Perverse Ethical Concerns: Online Misinformation and Offline Conflicts*

Dongkyu Chang† Allen Vong‡

(Click here for the latest version.)

December 6, 2021

Abstract

We study a model where a large number of citizens 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 citizens. We show that a platform that faces an ethical concern to internalize conflict costs due to misinformation could perversely aggravate conflicts. This cautionary observation highlights that societal efforts to mitigate conflicts, such as investments in ethical algorithm, public awareness campaigns, and government policies, are effective if and only if their implementations are sufficiently aggressive.

JEL codes: C72, D83, L86.

Keywords: platforms, social media, polarization, conflicts.

1 Introduction

Misinformation on social media platforms flame offline conflicts (see, e.g., Benkelman and Funke, 2019; the Sentinel Project, 2021). For example, in 2014, a fake story that

*This paper was previously circulated under the title "Perverse Ethical Concerns: Online Platforms and Offline Conflicts." We are grateful to Johannes Hörner, Larry Samuelson, and Marina Halac for their constant support. For helpful comments, we thank Yeon-Koo Che, Yi Chen, Syngjoo Choi, Liang Dai, Chris Edmond, Ilwoo Hwang, Shota Ichihashi, RC Lim, Bart Lipman, Delong Meng, Dmitry Shapiro, 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, and University of Macau. Allen Vong acknowledges the Asia-Pacific Academy of Economics and Management at the University of Macau for financial support.

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

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

a Buddhist woman was raped by two Muslim colleagues triggered clashes between Buddhists and Muslims in Myanmar, leading to deaths and injuries;[1] in 2020, 50 people armed with weapons attacked Roma people and set fire to their vans in Paris following false social media warnings that the vans were used for kidnappings.[2]

Societies respond with efforts that lead platforms to face ethical concerns to internalize the cost of these conflicts. These efforts include public awareness campaigns, such as the *Wall Street Journal*'s investigative podcast series and congress hearings.[3] These efforts also include an interdisciplinary research program on ethical algorithms (see, e.g., Wu, 2017; Kearns and Roth, 2019). Governments worldwide adopt policies to combat misinformation on platforms (see, e.g., Funke and Flamini, 2021).

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

We deliver our results in a model where a large number of citizens learn a hidden state, for instance, the change in vaccine efficacy against a virus variant, by using an online platform. The platform is an information intermediary. It receives news reports about the state and then creates a private, idiosyncratic signal for each citizen, based on an algorithm that filters misinformation in the reports. The algorithm is developed by the platform at a cost and is hidden from the citizens. Each citizen's signal summarizes the personalized contents that she reads on the platform. Upon receiving their signals, the citizens' beliefs about the state typically disagree. The disagreements trigger conflicts.

We begin with a baseline model where the platform is self-interested. The platform profits when citizens enjoy reading their contents on the platform; citizens like contents that are informative about the state as well as contents that conform to their individual

[1] See, e.g., "How a false rape allegation sparked violence in Myanmar," *Al Jazeera*, 27 October 2015.

[2] See, e.g., "Roma attacked in Paris after fake news reports," *The Guardian*, 27 March 2019.

[3] See, e.g., "Big tech CEOs face lawmakers in house hearing on social media's role in extremism, misinformation," *The Washington Post*, April 9, 2021.

biases. We next consider an alternative version of the model where the platform faces an additional ethical concern such that it finds the citizens' conflicts to be costly and internalizes the conflict cost when developing its filter. We contrast the equilibria in the two models and deliver our main result: the platform's ethical concern perversely aggravates conflicts unless the concern is sufficiently strong.

The presence of ethical concern boosts the platform's incentive to filter misinformation in equilibrium. The resulting more aggressive equilibrium filter improves the citizens' learning about the state. We call this phenomenon the learning effect. But the citizens also correctly anticipate the more aggressive filter, and thus become more confident about their own learning. We call this phenomenon the confidence effect. If the ethical concern is too weak to induce a sufficiently aggressive filter, the confidence effect dominates the learning effect and conflicts escalate. To be sure, the platform with ethical concern understands that more aggressive filtering could aggravate conflicts, but it fails to correctly internalize the conflict cost in equilibrium. Because citizens do not observe the filter, they view the signals as a credence good, assessing the signals based on their expectations of the filter but not the actual filter. Thus, given the citizens' expectations, the actual filter affects the distribution of the citizens' signals but not their inferences upon receiving the signals. When the platform best responds to the citizens' expectations, its ethical concern to mitigate conflicts then unambiguously leads to a more aggressive filter so as to reduce the dispersion of the signals.

We then apply this insight to draw policy implications. In Section 6, we return to the baseline model where the platform is self-interested and analyze popular government efforts that motivate platforms to mitigate conflicts driven by misinformation, such as legislation against misinformation. We find that when adopting these efforts, governments are confronted with the same challenge that platforms with ethical concern face: the efforts perversely aggravates conflicts unless their implementations are sufficiently aggressive.

More broadly, our results speak to debates concerning transparency of platform algorithms (see, e.g., MacCarthy, 2020). While a typical argument for transparency is to promote effective monitoring of platforms,[4] our result highlights alternatively that transparency allows platforms to correctly internalize their social responsibilities. We

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

show that if the filter were observable to the citizens, then the platform anticipates that the citizens perform inferences based on its actual choice of filter and no perverse outcome arises.

While government efforts to mitigate conflicts typically target platforms, some governments also adopt efforts that target the citizens, such as media literacy campaigns that educate "credulous" citizens whose abilities to process information are limited. In an extension, we model such credulous citizens as non-Bayesians who plainly believe that the state is equal to their received signals in the spirit of Kartik, Ottaviani and Squintani (2007) and Little (2017), and we model the campaign as a shock that turns the credulous citizens to "rational" citizens as in the baseline model. We show that the campaign could aggravate conflicts unless it is coupled with aggressive supply-side efforts that ensure sufficient filtering of misinformation. This is because the campaign disrupts the platform's filtering incentive by making it more difficult for the platform to influence the citizens' beliefs. If the learning effect associated with the disrupted filter dominates the confidence effect, then the campaign aggravates conflicts.

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

Within economics, our work contributes to the literature of disagreements among Bayesian agents driven by heterogeneous prior beliefs (see, e.g., Dixit and Weibull, 2007; Andreoni and Mylovanov, 2012; Sethi and Yildiz, 2012; Baliga, Hanany and Klibanoff, 2013; Zanardo, 2017; Kartik, Lee and Suen, 2021) or by competition among information providers (see, e.g., Perego and Yuksel, 2021). Departing from the literature, our analysis zooms in on conflicts driven by citizens' heterogeneous beliefs induced by platforms' optimizing algorithms: the perverse outcome in our model is precisely driven by the platform's best response to the citizens' expectations of its behavior.[5] To highlight this phenomenon, our setup considers citizens who share

[5] Thus, our analysis contrasts with Bayesian persuasion problems a là Kamenica and Gentzkow (2011) in which the information sender has commitment power.

a common prior belief about the state and learn individually on the platform.[6] In the extension that concerns media literacy campaigns, we depart from the literature by considering a society that consists of both Bayesians and non-Bayesians, namely rational and credulous citizens, and examining conflicts within and between the two groups. We view the analysis of credulous citizens and their conflicts with rational citizens as not only theoretically attractive but also important for policy prescriptions.

Broadly, our analysis contributes to the literature of media economics (see, e.g., Prat and Strömberg, 2013; Anderson, Waldfogel and Strömberg, 2015) and in particular the role of social media in political conflicts (see, e.g., Zhuravskaya, Petrova and Enikolopov, 2020). Our model departs by highlighting the credence nature of platform information and its policy implications. As discussed, the credence nature is key to driving the perverse outcome. The model also yields notable positive implications regarding platforms' filtering incentives that accord well with empirical findings. Specifically, our model predicts that self-interested platforms spend costs to filter misinformation to better provide contents that citizens enjoy. This prediction offers a reconciling perspective on platforms' significant investments in filtering misinformation despite often being criticized for catering to their users' preferences at the expense of filtering misinformation in practice.[7]

2 Model

A unit mass of citizens, indexed by $i \in [0, 1]$, learns a hidden state $\theta \in \mathbb{R}$ from an online platform. They share a common prior belief that θ is normally distributed with mean normalized to 0 and precision $p > 0$. Each citizen has a two-dimensional type

[6] Nonetheless, in Appendix A.1, we show that assuming heterogeneous prior beliefs do not alter our insights. Our baseline setup that citizens share a common prior belief and learn individually is reminiscent of models of common learning (e.g., Cripps, Ely, Mailath and Samuelson, 2008). While this literature focuses on asymptotic beliefs given an exogenous learning process, our analysis focuses on non-asymptotic beliefs given an endogenous learning process due to the platform's optimization.

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

$(b_i, s_i) \in \mathbb{R}^2$. As will be clear, the number b_i is citizen i's bias, capturing the value that citizen i would like the state to take; the number s_i represents the aggregate slant of the news sources to which citizen i subscribes on the platform. For our results, we require only that each citizen knows her own slant. Nonetheless, to avoid defining the players' beliefs on citizens' types for the ease of exposition, we assume that each citizen's type is commonly known.

As an overview of the model, the platform is an information intermediary that receives news reports about the state from external sources. The platform filters misinformation in the received reports according to a filtering algorithm that it develops at a cost. The platform then passes the filtered information to each citizen in a personalized manner, depending on the citizen's individual subscription of news sources on the platform. The citizens then infer the state given their received information. To be sure, in reality, platforms exhibit more flexibility than simply filtering misinformation in creating news contents for the citizens. For example, platforms could emphasize certain news reports over others or recommend certain reports to a citizen from sources that she does not subscribe to. Our analysis abstracts from these issues to focus on conflicts driven solely by misinformation.

Specifically, the platform chooses a filter $f \in \mathbb{R}_+$ and the citizens take no actions. The filter f is hidden from the citizens. We interpret a higher filter as a more aggressive filter. The filter f produces an idiosyncratic signal y_i for each citizen i, which summarizes the citizen's personalized contents on the platform and is given by

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

where ε_i is normally distributed with mean 0 and precision $q + f$ independently of the state θ and independently across citizens, representing misinformation in the contents that "escapes" the filter. The parameter $q > 0$ is exogenous and represents the default precision of the signal absent any filtering.[8] Thus, given a higher filter, each citizen's signal is more informative about the state. Finally, the slant $s_i \in \mathbb{R}$ captures that citizen i's signal is slanted in a personalized manner. The slant reflects the aggregate bias of the information supplied by the news sources to which citizen i subscribes on the platform.[9] To ease the exposition, we assume that $s_i = 0$ for each citizen i until

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

[9] Slanting by news outlets has been extensively studied in the literature (see, e.g., Mullainathan

Section 7, where we demonstrate that the restriction to zero slanting does not alter our main insights.

The signal y_i is private to citizen i. In practice, citizens might communicate their platform contents with their peers. Our results carry over to settings where the citizens observe not only their own signals, but also "a few" other citizens' signals. What is crucial to our results is that the unit mass of citizens do not observe the same signals, so that there is some posterior disagreement about the state among them. Alternatively, one can interpret the signal y_i as citizen i's acquired information about the state after reading her platform contents and communicating with her peers.[10]

Upon receiving signal y_i, citizen i forms an estimate of the state based on her expectation $f^{*,i}$ of the platform's (hidden) filter. The estimate is plainly her posterior mean of the state, denoted by $\mathbf{E}_i[\theta|y_i]$.

The platform's payoff is equal to its revenue minus its cost to develop the filter. Given a filter f, the platform incurs a quadratic cost $cf^2/2$, where $c > 0$ measures how costly it is for the platform to filter more aggressively.[11] The platform derives a higher revenue by attracting more citizens' activities on the platform. This higher revenue could result from, for instance, higher advertising revenue.[12] Citizens are more active if they enjoy the contents more. Specifically, given signals $y := (y_i)_{i\in[0,1]}$ and the citizens' expectations of the filter $f^* := (f^{*,i})_{i\in[0,1]}$, the platform's realized revenue is

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

where $\beta > 0$ and $\tau > 0$ are exogenous parameters. The parameter β measures how

and Shleifer, 2005; Gentzkow and Shapiro, 2006, 2010; Gentzkow, Shapiro and Stone, 2015; Che and Mierendorff, 2019). Moreover, our analysis takes each citizen's news subscription as exogenously given to focus on the platform's filtering problem. See, e.g., Jann and Schottmuller (2021) for an analysis of how citizens endogenously focus on certain news sources but forgo others, sorting themselves into different "echo chambers."

[10] In reality, citizens might acquire individual private signals about the state in addition to acquiring signals from the platform. We consider such a setting in Appendix A.1 and show that our main results carry over.

[11] For instance, the platform pays to hire and train their engineers to develop and maintain the filtering algorithm.

[12] For instance, *Facebook* makes money predominantly by showing advertisements from advertisers to its users. In a report provided by the SEC, advertising represented 98% of Facebook's $86 billion revenue in 2020. See https://www.sec.gov/ix?doc=/Archives/edgar/data/1326801/000132680121000014/fb-20201231.htm.

beneficial it is for the platform when the contents conform to the citizens' biases, as captured by the quadratic loss of the citizens' estimates from their biases. The parameter τ measures how beneficial it is for the platform to improve the quality of the citizens' learning, as captured by the citizens' negative posterior variances. The subscript i in the expectation and variance operators indicates that citizen i performs her inferences based on her expectation of the filter $f^{*,i}$, which need not coincide with the actual filter f.

To summarize, the platform's (expected) payoff given its filter f and the citizens' expectations f^* is

$$\mathbf{E}\left[v\left(y; f^*\right)\right] - \frac{cf^2}{2}, \tag{3}$$

where the expectation $\mathbf{E}$ is taken over the distribution of signal profiles y induced by filter f. We do not define payoffs for the citizens, as they take no actions and their payoffs are irrelevant for our analysis.

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 citizens; nonetheless, we allow the platform to contemplate deviations to arbitrary strategies. In any such equilibrium, the platform chooses its filter f to maximize its payoff (3) given the citizens' expectations f^*, such that their expectations are correct. Thus, the citizens' equilibrium expectations of the filter must be identical. Hereafter, when we say that the citizens' expectation is f^*, without loss, we refer to the event that they expect the same filter and we abuse notation to denote such filter by f^*. Moreover, throughout, we write $\mathbf{E}^*[\cdot]$ as each citizen's expectation by expecting filter f^* and write $\mathbf{E}[\cdot]$ as the platform's expectation by choosing filter f.

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

3 Equilibrium

Proposition 1 characterizes the unique equilibrium of the baseline model.

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

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

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

Equation (4) pins down the unique equilibrium filter f^{S} by equating the marginal benefit to filter on the left side and the marginal cost to filter on the right side. In the equilibrium, the platform filters solely to better provide bias-conforming contents: the filter strictly increases in the benefit β to provide bias-conforming contents and vanishes as β vanishes. To see why this is the case, observe that given any expectation f^*, the platform's expectation of the component of its revenue (2) that corresponds to improving citizens' learning is independent of the platform's actual filter f:

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

where the first equality follows from Bayesian updating.[13] Thus, given the expectation f^*, the platform's incentive to filter rely solely on the other component of its revenue, namely the bias-conforming component:

$$\mathbf{E}\left[\beta \int_0^1 -\left(\mathbf{E}^*\left[\theta|y_i\right] - b_i\right)^2 \mathrm{d}i\right], \tag{5}$$

where the posterior mean is

$$\mathbf{E}^*\left[\theta|y_i\right] = \frac{q+f^*}{p+q+f^*} y_i + \frac{p}{p+q+f^*} \mathbf{E}^*[\theta] = \frac{q+f^*}{p+q+f^*} y_i \tag{6}$$

by standard Bayesian updating. In view of (6), each citizen discounts her received signal by putting a weight short of unity on the signal (relative to the prior mean) when forming her state estimate.

From the platform's perspective, when it chooses the filter, the citizens' estimates

[13] This observation relies on the property of normal distributions that the posterior variance is independent of the signal realization. Our main result (Proposition 3) concerning perverse ethical concerns, nonetheless, does not hinge on this property. We provide a further discussion of our normal-quadratic specification at the end of Section 4.

are random (because their signals are random). By filtering more aggressively, the platform reduces the dispersion of the citizens' estimates and thus better caters to their biases in expectation, improving (5). Such reduction of the dispersion is more effective when the prior state precision p is smaller, as the citizens put more weight on the signals in their inferences. In addition, because of the diminishing returns to filtering, such reduction is more effective when the default precision q is smaller.

4 Offline Conflicts and Ethical Concern

We now introduce our notion of conflicts and then turn to an alternative version of the model where the platform faces an ethical concern to mitigate conflicts.

Given the citizens' expectation f^* and the realized signals y, the citizens' state estimates typically disagree. We measure such disagreement between any two citizens by the distance between their estimates. To provide a concrete context, consider, for example, that a government is contemplating a policy that affects the citizens' welfare and the best policy for their welfare is the one that matches the hidden state θ. After receiving their signals, each citizen believes that the optimal policy is precisely her own estimate, and the citizens disagree about the optimal policy.[14] The citizens' disagreements lead to conflicts.

We now consider an alternative version of the model where the platform faces ethical concern to mitigate conflicts. Such platform differs from a self-interested platform by internalizing the cost due to conflicts. The (realized) conflict cost, given citizens' expectation f^* and signals y, is measured by

$$\kappa(y; f^*) := \frac{1}{2} \int_0^1 \int_0^1 \left(\mathbf{E}^*[\theta|y_j] - \mathbf{E}^*[\theta|y_i]\right)^2 \mathrm{d}j \, \mathrm{d}i, \tag{7}$$

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

Then, in the presence of ethical concern, the platform's payoff by choosing f given

[14] Our measure of disagreement is familiar in the literature (see, e.g., Kartik et al., 2021). In general, such notion of disagreement is limiting when one is interested in comparing the citizens' posterior distributions. See, for example, Zanardo (2017) who examines the notion of disagreement between probability distributions axiomatically.

the citizens' expectation f^* is

$$\mathbf{E}\left[v\left(y;f^*\right)-h\cdot\kappa(y;f^*)\right]-\frac{cf^2}{2}, \tag{8}$$

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

Proposition 2 characterizes the unique equilibrium in this alternative version of the model.

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

$$\frac{(\beta+h)}{\left(p+q+f^{\mathrm{E}}\right)^2}=cf^{\mathrm{E}}. \tag{9}$$

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

As in the baseline model, the platform's filter does not affect the citizens' inferences given their expectation of the filter and their received signals. Unlike in the baseline model, the platform benefits by mitigating conflicts given its ethical concern. By filtering more aggressively, the platform reduces the dispersion of the citizens' estimates and mitigates conflicts. Thus, as depicted in Figure 1, the ethical concern causes the platform to filter more aggressively, and the filter strictly increases in the strength h. The comparative statics with respect to the other parameters concerning the filter f^{E} is analogous to that concerning the filter f^{S}.

5 Equilibrium Conflicts

In this section, we examine the equilibrium conflicts among the citizens and present our main result. In an equilibrium where the platform chooses filter f, we denote the

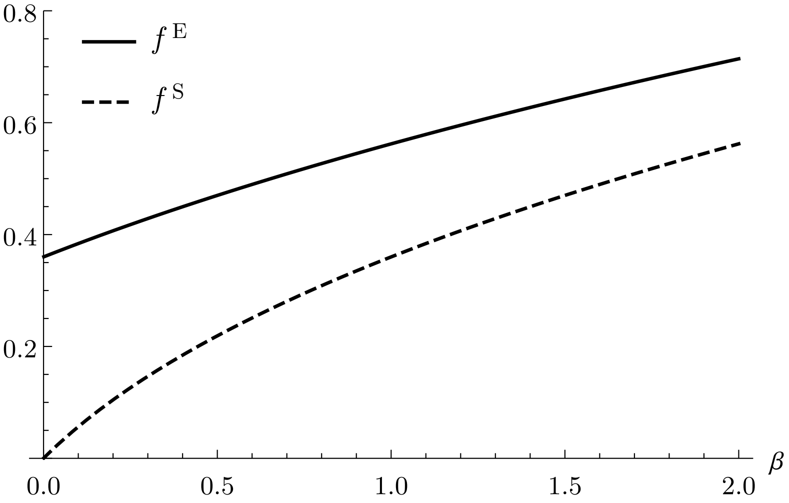

Figure 1: Filtering given $p = q = c = \beta = 1, h = 5$

conflict cost by

$$K(f) := \mathbf{E}\left[\kappa(y; f)\right], \tag{10}$$

where the expectation $\mathbf{E}$ is taken over the distribution of signals y induced by the filter f, and κ is defined in (7). Observe that in (10), we do not distinguish between the platform's actual filter and the citizens' expectation of its filter, because their expectation is correct in equilibrium.

Proposition 3 reports our main result, namely that the platform's ethical concern mitigates equilibrium conflicts if and only if the ethical concern is sufficiently strong.

Proposition 3. *There exists $\bar{h} \geq 0$ such that the ethical concern mitigates conflicts, namely $K(f^{\mathrm{S}}) > K(f^{\mathrm{E}})$, if and only if $h > \bar{h}$.*

As derived in the proof, the equilibrium conflict cost (10) is plainly

$$K(f) = \frac{q+f}{(p+q+f)^2} = \underbrace{\frac{q+f}{p+q+f}}_{\substack{\text{weight}\\ \text{on signal}}} \underbrace{\left(\frac{1}{p+q+f}\right)}_{\substack{\text{dispersion}\\ \text{of signals}}}. \tag{11}$$

More aggressive equilibrium filtering by the platform has two opposing effects for the conflicts. On the one hand, there is a learning effect that mitigates conflicts: it improves the citizens' learning about the state. On the other hand, there is a

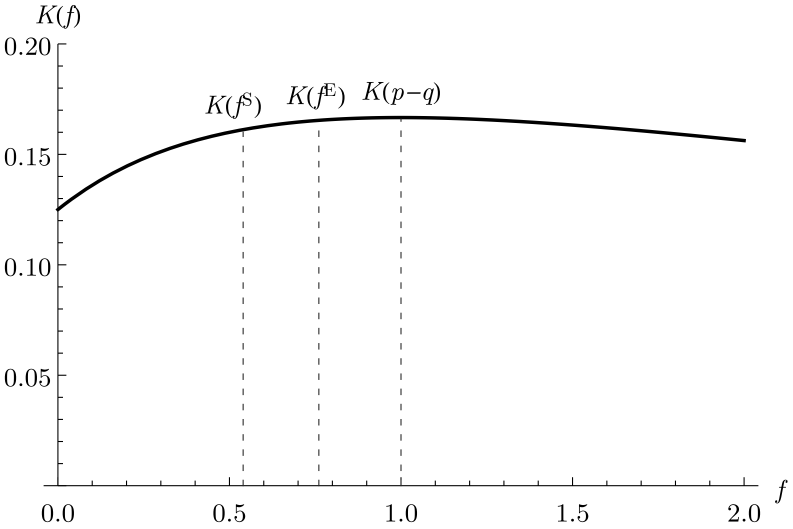

Figure 2: Equilibrium conflict cost given $\beta = c = k = h = 1, p = \frac{3}{2}, q = \frac{1}{2}$

confidence effect that aggravates conflicts: the citizens correctly anticipate the more aggressive filtering and thus put higher weights on their own signals in their inferences. As a result, the platform's ethical concern mitigates equilibrium conflicts if and only if the filter f^{E} given ethical concern is sufficiently larger than the self-interested filter f^{S}, or equivalently, if and only if the strength of the concern h is sufficiently large, so that the learning effect dominates the confidence effect.[15] Indeed, as Figure 2 depicts, the function (10) is single-peaked at $f = \max(0, p - q)$. If $f < p - q$ (resp., $f > p - q$), then the prior precision p exceeds (resp., falls short of) the signal precision $f + q$ so that an infinitesimal change in the equilibrium filter aggravates (resp., mitigates) conflicts as the confidence effect dominates (resp., is dominated by) the learning effect.

Why does the platform with ethical concern fail to correctly mitigate conflicts? The platform understands that a higher equilibrium filter could aggravate conflicts. But given the citizens' expectation of its filter, the platform also understands that its actual filter does not affect the citizens' inferences about the state given their signals, but only affects the distribution of the signals. That is, given citizens' expectation f^*,

[15] The reader who is familiar with the literature of global games may wish to compare the present result to a key takeaway of that literature in which citizens typically place "too much" weight (relative to the socially desirable level) on prior, public information because of the strategic complementarity of their actions. Here, citizens do not take actions, let alone strategic complementarity. In addition, Proposition 3 highlights that citizens put "too much" weight on their own private signals (relative to the case in Section 6 where the filter is observable to the citizens and no perverse outcome arises) in response to the platform's incentives.

the platform "incorrectly" internalizes the cost

$$\frac{q+f^*}{p+q+f^*}\left(\frac{1}{p+q+f}\right) \tag{12}$$

instead of (11). Hence, in the platform's best response to the citizens' expectation, the presence of ethical concern unambiguously boosts the platform's filtering incentive so as to reduce the dispersion of the signals. But then the citizens correctly anticipate the platform's such incentive, yielding the perverse outcome in equilibrium.

Finally, we analyze the structure of the threshold $\bar{h}$ in Proposition 3. Proposition 4 below shows that the threshold increases in the prior state precision p and that the threshold is vacuous if and only if p is small. The focus on varying the parameter p elucidates the race between the learning effect and the confidence effect.

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

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

Given a large p, the self-interested filter f^{S} is relatively small by Proposition 1 and the prior state precision is likely to exceed the signal precision. Unless the concern is strong enough to ensure a sufficiently large filter f^{E} driven by ethical concern, the learning effect is dominated by the confidence effect given the change of the filter from f^{S} to f^{E}. In contrast, given a small enough p, the filters f^{S} is relatively large and the signal precision exceeds the prior state precision. Given a change of the filter to f^{E} in the presence of ethical concern, the learning effect dominates the confidence effect no matter how weak the ethical concern is.

To close this section, we comment on our quadratic-normal specification. It has afforded a sharp characterization of the platform's incentives, the equilibrium conflicts, the implications of ethical concerns, as well as their associated comparative statics. Indeed, because of its tractability, the quadratic-normal specification is widely adopted in related signaling environments in the literature (see, e.g., Fischer and Verrecchia, 2000 and Bénabou and Tirole, 2006; see also Frankel and Kartik, 2019 for further discussion). We conjecture that our insights underlying Proposition 3 extend to environments with more general state and signal distributions and a more general cost structure, so long as the observation that the ethical concern causes a platform to filter more aggressively in response to the citizens' expectation carries over to these more general environments. A similar comment applies to Proposition 4. The critical insight underlying this result is that when the public information about the state has higher quality, citizens put a smaller weight on their signals relative to the public information for inferences, limiting the learning effect.

6 Government Efforts

As noted at the outset, governments worldwide adopt different efforts to mitigate offline conflicts driven by online misinformation. In this section, we return to our baseline version of the model where the platform is self-interested and cast several such efforts to analyze their effects on offline conflicts. The main takeaway is that these efforts echo our main result in Proposition 3, namely that their implementations must be aggressive enough in order to be effective and to not trigger any perverse outcomes. The proofs of the results in this section, except for Proposition 7, follow directly from Proposition 3 and are therefore omitted.

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

Proposition 5 below shows that introducing the floor mitigates conflicts if and only if the floor is sufficiently high. In the proposition, we write f^{S} as $f^{\mathrm{S}}(p)$ wherever appropriate to emphasize its dependence on p. Recall from Proposition 1 that $f^{S}(p)$ is strictly decreasing in p, so that the equilibrium signal precision exceeds the prior state precision, namely $q + f^{\mathrm{S}}(p) \geq p$, if and only if p is sufficiently small.

Proposition 5. *If p is small such that the signal precision exceeds the prior state precision, namely $q + f^{\mathrm{S}}(p) \geq p$, then the legislation mitigates conflicts: $K(f^{\mathrm{L}}) < K(f^{\mathrm{S}})$. Otherwise, there exists $F > f^{\mathrm{S}}$ such that $K(f^{\mathrm{L}}) < K(f^{\mathrm{S}})$ if and only if $\underline{f} > F$.*

The proposition follows because the learning effect is more likely to dominate the confidence effect given an increase in the equilibrium filter induced by the legislation when the prior state precision p is smaller. The proposition sheds light on policy discussions concerning the modification and elimination of platforms' immunity of Section 230 of the Communications Decency Act. Such immunity is commonly viewed as a "legal shield" that protects platforms from liability for third-party content that they host.[16] Plainly, modifying or eliminating platforms' immunity introduces a cost that platforms incur due to their insufficient filtering of misinformation, and motivate platforms to filter more aggressively. Our results highlight that for such changes to the platforms' immunity to effectively mitigate conflicts, they must be implemented sufficiently aggressively.

Arrests and cyber task forces. We next consider arrests of misinformation spreaders and cyber task forces against misinformation campaigns. We cast such efforts in our model as an increase of the default precision absent filtering from an initial value q^{B} to some $q^{\mathrm{A}} > q^{\mathrm{B}}$, and denote the corresponding unique equilibrium as characterized in Proposition 1 by f^{B} and f^{A}, respectively. Proposition 5 below shows that such efforts unambiguously mitigate conflicts if and only if their implementations are sufficiently aggressive. In the proposition, we write f^{B} as $f^{\mathrm{B}}(p)$ wherever appropriate to emphasize its dependence on p.

Proposition 6. *If p is small so that the initial signal precision exceeds the prior state precision, namely $q^{\mathrm{B}} + f^{\mathrm{B}}(p) \geq p$, then the increase in the default signal precision mitigates conflicts: $K(f^{\mathrm{A}}) < K(f^{\mathrm{B}})$. Otherwise, there exists $Q > q^{\mathrm{B}}$ such that $K(f^{\mathrm{A}}) < K(f^{\mathrm{B}})$ if and only if $q^{\mathrm{A}} > Q$.*

The intuition of Proposition 6 is analogous to that of Proposition 5, and so their statements share an analogous structure. 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, namely the sum of the default precision and the platform's filter, increases by direct application of the implicit function theorem on (4). The effect on the equilibrium conflicts given the efforts is thus identical that given a fixed default precision and a higher filter, which is the case in Proposition 5.

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

Transparency. Finally, we analyze a potential regulation effort on platform transparency that is commonly discussed in policy debates. Specifically, suppose that the platform's filter is publicly observable. Then the platform anticipates that the citizens perform inferences based on its actual filter. Contrary to (8), the payoff of a platform with ethical concern is

$$\mathbf{E}\left[v\left(y;f\right)-h\cdot\kappa(y;f)\right]-\frac{cf^2}{2}=\mathbf{E}\left[v\left(y;f\right)\right]-\frac{cf^2}{2}-h\cdot K(f) \tag{13}$$

where the expectation is taken with respect to the actual filter f. In view of (13), transparency allows the platform to correctly internalize its "social responsibility" and hence, as Proposition 7 below makes precise, no perverse outcome arises:

Proposition 7. *Suppose that the platform's filter is observable to the citizens. Then, given any equilibrium filter $\tilde{f}^{\mathrm{S}}$ absent ethical concern and any equilibrium filter $\tilde{f}^{\mathrm{E}}$ given ethical concern, $K(\tilde{f}^{\mathrm{S}})\geq K(\tilde{f}^{\mathrm{E}})$.*

In practice, calls for transparency are primarily motivated by the conventional wisdom that transparency is essential to accountability measures for platforms and consumer protection (see, e.g., MacCarthy, 2020). Proposition 7 offers an alternative case for transparency by highlighting its role to complement platforms' social responsibilities.

7 Extension

In this section, we consider an extended version of the model that nests the baseline model. In the extended model, each citizen's signal is slanted in a personalized manner, reflecting her individual subscription of potentially biased news sources on the platform. The presence of slanted signals allows us to consider two forms of ethical concern—one that internalizes conflict cost as in the baseline model and another one that removes the slants in the spirit of the FCC fairness doctrine. The extended model also features "credulous" citizens who lack the ability to interpret their signals, contrary to the "rational" citizens in the baseline model. The presence of credulous citizens allows us to study media literacy campaigns, which aim to improve credulous citizens' abilities to process information, and their implications for offline conflicts.

Specifically, we now relax the restriction that the slant $s_i=0$ for each citizen i

in (1) and allow for any $s_i \in \mathbb{R}$. Suppose further that citizens now differ not only in their biases, but also in their reading abilities: a mass $1-r \in (0,1)$ of the citizens are "credulous," who are non-Bayesians and simply take their received signals at face value, and the remaining mass r of citizens are "rational" as in the baseline model. Thus, upon receiving signal y_i, citizen i's state estimate is given by

$$\hat{\theta}_i\left(y_i\right) = \begin{cases} \mathbf{E}_i[\theta|y_i], & \text{if } i \text{ is rational,} \\ y_i, & \text{if } i \text{ is credulous.} \end{cases}$$

In this extended model, each citizen's type is three-dimensional, consisting of her bias, reading ability and slant. As in the baseline model, we assume that each citizen's type is commonly known for simplicity. For our results, we only require that each citizen knows her own slant and ability. Without loss, we assume that each citizen $i \in [0,r]$ is rational and each citizen $i \in (r,1]$ is credulous.

Given signals $y := (y_i)_{i\in[0,1]}$ and the rational citizens' expectations of the platform's filter $f^* := (f^{*,i})_{i\in[0,1]}$, the platform's realized revenue is now

$$v\left(y; f^*\right) := \beta \int_0^1 -(\hat{\theta}_i(y_i) - b_i)^2 \,\mathrm{d}i + \tau \int_0^r -\mathbf{Var}_i\left(\theta|y_i\right) \mathrm{d}i. \tag{14}$$

The revenue is independent of the quality of credulous citizens' learning about the state, as they are assumed to believe that their signals fully reveal the state. Moreover, because the credulous citizens do not form expectations about the filter, in equilibrium, the platform chooses its filter to maximize the payoff (14) given the rational citizens' expectation of its filter such that their expectation is correct. The model is otherwise identical.

In this extended model, contrary to (6), given rational citizens' expectation f^*, a rational citizen i's state estimate upon receiving signal y_i is given by

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

Thus, in addition to discounting the signal by putting a weight short of unity on the signal as in the baseline model, the citizen also correctly removes the slant in her inference.

Proposition 8 extends Proposition 1 to characterize the unique equilibrium in the present setting.

Proposition 8. *There exists a unique equilibrium. In the equilibrium, the platform chooses filter f_r^{S} which is positive and is characterized by*

$$\beta\left(\frac{r}{(p+q+f_r^{\mathrm{S}})^2}+\frac{1-r}{(q+f_r^{\mathrm{S}})^2}\right)=cf_r^{\mathrm{S}}. \tag{15}$$

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

The intuition behind the proposition is analogous to that behind Proposition 1. Different from the baseline model, the filter f_r^{S} depends on the mass of rational citizens r and is strictly decreasing in r. Intuitively, from the platform's perspective, the credulous citizens' (random) estimates are more dispersed than the rational citizens' estimates, as the credulous citizens do not discount their received signals. Thus, given a larger mass of rational citizens, the platform's marginal benefit to filter is smaller. Finally, observe that the filter is independent of the citizens' slants. This is because the platform's incentive to reduce the dispersion of the citizens' signals by filtering is undisturbed by the slants.

In Section 7.1, we turn to the implications of ethical concern for offline conflicts in this extended model. In Section 7.2, we examine the implications of media literacy campaigns for conflicts.

7.1 Ethical Concern

As mentioned, the presence of slanted signals allow us to examine two different forms of ethical concern.

Internalizing conflict cost. First, as in Section 5, we suppose that a platform who faces ethical concern internalizes the conflict cost among the citizens. That is, given filter f and rational citizens' expectation f^*, the platform's expected payoff given ethical concern remains to be given by (8), with the difference that the realized conflict cost is now given by

$$\kappa(y;f^*):=\frac{1}{2}\int_0^1\int_0^1\left(\hat{\theta}_j(y_j)-\hat{\theta}_i(y_i)\right)^2\mathrm{d}j\,\mathrm{d}i, \tag{16}$$

instead of (7).

Proposition 9 below extends Proposition 3 and characterizes the unique equilibrium when the platform faces ethical concern in the extended model.

Proposition 9. *Suppose that the platform faces ethical concern. Then there exists a unique equilibrium. In the equilibrium, the platform chooses filter f_r^{E} which is positive and is characterized by*

$$(\beta + h)\left(\frac{r}{(p+q+f_r^{\mathrm{E}})^2} + \frac{1-r}{(q+f_r^{\mathrm{E}})^2}\right) = cf_r^{\mathrm{E}}. \tag{17}$$

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

The special case where all citizens are rational so that $r = 1$, the filter f_r^{E} is equal to the filter f^{E} as characterized in (4). The intuition behind the proposition is analogous to that behind Proposition 3.

With Propositions 8 and 9 in place, we consider how the ethical concern affects the equilibrium conflict cost between 1. any two rational citizens, 2. any pair of rational and credulous citizens, and 3. any two credulous citizens. To state the result, for any two citizens i, j, we define

$$K_{ij}(f) := \mathbf{E}[(\hat{\theta}_i(y_i) - \hat{\theta}_j(y_j))^2] \tag{18}$$

as the equilibrium conflict cost between the two citizens given filter f. As in (10), (18) does not distinguish between the actual filter chosen by the platform and the rational citizens' expectation of the filter, since their expectation is correct in equilibrium.

Proposition 10. *The following holds.*

1. *There exists $\bar{h} \geq 0$ such that $K_{ij}(f_r^{\mathrm{S}}) > K_{ij}(f_r^{\mathrm{E}})$ for any two rational citizens i and j if and only if $h > \bar{h}$.*
2. *For any rational citizen i and credulous citizen j, $K_{ij}(f_r^{\mathrm{S}}) > K_{ij}(f_r^{\mathrm{E}})$.*
3. *For any two credulous citizens i and j, $K_{ij}(f_r^{\mathrm{S}}) > K_{ij}(f_r^{\mathrm{E}})$.*

Part 1 of the proposition says that the platform's ethical concern mitigates equilibrium conflicts between any two rational citizens if and only if the concern is sufficiently strong. Because the rational citizens correctly remove their slants in their inferences, the structure of equilibrium conflicts between any two rational citizens is identical to that in Proposition 3. Finally, part 2 and part 3 of the proposition say that the ethical concern mitigates equilibrium conflicts between any two citizens in which at least one of them is credulous. This is because unlike a rational citizen, a credulous citizen never discounts her signal. Hence, unlike in a rational citizen's inference, in a credulous citizen's inference the learning effect that mitigates conflicts is stronger and the confidence effect that aggravates conflicts is absent.

Fairness doctrine. In the presence of slanted information, several media scholars (see, e.g., Napoli, 2019) urge for introducing a version of the the FCC fairness doctrine for online media as an alternative form of ethical concern. The doctrine was originally applied to radio and television broadcasters, requiring that the broadcasters provide a fair and balanced presentation of information.[17]

To cast the doctrine in the model, suppose that the platform can overrule the citizens' individual subscriptions so that in determining each citizen i's signal y_i, the slant is $s_i = 0$. In Proposition 11 below, part 1 shows that that the doctrine does not affect affects among the rational citizens; part 2 shows that the doctrine unambiguously mitigates conflicts involving credulous citizens. To emphasize the (potential) dependence of the conflict cost (18) between citizens i and j given their slants s_i and s_j, we write $K_{ij}(\cdot)$ as $K_{ij}(\cdot; s_i, s_j)$. Since the platform's filter is determined independently of the citizens' slants in equilibrium, the equilibrium filter remains as f_r^{S} given the doctrine. In addition, the platform's slants affect only the conflicts between any pair of citizens involving at least one credulous citizen in equilibrium. Thus, the proposition follows:

Proposition 11. *The following holds.*

1. *For any rational citizens i and j, $K_{ij}(f_r^S; 0, 0) = K_{ij}(f_r^S; s_i, s_j)$.*
2. *For any citizens i, j where at least one of them is credulous, $K_{ij}(f_r^{\mathrm{S}}; 0, 0) \leq K_{ij}(f_r^{\mathrm{S}}; s_i, s_j)$; if citizen i is credulous with $s_i \neq 0$, then $K_{ij}(f_r^{\mathrm{S}}; 0, 0) < K_{ij}(f_r^{\mathrm{S}}; s_i, s_j)$.*

It is worth mentioning that the FCC eliminated the doctrine for broadcasters in 1987. The core justification of the elimination was that the doctrine was no longer

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

necessary, as the growing number of media outlets available facilitates consumers' access to diverse information. Our analysis highlights that such justification is limiting in the context of online media. While consumers' access to diverse information is also a defining feature of online media, the phenomenon of "echo chambers" where citizens choose to read certain contents and omit others is also prevalent. These citizens include those who are credulous and hence lack the sophistication to utilize the slanted information. As Proposition 11 highlights, the credulous citizens' slants flame conflicts and the doctrine is effective in mitigating conflicts caused by their slants.

7.2 Media Literacy Campaign

Finally, we turn to consider a media literacy campaign given which the credulous citizens become rational before the platform chooses its filter, and this event is common knowledge. The platform's unique equilibrium filter upon the campaign is plainly f^{S} as characterized in (4), or equivalently f_1^{S} as characterized in (15).

Proposition 12 below states that the effect of the campaign on equilibrium conflicts is ambiguous in general. In the proposition, we write f^{S} and f_1^{S} as $f^{\mathrm{S}}(p)$ and $f_1^{\mathrm{S}}(p)$ wherever appropriate to emphasize their dependence on p.

Proposition 12. *The following holds.*

1. *For any citizens i, j who were credulous before the campaign, the campaign mitigates their equilibrium conflicts: $K_{ij}(f_1^{\mathrm{S}}) < K_{ij}(f_r^{\mathrm{S}})$.*
2. *For any citizen i who was rational before the campaign and any citizen j who was credulous before the campaign, the campaign mitigates their equilibrium conflicts: $K_{ij}(f_1^{\mathrm{S}}) < K_{ij}(f_r^{\mathrm{S}})$.*
3. *Suppose that p is large so that the signal precision is strictly smaller than the prior state precision, namely $q + f_1^{\mathrm{S}}(p) < p$. Then, the campaign mitigates equilibrium conflicts between any two citizens i, j who were rational before the campaign if and only if the mass of these citizens is sufficiently large: there exists $\bar{r} \in [0, 1)$ such that*

$$K_{ij}(f_1^{\mathrm{S}}) < K_{ij}(f_r^{\mathrm{S}}) \iff r > \bar{r}.$$

Otherwise, the campaign aggravates their equilibrium conflicts: $K_{ij}(f_1^{\mathrm{S}}) > K_{ij}(f_r^{\mathrm{S}})$.

The first two parts of Proposition 12 show that the campaign unambiguously mitigates equilibrium conflicts between any two citizens involving at least one citizen that was credulous before the campaign. This is because the credulous citizens learn to discount the signals when forming their state estimates upon the campaign. Part 3 shows that the effect of the campaign on equilibrium conflicts between citizens who were rational before the campaign is ambiguous because both the learning effect and the confidence effect are present following the fall of the equilibrium filter due to the campaign. When the prior state precision p is small, the learning effect dominates the confidence effect and the fall in the equilibrium filter aggravates conflicts. In contrast, when p is large, the learning effect dominates the confidence effect so that the fall in the equilibrium filter aggravates conflicts if and only if the fall is large enough. This is the case when the mass of rational citizens before the campaign was small.

Mixed efforts. The above observations point to an appeal of performing a mix of demand-side and supply-side efforts, which is indeed a common practice. Let

$$\begin{aligned} K_A(f) := \frac{1}{2}\int_0^r\int_0^r K_{ij}(f)\,\mathrm{d}i\,\mathrm{d}j + \frac{1}{2}\int_r^1\int_r^1 K_{ij}(f)\,\mathrm{d}i\,\mathrm{d}j \\ + \int_0^r\int_r^1 K_{ij}(f)\,\mathrm{d}i\,\mathrm{d}j + \int_r^1\int_0^r K_{ij}(f)\,\mathrm{d}i\,\mathrm{d}j \end{aligned} \tag{19}$$

denote the aggregate conflict cost in an equilibrium with filter f, where K_{ij} is defined in (18) and the scalar $1/2$ accounts for the double-counting of the citizens in the double integrals. Then:

Corollary 1. *Given a filtering floor* $\underline{f} \geq f_r^{\mathrm{S}}$, *implementing a media literacy campaign unambiguously reduces aggregate equilibrium conflict cost* $K_A(\cdot)$.

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

8 Conclusion

Public concerns that misinformation on online platforms flames offline conflicts are paramount, so are societal efforts to address them. These efforts either lead platforms to face ethical concerns to mitigate conflicts or combat misinformation on the platforms. *Prima facie*, these efforts appear unambiguously effective. In this paper, we have offered a cautionary observation. We have elucidated the strategic implications of these efforts, highlighting their potential perverse consequences. In particular, we have shown that citizens who anticipate platforms' aggressive filtering of misinformation might become "too confident" of the individual learning and in turn become more hostile to others' disagreeing opinions; we have also highlighted that the platform with ethical concern fails to correctly internalize such phenomenon due to the credence nature of its information provision. A critical message that our results put forward is that for the societal efforts to mitigate conflicts to be effective, their implementations must be sufficiently aggressive.

We have restricted our attention of the strategic implications of ethical concerns to the context of offline conflicts incited by online misinformation. Thus, our analysis has deliberately limited the platform's available instruments to filtering algorithms in order to focus on these conflicts. While misinformation-related conflicts represent a pressing concern in policy discussions, there are other conflict sources that are documented empirically but 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 concerns in other contexts such as privacy, addiction, and fairness remain open. We leave it to future work to examine these issues in our framework.

Appendices

A Omitted Details

A.1 Additional Private Signals

Suppose that we extend the baseline model such that each citizen i receives an additional private signal x_i before she receives the signal y_i from the platform, where the signal x_i is given by $x_i = \theta + \eta_i$, with η_i being normally distributed with mean zero and precision $z > 0$ independently of θ and $(y_j)_{j\in[0,1]}$ and independently across citizens. to ease the exposition, we define $p^\dagger := p + z$ and $A^\dagger := \frac{z}{p+z}$.

After receiving the private signal x_i and before receiving the signal y_i from the platform, each citizen i's belief on the state is normally distributed with mean $A^\dagger x_i$ and precision $p^\dagger$. Moreover, each citizen i's state estimate upon receiving both signals x_i and y_i is given by

$$\mathbf{E}_i[\theta|x_i, y_i] = \frac{p^\dagger A^\dagger}{p^\dagger + q + f} x_i + \frac{q+f}{p^\dagger + q + f} y_i.$$

Proposition 13 below extends Propositions 1 and 2 in the present setting where each citizens receive an additional private signal.

Proposition 13. *The following holds.*

1. *Suppose that the platform is self-interested. In the unique equilibrium, the platform chooses filter $f^{\mathrm{S},\dagger}$ such that*

$$\frac{\beta}{(p+z+q+f^{\mathrm{S},\dagger})^2} = cf^{\mathrm{S},\dagger}.$$

2. *Suppose that the platform faces ethical concern. In the unique equilibrium, the platform chooses filter $f^{\mathrm{E},\dagger}$ such that*

$$\frac{\beta+h}{(p+z+q+f^{\mathrm{E},\dagger})^2} = cf^{\mathrm{E},\dagger}.$$

3. *The platform filters more aggressively given ethical concern: $f^{\mathrm{S},\dagger} < f^{\mathrm{E},\dagger}$. More-*

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

The proof of this proposition is analogous to that of Propositions 1 and 2 and is omitted. Finally, Proposition 14 below extends Propositions 3 and 4 to the present setting.

Proposition 14. *The following holds.*

1. *There exists $\bar{h}^{\dagger} \geq 0$ such that $K(f^{\mathrm{S}}) > K(f^{\mathrm{E}})$ if and only if $h > \bar{h}^{\dagger}$.*
2. *There exists $\bar{p}^{\dagger} > 0$ such that:*

   2a. If $p > \bar{p}^{\dagger}$, then $\bar{h}^{\dagger} \equiv \bar{h}^{\dagger}(p) > 0$ and is strictly increasing in p.

   2b. If $p \leq \bar{p}^{\dagger}$, then $\bar{h}^{\dagger} \equiv \bar{h}^{\dagger}(p) = 0$.

The statement of Proposition 14 and its intuition are analogous to those of Propositions 3 and 4. The presence of the additional private signal acquired by each citizen prior to interacting with the platform simply adjusts the weight she puts on the platform's signal. The learning effect and the confidence effect driven by the platform's ethical concern remains present.

B Proofs

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

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

B.1 Proof of Propositions 1 and 8

Here, we provide the proof of Proposition 8 in the extended setting in Section 7. The special case where $r = 1$ and $s_i = 0$ gives the proof of Proposition 1.

By choosing f given the rational citizens' expectation f^*, the platform's (expected) revenue is

$$\mathbf{E}\left[v(y; f^*)\right] = \mathbf{E}\left[\int_0^r -\beta\left(\mathbf{E}^*\left[\theta|y_i\right] - b_i\right)^2 - \tau\mathbf{Var}^*\left[\theta|y_i\right]\mathrm{d}i + \int_r^1 -\beta(y_i - b_i)^2\,\mathrm{d}i\right]. \tag{20}$$

By direct calculations,

$$\mathbf{E}\left[(\mathbf{E}^*[\theta|y_i]-b_i)^2\right] = \frac{A(f^*)^2}{pA(f)} + b_i^2, \tag{21}$$

$$\mathbf{E}\left[(y_i - b_i)^2\right] = \frac{1}{pA(f)} + (s_i - b_i)^2, \tag{22}$$

$$\mathbf{Var}^*[\theta|y_i] = \frac{1}{p+q+f^*}. \tag{23}$$

Substituting (21)—(23) into (20), the platform's payoff given filter f and expectations f^* is

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

The first order condition of this expression with respect to f, alongside the condition that the optimal f equals the expectation f^*, yields (15).

B.2 Proof of Propositions 2 and 9

Here, we provide the proof of Proposition 9 in the extended setting in Section 7. The special case where $r = 1$ and $s_i = 0$ gives the proof of Proposition 2. By direct calculations, the expected conflict costs are

$$\mathbf{E}\left[(y_j - y_i)^2\right] = (s_i - s_j)^2 + \frac{2}{q+f}, \tag{24}$$

$$\mathbf{E}\left[(\mathbf{E}^*[\theta|y_j] - \mathbf{E}^*[\theta|y_i])^2\right] = A(f^*)^2\left[\frac{2}{q+f}\right], \tag{25}$$

$$\mathbf{E}\left[(y_j - \mathbf{E}^*[\theta|y_i])^2\right] = s_j^2 + \frac{1+A(f^*)^2}{q+f} + \frac{p}{(p+q+f^*)^2}. \tag{26}$$

Hence, the expected conflict cost due to the disagreement between a rational citizen i and all other citizens given the platform's filter f and expectation f^* is

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

Similarly, the expected conflict cost due to the disagreement between a credulous citizen i and all other citizens given the platform's filter f and expectation f^* is

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

In the presence of ethical concern, the platform's expected payoff is

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

where $\mathbf{E}\left[v(y;f^*)\right]$ is derived above in the proof of Proposition 1.

Finally, the characterization (15) follows immediately from the first-order condition of this platform's expected payoff with respect to f, alongside the condition that the optimal f equals the expectation f^*.

B.3 Proof of Propositions 3 and 10

Here, we provide the proof of Proposition 10 in the extended setting in Section 7. The special case where $r = 1$ and $s_i = 0$ gives the proof of Proposition 3. Throughout this proof, to emphasize the dependence of f_r^{E} on h, we write f_r^{E} as $f_r^{\mathrm{E}}(h)$.

We first prove part 1 of Proposition 10, which implies Proposition 3. In an equilibrium with filter f, the expected conflict cost between any two rational citizens i and j is

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

in view of (25). This function is single-peaked at $f = \max(0, p-q)$, and is independent of the citizens' labels. If $p - q \leq f_r^{\mathrm{S}}$, then $K_{ij}(f_r^{\mathrm{E}}) \geq K_{ij}(f_r^{\mathrm{S}})$ for any $h > 0$. Thus, the statement holds with $\bar{h} = 0$. It remains to consider the case $p - q > f_r^{\mathrm{S}}$. It must then hold that $p-q > 0$. Let $h^* > 0$ denote the level of h such that $f^{\mathrm{S}} < f^{\mathrm{E}}(h^*) = p-q$. Also, define $\Delta K_{ij}(h) := K_{ij}(f_r^{\mathrm{E}}(h)) - K_{ij}(f_r^{\mathrm{S}})$. The function $\Delta K_{ij}(h)$ strictly increases over $[0, h^*)$ and strictly decreases over $[h^*, \infty)$. Also, $\Delta K_{ij}(0) = 0$ and $\lim_{h\to\infty} \Delta K_{ij}(h) < 0$ because $\lim_{h\to\infty} f_r^{\mathrm{E}}(h) = \infty$. Hence, there is $\bar{h} > 0$ such $\Delta K_{ij}(h) \geq 0$ if and only if

$h \in [0, \bar{h}]$.

Next, we turn to part 2. In an equilibrium with filter f, the expected conflict cost between any rational citizen i and any credulous citizen j is

$$s_j^2 + \frac{1 + A(f)^2}{q + f} + \frac{p}{(p + q + f)^2} = s_j^2 + \frac{1}{p + q + f} + \frac{1}{q + f}$$

by (26). The statement follows as this expression strictly decreases in f.

Finally, we prove part 3. In an equilibrium with filter f, the expected conflict cost between any two credulous citizens i and j is given by (24). The statement then follows as (24) strictly decreases in f^*.

B.4 Proof of Proposition 7

Suppose that the filter is publicly observable. Then, by choosing filter f, the self-interested platform's payoff is given by $\pi(f) := \mathbf{E}\left[v\left(y; f\right)\right] - cf^2/2$, where the expectation is taken with respect to filter f. On the other hand, by choosing filter f, the platform's payoff given ethical concern is $\pi(f) - h \cdot K(f)$ in view of (13). Then, given any equilibrium $\tilde{f}^{\text{S}}$ absent ethical concern and any equilibrium $\tilde{f}^{\text{E}}$ in the presence of ethical concern, it must hold that $\pi(\tilde{f}^{\text{E}}) \leq \pi(\tilde{f}^{\text{S}})$. To complete the proof, suppose, towards a contradiction, that $K(\tilde{f}^{\text{S}}) < K(\tilde{f}^{\text{E}})$. Then,

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

contradicting the assumption that $\tilde{f}^{\text{E}}$ is an equilibrium given ethical concern.

B.5 Proof of Proposition 11

The proposition directly follows from (24)-(26), as the doctrine reduces the conflict cost between any pair of citizens by changing the slants to zero without affecting the equilibrium filter.

B.6 Proof of Proposition 12

The proposition directly follows from (24)-(26), as $f_r^{\text{S}} < f^{\text{S}}$.

B.7 Proof of Proposition 14

Analogous to the proof of Proposition 3, the equilibrium conflict cost (among the rational citizens) is

$$K^{\dagger}(f) = \frac{p^{\dagger} A^{\dagger}}{(p^{\dagger} + q + f)^2} + \frac{q + f}{(p^{\dagger} + q + f)^2}.$$

This function is strictly increasing in f if and only if $f < p^{\dagger}(1 - 2A^{\dagger}) - q = p - z - q$. Thus, if $p \leq z + q$, then the proposition holds with $\bar{h}^{\dagger} = 0$. On the other hand, if $p > z + q$, then the proposition holds with $\bar{h}^{\dagger} > 0$. This proves the first statement.

Finally, we turn to the second statement. When p is sufficiently large, $p - z - q > 0$ and the function $K^{\dagger}(f)$ is strictly increasing over $[0, p - z - q)$ and strictly decreasing otherwise. The proposition then follows as the equilibrium filters $f^{\mathrm{S},\dagger}$ and $f^{\mathrm{E},\dagger}$ strictly decrease in p and vanish as p tends to infinity.

References

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

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

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

Bénabou, R. and Tirole, J. (2006). Incentives and Prosocial Behavior, *American Economic Review* **96**(5): 1652–1678.

Benkelman, S. and Funke, D. (2019). Misinformation is Inciting Violence Around the World. And Tech Platforms Don't Seem to Have a Plan to Stop It., *Poynter* .

Che, Y.-K. and Mierendorff, K. (2019). Optimal Dynamic Allocation of Attention, *American Economic Review* **109**(8): 2993–3029.

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

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

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

Fischer, P. E. and Verrecchia, R. E. (2000). Reporting Bias, *The Accounting Review* **75**(2): 229–245.

Frankel, A. and Kartik, N. (2019). Muddled Information, *Journal of Political Economy* **127**(4): 1739–1776.

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

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

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

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

Jann, O. and Schottmuller, C. (2021). Why Echo Chambers Are Useful.

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.

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

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

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

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

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

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

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

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

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

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

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

The Sentinel Project (2021). The Sentinel Project's Blog Series on Misinformation, Hate Speech, and Violence.

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

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

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