrm{E}})$ is an equilibrium given ethical concern.

B.10 Proof of Proposition 9

From (32)-(34) in the proof of Proposition 2, the implementation of the fairness doctrine (which only changes the slant for each citizen from s_i^{S} to zero) reduces the conflict cost between any pair of citizens. Moreover, the conflict cost is reduced strictly for any pair of citizens with at least one of them being credulous. Hence, the fairness doctrine unambiguously reduces the aggregate conflict cost.

References

Anderson, S. P., Waldfogel, J. and Strömberg, D. (2015). *Handbook of Media Economics*, Elsevier.

Andreoni, J. and Mylovanov, T. (2012). Diverging Opinions, *American Economic Journal: Microeconomics* **4**(1): 209–32.

Bakshy, E., Messing, S. and Adamic, L. A. (2015). Exposure to Ideologically Diverse News and Opinion on Facebook, *Science* **348**(6239): 1130–1132.

Baliga, S., Hanany, E. and Klibanoff, P. (2013). Polarization and Ambiguity, *American Economic Review* **103**(7): 3071–83.

Bursztyn, L., Egorov, G., Enikolopov, R. and Petrova, M. (2019). Social Media and Xenophobia: Evidence from Russia, *Technical report*, National Bureau of Economic Research.

Chen, H. and Suen, W. (2021). Competition for Attention and News Quality, *Technical report*, Working Paper, University of Hong Kong.

Cripps, M. W., Ely, J. C., Mailath, G. J. and Samuelson, L. (2008). Common Learning, *Econometrica* **76**(4): 909–933.

Cusumano, M., Gawer, A. and Yoffie, D. (2021). Social Media Companies Should Self-Regulate. Now., *Harvard Business Review* .

Dixit, A. K. and Weibull, J. W. (2007). Political Polarization, *Proceedings of the National Academy of Sciences* **104**(18): 7351–7356.

Funke, D. and Flamini, D. (2021). A Guide to Anti-misinformation Actions around the World, *Poynter* .

Gentzkow, M. and Shapiro, J. M. (2006). Media Bias And Reputation, *Journal of Political Economy* **114**(2): 280–316.

Gentzkow, M. and Shapiro, J. M. (2010). What Drives Media Slant? Evidence from US Daily Newspapers, *Econometrica* **78**(1): 35–71.

Gentzkow, M., Shapiro, J. M. and Stone, D. F. (2015). Media Bias in the Marketplace: Theory, *Handbook of Media Economics*, Vol. 1, Elsevier, pp. 623–645.

Holmström, B. (1999). Managerial Incentive Problems: A Dynamic Perspective, *The Review of Economic Studies* **66**(1): 169–182.

Karell, D. (2021). Online Extremism and Offline Harm, *Social Science Research Council* .

Kartik, N., Lee, F. X. and Suen, W. (2021). Information Validates the Prior: A Theorem on Bayesian Updating and Applications, *American Economic Review: Insights* **3**(2): 165–82.

Kartik, N., Ottaviani, M. and Squintani, F. (2007). Credulity, Lies, and Costly Talk, *Journal of Economic Theory* **134**(1): 93–116.

Kearns, M. and Roth, A. (2019). *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*, Oxford University Press.

Lang, J., Erickson, W. W. and Jing-Schmidt, Z. (2021). #MaskOn!#MaskOff! Digital Polarization of Mask-wearing in the United States during COVID-19, *PloS one* **16**(4): e0250817.

Little, A. T. (2017). Propaganda and Credulity, *Games and Economic Behavior* **102**: 224–232.

MacCarthy, M. (2020). Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry, *Working paper, Transatlantic Working Group* .

Mullainathan, S. and Shleifer, A. (2005). The Market for News, *American Economic Review* **95**(4): 1031–1053.

Müller, K. and Schwarz, C. (2018). Fanning the Flames of Hate: Social Media and Hate Crime, *Journal of the European Economic Association* .

Müller, K. and Schwarz, C. (2020). From Hashtag to Hate Crime: Twitter and Anti-minority Sentiment, *Working paper, Bocconi University* .

Napoli, P. M. (2019). *Social Media and the Public Interest*, Columbia University Press.

Perego, J. and Yuksel, S. (2021). Media Competition and Social Disagreement, *Working paper, Columbia Business School* .

Prat, A. and Strömberg, D. (2013). The Political Economy of Mass Media, *Advances in Economics and Econometrics* **2**: 135.

Sethi, R. and Yildiz, M. (2012). Public Disagreement, *American Economic Journal: Microeconomics* **4**(3): 57–95.

Settle, J. E. (2018). *Frenemies: How Social Media Polarizes America*, Cambridge University Press.

Suen, W. (2004). The Self-Perpetuation of Biased Beliefs, *The Economic Journal* **114**(495): 377–396.

Sunstein, C. R. (2001). *Republic.com*, Princeton University Press.

Sunstein, C. R. (2018). *# Republic*, Princeton University Press.

Wu, T. (2017). *The Attention Merchants: The Epic Scramble to Get Inside Our Heads*, Vintage.

Zanardo, E. (2017). How to Measure Disagreement?, *Working paper, Columbia University* .

Zhuravskaya, E., Petrova, M. and Enikolopov, R. (2020). Political Effects of the Internet and Social Media, *Annual Review of Economics* **12**: 415–438.