Perverse Ethical Concerns: Online Misinformation and Offline Conflicts*

(Click here for the latest version.)

Dongkyu Chang† Allen Vong‡

December 31, 2022

Abstract

We investigate a setting in which a large number of consumers learn a hidden state individually on an online platform. The platform receives news reports about the state and imperfectly filters misinformation in the reports, triggering conflicts about the value of the state among the consumers. We show that a platform with an ethical concern to internalize consumers' welfare could perversely reduce their welfare by aggravating conflicts among these consumers. We show that societal efforts that aim to improve consumers' welfare, such as investments in ethical algorithms, public awareness campaigns, and government policies, are effective if and only if their implementation is sufficiently aggressive.

JEL codes: C72, D83, L86.

Keywords: conflicts, misinformation, platform, social media.

*We are grateful to Johannes Hörner, Larry Samuelson, and Marina Halac for their constant support. For helpful comments, we thank Luís Cabral, Yeon-Koo Che, Yi Chen, Syngjoo Choi, Liang Dai, Chris Edmond, Sunny Huang, Ilwoo Hwang, Shota Ichihashi, Maxim Ivanov, RC Lim, Bart Lipman, Delong Meng, Dmitry Shapiro, Euncheol Shin, Wing Suen, Qianfeng Tang, Andy Zapechelnyuk, seminar participants in Fudan University, Korea University, SAET conference 2021, Seoul National University, Shanghai University of Finance and Economics, Shanghai Jiao Tong University, Sungkyunkwan University, TSE Online Economics of Platforms seminar, University of Macau, KAIST, Hiroshima University, Stevens Institute of Technology, Hong Kong University of Science and Technology, and Chinese University of Hong Kong. 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.

1 Introduction

Online platforms such as *Facebook*, *Instagram*, *Twitter*, and *YouTube* have become the main source for news consumption (see, e.g., Pew Research Center, 2021b). News consumers are, however, concerned with harmful misinformation on these platforms (see, e.g., Pew Research Center, 2021a). Misinformation causes allocative inefficiencies—misinformed consumers take sub-optimal actions. Another, arguably more pressing, concern is that misinformation causes harmful negative externalities—misinformed consumers have disagreeing worldviews and resolve their disagreements by means of conflicts and polarization via offline violence, online abuse, and social unrest. *Facebook*, for example, has been blamed for fueling the diffusion of misinformation that resulted in violence in the January 6 United States Capitol attack.[1]

In response to harmful misinformation on online platforms, societies adopt efforts that make these platforms ethically internalize consumers' welfare in providing news contents. These efforts include awareness campaigns such as the *Wall Street Journal*'s investigative podcast series, congress hearings, research programs on ethical algorithms (see, e.g., Wu, 2017; Kearns and Roth, 2019), and government policies (see, e.g., Funke and Flamini, 2021). The goal of this article is to understand the effectiveness of these societal efforts. In doing so, we focus on the role of online platforms as news suppliers and the role of platform users as news consumers, abstracting from other platform activities such as matching and trading.

Our analysis offers a caution concerning the said societal efforts. We show that platforms could perversely reduce consumers' welfare by ethically internalizing this welfare: consumers who anticipate this internalization could become too confident of the personalized content they read on the platforms and, in turn, more hostile to disagreeing worldviews; this increasing hostility aggravates harmful conflicts and reduces consumers' welfare. We show that these societal efforts improve consumers' welfare if and only if they are implemented sufficiently aggressively.

[1] See, e.g., "Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs" *The Washington Post*, October 22, 2021.

1.1 Overview of model and main result

We deliver our insights in a simple model where a unit mass of consumers must take an action, and their desirable actions depend on a hidden state. These actions could represent either physical actions or verbal opinions. Before taking actions, these consumers acquire information about the state from a strategic platform. This platform is an information intermediary. It receives news reports about the state from external sources and produces a noisy, private, personalized signal for each consumer. Each consumer's received signal depends on the news sources to which she subscribes on the platform as well as on the platform's filtering algorithm that filters misinformation in the news reports. The platform develops this filter at a cost and this filter is hidden from the consumers. The consumers use their signals to infer the state. Some consumers are rational and perform Bayesian inferences; the others are credulous non-Bayesians who believe that the state equals their signals. The signals that any two consumers receive typically disagree, triggering conflicts among their actions in the form of negative externalities that harm the consumers.

We begin with a basic model where the platform is self-interested and concerns only its profit. This platform profits when consumers enjoy reading its content, and the consumers enjoy content that is informative about the state as well as content that conforms to their own biases. We next consider an alternative model where the platform faces an additional ethical concern to internalize consumers' welfare—this platform maximizes a weighted sum of its profit and consumers' welfare. We then contrast the equilibria in these two models and deliver our main result: the platform's ethical concern could perversely reduce consumers' welfare, and this backfire must arise from an aggravation of consumers' conflicts. This backfire happens if and only if the population consists of sufficiently many rational consumers and the platform's ethical concern is not sufficiently strong.

In equilibrium, the ethical concern causes the platform to filter more aggressively, thereby improving the consumers' learning about the state. This learning effect mitigates conflicts. At the same time, the rational consumers correctly anticipate the more aggressive filter and thus become more confident about their own learning. This

confidence effect aggravates conflicts. Thus, the ethical concern mitigates conflicts between any two rational consumers if and only if the learning effect dominates the confidence effect or, equivalently, if and only if the ethical concern is sufficiently strong to induce a sufficiently aggressive filter. On the other hand, the ethical concern mitigates the conflict between any two credulous consumers. This is because, irrespective of the ethical concern, these consumers believe that the state equals their signals. Thus, in their inferences, the confidence effect is absent and the learning effect is stronger than in the rational consumers' inferences. Finally, the ethical concern mitigates the conflict between any rational consumer and any credulous consumer. Although the confidence effect that aggravates conflicts is present in the rational consumer's inference, this aggravation is dampened by the fact that this confidence effect leads the rational consumer to infer more similarly as the credulous consumer does. This observation, alongside the learning effect in both consumers' inferences, yields the result.

Of course, the platform knows that a more aggressive filter could aggravate conflicts and in turn could perversely reduce consumers' welfare. Why, then, would the aggravating conflicts among the rational consumers arise? This is due to the credence nature of the platform's signals, which prevents the platform from correctly internalizing equilibrium conflicts. Since the filter is hidden from the consumers, the rational consumers assess their signals based on their conjectures of the filter but not the actual filter. The platform's actual filter thus affects only the distribution of the consumers' signals but not the consumers' state inferences given the signals. In turn, when the platform best responds to the consumers' conjectures, the ethical concern boosts its filtering incentives to reduce the dispersion of the signals, thereby reducing the disagreement in the consumers' inferences. But then in equilibrium, the rational consumers correctly anticipate the platform's such incentive when forming their conjectures, yielding the perverse outcome.

1.2 Policy implications

We apply our main result to draw policy implications. We examine misinformation legislation, arrests of misinformation spreaders, and cyber task forces which are designed to reduce misinformation on platforms. We find that these efforts mitigate conflicts if and only if their implementation is sufficiently aggressive that the learning effect dominates the confidence effect. We then analyze media literacy campaigns that educate credulous consumers and turn them into rational consumers. We show that these campaigns disrupt the platform's filtering incentives. Thus, these campaigns aggravate conflicts unless they are coupled with aggressive efforts that boost the platform's filter.

Our results also speak to debates over the transparency of algorithms (see, e.g., MacCarthy, 2020). While a typical case for transparency is to improve the monitoring of platforms,[2] we show alternatively that transparency allows platforms to correctly internalize their social responsibilities: if the filter is not hidden, then the consumers would draw inferences based on the actual filter and no perverse outcome would arise.

Although our analysis focuses on filtering of misinformation, our model is flexible and can encompass other platform instruments. We demonstrate this flexibility in an extension where the platform manipulates the media slants of consumers' news subscriptions in addition to filtering. Notably, we show that the credence nature of the platform's signals causes the platform to manipulate not only the slants of the news subscriptions of the credulous consumers, but also those of the rational consumers, irrespective of whether the platform faces an ethical concern. This prediction contrasts with familiar findings in the literature on media bias (e.g., Mullainathan and Shleifer, 2005; see also Gentzkow, Shapiro and Stone, 2015 for a survey) and sheds light on evidence that social media users tend to encounter content 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). Moreover, while we interpret our model as a model of news transmission by an online platform,

[2]See "Whistle-blower unites democrats and republicans in calling for regulation of Facebook," *The New York Times*, October 5, 2021.

our insights can be applied to other information providers such as traditional media, the government, or social groups, in which the consideration of whether the provider should be "ethical" is relevant.

1.3 Related literature

In motivation, the first part of our analysis, namely the equilibrium characterizations with and without ethical concern, is related to the literature on information design by platforms (e.g., Candogan and Drakopoulos, 2020; Chen and Papanastasiou, 2021). This literature studies variants of the Bayesian persuasion problem à la Kamenica and Gentzkow (2011) and considers platforms that commit to their chosen information transmission mechanisms. Our model differs by assuming that the platform has no commitment power and that it strategically responds to consumers' conjectures of its algorithm. Our main insight, namely the perverse outcome, is precisely driven by the platform's such strategic response.

The second part of our analysis, namely the policy implications, is most closely related to Mostagir and Siderius (2022) and Acemoglu, Ozdaglar and Siderius (2022). These papers examine consumers' learning of a binary state and derive conditions given which several exogenous misinformation policies could backfire for their learning. Our analysis differs in focusing on the policy implications for consumers' conflicts that result from their learning, but not for their learning per se. Indeed, in our Gaussian environment, the misinformation policies that we consider always improve the consumers' learning. The fundamental mechanisms that drive the perverse outcome in our analysis and in theirs are also different. The mechanism in ours is the inevitable failure by a platform to correctly internalize the equilibrium conflict cost, as it best replies to the consumers' conjectures of its filter. In theirs, the mechanism is the excessive (resp., insufficient) weight that consumers put on their own prior belief relative to platform information. For example, Mostagir and Siderius (2022) show that Bayesian agents, who know that misinformation policies are at work, might rely too much (resp., too little) on platform information and in turn, are more vulnerable to misinformation (resp., under-utilize the platform information). Acemoglu et al. (2022)

further show that this latter effect could be amplified because consumers are more willing to share their information with their peers in the presence of misinformation policies, facilitating the spread of misinformation.

In terms of modeling, our model is most closely related to Little (2012, 2015) and Edmond and Lu (2021). Like our model, these models study strategic information transmission from a sender without commitment power to many receivers (i.e., consumers) in a Gaussian environment. Unlike our model, the sender's only instrument in these models is to manipulate the mean of the consumers' signals; these models do not allow the sender to manipulate the noise in the consumers' signals and do not examine the equilibrium structure of the consumers' disagreement about the state. In this latter regard, our work also speaks to the literature on disagreement. This literature typically focuses on Bayesian agents whose disagreements are due to their heterogeneous prior beliefs (e.g., Dixit and Weibull, 2007; Andreoni and Mylovanov, 2012; Sethi and Yildiz, 2012; Kartik, Lee and Suen, 2021) or competition among information senders (e.g., Perego and Yuksel, 2021). In contrast, in our model, disagreements are due to consumers' heterogeneous posterior beliefs induced endogenously by the platform's optimizing algorithm. We also depart by examining disagreements within and between Bayesian consumers and non-Bayesian consumers. We view these departures as not only theoretically attractive but also policy-relevant. Finally, as we shall discuss in detail, our analysis also sheds light on empirical evidence that overconfidence exacerbates disagreements (see, e.g., Ortoleva and Snowberg, 2015).

More broadly, the literature on misinformation has examined consumers' strategic sharing of news articles containing misinformation on social media (e.g., Papanastasiou, 2020; Acemoglu et al., 2022) and their strategic subscriptions of biased news sources (e.g., Jann and Schottmuller, 2021). We view our work as complementary to these papers: to focus on the platform's perverse incentives, our model abstracts from such strategic behavior of the consumers and simply takes the presence of misinformation and biased subscriptions as given. Finally, this paper speaks to interdisciplinary research programs on ethical algorithms, as noted at the outset, which covers topics beyond conflicts, such as privacy and addiction. We contribute by elucidating the

strategic implications of platforms' ethical concerns. Limiting attention to conflicts incited by misinformation, we offer a caution against the conventional wisdom that arguably underlies this program, namely that ethical concerns are unambiguously socially desirable.

2 The basic model

There are a unit mass of consumers, indexed by $i \in [0,1]$, and a platform. There is a hidden state $\theta \in \mathbf{R}$ that is distributed normally with mean normalized to 0 and precision $p > 0$.

Consumers. Each consumer i has a type $(b_i, s_i, l_i) \in \mathbf{R}^2 \times \{l^R, l^C\}$. Here, b_i represents this consumer's bias, capturing her preferred value of the state; s_i represents the media slant of this consumer's news subscriptions on the platform; finally, l_i represents this consumer's literacy, which is either rational ($l_i = l^R$) or credulous ($l_i = l^C$). For our results, we only need each rational consumer to know her type. Nonetheless, to ease the exposition, we avoid defining beliefs on types by assuming that consumers' types are commonly known and in turn, without loss, that each consumer $i \in [0, r]$ is rational and each consumer $i \in (r, 1]$ is credulous for some $r \in [0, 1]$.

Each consumer i must take an action $a_i \in \mathbf{R}$, which could be interpreted as either a physical action that this consumer must take or as an opinion that this consumer must express. This consumer's desirable action depends on the state θ. Before this consumer takes her action, she acquires information about this state from the platform. This information is summarized by a signal $y_i \in \mathbf{R}$ that is described below in (2).[3] The difference between rational consumers and credulous consumers lies in their inferences about the state. Upon receiving a signal, a rational consumer forms a Bayesian posterior distribution about the state. On the other hand, a credulous consumer takes her received signal at its face value: she forms a Dirac distribution about the state

[3]Our results extend if the consumers observe "a few" other consumers' signals. What is crucial to our results is that the consumers do not observe the same signals, ensuring that the consumers have some posterior disagreements about the state.

that this state is equal to her received signal with probability one.

Given state θ and a profile of actions $a := (a_i)_{i\in[0,1]}$, consumer i's realized utility is

$$-\tau(a_i - \theta)^2 - \frac{\xi}{2}\int_0^1 \int_0^1 (a_k - a_j)^2 \,\mathrm{d}k \,\mathrm{d}j. \tag{1}$$

This "action utility" has two components. The first component captures the consumer's desire to match her action with the state. The second component captures negative externalities that harm this consumer given the consumers' conflicts, i.e., the disagreement in the consumers' actions, as motivated at the outset.[4]

Platform. At the outset, the platform chooses a filter $f \in \mathbf{R}_+$, hidden from the consumers, that determines the precision of the consumers' signals about the state. Developing this filter is costly because, for instance, the platform needs to hire and train engineers to do so. By choosing filter f, the platform incurs a cost $cf^2/2$, where $c > 0$ is an exogenous parameter. Consumer i's signal is given by

$$y_i = \theta + \varepsilon_i, \tag{2}$$

where the noise ε_i is normally distributed with mean equal to the consumer's slant s_i and precision $q + f$, with $q > 0$, independently of the state θ and independently across consumers. This noise represents the misinformation in the consumer's potentially slanted news subscription that "escapes" the filter and is read by the consumer. If the consumer's slant is positive (resp., negative), then the news sources to which the consumer subscribes tend to report more positive (resp., negative) news about the state.[5] The parameter q represents the default precision of the signal absent any

[4] Our notion of disagreement is familiar from the literature (see, e.g., Kartik et al., 2021). This notion is restrictive if one is interested in comparing the consumers' posterior distributions. See, e.g., Zanardo (2017) who axiomatically examines disagreement between probability distributions.

[5] Thus, two consumers i and j who share the same slant could receive different misinformation ε_i and ε_j from the news reports and thus different signals y_i and y_j. This assumption is natural as the news sources to which each of these consumers subscribes could differ even though the aggregate slants of their sources are identical.

filtering.[6] We interpret a higher filter as a more aggressive filter; given a higher filter, the consumer's signal is more informative about the state.

The platform derives advertising revenue that is proportional to the time that consumers spend on the platform.[7] Each consumer spends a duration of time, normalized to one, to acquire her signal about the state. This consumer spends an extended duration of time on the platform, depending on how much she enjoys the information. Specifically, let

$$-\beta \mathbf{E}_i \left[(\theta - b_i)^2 \middle| y_i \right]$$

denote consumer i's "psychological utility" upon receiving signal y_i, where $\beta > 0$ is an exogenous parameter and $\mathbf{E}_i$ denotes this consumer's expectation about the state. This psychological utility is higher if the consumer's inferred state is closer to what she would like it to be. Note that if this consumer i is rational, then her expectation $\mathbf{E}_i$ is taken with respect to her conjectured filter f_i^* that the platform has chosen. Because the actual filter f chosen by the platform is hidden, *a priori*, the consumer's conjectured filter can be different from the actual filter. Thus, up to a positive transformation, given a profile of signals $y := (y_i)_{i \in [0,1]}$ and the rational consumers' conjectures $f^* := (f_i^*)_{i \in [0,r]}$, the platform's realized revenue is given by

$$R(y, f^*) := \int_0^1 \mathbf{E}_i \left[-\beta (\theta - b_i)^2 \middle| y_i \right] \mathrm{d}i. \tag{3}$$

To be sure, in reality, platforms exhibit greater flexibility than simply filtering misinformation when creating content for the consumers. For instance, platforms could manipulate each consumer's slant by recommending certain (biased) news sources for the consumer to subscribe to. Consumers might also acquire private signals, for example, by communicating with others, in addition to signals from the platform.

[6] The assumption that q is positive plainly serves to ease the exposition. It simply rules out a trivial equilibrium with zero filtering.

[7] For example, *Facebook* makes money primarily by showing its users advertiser content. In a report by the SEC, advertising represented 98% of *Facebook*'s revenue in 2020. See https://www.sec.gov/ix?doc=/Archives/edgar/data/1326801/000132680121000014/fb-20201231.htm.

We demonstrate that these extensions do not alter our main insights in Section 7. Moreover, in terms of our formal analysis, it is without loss for most of our results, except in Section 6 where media literacy campaigns are concerned, to assume that each consumer's slant is equal to zero. We introduce the possibility of non-zero slants to make it transparent that these slants do not drive our results and to avoid repetitively introducing some of our model elements in Section 6.

Payoffs. Given state θ, action profile a, a signal profile $y := (y_i)_{i\in[0,1]}$, the platform's filter f, and rational consumers' conjecture of its filter $f^* = (f_i^*)_{i\in[0,r]}$, each consumer i's realized payoff is equal to the sum of her psychological utility and her action utility:

$$\mathbf{E}_i\left[-\beta(\theta - b_i)^2\middle|y_i\right] - \tau(a_i - \theta)^2 - \frac{\xi}{2}\int_0^1\int_0^1 (a_k - a_j)^2 \mathrm{d}k\mathrm{d}j.$$

On the other hand, the platform's realized payoff is given by

$$R(y, f^*) - \frac{cf^2}{2}. \tag{4}$$

In this basic model, we say that the platform is self-interested as its payoff (4) is plainly its profit.

Solution concept. The solution concept that we use is Bayesian Nash equilibrium in pure strategies, henceforth equilibrium. We focus on equilibria in pure strategies to facilitate tractable belief updating by the consumers; nonetheless, we allow the platform to contemplate deviations to arbitrary strategies. In any such equilibrium, the platform chooses a filter f to maximize its payoff (4) given the rational consumers' conjectures f^*, such that their conjectures are correct. Thus, their equilibrium conjectures must be identical. Hereafter, when we say that the rational consumers' conjecture is f^*, we mean that they conjecture the same filter and abuse notation to denote this filter by $f^* \in [0, \infty)$.

Hereafter, to ease the exposition, we write $\mathbf{E}^*[\cdot]$ as each rational consumer's expectation when they conjecture filter f^* and write $\mathbf{E}[\cdot]$ as the platform's expectation

by choosing filter f. In addition, we will use the notations $\mathbf{Var}^*[\cdot]$, $\mathbf{Var}[\cdot]$, and $\mathbf{Var}_i[\cdot]$ to denote the variance operators corresponding $\mathbf{E}^*[\cdot]$, $\mathbf{E}[\cdot]$, and $\mathbf{E}_i[\cdot]$, respectively.

The next section characterizes the unique equilibrium in this basic model. We then turn to define our notion of conflicts and introduce a platform with ethical concern. We report our main results in Section 5. All proofs are in the Appendix.

3 Equilibrium

In this section, we characterize a unique equilibrium when the platform is self-interested. By choosing filter f, the platform's expected revenue can be written in the following form:

$$\begin{aligned}\mathbf{E}[R(y, f^*)] &= \mathbf{E}\left[\int_0^1 \mathbf{E}_i\left[-\beta(\theta - b_i)^2 | y_i\right] \mathrm{d}i\right] \\ &= \beta\mathbf{E}\left[\int_0^1 -(\mathbf{E}_i[\theta|y_i] - b_i)^2 - \mathbf{Var}_i[\theta|y_i]\mathrm{d}i\right]. \end{aligned} \tag{5}$$

This expression reveals that the platform can generate revenue via two channels. The first channel is to ensure that the consumers' inferred states given their signals conforms to their biases, as captured by the quadratic loss of their estimates from biases. The second channel is to improve the consumers' (perceived) quality of learning, as captured by their negative posterior variances of the state.

Proposition 1 below characterizes the equilibrium given a self-interested platform.

Proposition 1. *There is a unique equilibrium. In this equilibrium, the platform chooses filter $f^{\mathrm{S}} \equiv f^{\mathrm{S}}(\beta, c, p, q, r) > 0$ characterized by*

$$\beta\left(\frac{r}{(p + q + f^{\mathrm{S}})^2} + \frac{1 - r}{(q + f^{\mathrm{S}})^2}\right) = cf^{\mathrm{S}}. \tag{6}$$

This filter f^{S} is strictly increasing in β and is strictly decreasing in (r, c, p, q).

Equation (6) pins down the equilibrium filter f^{S} by equating the marginal benefit of filtering on the left side and the marginal cost of filtering on the right side.

It is instructive to consider a sketch of the proof of this proposition. In equilibrium, the platform chooses its filter to best reply to the rational consumers' conjecture. Given any conjecture f^*, the component of the platform's revenue (5) corresponding to the rational consumers' quality of learning, by standard Bayesian updating, is[8]

$$\mathbf{E}\left[\mathbf{Var}_i\left[\theta|y_i\right]\right] = \begin{cases} \mathbf{E}\left[\dfrac{1}{p+q+f^*}\right] = \dfrac{1}{p+q+f^*}, & \text{if } i \text{ is rational,} \\ 0, & \text{if } i \text{ is credulous.} \end{cases}$$

This expression is independent of the actual filter f. As a result, the platform filters only to maximize the bias-conforming component of its revenue, which is given by

$$\beta\mathbf{E}\left[\int_0^1 -(\mathbf{E}_i[\theta|y_i]-b_i)^2\,\mathrm{d}i\right] = -\beta\mathbf{E}\left[\int_0^r (\mathbf{E}^*\left[\theta|y_i\right]-b_i)^2\,\mathrm{d}i + \int_r^1 (y_i-b_i)^2\,\mathrm{d}i\right]. \quad (7)$$

By standard Bayesian updating, each rational consumer i's state estimate in (7) is

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

Thus, to form an estimate, this consumer discounts her signal by removing her slant and assigning a weight less than unity on the unslanted signal.

From the platform's perspective, when it chooses the filter, the consumers' signals, and hence their estimates given the conjecture f^*, are random. Choosing a higher filter raises its cost but, at the same time, improves (7) by reducing the dispersion of both the rational consumers' received signals and the credulous consumers' received signals, and in turn reducing the dispersion of their state estimates away from their biases. The equilibrium filter f^{S} is precisely the rational consumers' conjecture given which the platform's best response is to pick filter f^{S}, leading to Proposition 1.[9]

The platform's desire to reduce signal dispersion yields the comparative statics in the proposition. This reduction is more effective given a smaller prior state precision

[8]This observation relies on the property of normal distribution that the posterior variance does not depend on the signal. Our main results (Propositions 4 and 5) do not hinge on this property. We provide a further discussion at the end of Section 7.

[9]This equilibrium phenomenon, where the platform's best response to the rational consumers' conjecture is precisely the consumers' conjecture, is reminiscent of Holmström (1999).

p, as the rational consumers place a higher weight on their signals in their inferences. On the other hand, there is diminishing returns to filtering. Thus, the reduction is more effective when the default signal precision q is smaller. Further, the credulous consumers' estimates are more dispersed than the rational consumers' estimates, as the credulous consumers do not discount their signals. Thus, given a larger mass r of rational consumers, the platform's marginal benefit of filtering is smaller. Finally, and intuitively, a higher marginal cost c leads to a smaller equilibrium filter.

4 Ethical concern

In this section, we define a platform with ethical concern. Different from a self-interested platform that maximizes only its profit, a platform with ethical concern maximizes a weighted sum of its profit and consumers' welfare.

4.1 Consumers' welfare

We begin by defining consumers' welfare, given in (11) below. To this end, we introduce some notations. Given the platform's filter f and the rational consumers' conjecture f^*, let $\alpha_i^{f^*}(y_i) \in \mathbf{R}$ denote consumer i's optimal action upon receiving signal y_i. In view of (1), this action maximizes her interim action utility:

$$\alpha_i^{f^*}(y_i) \in \operatorname*{argmax}_{a_i \in \mathbf{R}} \mathbf{E}_i \left[-\tau(a_i - \theta)^2 - \frac{\xi}{2} \int_0^1 \int_0^1 (a_k - a_j)^2 \,\mathrm{d}k \,\mathrm{d}j \middle| y_i \right].$$

Consumer i's optimal action is therefore

$$\alpha_i^{f^*}(y_i) = \begin{cases} \mathbf{E}^*[\theta|y_i], & \text{if } i \text{ is rational,} \\ y_i, & \text{if } i \text{ is credulous.} \end{cases}$$

In addition, let

$$\kappa_{ij}(y, f^*) := \frac{1}{2} \left(\alpha_i^{f^*}(y_i) - \alpha_j^{f^*}(y_j) \right)^2 \tag{9}$$

denote the realized conflict cost resulting from the disagreeing actions between consumers i and j, given signal profile y, with normalizing constant $1/2$.

From the platform's perspective, given the platform's actual filter f and the rational consumers' conjecture f^*, consumer i's *ex ante* optimal action utility is therefore

$$U_i(f, f^*) := \mathbf{E}\left[-\tau(\alpha_i^{f^*}(y_i) - \theta)^2 - \xi \int_0^1 \int_0^1 \kappa_{kj}(y, f^*)\,\mathrm{d}k\,\mathrm{d}j\right].$$

Note that this is the platform's evaluation of consumer i's action utility: even if this consumer is credulous, the expectation $\mathbf{E}[\cdot]$ is taken with respect to the true probability distribution of the signals given the platform's filter f. The (aggregate) consumers' *ex ante* optimal action utility is given by

$$U(f, f^*) := \int_0^1 U_i(f, f^*)\,\mathrm{d}i. \tag{10}$$

Finally, from the platform's perspective, given the platform's actual filter f and the rational consumers' conjecture f^*, consumers' welfare is defined as a sum of their psychological utilities and their action utilities:

$$W(f, f^*) := \mathbf{E}[R(y, f^*)] + U(f, f^*). \tag{11}$$

4.2 Equilibrium with ethical concern

We next define the platform's ethical concern and establish the unique equilibrium in the presence of this concern. The payoff of a platform with ethical concern is given by a weighted sum of its revenue and the consumers' welfare net of its filtering cost:

$$(1 - \phi)\left(\mathbf{E}\left[R(y, f^*)\right] - \frac{cf^2}{2}\right) + \phi W(f, f^*), \tag{12}$$

where $\phi \in (0, 1)$ is an exogenous parameter that captures the strength of this platform's ethical concern, $R(y, f^*)$ is given in (3), and $W(f, f^*)$ is given in (11).

Note that this payoff (12) is proportional to

$$\mathbf{E}\left[R(y,f^*)\right]+\frac{\phi}{1-\phi}W(f,f^*)-\frac{cf^2}{2}.$$

Because our results concern how the strength of the platform's ethical concern affects equilibrium outcomes, without loss of generality, we work with the following normalization of this platform's payoff in what follows:

$$\mathbf{E}\left[R(y,f^*)\right]+\frac{h}{\beta+\tau+\xi}W(f,f^*)-\frac{cf^2}{2}, \tag{13}$$

where $h>0$ is an exogenous parameter that captures the (normalized) strength of this platform's ethical concern, and the fraction $1/(\beta+\tau+\xi)$ is a normalizing constant. The parameter h can alternatively be interpreted as capturing the strength of the platform's reputation concern for performing its social responsibilities. The model is otherwise identical to the basic model in Section 2. As will be evident in (14) below, this normalized payoff (13) permits a sharp comparison of equilibrium outcomes in case of a self-interested platform and in case of a platform with ethical concern.

Proposition 2 below characterizes the unique equilibrium.

Proposition 2. *There is a unique equilibrium. In this equilibrium, the platform chooses filter* $f^{\mathrm{E}}\equiv f^{\mathrm{E}}(\beta,c,p,q,r,h)>0$ *characterized by*

$$(\beta+h)\left(\frac{r}{(p+q+f^{\mathrm{E}})^2}+\frac{1-r}{(q+f^{\mathrm{E}})^2}\right)=cf^{\mathrm{E}}, \tag{14}$$

The filter f^{E} *strictly exceeds* f^{S}*, is strictly increasing in* (β,h)*, and is strictly decreasing in* (r,c,p,q)*.*

As in the basic model, given the rational consumers' conjecture and their signals, the platform's actual filter does not affect the consumers' inferences. Unlike in the basic model, the platform benefits from choosing a higher filter to reduce the dispersion of the consumers' signals, thereby improving the consumers' psychological utilities and action utilities. Thus, the ethical concern boosts the platform's filter, and the

filter strictly increases in the strength h. The results of the comparative statics with respect to the other parameters concerning filter f^{E} are analogous to those concerning filter f^{S}.

5 Consumers' welfare and conflicts

This section reports our main results. We demonstrate how the platform's ethical concern to internalize consumers' welfare could perversely harm consumers' welfare, and show that this backfire occurs only because the platform's ethical concern aggravates consumers' equilibrium conflicts.

5.1 Equilibrium consumers' welfare

To begin, we consider the structure of consumers' equilibrium welfare. Define the aggregate consumers' psychological utilities in an equilibrium with filter f as:

$$W_P(f) := \mathbf{E}\left[-\beta \int_0^1 \mathbf{E}_i[(\theta - b_i)^2 | y_i] \,\mathrm{d}i\right]. \tag{15}$$

In (15), we do not distinguish between the rational consumers' conjecture and the platform's actual filter, as their conjecture is correct in equilibrium. Similarly, in this equilibrium, define the aggregate consumers' benefit from matching their actions with the state and the negative externalities as:

$$W_A(f) := \mathbf{E}\left[-\tau \int_0^1 (\alpha_i^f(y_i) - \theta)^2 \,\mathrm{d}i\right] \quad \text{and} \quad W_C(f) := \mathbf{E}\left[-\xi \int_0^1 \int_0^1 \kappa_{kj}(y, f) \,\mathrm{d}k \,\mathrm{d}j\right].$$

Proposition 3. *The following holds.*

1. *$W_P(f)$ and $W_A(f)$ are strictly increasing in f.*

2. *For any r sufficiently close to one, $W_C(f)$ is single-dipped: for some cutoff $\bar{f}_C \geq 0$, $W_C(f)$ is strictly decreasing on $[0, \bar{f}_C)$ and is strictly increasing on $[\bar{f}_C, \infty)$.*

Part 1 of this proposition is intuitive. Given a more aggressive filter, the consumers' signals are less dispersed (around the true state), thereby improving the consumers' psychological utilities and their utilities from matching actions with the state.

Part 2 of this proposition hints at our main message: the platform's ethical concern, by boosting the platform's filter, could reduce consumers' welfare by perversely aggravating consumers' conflicts. Ironically, by using (11) to define aggregate consumers' welfare in an equilibrium with filter f as

$$W(f) := W(f, f) = W_P(f) + W_A(f) + W_C(f),$$

the perverse welfare reduction happens precisely when consumers are sufficiently worried about negative externalities driven by social conflicts, i.e., when ξ is sufficiently high, provided that most consumers are rational:

Corollary 1. *There exists $\underline{\xi} \equiv \underline{\xi}(\tau, p, q) \in [0, \infty]$ such that the following holds.*

1. *If $\xi \leq \underline{\xi}$, then $W(f)$ is strictly increasing.*
2. *If $\xi > \underline{\xi}$, then there exists $\underline{r}_{\xi,\tau} \in (0, 1)$ such that for every $r \geq \underline{r}_{\xi,\tau}$, $W(f)$ is single-dipped: for some cutoff $\bar{f}_{\xi,\tau,r} > 0$, $W(f)$ is strictly decreasing on $[0, \bar{f}_{\xi,\tau,r})$ and is strictly increasing on $[\bar{f}_{\xi,\tau,r}, \infty)$.*

This cutoff $\underline{\xi} \equiv \underline{\xi}(\tau, p, q)$ satisfies the following conditions:

1. *If $p \leq q$, then $\underline{\xi} = \infty$.*
2. *If $p > q$, then $\underline{\xi} < \infty$. Moreover, $\underline{\xi}(\tau, p, q)$ is strictly increasing is τ, strictly decreasing in p, and strictly increasing in q.*

The reason underlying this corollary is intuitive. If ξ is high, then conflicts play an important role in determining consumers' welfare; this welfare thus reflects the aggravating conflicts driven by the platform's ethical concern.

In view of Proposition 3 and Corollary 1, hereafter, we characterize the structure of equilibrium conflicts. This characterization elucidates when the platform's ethical concern would perversely aggravate these conflicts.

5.2 Equilibrium conflicts

Proposition 4 below elucidates the structure of the equilibrium conflict cost and the implications of the platform's ethical concern on this cost. It is useful to define, for any two consumers i and j,

$$K_{ij}(f) := \mathbf{E}[\kappa_{ij}(y, f)] \tag{16}$$

as the (expected) conflict cost between the two consumers in an equilibrium where the filter is f. Note that in (16), we again do not distinguish between the rational consumers' conjecture and the platform's actual filter, as their conjecture is correct in equilibrium.

Proposition 4. *The following holds.*

1. *The ethical concern reduces the equilibrium conflict cost between any two rational consumers if and only if the concern is sufficiently strong: there exists $\bar{h} \geq 0$ such that for any two rational consumers i and j, $K_{ij}(f^{\mathrm{S}}) > K_{ij}(f^{\mathrm{E}})$ if and only if $h > \bar{h}$.*

2. *The ethical concern unambiguously reduces the equilibrium conflict cost between any rational consumer i and any credulous consumer j: $K_{ij}(f^{\mathrm{S}}) > K_{ij}(f^{\mathrm{E}})$.*

3. *The ethical concern unambiguously reduces the equilibrium conflict cost between any two credulous consumers i and j: $K_{ij}(f^{\mathrm{S}}) > K_{ij}(f^{\mathrm{E}})$.*

Thus, on an aggregate level, the platform's ethical concern mitigates equilibrium conflicts if and only if this concern is sufficiently strong, provided that the proportion of rational consumers is sufficiently large.

Corollary 2. *There exists $\bar{r} \in (0, 1)$ such that if the mass of rational consumers is $r \geq \bar{r}$, then there exists $\hat{h} \geq 0$ such that the platform's ethical concern reduces the overall equilibrium conflict cost, i.e.,*

$$\int_0^1 \int_0^1 K_{ij}(f^{\mathrm{E}}) \, \mathrm{d}i \, \mathrm{d}j < \int_0^1 \int_0^1 K_{ij}(f^{\mathrm{S}}) \, \mathrm{d}i \, \mathrm{d}j,$$

if and only if the strength of the platform's ethical concern h satisfies $h > \hat{h}$.

To understand Proposition 4, let us write each consumer i's state estimate $\mathbf{E}_i[\theta|y_i]$ given her signal y_i and equilibrium filter f as $\mathbf{E}_i[\theta|y_i] = w_i(y_i - r_i s_i)$, where

$$w_i = \begin{cases} \dfrac{q+f}{p+q+f}, & \text{if } i \text{ is rational,} \\ 1, & \text{if } i \text{ is credulous,} \end{cases}$$

represents the weight that the consumer places on her signal and

$$r_i = \begin{cases} 1, & \text{if } i \text{ is rational,} \\ 0, & \text{if } i \text{ is credulous,} \end{cases} \tag{17}$$

captures the consumer's ability to remove the slant in her signal. Then, the equilibrium conflict cost (16) between any two consumers i and j simplifies to

$$K_{ij}(f) = \frac{1}{2} \cdot \left[\underbrace{((1-r_i)s_i - (1-r_j)s_j)^2}_{\text{term A}} + \underbrace{(w_i^2 + w_j^2)\frac{1}{q+f}}_{\text{term B}} + \underbrace{(w_i - w_j)^2 \frac{1}{p}}_{\text{term C}} \right]. \tag{18}$$

This cost is driven by the disagreeing misinformation ε_i and ε_j that the two consumers receive in their signals as well as the (potentially) disagreeing weights they place on the signals. The cost due to the disagreement in misinformation arises from the different slants of the consumers' news subscriptions, as captured by term A in (18), as well as the dispersion of misinformation, as captured by term B. Finally, the cost due to the disagreement in weights is captured by term C. This disagreement implies that the two consumers place different weights on the prior state distribution; this disagreement aggravates their conflict because the dispersion of their signals is partly driven by the dispersion of the state, as measured by $1/p$.

Let us first consider part 1 of the proposition. As the two rational consumers i and j remove their slants and place identical weights on their signals in equilibrium,

their conflict cost is driven only by term B in (18):

$$K_{ij}(f) = \left(\frac{q+f}{p+q+f}\right)^2 \frac{1}{q+f}. \tag{19}$$

A higher equilibrium filter induced by the platform's ethical concern has two opposing effects on their conflict cost. It improves their learning about the state. This learning effect mitigates their conflict. There is also a confidence effect that aggravates their conflict: the consumers correctly anticipate the higher filter and thus place higher weights on their own signals. Thus, the platform's ethical concern reduces the conflict cost if and only if the learning effect dominates the confidence effect or, equivalently, if and only if the strength of the concern h is sufficiently large that the filter f^{E} given ethical concern is sufficiently larger than the self-interested filter f^{S}. More precisely, part 1 follows because (19) is single-peaked at $f = \max(0, p-q)$. If $f < p-q$ (resp., $f > p-q$), then the prior state precision p exceeds (resp., falls short of) the signal precision $f+q$ such that an infinitesimal change in the equilibrium filter aggravates (resp., mitigates) their conflicts, as the confidence effect dominates (resp., is dominated by) the learning effect.[10]

Why could this perverse outcome happen? The platform understands that a higher equilibrium filter could aggravate conflicts. But given the rational consumers' conjecture, the platform also understands that its actual filter affects the dispersion of the signals but does not affect the weights that the consumers place on their signals. Thus, given any conjecture f^*, the platform internalizes the cost

$$\left(\frac{q+f^*}{p+q+f^*}\right)^2 \left(\frac{1}{q+f}\right)$$

instead of (19). This causes the platform to filter more aggressively relative to the setting absent ethical concern so as to reduce the signal dispersion. But then the rational consumers correctly anticipate this incentive of the platform and form their

[10]Notably, this race between the learning effect and the confidence effect, as well as the possibility that the confidence effect dominates the learning effect, are well documented in experimental findings (see, e.g., Hall, Ariss and Todorov, 2007; Tsai, Klayman and Hastie, 2008).

conjecture f^* accordingly, yielding the perverse outcome.[11]

Consider then part 2. Unlike the rational consumers, the credulous consumers do not remove their slants and they place a full weight on their signals for inferences. The equilibrium conflict cost between a rational consumer i and a credulous consumer j is therefore driven by all three terms in (18):

$$K_{ij}(f) = \frac{1}{2}\left[s_j^2 + \frac{1}{q+f} + \frac{1}{p+q+f}\right]. \tag{20}$$

Part 2 then follows because (20) is strictly decreasing in f. The underlying intuition, nonetheless, is not straightforward. Indeed, the rational consumer correctly conjectures the higher filter given the platform's ethical concern and so the confidence effect that aggravates conflicts remains present. This is captured by an increase in her weight in term B in (18). Nonetheless, contrary to the conflict between any two rational consumers, a boost of the rational consumer's confidence here reduces her disagreement with the credulous consumer concerning how much weight to place on their signals. This latter effect is captured by term C in (18) and mitigates the conflict aggravation caused by the confidence effect. Finally, because the credulous consumer assigns a full weight on her signal, the learning effect is stronger and the confidence effect is absent in her inference. Overall, the ethical concern mitigates these two consumers' conflict.

Consider finally part 3. The equilibrium conflict cost between any two credulous consumers i and j is driven only by terms A and B in (18), because these two consumers place the same weight on their signals:

$$K_{ij}(f) = \frac{1}{2}(s_i - s_j)^2 + \frac{1}{q+f}. \tag{21}$$

Part 3 then follows because (18) is strictly decreasing in f. The underlying intuition

[11] Readers familiar with the literature on global games may wish to compare the present result to a key takeaway of that literature where consumers place "too much" weight (relative to the socially desirable level) on prior, public information because of the strategic complementarity of their actions. Here, consumers take no actions, let alone exhibit strategic complementarity, and Proposition 4 highlights that consumers put "too much" weight on their own signals (relative to the case in Section 6 where the filter is publicly observable and no perverse outcome arises) in response to the platform's incentives.

here is straightforward. There is only learning effect but no confidence effect; moreover, the two consumers do not disagree on the weight to place on their signals.

Let us rank these three equilibrium conflict costs (19), (20), and (21). As the credulous consumers do not discount their signals, (20) and (21) are strictly higher than (19). Thus, if we interpret the credulous consumers as suffering from an overconfidence bias by believing that the signals are more precise about the state than they actually are (see, e.g., Moore and Healy, 2008), our ranking of the conflict costs sheds light on empirical evidence that overconfidence exacerbates conflicts (see, e.g., Ortoleva and Snowberg, 2015). Moreover, by showing that the ethical concern unambiguously reduces (20) and (21), Proposition 4 highlights that the ethical concern mitigates such exacerbation. On the other hand, the ranking between (20) and (21) is ambiguous. The conflict cost due to signal dispersion is smaller in (20) as the rational consumer discounts her signal, but the conflict cost due to slanting could be smaller in (21) if both credulous consumers' news subscriptions are slanted similarly.

Finally, to understand how strong the ethical concern needs to be to preempt the perverse outcome among the rational consumers, Proposition 5 below analyzes the threshold $\bar{h}$ as determined in Proposition 4. In the remainder of this paper, we write f^{S} as $f^{\mathrm{S}}(p)$ wherever appropriate to emphasize the filter's dependence on p. Recall that $f^{S}(p)$ is strictly decreasing in p by Proposition 1. Thus, the signal precision exceeds the prior (state) precision, i.e., $q + f^{\mathrm{S}}(p) \geq p$, if and only if p is small enough.

Proposition 5. *There exists $\bar{p} > 0$ such that:*

1. *If the prior precision is $p > \bar{p}$, then $\bar{h} \equiv \bar{h}(p) > 0$ and is strictly increasing in p.*

2. *If the prior precision is $p \leq \bar{p}$, then $\bar{h} \equiv \bar{h}(p) = 0$.*

This proposition reflects the race between the learning effect and the confidence effect. Given a large p, the prior precision p exceeds the signal precision $q + f^{\mathrm{S}}$. Unless the ethical concern is strong enough to ensure a large enough filter f^{E}, the learning effect is dominated by the confidence effect given the change in the filter from f^{S} to f^{E}. In contrast, given a small p, the signal precision exceeds the prior precision. Upon

a change in the filter from f^{S} to f^{E}, the learning effect dominates the confidence effect regardless of how weak the ethical concern is.

6 Government efforts

In this section, we return to our basic model with a self-interested platform. We apply our main results in Section 5 to examine several popular government efforts to mitigate conflicts. Funke and Flamini (2021) survey these efforts worldwide. Propositions 6 and 7 below begin with efforts that target the platform. The main takeaway is that these efforts echo Proposition 4, namely that these efforts must be aggressive enough to not perversely aggravate equilibrium conflicts.

Legislation. We first consider legislation or proposals of legislation that holds the platform accountable for the misinformation that it fails to censor from the news reports, ensuring sufficient filtering. One example is the modification and elimination of platforms' immunity under Section 230 of the Communications Decency Act; this immunity is commonly viewed as a "legal shield" that protects platforms from liability for third-party content that they host.[12]

To capture such legislation, we consider a filter floor $\underline{f} > f^{\mathrm{S}}$, where the filter f^{S} is characterized by (6), such that the platform's choice of filter must exceed $\underline{f}$. There is then a unique equilibrium where the platform chooses filter $f^{\mathrm{L}} = \underline{f}$. Proposition 6 below is a direct application of Propositions 4 and 5.

Proposition 6. *The following holds.*

1. *If the prior precision p is small, then the floor reduces the equilibrium conflict cost between any two rational consumers i and j: if $q + f^{\mathrm{S}}(p) \geq p$, then $K_{ij}(f^{\mathrm{L}}) < K_{ij}(f^{\mathrm{S}})$. Otherwise, the floor reduces their equilibrium conflict cost if and only if the floor is sufficiently high: there exists $F > f^{\mathrm{S}}$ such that $K_{ij}(f^{\mathrm{L}}) < K_{ij}(f^{\mathrm{S}})$ if and only if $\underline{f} > F$.*

[12] See, "Legal shield for social media is targeted by lawmakers," *The New York Times*, May 28, 2020.

2. *The floor reduces the equilibrium conflict cost between any rational consumer i and any credulous consumer j:* $K_{ij}(f^{\mathrm{L}}) < K_{ij}(f^{\mathrm{S}})$.

3. *The floor reduces the equilibrium conflict cost between any two credulous consumers i and j:* $K_{ij}(f^{\mathrm{L}}) < K_{ij}(f^{\mathrm{S}})$.

Arrests and cyber task forces. We next consider arrests of misinformation spreaders and cyber task forces against misinformation campaigns. For example, in 2018, Thai authorities issued warrants for the arrest of 29 people for sharing or liking false claims on *Facebook*;[13] in the same year, the British government set up the National Security Communications Unit against misinformation campaigns.[14]

We model these efforts as an increase in the default precision absent filtering from an initial value q^{B} to some $q^{\mathrm{A}} > q^{\mathrm{B}}$ and denote the corresponding equilibrium filters as characterized in Proposition 1 by f^{B} and f^{A}. Proposition 7 below, again, is a direct application of Propositions 4 and 5.

Proposition 7. *The following holds.*

1. *If the prior precision p is small, then the increase in default signal precision reduces the equilibrium conflict cost between any two rational consumers i and j: if $q^{\mathrm{B}} + f^{\mathrm{B}}(p) \geq p$, then $K_{ij}(f^{\mathrm{A}}) < K_{ij}(f^{\mathrm{B}})$. Otherwise, it reduces their equilibrium conflict cost if and only if the increase is sufficiently large: there exists $Q > q^{\mathrm{B}}$ such that $K_{ij}(f^{\mathrm{A}}) < K_{ij}(f^{\mathrm{B}})$ if and only if $q^{\mathrm{A}} > Q$.*

2. *The increase in default signal precision reduces the equilibrium conflict cost between any rational consumer i and any credulous consumer j:* $K_{ij}(f^{\mathrm{A}}) < K_{ij}(f^{\mathrm{B}})$.

3. *The increase in default signal precision reduces the equilibrium conflict cost between any two credulous consumers i and j:* $K_{ij}(f^{\mathrm{A}}) < K_{ij}(f^{\mathrm{B}})$.

[13] See "Thai government steps up efforts to crack down on fake news," *The South China Morning Post*, 14 June, 2018.

[14] See "Government announces anti-fake news unit," *BBC*, 23 January, 2018.

The intuition of Proposition 7 is analogous to that of Proposition 6, and so their statements share an analogous structure. While a higher default precision q undermines the platform's filtering incentives in view of Proposition 1, the overall precision of the platform's signal increases by a direct application of the implicit function theorem on (6), i.e., $q^B + f^B < q^A + f^A$. The effect on the equilibrium conflicts given a higher default precision is thus identical to that given a fixed default precision and a higher filter, which is the case in Proposition 6.

Transparency. We next analyze a potential regulatory effort on platform transparency that is commonly discussed in policy debates (see, e.g., MacCarthy, 2020). Calls for transparency are typically motivated by the conventional wisdom that transparency is essential to accountability measures for platforms and consumer protection. Proposition 8 below provides an additional case for transparency by highlighting that transparency allows the platform to correctly internalize its social responsibility.

In our model, if the platform's filter is publicly observable, then the rational consumers draw inferences based on its actual filter instead of their conjectured filter. Thus, by choosing filter f, contrary to (4), the payoff of a self-interested platform is

$$\mathbf{E}\left[R\left(y;f\right)\right] - \frac{cf^2}{2}, \tag{22}$$

and contrary to (12), the payoff of a platform with ethical concern is

$$\mathbf{E}\left[R(y,f)\right] + \frac{h}{\beta+\tau+\xi}W(f,f) - \frac{cf^2}{2}, \tag{23}$$

where the expectation $\mathbf{E}$ is taken with respect to the actual filter f. Observability of the platform's filter removes the credence nature of the platform's signals and thus rules out the perverse outcome:

Proposition 8. *Suppose that the platform's filter is observable to the consumers. Given any equilibrium filter $\tilde{f}^{\mathrm{S}}$ in the model with a self-interested platform and any equilibrium filter $\tilde{f}^{\mathrm{E}}$ in the model where the platform faces ethical concern, $W(\tilde{f}^{\mathrm{E}}) \geq W(\tilde{f}^{\mathrm{S}})$.*

Fairness doctrine. In the presence of slanted information, some media scholars advocate a version of the FCC fairness doctrine for online media (see, e.g., Napoli, 2019, 2021). The doctrine was originally applied to radio and television broadcasters, requiring that the broadcasters provide a fair and balanced presentation of information. We model this doctrine by supposing that the platform intervenes the consumers' news subscriptions so that each consumer i's slant s_i is fixed at zero. To emphasize the dependence of slants on conflict cost, we write $K_{ij}(f; s_i, s_j)$ as the conflict cost between consumers i and j with slants s_i and s_j in an equilibrium with filter f. By (6), the filter is determined independently of the slants in equilibrium. Thus, the equilibrium filter given the doctrine remains as f^{S} and the doctrine changes the equilibrium conflict cost between consumers i and j from $K_{ij}(f^{\mathrm{S}}; s_i, s_j)$ to $K_{ij}(f^{\mathrm{S}}; 0, 0)$.

Proposition 9. *The following holds.*

1. *The doctrine does not affect the equilibrium conflict cost between any two rational consumers i and j: $K_{ij}(f^{\mathrm{S}}; s_i, s_j) = K_{ij}(f^{S}; 0, 0)$.*

2. *The doctrine unambiguously reduces the equilibrium conflict cost between any two consumers i and j in which at least one of them is credulous: $K_{ij}(f^{\mathrm{S}}; 0, 0) \leq K_{ij}(f^{\mathrm{S}}; s_i, s_j)$; in addition, if $(1 - r_i)s_i - (1 - r_j)s_j \neq 0$, where r_i, r_j are defined in (17), then $K_{ij}(f^{\mathrm{S}}; 0, 0) < K_{ij}(f^{\mathrm{S}}; s_i, s_j)$.*

This proposition follows from our discussion of Proposition 4 in Section 5: in equilibrium, a consumer's slant adds to the conflict cost only if the consumer is credulous. Note that the FCC eliminated the doctrine for broadcasters in 1987. The core justification for the elimination was that the doctrine was no longer necessary, as the growing number of media outlets available facilitated consumers' access to diverse information. Proposition 9 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, the phenomenon of "echo chambers" where consumers choose to read certain (slanted) content and omit others is prevalent. These consumers include those who are credulous and hence lack the sophistication to utilize the slanted information.

Media literacy campaign. Finally, we turn to a government effort that targets the consumers rather than the platform. Specifically, we consider a media literacy campaign under which the credulous consumers become rational before the platform chooses its filter, and this event is common knowledge. The platform's equilibrium filter given the campaign is plainly characterized by (6) in Proposition 1 with $r = 1$ being imposed; we denote by f^{M} this filter given a campaign. We continue to denote by f^{S} the equilibrium filter absent a campaign. By Proposition 1, the platform filters less aggressively when the mass of rational consumers is larger, i.e., $f^{\mathrm{M}} < f^{\mathrm{S}}$. In what follows, we write f^{M} as $f^{\mathrm{M}}(p)$ wherever appropriate to emphasize its dependence on the prior precision p.

Proposition 10. *The following holds.*

1. *If the prior precision p is small, then the media literacy campaign increases the equilibrium conflict cost between any two consumers i and j who are rational absent the campaign: if $q + f^{\mathrm{M}}(p) \geq p$, then $K_{ij}(f^{\mathrm{M}}) > K_{ij}(f^{\mathrm{S}})$. Otherwise, the campaign reduces the equilibrium conflict cost if and only if the mass of rational consumers absent the campaign is large: there exists $\bar{r} \in [0, 1)$ such that $K_{ij}(f^{\mathrm{M}}) \leq K_{ij}(f^{\mathrm{S}})$ if and only if $r \geq \bar{r}$.*

2. *The media literacy campaign reduces the equilibrium conflict cost between any consumers i and j in which at least one of them is credulous in the absence of the campaign: $K_{ij}(f^{\mathrm{M}}) < K_{ij}(f^{\mathrm{S}})$.*

Consider part 1 of the proposition. When the prior precision p is small, the learning effect dominates the confidence effect; thus, the fall in the equilibrium filter due to the campaign unambiguously aggravates conflicts. In contrast, when p is large, the confidence effect dominates the learning effect such that the fall in the equilibrium filter aggravates conflicts if and only if the fall is large enough, which is the case when the mass of rational consumers before the campaign was small. Notably, this part of the proposition contrasts with the implications of the aforementioned efforts targeting the platform. As we have seen, the efforts targeting the platform which unambiguously mitigate conflicts when p is small but possibly aggravates conflicts when p is large.

Finally, part 2 follows because the credulous consumers learn to discount their signals in their inferences given the campaign.

Combined efforts. The preceding discussion points to an appeal of implementing a media literacy campaign when efforts targeting the platform's filtering are already in place, which is indeed a common practice:

Corollary 3. *Given a filtering floor $\underline{f} \geq f^{\mathrm{S}}$, where f^{S} is characterized by (6), implementing a media literacy campaign unambiguously reduces the equilibrium conflict cost between any two consumers i and j.*

This result follows because absent a campaign, the platform filters at the binding level $\underline{f}$; given a campaign, the equilibrium filter remains as $\underline{f}$ and all consumers discount their signals.

7 Discussion

In this section, we briefly describe several ways in which our model can be enriched and identify some open questions.

Private information acquisition. The basic model has assumed that consumers only learn about the state from the platform's signals. We can extend the basic model by assuming that each consumer i receives a private signal x_i about the state in addition to the platform's signal y_i, where $x_i = \theta + \eta_i$ and η_i is normally distributed with mean 0 and precision $z > 0$. This noise η_i is drawn independently of θ and $(y_j)_{j\in[0,1]}$, and independently across consumers. We continue to assume that each credulous consumer's state estimate is given by the signal she receives.[15] In contrast, each rational consumer i's state estimate, given her signals (x_i, y_i) and conjecture f^*,

[15]The results go through if the credulous consumer believe that the state is a weighted average of her two received signals. We do not consider this case plainly for simplicity. The purpose of this extension is to show that the platform's incentive to reduce signal dispersion in the basic model carries over.

is

$$\mathbf{E}^*[\theta|x_i, y_i] = \frac{z}{p+z+q+f^*}x_i + \frac{q+f^*}{p+z+q+f^*}(y_i - s_i),$$

by standard Bayesian updating. Different from (8) in the basic model, the consumer incorporates the signal x_i in her estimate. Nonetheless, the platform's filtering incentives to reduce the dispersion of its signals $(y_i)_{i\in[0,1]}$ remain present. Thus, our results extend in a straightforward manner.

Slant manipulation. The basic model has abstracted from the platform's ability to manipulate each consumer's slant by recommending certain (biased) news sources for the consumer to subscribe to, such as *Twitter*'s "Suggested Follows" listings. We explore this extension thoroughly in the online Appendix. Specifically, we assume that in addition to choosing a filter f, the platform chooses a slant manipulation $m_i \in \mathbf{R}$ for each consumer i at a cost. The consumer's signal remains as given by $y_i = \theta + \varepsilon_i$, but ε_i is now normally distributed with mean $s_i + m_i$ and precision $q + f$.

We show in the online Appendix that in this extension, our main insights extend. Briefly, we show that the equilibrium filter with or without ethical concern is unaffected by the slant manipulation. This is because the platform's filtering incentive to reduce signal dispersion is orthogonal to manipulating the mean of each consumer's signal. On the other hand, irrespective of whether the platform faces an ethical concern, the platform typically manipulates the slants for all consumers towards their individual biases. In equilibrium, the rational consumers correctly anticipate the manipulations on their slants and thus correctly remove their manipulated slants in their inferences. Thus, the equilibrium conflict cost between any two rational consumers is determined by the platform's filter as in the main analysis. Moreover, as in the main analysis, the more aggressive equilibrium filter due to the ethical concern reduces the aggregate cost of conflicts involving the credulous consumers. Finally, as the credulous consumers fail to eliminate their manipulated slants in their inferences, the ethical concern causes the platform to adjust its slant manipulation to further reduce the aggregate cost of conflicts involving these consumers.

Beyond the quadratic-normal setup. The quadratic-normal specification of our basic model has afforded much tractability, as well as a sharp characterization of the equilibrium conflict cost and the corresponding implications of ethical concern. We conjecture that our main results, Propositions 4 and 5, extend qualitatively to more general environments. After all, the two drivers of our main results are as follows. First, the platform's signals have a credence nature so that rational consumers form conjectures about the filter and the ethical concern causes the platform to boost its filter in response to the conjectures. Second, the rational consumers place a smaller weight on the signals in their inferences given a higher prior state precision.

Other forms of ethical concern. We have limited our attention of ethical concern to the context of conflicts incited by online misinformation. There are other conflict sources from which our analysis abstracts, such as the role of platforms' recommendation algorithms in spreading hate speeches or controversial information (see, e.g., Müller and Schwarz, 2018, 2020; Karell, 2021) and in coordinating protests (see, e.g., Enikolopov, Makarin and Petrova, 2020). Further, the strategic implications of ethical concern in other contexts such as privacy, addiction, and fairness remain open. We leave these issues to future work.

Appendices

A Proofs

Throughout Appendix A, given (p, q), we define $A : \mathbf{R}_+ \to (0, 1)$ such that

$$A(f) := \frac{q + f}{p + q + f}.$$

Note that $A(f)$ is the weight that each rational consumer places on her signal when forming a state estimate, given a conjecture f of the platform's filter.

A.1 Proof of Proposition 1

Given the platform's actual choice f and the rational consumers' conjecture f^*,

$$\mathbf{E}\Big[\mathbf{E}^*[(\theta - b_i)^2|y_i]\Big] = \mathbf{E}\Big[(\mathbf{E}^*[\theta|y_i] - b_i)^2 + \mathbf{Var}^*[\theta|y_i]\Big] = \frac{A(f^*)^2}{pA(f)} + b_i^2 + \frac{1}{p+q+f^*},$$
$$\mathbf{E}\left[(y_i - \theta)^2\right] = \frac{1}{pA(f)} + (s_i - b_i)^2.$$

The platform's expected revenue is therefore

$$\mathbf{E}\left[R(y, f^*)\right] = -\beta\left[\frac{1 - r + rA(f^*)^2}{pA(f)} + \frac{r}{p + q + f^*} + \int_0^r b_i^2 \,\mathrm{d}i + \int_r^1 (s_i - b_i)^2 \,\mathrm{d}i\right]. \quad (24)$$

The first-order condition of the platform's payoff with respect to f is

$$\beta\left[\frac{rA(f^*)^2 A'(f)}{pA(f)^2} + \frac{(1 - r)A'(f)}{pA(f)^2}\right] = cf.$$

In equilibrium, this first-order condition must hold with $f = f^*$. Thus, f^* must satisfy

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

When the rational citizens' conjecture is $f^* = f^{\mathrm{S}}$, we obtain the equilibrium filter f^{S} as characterized in the proposition.

A.2 Proof of Proposition 2

By direct calculations,

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

Hence, the aggregate expected cost of the conflict of each rational consumer i against all other consumers is

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

Similarly, the aggregate expected cost of the conflict of each credulous consumer i against all other consumers is

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

Therefore,

$$\mathbf{E}\left[\int_0^1 \int_0^1 \kappa_{kj}(y, f^*)\,\mathrm{d}k\,\mathrm{d}j\right] = \frac{rA(f^*)^2 + 1 - r}{q+f} + \frac{r(1-r)p}{(p+q+f^*)^2} + C \tag{25}$$

where C represents a collection of terms that are independent of f and f^*. On the other hand,

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

and thus

$$\mathbf{E}\left[\int_0^1(\alpha_i^{f^*}(y_i)-\theta)^2\,\mathrm{d}i\right]=\frac{r}{p+q+f^*}+\frac{1-r}{q+f}+\int_r^1 s_i^2\,\mathrm{d}i.$$

Summing up,

$$\begin{aligned}U(f,f^*)&=-\tau\mathbf{E}\left[\int_0^1(\alpha_i^{f^*}(y_i)-\theta)^2\,\mathrm{d}i\right]-\xi\mathbf{E}\left[\int_0^1\int_0^1\kappa_{kj}(y,f^*)\,\mathrm{d}k\,\mathrm{d}j\right]\\&=-(\tau+\xi)\frac{rA(f^*)^2+1-r}{q+f}-\xi C.\end{aligned}$$

The platform's expected payoff can therefore be written as

$$\mathbf{E}\left[R(y,f^*)\right]-\frac{cf^2}{2}+\frac{h}{\beta+\tau+\xi}\left[-(\beta+\tau+\xi)\frac{rA(f^*)^2+1-r}{q+f}\right]-\frac{h\xi}{\beta+\tau+\xi}C,$$

The first-order condition of this payoff function with respect to f, together with the equilibrium condition that the first-order condition must hold at $f=f^*$ as in the proof of Proposition 1, yields the equilibrium filter f^{E} as characterized in the proposition.

A.3 Proof of Proposition 3

To prove Proposition 3-(i), note first that, by the definition of $W_A(f)$ and $\alpha_i^f(y)$,

$$\begin{aligned}W_A(f)&=\mathbf{E}\left[-\tau\int_0^1(\alpha_i^f(y_i)-\theta)^2\,\mathrm{d}i\right]=-\tau\mathbf{E}\left[\int_0^r(\mathbf{E}[\theta|y_i]-\theta)^2\,\mathrm{d}i+\int_r^1(y_i-\theta)^2\,\mathrm{d}i\right]\\&=-\tau\left[\int_0^r\mathbf{Var}[\theta|y_i]\,\mathrm{d}i+\int_r^1\mathbf{Var}[\varepsilon_i]\,\mathrm{d}i\right]=-\tau\left[\frac{r}{p+q+f}+\frac{1-r}{q+f}\right].\end{aligned}$$

Hence, $W_A(f)$ strictly increases in f. Next, from (24),

$$W_P(f) = \mathbf{E}[R(y,f)] = -\beta\left[\frac{1-r}{pA(f)} + \frac{rA(f)}{p} + \frac{r}{p+q+f} + \int_0^r b_i^2\,\mathrm{d}i + \int_r^1 (s_i - b_i)^2\,\mathrm{d}i\right]$$
$$= -\beta\left[\frac{1-r}{pA(f)} + \frac{r}{p} + \int_0^r b_i^2 \mathrm{d}i + \int_r^1 (s_i - b_i)^2 \mathrm{d}i\right]. \tag{26}$$

$A(q) = (q+f)/(p+q+f)$ strictly increases in f, so does $W_P(f)$.

Next, to prove Proposition 3-(ii), note that from (25),

$$W_C(f) = \mathbf{E}\left[-\xi \int_0^1 \int_0^1 \kappa_{kj}(y,f)\,\mathrm{d}k\,\mathrm{d}j\right] = -\xi\frac{rA(f)^2 + 1 - r}{q+f} - \xi\frac{r(1-r)p}{(p+q+f)^2} - \xi C$$

where C is a collection of terms that are independent of f. Define

$$T_R(f) := \frac{rA(f)^2}{q+f} = \frac{r(q+f)}{(p+q+f)^2} \quad \text{and} \quad T_{-R}(f) := \frac{1-r}{q+f} + \frac{r(1-r)p}{(p+q+f)^2}$$

so that

$$W_C(f) = -\xi T_R(f) - \xi T_{-R}(f) - \xi C.$$

For any $r \in [0,1]$, $T_R(f)$ is strictly decreasing on $[\max\{p-q, 0\}, \infty)$ and $T_{-R}(f)$ is strictly decreasing on $[0,\infty)$. Hence, if $p - q \leq 0$, $W_C(f)$ is strictly increasing and therefore trivially single-peaked.

Let us turn to the case where $p - q > 0$. Observe first that both $T_R(f)$ and $T_{-R}(f)$ are strictly decreasing on $(\frac{3}{2}p - q, \infty)$. Thus, to complete the proof, it suffices to show that $T_R(f) + T_{-R}(f)$ is strictly concave on $[0, \frac{3}{2}p - q]$ whenever r is sufficiently close to 1. For each $r \in (0,1)$, the second derivative of $T_R(f)$, on $[0, \frac{3}{2}p - q]$, is negative and bounded as follows:

$$\frac{\mathrm{d}^2 T_R(f)}{\mathrm{d}f^2} = 2r\frac{f+q-2p}{(p+q+f)^4} \leq 2r\frac{\left(\frac{3}{2}p - q\right) + q - 2p}{(p+q+f)^4} = -\frac{rp}{(p+q+f)^4}$$
$$\leq -\frac{rp}{(p+q+\left(\frac{3}{2}p - q\right))^4} < 0.$$

On the other hand, as $r \to 1$, the second derivative of $T_{-R}(f)$ uniformly vanishes on

$[0, \frac{3}{2}p - q]$. Hence, $T_R(f) + T_{-R}(f)$ is indeed strictly concave on $[0, \frac{3}{2}p - q]$ whenever r is sufficiently close to 1.

A.4 Proof of Corollary 1

In this proof, we use the notation $W_r(f)$ to denote the expected consumer welfare (i.e., we add the subscript r) to emphasize that its value also depends on r. From the observation in Section A.3,

$$W_r(f) = -\frac{\beta(1-r)}{pA(f)} - \tau\left[\frac{r}{p+q+f} + \frac{1-r}{q+f}\right] - \xi\left[\frac{r(1-r)}{p+q+f} + \frac{1-r+r^2A(f)^2}{q+f}\right] + D$$

where $A(q) \equiv (q+f)/(p+q+f)$, and D stands for a collection of terms that are independent of f. Define

$$\begin{aligned} G_r(f) &:= -r\left[\frac{\tau}{p+q+f} + r\xi\frac{q+f}{(p+q+f)^2}\right] \\ H_r(f) &:= -(1-r)\left[\beta\frac{p+q+f}{p(q+f)} + \tau\frac{1}{q+f} + \xi\frac{r}{q+f}\right] \end{aligned}$$

so that

$$W_r(f) = G_r(f) + H_r(f) + D.$$

The first and second derivatives of $G_r(f)$ with respect to f are given by

$$\begin{aligned} G_r'(f) &= -\xi\frac{r^2(p-q-f)}{(p+q+f)^3} + r\frac{\tau}{(p+q+f)^2} \\ G_r''(f) &= 2r\frac{(-\tau-\xi r)f - \tau(p+q) + \xi r(2p-q)}{(p+q+f)^4}, \end{aligned}$$

and therefore

$$\begin{aligned} G_r'(f) < 0 \quad &\Longleftrightarrow \quad f < \underline{f}_r := \frac{\xi r - \tau}{\xi r + \tau}p - q \\ G_r''(f) > 0 \quad &\Longleftrightarrow \quad f < \bar{f}_r := \frac{2\xi r - \tau}{\xi r + \tau}p - q \end{aligned}$$

where $\underline{f}_r < \bar{f}_r$. On the other hand,

$$H_r'(f) = (1-r)\frac{\beta+\tau+\xi r}{(q+f)^2} > 0 \quad \text{and} \quad H_r''(f) = -2(1-r)\frac{\beta+\tau+\xi r}{(q+f)^3} < 0.$$

There are two subcases. First, consider the case $\xi(p-q) > \tau(p+q)$. In this case, both $\underline{f}_r$ and $\bar{f}_r$ are strictly positive for any r sufficiently close to one. Also, note that $H_r'(f)$ and $H_r''(f)$ uniformly converge to 0 over the interval $(0, 2\max_{r\in[0,1]} \bar{f}_r) = (0, 2\bar{f}_1)$. Hence, there is a threshold $\bar{r} \in (0,1)$ such that the following property holds whenever $r \in (\bar{r}, 1]$: There is $\varepsilon_r > 0$ such that the following statements hold true.

(a) $W'(f) < 0$ at any f such that $0 < f < \underline{f}_r - \varepsilon_r$.

(b) $W'(f) > 0$ at any f such that $\bar{f}_r < f < \infty$.

(c) $W''(f) > 0$ at any f such that $\underline{f}_r - \varepsilon_r < f < \bar{f}_r$.

In conclusion, for any r sufficiently close to 1, there is a threshold $f_r \in (\underline{f}_r - \varepsilon_r, \bar{f}_r)$ such that $W(f)$ decreases in f over $(0, f_r)$ and increases in f over (f_r, ∞) where both intervals $(0, f_r)$ and (f_r, ∞) are non-empty.

Next, suppose $\xi(p-q) \leq \tau(p+q)$. Then, $\underline{f}_r \leq 0$ for all $r \in [0,1]$, and hence, $W_r'(f) = G_r'(f) + H_r(f) > 0$ for any $f > 0$ and $r \in [0,1]$.

A.5 Proof of Proposition 4

Proof of Proposition 4 Recall from Proposition 2 that f^{E} is continuous and strictly increasing in h. In this proof, we will often denote f^{E} by $f^{\mathrm{E}}(h)$ to emphasize its dependence on h. Note that $f^{\mathrm{E}}(0) = f^{\mathrm{S}}$. By direct calculations of (16),

$$K_{ij}(f) = \frac{(s_i - s_j)^2}{2} + \frac{1}{q+f} \qquad \forall i,j \in (r,1], \tag{27}$$

$$K_{ij}(f) = \frac{q+f}{(p+q+f)^2} \qquad \forall i,j \in [0,r], \tag{28}$$

$$K_{ij}(f) = \frac{s_j^2}{2} + \frac{1}{2(q+f)} + \frac{1}{2(p+q+f)} \qquad \forall i \in [0,r],\ j \in (r,1]. \tag{29}$$

Parts 2 and 3 of Proposition 4 are immediate because $K_{ij}(f)$ strictly decreases in f whenever either citizen i or citizen j is credulous. To prove part 1, note first that $K_{ij}(f)$ is single-peaked at $f = \max(0, p-q)$. If $p-q \leq f^{\mathrm{S}}$, then $K_{ij}(f^{\mathrm{E}}(h)) < K_{ij}(f^{\mathrm{S}})$ for any $h > 0$. Thus, part 1 holds with $\bar{h} = 0$ for the case of $p - q \leq f^{\mathrm{S}}$. Next, consider the case of $p - q > f^{\mathrm{S}} \geq 0$. Let $h^* > 0$ be such that $f^{\mathrm{S}} < f^{\mathrm{E}}(h^*) = p - q$. Furthermore, define $\Delta K_{ij}(h) := K_{ij}(f^{\mathrm{E}}(h)) - K_{ij}(f^{\mathrm{S}})$. The function $\Delta K_{ij}(h)$ strictly increases over $[0, h^*)$ and strictly decreases over $[h^*, \infty)$. Additionally, $\Delta K_{ij}(0) = 0$ and $\lim_{h\to\infty} \Delta K_{ij}(h) < 0$ because $\lim_{h\to\infty} f^{\mathrm{E}}(h) = \infty$. Hence, there is $\bar{h} > 0$ such $\Delta K_{ij}(h) \geq 0$ if and only if $h \in [0, \bar{h}]$. Finally, because K_{ij} is independent of i and j, so is $\bar{h}$.

A.6 Proof of Corollary 2

By Proposition 3-(ii),

$$K_A(f) := -W_C(f) = \int_0^1 \int_0^1 K_{ij}(f^{\mathrm{E}})\, \mathrm{d}i$$

is single-peaked: for some cutoff $\bar{f} \geq 0$, $K_A(f)$ is strictly increasing on $[0, \bar{f}]$ and is strictly decreasing on $[\bar{f}, \infty)$. The corollary follows because $f^{\mathrm{E}} \geq f^{\mathrm{S}}$ and f^{E} strictly increases in h.

A.7 Proof of Proposition 5

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 continuous and increasing in h, where $f^{\mathrm{E}} \uparrow \infty$ as $h \uparrow \infty$. 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. The monotonicity of $f^{\mathrm{S}}(p)$ with respect to p guarantees that there exists $\bar{p} > 0$ such that $f^{\mathrm{S}}(p) < p - q$ if and only if $p > \bar{p}$.

To prove part 1, consider the case such that $p > \bar{p}$, thereby $f^{\mathrm{S}}(p) < p - q$. Note that $f^{\mathrm{S}}(p) = f^{\mathrm{E}}(0|p)$ and $f^{\mathrm{S}}(p) < f^{\mathrm{E}}(h|p)$ for any $h > 0$. Also, recall that the conflict

cost between any two rational citizens i and j is

$$K_{ij}(f) = \frac{q+f}{(p+q+f)^2}$$

given the equilibrium filter f. Because $K_{ij}(f)$ is single-peaked at $f = \max(0, p-q)$ and $f^{\mathrm{S}}(p) < p-q$, there is $\bar{h}(p) > 0$ such that $f^{\mathrm{E}}(\bar{h}(p)|p) > p-q > f^{\mathrm{S}}(p)$,

$$\frac{q+f^{\mathrm{E}}(\bar{h}(p)|p)}{(p+q+f^{\mathrm{E}}(\bar{h}(p)|p))^2} = \frac{q+f^{\mathrm{S}}(p)}{(p+q+f^{\mathrm{S}}(p))^2}, \tag{30}$$

and, therefore,

$$\frac{q+f^{\mathrm{E}}(h|p)}{(p+q+f^{\mathrm{E}}(h|p))^2} < \frac{q+f^{\mathrm{S}}(p)}{(p+q+f^{\mathrm{S}}(p))^2} \quad \text{if and only if} \quad h > \bar{h}(p).$$

It remains to show that $\bar{h}(p)$ increases in p. Differentiating both sides of (30) with respect to p, and then rearranging terms, we obtain

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

The right side of the last equation is negative:

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

where the last equality follows from (30). Hence,

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

Finally, to prove part 2, consider the case such that $p \leq \bar{p}$, thereby $f^{\mathrm{E}}(h|p) \geq f^{\mathrm{S}}(p) \geq p-q$ for all $h \geq 0$. Because $K_{ij}(f)$ is strictly decreasing on $(p-q, \infty)$, $K_{ij}(f^{\mathrm{S}}(p)) > K_{ij}(f^{\mathrm{E}}(h|p))$ at all $h \geq 0 = \bar{h}(p)$.

A.8 Proof of Proposition 6

Part 1 directly follows the fact that for any two rational citizens i and j, the mapping $f \mapsto K_{ij}(f)$ is single-peaked at $f = \max(0, p-q)$, as shown in the proof of Proposition 4. We next prove parts 2 and 3. Note that $f^{\mathrm{L}} > f^{\mathrm{S}}$ as discussed in the main text. Recall from (27) and (29) 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. In conclusion, $K_{ij}(f^{\mathrm{L}}) < K_{ij}(f^{\mathrm{S}})$ for any rational citizen i and credulous citizen j, and $K_{ij}(f^{\mathrm{L}}) < K_{ij}(f^{\mathrm{S}})$ for any two credulous citizens i and j.

A.9 Proof of Proposition 7

By Proposition 1, it holds that

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

Because $q^A > q^B$, it follows that $f^{\mathrm{B}} > f^{\mathrm{A}}$ but $q^{\mathrm{B}} + f^{\mathrm{B}} < q^{\mathrm{A}} + f^{\mathrm{A}}$.

Part 1 directly follows the fact that for any two rational citizens i and j, the mapping $z \equiv f + q \mapsto K_{ij}(z - q)$ is single-peaked at $z \equiv q + f = p$, as shown in the proof of Proposition 4. To prove the parts 2 and 3, recall from (27) and (29) 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, $K_{ij}(f^{\mathrm{A}}) < K_{ij}(f^{\mathrm{B}})$ for any rational citizen i and credulous citizen j, and $K_{ij}(f^{\mathrm{A}}) < K_{ij}(f^{\mathrm{B}})$ for any two credulous citizens i and j.

A.10 Proof of Proposition 8

Given equilibrium filter f, let $R(f) := \mathbf{E}[R(y, f)]$. Because the self-interested platform maximizes its payoff in equilibrium:

$$R(\tilde{f}^{\mathrm{S}}) - \frac{c}{2}(\tilde{f}^{\mathrm{S}})^2 \geq R(\tilde{f}^{\mathrm{E}}) - \frac{c}{2}(\tilde{f}^{\mathrm{E}})^2. \tag{31}$$

Because the platform with ethical concern also maximizes its payoff in equilibrium:

$$\begin{aligned} & R(\tilde{f}^{\mathrm{E}}) + \frac{h}{\beta+\tau+\xi} W(\tilde{f}^{\mathrm{E}}) - \frac{c}{2}(\tilde{f}^{\mathrm{E}})^2 \\ \geq\ & R(\tilde{f}^{\mathrm{S}}) + \frac{h}{\beta+\tau+\xi} W(\tilde{f}^{\mathrm{S}}) - \frac{c}{2}(\tilde{f}^{\mathrm{S}})^2 \\ \geq\ & R(\tilde{f}^{\mathrm{E}}) + \frac{h}{\beta+\tau+\xi} W(\tilde{f}^{\mathrm{S}}) - \frac{c}{2}(\tilde{f}^{\mathrm{E}})^2, \end{aligned}$$

where the last line follows from (31). These inequalities imply that $W(\tilde{f}^{\mathrm{E}}) \geq W(\tilde{f}^{\mathrm{S}})$, as desired.

A.11 Proof of Proposition 9

Part 1 is immediate because $K_{ij}(f)$ is independent of s_i and s_j if both citizens i and j are rational. To prove Part 2, note that by (27)—(29), the implementation of the fairness doctrine (which sets the manipulated slant for each citizen to zero) reduces the conflict cost between any pair of citizens. Moreover, the conflict cost is reduced strictly whenever $(1-r_i)s_i - (1-r_j)s_j \neq 0$.

A.12 Proof of Proposition 10

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

We prove part 2 first and then part 1. To show part 2, note that the first-order condition and the observation that $f^{\mathrm{M}} < f^{\mathrm{S}}$ together imply

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

and thus

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

where the first inequality in the last line holds as $(f+q)/(p+q+f)$ is strictly increasing in f. By combining these two inequalities and (27)—(29), part 2 follows.

We now prove part 1. Fix two citizens i and j who are rational without the campaign. 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 $K_{ij}(f)$ is single-peaked at $f = \max(0, p-q)$, there exists $\bar{r} \in [0,1)$ such that

$$K_{ij}(f^{\mathrm{M}}) = K_{ij}(f^{\mathrm{S}}|_{r=1}) > K_{ij}(f^{\mathrm{S}}|_{r=r_B}) = K_{ij}(f^{\mathrm{S}}) \tag{32}$$

for any $i, j \leq r_B$ if and only if $r_B < \bar{r}$. Finally, if either $f^M \geq p-q > 0$ or $p-q \leq 0$ holds, then $K_{ij}(f^{\mathrm{S}}|_{r=r_B}) = K_{ij}(f^{\mathrm{S}})$ is strictly increasing in r_B, so that (32) always holds.

A.13 Proof of Corollary 3

Let f^{S} denote the algorithm that the platform would choose without any regulations and without the media literacy campaign. Let f^{B} and f^{M} denote the filters that the platform chooses before and given a campaign, respectively, where a filter floor $\underline{f} \geq f^{\mathrm{S}}$ is imposed in both cases.

First, from Proposition 1, the platform's choice of f would strictly decrease in r if there were no filter floor. Hence, any filter floor $\underline{f} \geq f^{\mathrm{S}}$ binds both before and given a campaign, and therefore, $f^{\mathrm{B}} = f^{\mathrm{M}} = \underline{f}$. In other words, with a filter floor $\underline{f} > f^{\mathrm{S}}$, the platform chooses $f = \underline{f}$ before and given a campaign.

Next, recall from (27)—(29) that, with the filter $f = \underline{f}$ being fixed, the conflict cost between two rational citizens is strictly smaller than the conflict cost between any other possible pair of citizens. Hence, a campaign only reduces the conflict cost

between any pair of citizens.

References

Acemoglu, D., Ozdaglar, A. and Siderius, J. (2022). A Model of Online Misinformation, *Working paper, Massachusetts Institute of Technology* .

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.

Candogan, O. and Drakopoulos, K. (2020). Optimal Signaling of Content Accuracy: Engagement vs. Misinformation, *Operations Research* **68**(2): 497–515.

Chen, L. and Papanastasiou, Y. (2021). Seeding the Herd: Pricing and Welfare Effects of Social Learning Manipulation, *Management Science* **67**(11): 6734–6750.

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

Edmond, C. and Lu, Y. K. (2021). Creating Confusion, *Journal of Economic Theory* **191**: 105–145.

Enikolopov, R., Makarin, A. and Petrova, M. (2020). Social Media and Protest Participation: Evidence from Russia, *Econometrica* **88**(4): 1479–1514.

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

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.

Hall, C. C., Ariss, L. and Todorov, A. (2007). The Illusion of Knowledge: When More Information Reduces Accuracy and Increases Confidence, *Organizational Behavior and Human Decision Processes* **103**(2): 277–290.

Holmström, B. (1999). Managerial Incentive Problems: A Dynamic Perspective, *The Review of Economic Studies* **66**(1): 169–182 (Originally published in 1982 in Essays in Honor of Professor Lars Wahlbeck).

Jann, O. and Schottmuller, C. (2021). Why Echo Chambers Are Useful, *Working paper, CERGE-EI* .

Kamenica, E. and Gentzkow, M. (2011). Bayesian Persuasion, *American Economic Review* **101**(6): 2590–2615.

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.

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. (2012). Elections, Fraud, and Election Monitoring in the Shadow of Revolution, *Quarterly Journal of Political Science* **7**(3): 249–283.

Little, A. T. (2015). Fraud and Monitoring in Non-competitive Elections, *Political Science Research and Methods* **3**(1): 21–41.

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

Moore, D. A. and Healy, P. J. (2008). The Trouble with Overconfidence, *Psychological review* **115**(2): 502.

Mostagir, M. and Siderius, J. (2022). Naive and Bayesian Learning with Misinformation Policies, *Working paper, University of Michigan* .

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.

Napoli, P. M. (2021). Back from the dead (again): The Specter of the Fairness Doctrine and Its Lesson for Social Media Regulation, *Policy & Internet* .

Ortoleva, P. and Snowberg, E. (2015). Overconfidence in Political Behavior, *American Economic Review* **105**(2): 504–35.

Papanastasiou, Y. (2020). Fake News Propagation and Detection: A Sequential Model, *Management Science* **66**(5): 1826–1846.

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

Pew Research Center (2021a). More Americans Now Say Government Should Take Steps to Restrict False Information Online Than in 2018.

Pew Research Center (2021b). News Consumption Across Social Media in 2021.

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

Tsai, C. I., Klayman, J. and Hastie, R. (2008). Effects of Amount of Information on Judgment Accuracy and Confidence, *Organizational Behavior and Human Decision Processes* **107**(2): 97–105.

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* .