Perverse Ethics

Dongkyu Chang[1] Allen Vong[2]

[1]CityU of HK

[2]UM

Motivation

Social media platforms create diverse beliefs.

Diverse beliefs create conflicts.

Platforms are urged to “ethically” internalize the conflict costs.

This paper: a cautionary tale.

- By internalizing conflict costs, platforms may aggravate conflicts.

Roadmap

1. Model and equilibrium analysis
   - Baseline version: platform is self-interested.
   - Alternative version: platform is ethical.
2. Main result: Ethicality may aggravates conflict costs.
3. Regulations
4. Literature

Model

Setup

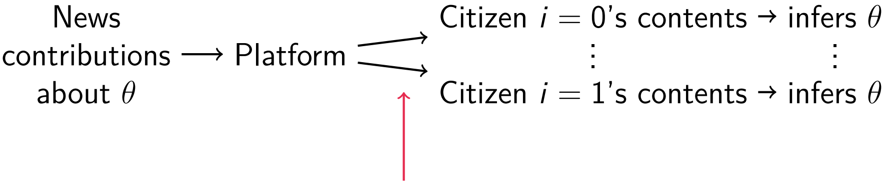
One-shot game.

Players:

- Platform.
- Rational citizens, $i \in [0, r]$ where $0 < r \leq 1$.
- Credulous citizens, $i \in (r, 1]$.

Hidden state $\theta \sim N(0, 1/p)$, where $p > 0$.

Overview



News contributions about θ ⟶ Platform

Citizen $i = 0$'s contents → infers θ

⋮ ⋮

Citizen $i = 1$'s contents → infers θ

Depends on platform's upfront investment in <u>two</u> algorithms

One algorithm filters misinformation.

Another algorithm determines a personalized slant for each citizen.

Overview

Delta arrived in Israel. Some vaccinated people are infected.

- The state θ captures the change in Pfizer vaccine's effectiveness.

News contribution 1: (filtering: flagged on Facebook.)

"Vaccines are worthless. $>$ 80% of the infected people are vaccinated."

News contribution 2:

"Alarming news: half of the infected people are vaccinated."

News contribution 3:

"Data shows that vaccine is 80% effective against infection."

(Slanting towards negative news: omit #3 in citizen *i*'s content.)

Algorithms

The platform chooses a pair of algorithms (f, s):

- Filter $f \in [0, \infty)$.
- Slant $s : [0, 1] \to \mathbf{R}$, where s_i is citizen i's personalized slant.

The algorithms are hidden from the citizens.

But the (rational) citizens have rational expectation about (f, s)

The citizens take no actions.

Signals

Given (f, s), each citizen i receives a private signal

$$y_i = \theta + s_i + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \frac{1}{q+f})$ represents misinformation ($q > 0$: default precision).

Suppose that the citizens expect that the platform chooses (f^*, s^*) (possibly, $(f^*, s^*) \neq$ the platform's actual choice).

Each citizen i's estimate of θ upon receiving y_i:

$$\hat{\theta}_i(y_i) = \begin{cases} \mathbf{E}^*[\theta|y_i] = \overbrace{\frac{q+f^*}{p+q+f^*}}^{\text{weight on signal}} \cdot \overbrace{(y_i - s_i^*)}^{\theta+\epsilon_i+s_i-s_i^*} & \text{if } i \text{ is rational} \\ y_i & \text{if } i \text{ is credulous} \end{cases}$$

Platform's payoff

Revenue from rational citizens:

$$v_R\left(f,s;f^*,s^*\right) := \mathbf{E}\left[\int_0^r \overbrace{-\beta(\hat{\theta}_i(y_i)-b_i)^2-\tau\mathbf{Var}_i^*\left[\theta\middle|y_i\right]}^{\text{revenue from citizen } i\text{'s activities on the platform}}\,\mathrm{d}i\right],$$

citizen i's estimate of θ

$\in \mathbf{R}$; citizen i's bias

Revenue from credulous citizens:

$$v_C(f,s) := \mathbf{E}\left[\int_r^1 -\beta(\hat{\theta}(y_i)-b_i)^2\,\mathrm{d}i\right] = \mathbf{E}\left[\int_r^1 -\beta(y_i-b_i)^2\,\mathrm{d}i\right]$$

Cost to develop the algorithms:

$$\frac{c}{2}f^2+\frac{k}{2}\int_0^1 s_i^2\,\mathrm{d}i.$$

Solution concept

Pure-strategy Bayesian Nash Equilibria, henceforth equilibria.

In equilibrium, users' expectations are correct.

(f^*, s^*) is an equilibrium if and only if

$$(f^*, s^*) \in \arg\max_{f,s} \left\{ \underbrace{v_R(f, s; f^*, s^*) + v_C(f, s)}_{= v(f,s;f^*,s^*)} - \frac{c}{2} f^2 - \frac{k}{2} \int_0^1 s_i^2 \, \mathrm{d}i \right\}$$

Equilibrium

Equilibrium

Proposition. *There exists an essentially unique equilibrium. In the equilibrium, the platform chooses* $(f, s) = (f^S, s^S)$ *where:*

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

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

2. *For every citizen* i, s_i^S *is characterized by*

$$s_i^S = \begin{cases} \frac{2\beta}{k}\left(\frac{q+f^S}{p+q+f^S}\right) b_i & \text{if } i \in [0, r] \\ \frac{2\beta}{2\beta+k} b_i & \text{if } i \in (r, 1] \end{cases}$$

Proof sketch

Focus on the platform's profit from rational citizens.

Platform's incentives depend on the (rational) citizens' inferences:

Lemma. *Suppose that the rational citizens expect that the platform plays (f^*, s^*). Then each rational citizen i's posterior belief about the state θ, upon receiving signal y_i, is Gaussian:*

$$\theta | y_i \sim N\Bigg(\underbrace{\frac{q+f^*}{p+q+f^*}(y_i - s_i^*)}_{=\hat{\theta}_i(y_i)}, \ \underbrace{\frac{1}{p+q+f^*}}_{=\mathbf{Var}_i^*[\theta|y_i]} \Bigg).$$

Proof sketch

Platform's (expected) revenue from the rational citizens:

$$\int_0^r \mathbf{E}\left[-\beta\left(\frac{q+f^*}{p+q+f^*}(y_i - s_i^*) - b_i\right)^2\right] - \frac{\tau}{p+q+f^*}\,\mathrm{d}i$$

Thus, for each user i, platform wishes to

1. "minimize the spread" of $y_i = \theta + s_i + \epsilon_i$,
2. "pull" $y_i = \theta + s_i + \epsilon_i$ closer to bias b_i.

Proof sketch

From the platform's perspective:

$$\hat{\theta}_i(y_i) \sim N\left(\frac{q+f^*}{p+q+f^*}\left(s_i - s_i^*\right), \left(\frac{q+f^*}{p+q+f^*}\right)^2 \frac{p+q+f}{p(q+f)}\right).$$

To reduce spread, platform chooses higher f.

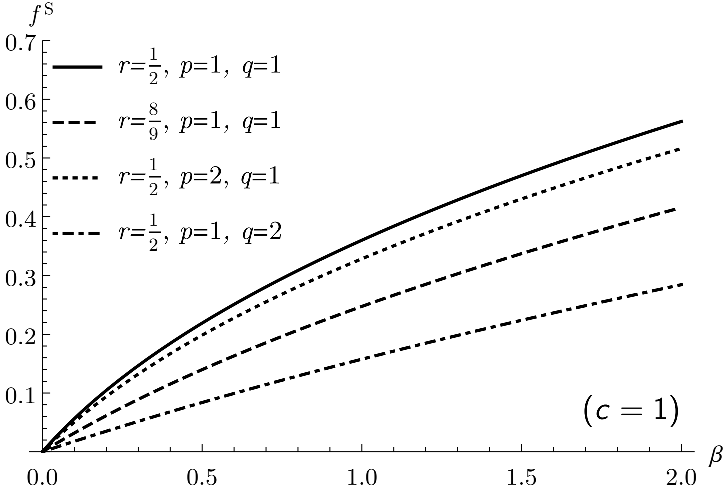
To pull $\hat{\theta}_i(y_i)$, platform chooses more extreme s_i.

In equilibrium, no incentive to reduce f or pull s_i further.

Finally, given higher p, users put less weight on signals for inference.

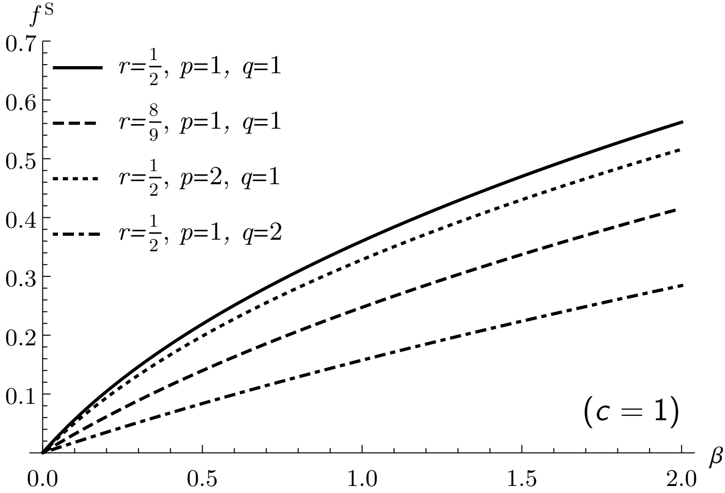
- Thus, platform has less incentives to filter.
- Platform's marginal return from slanting also decreases. □

Filter



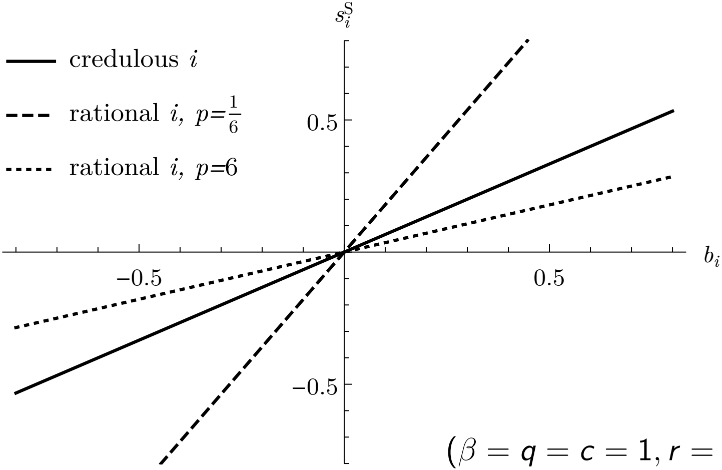
Filtering to cater to biases (β and r).

Filter



Higher p and q crowd out filtering incentives: $\frac{\mathrm{d}f^S}{\mathrm{d}p}, \frac{\mathrm{d}f^S}{\mathrm{d}q} \in (-1, 0)$..

Slants



$(\beta = q = c = 1, r = 1/2)$

$\mathrm{d}|s_i^S|/\,\mathrm{d}p < 0$ and $\partial|s_i^S|/\partial f^S > 0$ for rational citizens.

Social Conflict and Ethics

Social Conflicts

The citizens' estimates of θ typically disagree.

A regulator wishes to minimize conflicts due to disagreements.

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

Ethicality

An ethical platform's payoff is

$$v\left(f, s; f^{*}, s^{*}\right) - \kappa(f, s; f^{*}, s^{*}) - \frac{c}{2} f^{2} - \int_{0}^{1} \frac{k}{2} s_{i}^{2} \, \mathrm{d}i.$$

The model is otherwise identical.

Equilibrium (with Ethical Platform)

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

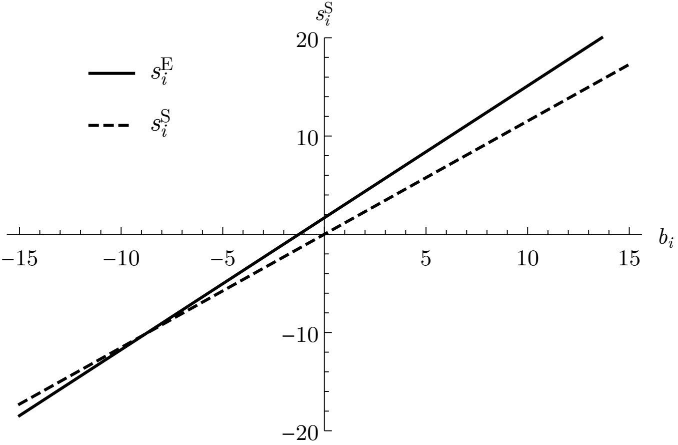
1. f^{E} *strictly exceeds* f^{S} *and is uniquely characterized by*

$$(\beta + h)\left[\frac{r}{(p+q+f^{\mathsf{E}})^2} + \frac{1-r}{(q+f^{\mathsf{E}})^2}\right] = c'\left(f^{\mathsf{E}}\right),$$

2. *For almost every* i, s_i^{E} *is characterized by*

$$s_i^E = \begin{cases} \frac{2\beta}{k}\left(\frac{q+f^E}{p+q+f^E}\right)\left(b_i + \frac{2h}{k+2\beta+2hr}\int_r^1 b_j \,\mathrm{d}j\right) & \text{if } i \text{ is rational} \\ \frac{2\beta}{2\beta+k+2h}\left(b_i + \frac{2h}{k+2\beta+2hr}\int_r^1 b_j \,\mathrm{d}j\right) & \text{if } i \text{ is credulous} \end{cases}$$

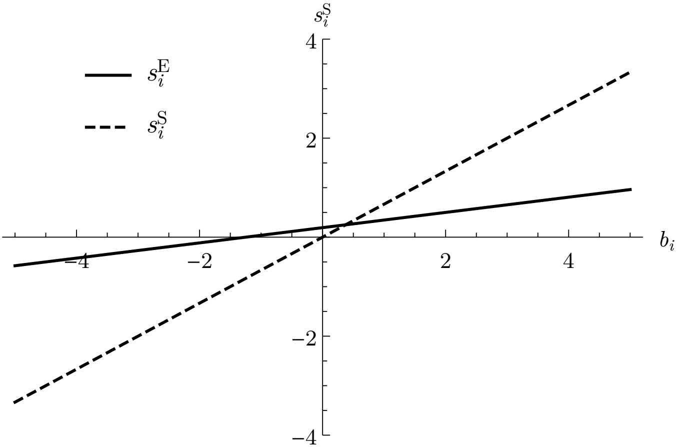
Slants



More personalized contents for rational citizens

$$\left(p = q = k = c = 1, r = .5, h = 5, \int_r^1 b_i \, \mathrm{d}i = 1\right)$$

Slants



Less personalized contents for credulous citizens

$$\left(p = q = k = c = 1, r = .5, h = 5, \int_r^1 b_i \,\mathrm{d}i = 1\right)$$

Proof sketch

Additional filtering incentives to reduce social conflict

$$\mathbf{E}\left[(\hat{\theta}_i(y_i) - \hat{\theta}_j(y_j))^2\right]$$

$$= \begin{cases} \mathbf{E}\left[(\mathbf{E}^*[\theta|y_j] - \mathbf{E}^*[\theta|y_i])^2\right] & \text{between rational } i \text{ and } j \\ \mathbf{E}\left[(y_j - y_i)^2\right] & \text{between credulous } i \text{ and } j \\ \mathbf{E}\left[(y_j - \mathbf{E}^*[\theta|y_i])^2\right] & \text{btw rational } i \text{ and credulous } j \end{cases}$$

Additional incentives to slant less for credulous citizens.

Given $f^E > f^S$, higher MR from slanting rational citizens' signals.

Perverse Ethics

Equilibrium Conflict Cost

Equilibrium conflict cost

$$\begin{array}{rl} \text{among} \\ \text{rational citizens} & : K_R(f,s) = \mathbf{E}\left[\frac{h}{2}\int_0^r \int_0^r \left(\hat{\theta}_j(y_j) - \hat{\theta}_i(y_i)\right)^2 \mathrm{d}j\,\mathrm{d}i\right] \\ \text{among} \\ \text{credulous citizens} & : K_C(f,s) = \mathbf{E}\left[\frac{h}{2}\int_r^1 \int_r^1 \left(\hat{\theta}_j(y_j) - \hat{\theta}_i(y_i)\right)^2 \mathrm{d}j\,\mathrm{d}i\right] \\ \text{between} \\ \text{the two groups} & : K_B(f,s) = \mathbf{E}\left[h\int_r^1 \int_1^r \left(\hat{\theta}_j(y_j) - \hat{\theta}_i(y_i)\right)^2 \mathrm{d}j\,\mathrm{d}i\right] \end{array}$$

Proposition. *The following holds.*

1. $\exists \bar{p} > 0$ *such that* $K_R(f^S, s^S) < K_R(f^E, s^E)$ *iff* $p > \bar{p}$.
2. $K_C(f^S, s^S) > K_C(f^E, s^E)$ *and* $K_B(f^S, s^S) > K_B(f^E, s^E)$.

Main result

The aggregate conflict cost in equilibrium (f, s):

$$K(f,s) := K_R(f,s) + K_B(f,s) + K_C(f,s) \quad (= \kappa(f,s;f,s)\ )$$

Corollary (Perverse ethics).

There is $r \in (0,1)$ *such that the following holds for every* $r \in [\bar{r}, 1)$*:*

$$\exists p' > 0 \quad \textit{such that} \quad K(f^S, s^S) < K(f^E, s^E) \quad \textit{whenever } p > p'.$$

"Unless the state is sufficiently uncertain, ethicality backfires."

Proof sketch

For all rational citizens $i \in [0, r]$,

$$\hat{\theta}_i(y_i) = A(y_i - s_i^*) \quad \text{where} \quad A = \frac{q + f^*}{p + q + f^*}.$$

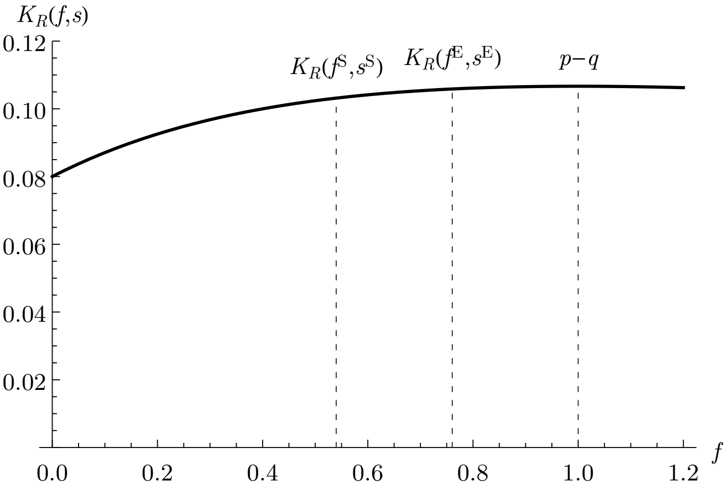
For now, suppose that the weight $A > 0$ is exogenously given.

$$K_R(f, s^*) = \frac{h}{2} \int_0^r \int_0^r A^2 \left[\frac{2}{q + f} \right] \mathrm{d}i \, \mathrm{d}j$$

In equilibrium,

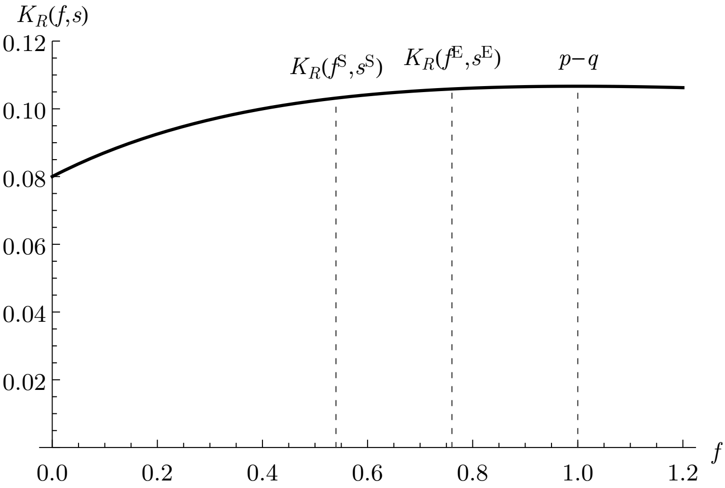
$$A = \frac{q + f^*}{p + q + f^*} = \frac{q + f}{p + q + f} \quad \text{increases in } f.$$

Proof sketch



$$f\uparrow \quad \Longrightarrow \quad \begin{cases} \text{less misinformation} \\ \text{but more weight on } y_i\text{'s} \end{cases}$$

Proof sketch



$$p\uparrow \implies \left\{ \begin{array}{l} \text{smaller learning benefit by filtering} \\ \text{\& crowding out filtering incentives} \end{array} \right. \implies f^E \downarrow \; f^S \downarrow .$$

Regulations

Filtering Floor

Consider legislation that ensures $f \geq \underline{f} > 0$.

Let (f^L, s^L) denote the algorithms with the filtering floor $\underline{f}$. Then,

$$f^L = \begin{cases} f^{\mathsf{S}} & \text{if } \underline{f} < f^{\mathsf{S}} \\ \underline{f} & \text{if } \underline{f} \geq f^{\mathsf{S}}. \end{cases}$$

where f^S denotes the status quo filtering level.

The legislation should be sufficiently aggressive to guarantee a success.

Proposition.

1. *$K_C^L < K_C(f^S, s^S)$ and $K_B^L < K_B(f^S, s^S)$ whenever $\underline{f} > f^S$.*
2. *$K_R^L < K_R(f^S, s^S)$ whenever $\underline{f} > f^S \geq p - q$.*
3. *Suppose $f^S < p - q$. Then, there is $F > f^S$ such that*

$$K_R^L < K_R(f^S, s^S) \quad \textit{if and only if} \quad \underline{f} > F.$$

Arrest of Misinformation

Increase the default precision level from q to q^A.

The equilibrium filtering level increases from f^S to f^A.

$f^A < f^S$ but $q^A + f^A > q + f^S$.

Proposition.

1. $K_C^A < K_C(f^S, s^S)$ *and* $K_B^A < K_B(f^S, s^S)$.
2. $K_R^A < K_R(f^S, s^S)$ *whenever* $f^S \geq p - q$.
3. *Suppose* $f^S < p - q$, *there is* $Q > q$ *such that*

$$K_R^A < K_R(f^S, s^S) \quad \textit{if and only if} \quad q^A > Q.$$

Fairness Doctrine

Originally a regulation on radio and television news.

The platform should deliver all contrasting views on pubic issues.

Effectively, the platform cannot slant (i.e., $s^F = 0$).

Proposition. *$K_R^F = K_R(f^S, s^S)$, $K_C^F < K_C(f^S, s^S)$, and $K_B^F < K_B(f^S, s^S)$.*

Media Literacy Campaign

Consider a media literacy campaign that increases r to 1.

$\mathbf{E}[(\hat{\theta}_i(y_i) - \hat{\theta}_j(y_i))^2]$ is lower when both i and j are rational, compared to the case at least one of i and j is credulous.

Hence, the media literacy campaign decreases the conflict, *provided that the platform does not respond to the campaign.*

Proposition. *There is* $\bar{\beta} \geq \underline{\beta} > 0$ *such that*

1. $p - q < f^M < f^S$ *and* $K^M > K(f^S, s^S)$ *whenever* $\beta > \bar{\beta}$*.*
2. $K^M < K(f^S, s^S)$ *whenever* $\beta < \underline{\beta}$*.*

The regulator may combine media literacy campaign with a regulation on the supply side (e.g., filtering floor).

Transparency

Proposition (Transparency).

Suppose that the platform's algorithms are publicly observable. Then

$$K(f^S, s^S) \leq K(f^E, s^E).$$

With transparency,

- The citizens would infer the state from their personalized signals based on the platform's actual choice of the algorithms.
- The ethical platform correctly internalizes the conflict costs.
- MacCarthy (2020)

Literature

Literature

Ethical design in computer science

Wu (2017), Kearns & Roth (2019).

Here: cautionary tale against conventional wisdom about ethicality.

Traditional media

Mullainathan & Shleifer (2005),
Anderson, Waldfogel & Strömberg (2015), Perego & Yuksel (2021).

Here: individual signals enable conflicts; slanting given rational users.

Internet media

Peitz & Reisinger (2015), MacCarthy (2020).

Here: platforms' incentives to fight spams; implications of ethicality.

Measure for disagreement

Kartik, Lee, & Suen (2021), Zanardo (2017).