Supplementary Material: Disinformation, Stochastic Harm, and Costly Effort: A Principal-Agent Analysis of Regulating Social Media Platforms

SUBMISSION 3063

ACM Reference Format:
Submission 3063. 2022. Supplementary Material: Disinformation, Stochastic Harm, and Costly Effort: A Principal-Agent Analysis of Regulating Social Media Platforms. In . ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 FUTURE WORK

Our model makes a number of simplifying assumptions. Treating public standards as explicit implies that a platform can guarantee a given probability of escaping punishment if it conforms to an explicit standard, which is an oversimplification of reality. The assumption that the platform can perfectly tune its proprietary model to flag toxic content is also unrealistic; technical challenges, although they may not pose the main obstacle to the practical control of disinformation, are nevertheless a real issue [2]. Extending the model to more richly model these aspects are important directions for future work.

Homogeneous Harm. Another simplifying assumption of our setup the expression of the harm from disinformation as a binary event. This binary notion of harm might seem restrictive, especially because the harm from disinformation can manifest in many forms: rare events such as the Capitol Hill riots or the Pizzagate shooting are dramatic and immediately observable, in comparison to harm from the degradation of public discourse or from the spread of climate change denial or anti-vaccine propaganda, which are more subtle manifestations.

Regardless, our setup is without loss of generality: recall that our social welfare expression (Equation **??**) quantifies the expected societal costs of harm from disinformation, $h(e)D$ should a harmful event occur for a platform's given level of effort. This expression can be augmented to capture different types of harm: we will simply substitute our harm function with different probability distribution functions for the different kinds of harm and include the associated societal costs. This practice straightforwardly preserves our model's notion of quantifying expected harm.

Heterogeneous Content. While our notion of measuring content harmfulness via a binary harmful event is without loss of generality, it is meaningfully different to consider the heterogeneity of content in terms of how harmful a particular piece of content is and how much benefit it brings to a social platform. Our simple model of the platform's costs of effort $c(e)$ implies the homogeneity of all content with respect to the value it brings to the platform, since it suggests each content item attains the same amount of engagement from users.

Yet, in reality, just as content is not homogeneous in terms of the varying degrees of harmfulness of each item, content will also differ in the levels of user engagement attained. Therefore, modeling this heterogeneous relationship of the harm and benefit of content in future work will likely drive different conclusions. For example, with such explicit modeling, one question we might hope to answer is whether highly toxic content is more likely to produce high levels

of user engagement (in the form of likes, shares, retweets, comments etc.) than less toxic content, thereby being more valuable to the platform. This will shed light on the degree to which the incentives of social platforms relating to the control of disinformation are misaligned with those of society, which, in turn, will inform the nature of any regulatory interventions required to realign these incentives.

Taxing toxicity. A *Pigouvian tax* is a tax on a market transaction that generates a negative externality borne by individuals not directly involved in the transaction [4]. Social platforms exhibit the precise criterion of generating negative externalities that calls for the levying of this tax: the more users a platform has, the more lucrative it is for advertisers to pay for the platform's services to target them with ads; and furthermore, the more time these users spend engaging with other users and content on the platform, the greater the opportunity for the platform to cater to the precise needs of advertisers. Thus, a platform benefits from more engagement than less, irrespective of whether such engagement is induced from harmful or benign content. But because it is only society that incurs the costs of harmful content, the idea behind taxation is to internalize the costs of toxicity to the original transaction between the platform and an advertiser.

Devising a taxation scheme requires a good harm model to measure content toxicity, using an access to data and expertise that is only available to platforms. Therefore, a mechanism designer (regulator) might instead impose taxation in a more crude manner. For example, the regulator can ask the platform report its cost function for moderating content and then tax the platform based on its report. Since the indirect costs of effort essentially capture the value of engagement, this mechanism levies a tax on user engagement on the platform more generally, rather than on the harmfulness of hosted content, which is the entity we wish to control on social platforms.

As shown by Proposition 5, the platform's costs of effort underpin the incentive problem for disinformation mitigation; thus, any effective mechanism must in some way be responsive to these costs. Naturally, such a mechanism must also factor in incentives that might prevent the platform from misreporting its true cost function for moderating content, in the hopes of attaining a lower tax rate, for instance.

2 PROOFS

Here we recall our main results and include proofs omitted from the main body of the paper.

Proposition 1. *Given a fixed adequate effort level e', there exists no fine scheme F that can incentivize the platform to exert more effort than e'.*

Proof. By contradiction. Suppose the platform prefers to exert effort $e > e'$. Thus, the following must hold:

$$\begin{aligned} & EU(e|e_c) > EU(e'|e_c) \\ \iff & -c(e) - rP_f(e|e_c)F > -c(e') - rP_f(e'|e_c)F \end{aligned}$$

By definition, $P_f(e'|e_c) = 0$ and therefore $P_f(e|e_c) = 0$. Thus,

$$\begin{aligned} & -c(e) - rP_f(e|e_c)F > -c(e') - rP_f(e'|e_c)F \\ \iff & -c(e) > -c(e') \\ \iff & c(e') > c(e) \end{aligned}$$

which does not hold for $e > e'$ because by assumption $c'(e) > 0$ for all e (contradiction). □

Lemma 1. *Fix a state e_c representing the current effort required by the public standard, and an arbitrary policy π, and let $e_h > e_c$ be the effort that the public standard will require if harm occurs. For all $e_2 > e_1 \geq e_c$,*

$$q_\pi(e_c, e_2) > q_\pi(e_c, e_1) \iff d(\pi, e_c) - v_\pi(e_h) > \frac{c(e_2) - c(e_1)}{\gamma(h(e_1) - h(e_2))}, \tag{1}$$

where $d(\pi, e_c) = g(e_c)v_\pi(\chi(e_c)) + (1 - g(e_c))v_\pi(e_c)$.

Proof. At e_c, the state-action value function for some effort e is given by:

$$\begin{aligned} q_\pi(e_c, e) &= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = e_c, A_t = e] \\ &= \sum_{e_c'} P(e_c' \mid s = e_c, a = e)[-c(e) + \gamma v_\pi(e_c')] \\ &= -c(e) + \gamma[h(e)v_\pi(e_h) + (1 - h(e))(g(e_c)v_\pi(\chi(e_c)) + (1 - g(e_c))v_\pi(e_c))] \end{aligned}$$

By substituting in $d(\pi, e_c) = g(e_c)v_\pi(\chi(e_c)) + (1 - g(e_c))v_\pi(e_c)$ we have:

$$q_\pi(e_c, e) = -c(e) + \gamma[h(e)v_\pi(e_h) + (1 - h(e))d(\pi, e_c)]. \tag{2}$$

Thus, for $q_\pi(e_c, e_2) > q_\pi(e_c, e_1)$, we have:

$$\begin{aligned} & -c(e_2) + \gamma[h(e_2)v_\pi(e_h) + (1 - h(e_2))d(\pi, e_c)] > \\ & -c(e_1) + \gamma[h(e_1)v_\pi(e_h) + (1 - h(e_1))d(\pi, e_c)] \\ \iff \quad & -c(e_2) + \gamma[h(e_2)v_\pi(e_h) + d(\pi, e_c) - h(e_2)d(\pi, e_c)] > \\ & -c(e_1) + \gamma[h(e_1)v_\pi(e_h) + d(\pi, e_c) - h(e_1)d(\pi, e_c)] \\ \iff \quad & -c(e_2) + \gamma h(e_2)v_\pi(e_h) + \gamma d(\pi, e_c) - \gamma h(e_2)d(\pi, e_c) > \\ & -c(e_1) + \gamma h(e_1)v_\pi(e_h) + \gamma d(\pi, e_c) - \gamma h(e_1)d(\pi, e_c) \\ \iff \quad & -c(e_2) + \gamma h(e_2)v_\pi(e_h) - \gamma h(e_2)d(\pi, e_c) > \\ & -c(e_1) + \gamma h(e_1)v_\pi(e_h) - \gamma h(e_1)d(\pi, e_c) \\ \iff \quad & -c(e_2) - \gamma h(e_2)(d(\pi, e_c) - v_\pi(e_h)) > \\ & -c(e_1) - \gamma h(e_1)(d(\pi, e_c) - v_\pi(e_h)) \\ \iff \quad & \gamma h(e_1)(d(\pi, e_c) - v_\pi(e_h)) - \gamma h(e_2)(d(\pi, e_c) - v_\pi(e_h)) > c(e_2) - c(e_1) \\ \iff \quad & \gamma(h(e_1) - h(e_2))(d(\pi, e_c) - v_\pi(e_h)) > c(e_2) - c(e_1) \\ \iff \quad & d(\pi, e_c) - v_\pi(e_h) > \frac{c(e_2) - c(e_1)}{\gamma(h(e_1) - h(e_2))}. \end{aligned}$$

□

Lemma 2. *Given a threshold strategy π^τ, the state value function $v_{\pi^\tau}(e_c)$ is fixed for all $e_c \leq \tau$.*

Proof. The state value function for some arbitrary $e_c \leq \tau$ is given by,

$$v_{\pi^\tau}(e_c) = -c(\tau) + \gamma[h(\tau)v_{\pi^\tau}(e_h) + (1 - h(\tau))(g(e_c)v_{\pi^\tau}(\chi(e_c)) + (1 - g(e_c))v_{\pi^\tau}(e_c))]. \tag{3}$$

Note that the platform's policy specifying effort for all $e_c \leq \tau$ is fixed by definition; that is, $\pi^\tau(e_c) = \tau$ for all $e_c \leq \tau$. Thus, the transition to state e_h is also fixed because the transition probability $h(\tau)$ is fixed. And similarly, the probability that harm does not occur is also fixed at $(1 - h(\tau))$.

Let $e_0 = \min \mathcal{S}$. We prove inductively that $v_{\pi^\tau}(e_k) = v_{\pi^\tau}(e_0)$ for all $e_0 \le e_k \le \tau$. The base case ($v_{\pi^\tau}(e_0) = v_{\pi^\tau}(e_0)$) is immediate. For the inductive step, assume that $v_{\pi^\tau}(e_{k-1}) = v_{\pi^\tau}(e_0)$. Then

$$\begin{aligned} v_{\pi^\tau}(e_k) &= -c(\tau) + \gamma[h(\tau)v_{\pi^\tau}(e_h) + (1-h(\tau))(1-g(e_k))v_{\pi^\tau}(e_k) + (1-h(\tau))g(e_k)v_{\pi^\tau}(e_{k-1})] \\ &= -c(\tau) + \gamma[h(\tau)v_{\pi^\tau}(e_h) + (1-h(\tau))(1-g(e_k))v_{\pi^\tau}(e_k) + (1-h(\tau))g(e_k)v_{\pi^\tau}(e_0)]. \end{aligned}$$

Thus, $v_{\pi^\tau}(e_k) = g(e_k)V_1 + (1-g(e_k))V_0$, where

$$\begin{aligned} V_1 &= -c(\tau) + \gamma[h(\tau)v_{\pi^\tau}(e_h) + (1-h(\tau))v_{\pi^\tau}(e_{k-1})] \\ &= -c(\tau) + \gamma[h(\tau)v_{\pi^\tau}(e_h) + (1-h(\tau))v_{\pi^\tau}(e_0)] \\ &= v_{\pi^\tau}(e_0) \end{aligned}$$

and

$$\begin{aligned} V_0 &= -c(\tau) + \gamma h(\tau)v_{\pi^\tau}(e_h) + \gamma(1-h(\tau))v_{\pi^\tau}(e_k) \\ &= -c(\tau) + \gamma h(\tau)v_{\pi^\tau}(e_h) + \gamma(1-h(\tau))[g(e_k)V_1 + (1-g(e_k))V_0]. \end{aligned}$$

Note that the following is also true for V_1:

$$\begin{aligned} V_1 &= -c(\tau) + \gamma[h(\tau)v_{\pi^\tau}(e_h) + (1-h(\tau))v_{\pi^\tau}(e_0)] \\ &= -c(\tau) + \gamma[h(\tau)v_{\pi^\tau}(e_h) + (1-h(\tau))g(e_k)v_{\pi^\tau}(e_0) + (1-h(\tau))(1-g(e_k))v_{\pi^\tau}(e_0)] \\ &= -c(\tau) + \gamma h(\tau)v_{\pi^\tau}(e_h) + \gamma(1-h(\tau))g(e_k)v_{\pi^\tau}(e_0) + \gamma(1-h(\tau))(1-g(e_k))v_{\pi^\tau}(e_0) \\ &= \sum_{j=0}^{\infty} \gamma^j (1-h(\tau))^j (1-g(e_k))^j [-c(\tau) + \gamma h(\tau)v_{\pi^\tau}(e_h) + \gamma(1-h(\tau))g(e_k)v_{\pi^\tau}(e_0)] \\ &= v_{\pi^\tau}(e_0) \end{aligned}$$

for all $g(e_k) \in [0,1]$.

Thus, for V_0:

$$\begin{aligned} V_0 &= -c(\tau) + \gamma h(\tau)v_{\pi^\tau}(e_h) + \gamma(1-h(\tau))[g(e_k)V_1 + (1-g(e_k))V_0] \\ &= -c(\tau) + \gamma h(\tau)v_{\pi^\tau}(e_h) + \gamma(1-h(\tau))[g(e_k)v_{\pi^\tau}(e_0) + (1-g(e_k))V_0] \\ &= -c(\tau) + \gamma h(\tau)v_{\pi^\tau}(e_h) + \gamma(1-h(\tau))g(e_k)v_{\pi^\tau}(e_0) + \gamma(1-h(\tau))(1-g(e_k))V_0 \\ &= \sum_{j=0}^{\infty} \gamma^j (1-h(\tau))^j (1-g(e_k))^j [-c(\tau) + \gamma h(\tau)v_{\pi^\tau}(e_h) + \gamma(1-h(\tau))g(e_k)v_{\pi^\tau}(e_0)] \\ &= V_1 \\ &= v_{\pi^\tau}(e_0). \end{aligned}$$

But then

$$\begin{aligned} v_{\pi^\tau}(e_k) &= g(e_k)V_1 + (1-g(e_k))V_0 \\ &= g(e_k)v_{\pi^\tau}(e_0) + (1-g(e_k))v_{\pi^\tau}(e_0) \\ &= v_{\pi^\tau}(e_0) \end{aligned}$$

for all $g(e_k) \in [0,1]$, and we are done. □

Proposition 2. *For all threshold strategies π^τ, we have that $v_{\pi^\tau}(e_h) \leq v_{\pi^\tau}(e_c)$ holds for all $e_c \in S$.*

Proof. For ease of notation, let $S = \{e^0, e^1, \ldots, e^h\}$ denote the set of all states with $e^0 < \cdots < e^h$, and let $\pi = \pi^\tau$ with $\tau = 0$. Note that this specification is w.l.o.g.; for $\tau > 0$, we will consider a subset of S such that the first state of this subset $e^0 = \sup\{e \in S \mid e \leq \tau\}$, since from Lemma 2 we know that the state value function for all $e \leq \tau$ is fixed.

Now we move on to the proof. Suppose the claim is false. Then $\{e \mid v_\pi(e) < v_\pi(e^h)\} \neq \emptyset$. Let $e^z = \min\{e \mid v_\pi(e) < v_\pi(e^h)\}$ and $d(e^j) = (1 - g(e^j))v_\pi(e^j) + g(e^j)v_\pi(e^{j-1})$ for all $0 \leq j \leq h$.

First, observe that

$$\begin{aligned} d(e^z) &= (1 - g(e^z))v_\pi(e^z) + g(e^z)v_\pi(e^{z-1}) \\ &\geq (1 - g(e^z))v_\pi(e^z) + g(e^z)v_\pi(e^z) \\ &= v_\pi(e^z), \end{aligned}$$

where the inequality follows from combining the assumptions $v_\pi(e^h) > v_\pi(e^z)$ with $v_\pi(e^{z-1}) \geq v_\pi(e^h)$, both from the definition of e^z. Note that if $e^z = e^0$, then the same result holds, since $g(e^0) = 0$.

It then follows that

$$\begin{aligned} v_\pi(e^z) &= -c(e^z) + \gamma[h(e^z)v_\pi(e^h) + (1 - h(e^z))d(e^z)] \\ &\geq -c(e^z) + \gamma[h(e^z)v_\pi(e^h) + (1 - h(e^z))v_\pi(e^z)] \\ &> -c(e^z) + \gamma[h(e^z)v_\pi(e^z) + (1 - h(e^z))v_\pi(e^z)] \\ &= -c(e^z) + \gamma v_\pi(e^z) \\ &\geq \sum_{j=0}^{\infty} \gamma^j(-c(e^z)). \end{aligned}$$

We now show inductively that $v_\pi(e^k) \geq v_\pi(e^h)$ for all $z \leq k < h$.

The base case is e^{h-1}. Suppose the contrary that $v_\pi(e^{h-1}) < v_\pi(e^h)$. Then we have

$$\begin{aligned} d(e^h) &= (1 - g(e^h))v_\pi(e^h) + g(e^h)v_\pi(e^{h-1}) \\ &\leq v_\pi(e^h), \end{aligned}$$

because $0 \leq g(e^h) \leq 1$, which gives

$$\begin{aligned} v_\pi(e^h) &= -c(e^h) + \gamma[h(e^h)v_\pi(e^h) + (1 - h(e^h))d(e^h)] \\ &\leq -c(e^h) + \gamma[h(e^h)v_\pi(e^h) + (1 - h(e^h))v_\pi(e^h)] \\ &= -c(e^h) + \gamma v_\pi(e^h) \\ &\leq \sum_{j=1}^{\infty} \gamma^j(-c(e^h)) \\ &< \sum_{j=1}^{\infty} \gamma^j(-c(e^z)) \\ &< v_\pi(e^z), \end{aligned}$$

contradicting the definition of e^z.

For the inductive step, assume that $v_\pi(e^k) \geq v_\pi(e^h)$, for some $z < k < h$. Then we show that $v_\pi(e^{k-1}) \geq v_\pi(e^h)$. Assume not; then similarly we have

$$\begin{aligned} d(e^k) &= (1 - g(e^k))v_\pi(e^k) + g(e^k)v_\pi(e^{k-1}) \\ &\leq (1 - g(e^k))v_\pi(e^k) + g(e^k)v_\pi(e^h) \\ &\leq (1 - g(e^k))v_\pi(e^k) + g(e^k)v_\pi(e^k) \\ &= v_\pi(e^k) \end{aligned}$$

and thus

$$\begin{aligned} v_\pi(e^k) &= -c(e^k) + \gamma[h(e^k)v_\pi(e^h) + (1 - h(e^k))d(e^k)] \\ &\leq -c(e^k) + \gamma[h(e^k)v_\pi(e^h) + (1 - h(e^k))v_\pi(e^k)] \\ &\leq -c(e^k) + \gamma[h(e^k)v_\pi(e^k) + (1 - h(e^k))v_\pi(e^k)] \\ &= -c(e^k) + \gamma v_\pi(e^k) \\ &\leq \sum_{j=1}^{\infty} \gamma^j(-c(e^k)) \\ &< \sum_{j=1}^{\infty} \gamma^j(-c(e^z)) \\ &< v_\pi(e^z) \\ &< v_\pi(e^h) \\ &\leq v_\pi(e^k), \end{aligned}$$

again yielding a contradiction.

Therefore, $v_\pi(e^k) \geq v_\pi(e^h)$ is true for all $z \leq k < h$, which in particular implies that the initial claim $\{e \mid v_\pi(e) < v_\pi(e^h)\} \neq \emptyset$ must be false, thus completing the proof. □

The following existing results support our main result in Theorem 1.

Lemma 3 ([1]). *Suppose f is a differentiable function of one variable in dom(f). Then f is convex if and only if*

$$f(y) - f(x) \geq f'(x)(y - x)$$

holds for all $x, y \in$ dom(f). And analogously for strict convexity,

$$f(y) - f(x) > f'(x)(y - x) \tag{4}$$

for all $x \neq y$.

Lemma 4 ([3]). *Given a pair of deterministic policies π and π' such that for all states $s \in S$*

$$q_\pi(s, \pi'(s)) \geq v_\pi(s),$$

then $v_{\pi'}(s) \geq v_\pi(s)$.

Theorem 1. *The optimal strategy π^* for the platform is a threshold strategy $\pi^* = \pi^{\hat{e}}$, with threshold*

$$\hat{e} = \sup\left\{ e \in [0, e_h] \,\middle|\, q_{\pi^e}(s^{-1}(e), e) - q_{\pi^e}(e_h, \pi^e(e_h)) \geq -\frac{c'(e)}{\gamma h'(e)} \right\}, \tag{5}$$

where $s^{-1}(e) = \sup\{e_c \in S \mid e_c \leq e\}$.

Proof. By contradiction. Suppose $\pi^{\hat{e}}$ is suboptimal. Then by the process of policy improvement, there must exist a state e_c where some effort $e \neq \pi^{\hat{e}}(e_c)$ guarantees a higher expected reward than $\pi^{\hat{e}}(e_c)$. Thus, we apply the policy improvement theorem (Lemma 4) to find any such e_c where $q_{\pi^{\hat{e}}}(e_c, e) > v_{\pi^{\hat{e}}}(e_c)$ holds, which would imply that a greedy deviation from $\pi^{\hat{e}}$ exists as the better policy.

Case 1 ($\forall e_c$): Less aggressive effort than e_c. The first deviation from $\pi^{\hat{e}}$ at any e_c might be to exert less aggressive effort $e < e_c$. Suppose that less aggressive effort e guarantees a higher expected reward than the required effort e_c. However, we know that lower effort than e_c does not guarantee a higher expected reward for all e_c because e_c by definition is the platform's individually-optimal level of effort. Thus, we have a contradiction and this deviation does not work.

Case 2 ($e_c \leq \hat{e}$): Less aggressive effort than $\hat{e}$. Suppose that the platform prefers to exert less aggressive effort e_1 such that $e_c \leq e_1 < \hat{e}$. Then $q_{\pi^{\hat{e}}}(e_c, e_1) > q_{\pi^{\hat{e}}}(e_c, \hat{e})$ must be true.

Thus, $q_{\pi^{\hat{e}}}(e_c, \hat{e}) > q_{\pi^{\hat{e}}}(e_c, e_1)$ *must not* be true (contrapositive); or, by substituting in equation (1) from Lemma 1, the following must not hold:

$$d(\pi^{\hat{e}}, e_c) - v_{\pi^{\hat{e}}}(e_h) > \frac{c(\hat{e}) - c(e_1)}{\gamma(h(e_1) - h(\hat{e}))}. \tag{6}$$

From the definition in (5), note that because $\hat{e}$ is the supremum taken over a closed interval, it satisfies the following equation (*intermediate value theorem*):

$$q_{\pi^{\hat{e}}}(s^{-1}(\hat{e}), \hat{e}) - q_{\pi^{\hat{e}}}(e_h, e_h) = -\frac{c'(\hat{e})}{\gamma h'(\hat{e})}. \tag{7}$$

Now consider the L.H.S of (6) and of (7). Recall that $d(\pi^{\hat{e}}, e_c) = g(e_c) v_{\pi^{\hat{e}}}(\chi(e_c)) + (1 - g(e_c)) v_{\pi^{\hat{e}}}(e_c)$. Since $\pi^{\hat{e}}(e_c) = \hat{e}$ is fixed for all $e_c < \hat{e}$, the value functions $v_{\pi^{\hat{e}}}(\chi(e_c))$ and $v_{\pi^{\hat{e}}}(e_c)$ must be equal (Lemma 2). Thus, $d(\pi^{\hat{e}}, e_c) = v_{\pi^{\hat{e}}}(e_c)$ as $0 \leq g(e_c) \leq 1$. Furthermore, because $s^{-1}(\hat{e}) < \hat{e}$ by definition, $q_{\pi^{\hat{e}}}(s^{-1}(\hat{e}), \hat{e}) = v_{\pi^{\hat{e}}}(e_c)$ must be true. Moreover, $v_{\pi^{\hat{e}}}(e_h) = q_{\pi^{\hat{e}}}(e_h, e_h)$ as $e_h \geq \hat{e}$. Thus, the L.H.S of (6) and of (7) are equal, or

$$d(\pi^{\hat{e}}, e_c) - v_{\pi^{\hat{e}}}(e_h) = q_{\pi^{\hat{e}}}(s^{-1}(\hat{e}), \hat{e}) - q_{\pi^{\hat{e}}}(e_h, e_h). \tag{8}$$

Suppose that the following is true of the R.H.S of (6) and (7):

$$-\frac{c'(\hat{e})}{\gamma h'(\hat{e})} > \frac{c(\hat{e}) - c(e_1)}{\gamma(h(e_1) - h(\hat{e}))} \tag{9}$$

Thus,

$$\begin{aligned} & -\frac{c'(\hat{e})}{\gamma h'(\hat{e})} > \frac{c(\hat{e}) - c(e_1)}{\gamma(h(e_1) - h(\hat{e}))} \\ \iff & -\frac{c'(\hat{e})}{h'(\hat{e})} > \frac{c(\hat{e}) - c(e_1)}{h(e_1) - h(\hat{e})} \\ \iff & -\frac{c'(\hat{e})(e_1 - \hat{e})}{h'(\hat{e})(e_1 - \hat{e})} > -\frac{c(e_1) - c(\hat{e})}{h(e_1) - h(\hat{e})} \\ \iff & \frac{c(e_1) - c(\hat{e})}{h(e_1) - h(\hat{e})} > \frac{c'(\hat{e})(e_1 - \hat{e})}{h'(\hat{e})(e_1 - \hat{e})} \end{aligned}$$

Notice that the final inequality is always true: we know by assumption that c is strictly convex ($c''(e) > 0$) and so from equation (4) in Lemma 3 it follows that the numerator of the L.H.S must be strictly greater than the numerator of the R.H.S, i.e., $c(e_1) - c(\hat{e}) > c'(\hat{e})(e_1 - \hat{e})$; similarly, because h is convex ($h''(e) \geq 0$), the denominator of the L.H.S must be be weakly greater than the denominator of the R.H.S, i.e., $h(e_1) - h(\hat{e}) \geq h'(\hat{e})(e_1 - \hat{e})$. Since $h'(e) < 0$ and $e_1 < \hat{e}$, it follows that the L.H.S fraction overall is strictly greater (*less negative*) than the R.H.S fraction (*more negative*).

Therefore, if (9) holds, then condition (6) must also hold because:

$$\begin{aligned} & q_{\pi^{\hat{e}}}(s^{-1}(\hat{e}), \hat{e}) - q_{\pi^{\hat{e}}}(e_h, e_h) = -\frac{c'(\hat{e})}{\gamma h'(\hat{e})} \\ \iff & d(\pi^{\hat{e}}, e_c) - v_{\pi^{\hat{e}}}(e_h) = -\frac{c'(\hat{e})}{\gamma h'(\hat{e})} \\ & > \frac{c(\hat{e}) - c(e_1)}{\gamma(h(e_1) - h(\hat{e}))}. \end{aligned}$$

If (6) holds, the contrapositive statement is false, and so the original statement must also be false; thus, the platform instead prefers to exactly exert effort $\hat{e}$, and no less, for all $e_c < \hat{e}$, a contradiction.

Case 3 ($\boldsymbol{e_c \leq \hat{e}}$): More aggressive effort than $\boldsymbol{\hat{e}}$. Suppose the platform prefers to exert excessive effort at some $e_2 > \hat{e}$. It follows that $q_{\pi^{\hat{e}}}(e_c, e_2) > q_{\pi^{\hat{e}}}(e_c, \hat{e})$ must hold, and so we have (Lemma 1):

$$d(\pi^{\hat{e}}, e_c) - v_{\pi^{\hat{e}}}(e_h) > \frac{c(e_2) - c(\hat{e})}{\gamma(h(\hat{e}) - h(e_2))}, \tag{10}$$

must also hold. Recall from (8) that,

$$d(\pi^{\hat{e}}, e_c) - v_{\pi^{\hat{e}}}(e_h) = q_{\pi^{\hat{e}}}(s^{-1}(\hat{e}), \hat{e}) - q_{\pi^{\hat{e}}}(e_h, e_h).$$

Thus,

$$\begin{aligned} & d(\pi^{\hat{e}}, e_c) - v_{\pi^{\hat{e}}}(e_h) > \frac{c(e_2) - c(\hat{e})}{\gamma(h(\hat{e}) - h(e_2))} \\ \iff & q_{\pi^{\hat{e}}}(s^{-1}(\hat{e}), \hat{e}) - q_{\pi^{\hat{e}}}(e_h, e_h) > \frac{c(e_2) - c(\hat{e})}{\gamma(h(\hat{e}) - h(e_2))} \end{aligned}$$

We know from (7) that,

$$q_{\pi^{\hat{e}}}(s^{-1}(\hat{e}), \hat{e}) - q_{\pi^{\hat{e}}}(e_h, e_h) = -\frac{c'(\hat{e})}{\gamma h'(\hat{e})}.$$

Thus, in order to guarantee that (10) holds, the following must be true:

$$\begin{aligned} & -\frac{c'(\hat{e})}{\gamma h'(\hat{e})} > \frac{c(e_2)-c(\hat{e})}{\gamma(h(\hat{e})-h(e_2))} \\ \iff & -\frac{c'(\hat{e})}{h'(\hat{e})} > \frac{c(e_2)-c(\hat{e})}{h(\hat{e})-h(e_2)} \\ \iff & -\frac{c'(\hat{e})(e_2-\hat{e})}{h'(\hat{e})(e_2-\hat{e})} > -\frac{c(e_2)-c(\hat{e})}{h(e_2)-h(\hat{e})} \\ \iff & \frac{c(e_2)-c(\hat{e})}{h(e_2)-h(\hat{e})} > \frac{c'(\hat{e})(e_2-\hat{e})}{h'(\hat{e})(e_2-\hat{e})} \end{aligned}$$

However, notice that this inequality *does not* hold for $e_2 > \hat{e}$: since c is strictly convex, we know from Lemma 3 that the numerator of the L.H.S is strictly greater than that of the R.H.S, i.e., $c(e_2) - c(\hat{e}) > c'(\hat{e})(e_2 - \hat{e})$; and similarly, because h is convex, the denominator of the L.H.S is weakly greater than that of the R.H.S, i.e., $h(e_2) - h(\hat{e}) \geq h'(\hat{e})(e_2 - \hat{e})$. Since $h'(e) < 0$ and $\hat{e} < e_2$, it follows that the L.H.S fraction overall must be strictly smaller (*more negative*) than the R.H.S fraction (*less negative*), that is,

$$\frac{c(e_2)-c(\hat{e})}{h(e_2)-h(\hat{e})} < \frac{c'(\hat{e})(e_2-\hat{e})}{h'(\hat{e})(e_2-\hat{e})}$$

must be true, a contradiction.

Case 4 ($\boldsymbol{e_c > \hat{e}}$): More aggressive effort than $\boldsymbol{e_c}$. We prove an intermediate result to arrive at our contradiction for this case. We first show that $\hat{e}$ is the optimal effort threshold for all threshold strategies.

Let τ be the smallest $\tau > \hat{e}$ satisfying $q_{\pi^\tau}(e, \pi^\tau(e)) \geq q_{\pi^{\hat{e}}}(e, \pi^{\hat{e}}(e))$ for all $e \in S$. Let $s^{-1}(\tau) = e_1 < \tau$. Then following the definition of $\hat{e}$ in (5), we have

$$\begin{aligned} & q_{\pi^\tau}(e_1, \tau) - v_{\pi^\tau}(e^h) < \frac{-c'(\tau)}{\gamma h'(\tau)} \\ \iff & q_{\pi^\tau}(e_1, \tau) - v_{\pi^\tau}(e^h) < \frac{-c'(\sigma)}{\gamma h'(\sigma)} \end{aligned}$$

for $e_1 < \sigma < \tau$ and $\tau - \sigma$ sufficiently small. But since $d(\pi^\tau, e_1) = q_{\pi^\tau}(e_1, \tau)$ (Lemma 2), we have

$$\begin{aligned} d(\pi^\tau, e_1) - v_{\pi^\tau}(e^h) &< \frac{-c'(\sigma)}{\gamma h'(\sigma)} \\ &< \frac{c(\tau)-c(\sigma)}{\gamma(h(\sigma)-h(\tau))}, \end{aligned}$$

which implies by Lemma 1 that $q_{\pi^\tau}(e_1, \sigma) \geq q_{\pi^\tau}(e_1, \tau)$, and hence by the policy improvement theorem, $v_{\pi^\sigma}(e) \geq v_{\pi^\tau}(e)$ for all $e \in S$, contradicting the definition of τ. Hence there is no such threshold $\tau > \hat{e}$, and so $\hat{e}$ is the optimal threshold among all threshold strategies.

Now suppose the platform prefers to exert more aggressive effort at $e_2 > e_c$ for some $e_c > \hat{e}$. Thus, by Lemma 1, the following must hold:

$$\begin{aligned} & d(\pi^{\hat{e}}, e_c) - v_{\pi^{\hat{e}}}(e_h) > \frac{c(e_2)-c(e_c)}{\gamma(h(e_c)-h(e_2))} \\ \iff & g(e_c) v_{\pi^{\hat{e}}}(\chi(e_c)) + (1-g(e_c)) v_{\pi^{\hat{e}}}(e_c) - v_{\pi^{\hat{e}}}(e_h) > \frac{c(e_2)-c(e_c)}{\gamma(h(e_c)-h(e_2))}. \end{aligned}$$

Thus, we have:

$$\begin{aligned} g(e_c)v_{\pi^{\hat{e}}}(\chi(e_c)) + (1-g(e_c))v_{\pi^{\hat{e}}}(e_c) - v_{\pi^{\hat{e}}}(e_h) &> \frac{c(e_2)-c(e_c)}{\gamma(h(e_c)-h(e_2))} \\ &> -\frac{c'(e_c)}{\gamma h'(e_c)} \\ &> -\frac{c'(\hat{e})}{\gamma h'(\hat{e})} \\ &= q_{\pi^{\hat{e}}}(s^{-1}(\hat{e}),\hat{e}) - q_{\pi^{\hat{e}}}(e_h,e_h) \\ &= v_{\pi^{\hat{e}}}(s^{-1}(\hat{e})) - v_{\pi^{\hat{e}}}(e_h). \end{aligned}$$

Thus,

$$\begin{aligned} & g(e_c)v_{\pi^{\hat{e}}}(\chi(e_c)) + (1-g(e_c))v_{\pi^{\hat{e}}}(e_c) - v_{\pi^{\hat{e}}}(e_h) > v_{\pi^{\hat{e}}}(s^{-1}(\hat{e})) - v_{\pi^{\hat{e}}}(e_h) \\ \iff \quad & g(e_c)v_{\pi^{\hat{e}}}(\chi(e_c)) + (1-g(e_c))v_{\pi^{\hat{e}}}(e_c) > v_{\pi^{\hat{e}}}(s^{-1}(\hat{e})), \end{aligned}$$

which implies that $v_{\pi^{\hat{e}}}(e_c) > v_{\pi^{\hat{e}}}(s^{-1}(\hat{e}))$ and/or $v_{\pi^{\hat{e}}}(\chi(e_c)) > v_{\pi^{\hat{e}}}(s^{-1}(\hat{e}))$. Thus, it follows that a new threshold strategy with threshold strictly greater than $\hat{e}$ will be preferable to $\hat{e}$, since exerting more aggressive effort e_2 in state e_c such that $e_2 > e_c > \hat{e}$ yields a better value. However, this implication contradicts our intermediate result because no threshold greater than $\hat{e}$ is optimal and we are done.

The process of policy improvement must give us a strictly better policy except when the original policy is already optimal [3]. Since there exists no greedy deviation $e \neq \pi^{\hat{e}}(e_c)$ such that $q(e_c,e) > q(e_c,\pi^{\hat{e}}(e_c))$ is true for any e_c, the proposed policy $\pi^{\hat{e}}$ must be optimal, thus completing the proof. □

Proposition 3. *For any given socially optimal level of effort e^*, there exists a MDP consistent with our given conditions such that the optimal policy for the platform is a threshold strategy with threshold $\tau = e^*$.*

Proof. We know from Theorem 1 that under the specified conditions, the platform's optimal effort at any state e_c is a threshold strategy with threshold $\tau = \hat{e}$. In order to induce e^* as the optimal threshold, $\hat{e}$ must equal e^*; thus, from the defining constraint in (5), there must exist some e_h such that the following holds:

$$\begin{aligned} & q_{\pi^{e^*}}(s^{-1}(e^*),e^*) - q_{\pi^{e^*}}(e_h,\pi^{e^*}(e_h)) = -\frac{c'(e^*)}{\gamma h'(e^*)} \\ \iff \quad & v_{\pi^{e^*}}(e_0) - v_{\pi^{e^*}}(e_h) = -\frac{c'(e^*)}{\gamma h'(e^*)}. \end{aligned} \tag{11}$$

where $e_0 = s^{-1}(e^*) \leq e^*$ (by definition).

Thus, we have

$$\begin{aligned} v_{\pi^{e^*}}(e_0) &= -c(e^*) + \gamma[h(e^*)v_{\pi^{e^*}}(e_h) + (1-h(e^*))v_{\pi^{e^*}}(e_0)] \\ v_{\pi^{e^*}}(e_0) &= -c(e^*) + \gamma h(e^*)v_{\pi^{e^*}}(e_h) + \gamma(1-h(e^*))v_{\pi^{e^*}}(e_0) \\ v_{\pi^{e^*}}(e_0) - \gamma(1-h(e^*))v_{\pi^{e^*}}(e_0) &= -c(e^*) + \gamma h(e^*)v_{\pi^{e^*}}(e_h) \\ v_{\pi^{e^*}}(e_0)[1-\gamma(1-h(e^*))] &= -c(e^*) + \gamma h(e^*)v_{\pi^{e^*}}(e_h), \end{aligned}$$

and finally

$$v_{\pi^{e^*}}(e_0) = \frac{-c(e^*) + \gamma h(e^*)v_{\pi^{e^*}}(e_h)}{1-\gamma(1-h(e^*))}. \tag{12}$$

By substituting (12) in (11), we have

$$v_{\pi^{e^*}}(e_0) - v_{\pi^{e^*}}(e_h) = -\frac{c'(e^*)}{\gamma h'(e^*)}$$

$$\frac{-c(e^*) + \gamma h(e^*) v_{\pi^{e^*}}(e_h)}{1-\gamma(1-h(e^*))} - v_{\pi^{e^*}}(e_h) = -\frac{c'(e^*)}{\gamma h'(e^*)}$$

$$-c(e^*) + \gamma h(e^*) v_{\pi^{e^*}}(e_h) - (1-\gamma(1-h(e^*)))v_{\pi^{e^*}}(e_h) = -\frac{c'(e^*)(1-\gamma(1-h(e^*)))}{\gamma h'(e^*)}$$

$$-v_{\pi^{e^*}}(e_h)[-\gamma h(e^*) + 1 - \gamma(1-h(e^*))] = c(e^*) - \frac{c'(e^*)(1-\gamma(1-h(e^*)))}{\gamma h'(e^*)}$$

$$-v_{\pi^{e^*}}(e_h)[-\gamma h(e^*) + 1 - \gamma + \gamma h(e^*))] = c(e^*) - \frac{c'(e^*)(1-\gamma(1-h(e^*)))}{\gamma h'(e^*)}$$

$$-v_{\pi^{e^*}}(e_h)(1-\gamma) = c(e^*) - \frac{c'(e^*)(1-\gamma(1-h(e^*)))}{\gamma h'(e^*)},$$

and finally

$$-v_{\pi^{e^*}}(e_h) = \left(\frac{1}{1-\gamma}\right)\left(c(e^*) - \frac{c'(e^*)(1-\gamma(1-h(e^*)))}{\gamma h'(e^*)}\right). \tag{13}$$

We show by the *intermediate value theorem* (IVT) that (13) holds for some $e_h \in (e_{min}, e_{max})$ where e_{min} and e_{max} correspond to the lowest and highest possible levels of effort, respectively.

Let $G(e_h) = -v_{\pi^{e^*}}(e_h) - K$ where $K = \left(\frac{1}{1-\gamma}\right)\left(c(e^*) - \frac{c'(e^*)(1-\gamma(1-h(e^*)))}{\gamma h'(e^*)}\right)$. First, observe that $-v_{\pi^{e^*}}(e_h) \in (c(e_h), \frac{c(e_h)}{1-\gamma})$ by construction. Moreover, note that $K > 0$. Thus, we have the following at the lower bound of e_h:

$$\begin{aligned} G(e_{min}) &= -v_{\pi^{e^*}}(e_{min}) - K \\ &< \frac{c(e_{min})}{1-\gamma} - \left(\frac{1}{1-\gamma}\right)\left(c(e^*) - \frac{c'(e^*)(1-\gamma(1-h(e^*)))}{\gamma h'(e^*)}\right) \\ &= \frac{c(e_{min})}{1-\gamma} - \frac{c(e^*)}{1-\gamma} + \left(\frac{1}{1-\gamma}\right)\left(\frac{c'(e^*)(1-\gamma(1-h(e^*)))}{\gamma h'(e^*)}\right) \\ &< 0, \end{aligned}$$

for all $e^* \geq e_{min}$.

And we have the following at the upper bound of e_h:

$$\begin{aligned} G(e_{max}) &= -v_{\pi^{e^*}}(e_{max}) - K \\ &> c(e_{max}) - K \\ &> 0, \end{aligned}$$

which holds because the cost of exerting maximum possible effort, or $c(e_{max})$, is sufficiently large (by assumption).

Hence, because $G(e_{min}) < 0 < G(e_{max})$, it follows by the IVT that there exists some $e_h \in (e_{min}, e_{max})$ such that

$$\begin{aligned} & G(e_h) = 0 \\ \iff & -v_{\pi^{e^*}}(e_h) - K = 0 \\ \iff & -v_{\pi^{e^*}}(e_h) = K \\ \iff & -v_{\pi^{e^*}}(e_h) = \left(\frac{1}{1-\gamma}\right)\left(c(e^*) - \frac{c'(e^*)(1-\gamma(1-h(e^*)))}{\gamma h'(e^*)}\right), \end{aligned}$$

and we are done. □

Proposition 4. *The platform's optimal stable effort is guaranteed to be socially suboptimal unless the public standard becomes excessive by requiring effort $e_h > e^*$ if harm occurs.*

Proof. This result directly follows from the defining condition of the platform's stable effort $\hat{e}$ in (5). For $\hat{e}$ to equal e^*, the required effort e_h must be strictly greater than e^*. By definition, $\hat{e}$ is the supremum over the closed interval $[0, e_h]$ and so if $e_h < e^*$, then $\hat{e} < e^*$ is also true.

If $e_h = e^*$, then $\hat{e} < e^*$ is also true; the L.H.S of (5) equals zero for $e = e_h$, or

$$q_{\pi^{e_h}}(s^{-1}(e_h), e_h) - q_{\pi^{e_h}}(e_h, \pi^{e_h}(e_h)) = 0,$$

while the R.H.S is always positive, or

$$-\frac{c'(e_h)}{\gamma h'(e_h)} > 0,$$

since $c'(e) > 0$ and $h'(e) < 0$ for all $e \in [0, 1]$, and therefore the inequality is not satisfied. Thus, $e_h > e^*$ must be true in order for the platform's stable effort $\hat{e}$ to equal e^*. □

Proposition 5. *There is no way of adjusting the effort e_c required by the public standard, purely as a function of the harm function h and the cost of damages D, without regard to the cost function c, such that the platform is always incentivized to exert the socially-optimal level of effort.*

Proof. Consider the simplest possible case where we assume there exist only two possible cost functions, c_1 and c_2. Let e_1^* be the socially-optimal effort induced by cost function c_1 and e_2^* be that induced by cost function c_2, and let $e_2^* > e_1^*$. Note that the e_1^* and e_2^* can be trivially computed from the expected social welfare equation, $EW(e) = -h(e)D - c(e)$, by equating the marginal social welfare to zero.

Suppose that the actual socially-optimal effort is e_1^*. Note first that if the regulator sets the public standard to require effort $e_c = e_1^*$, there should not be any increase in this level of effort because a transition to some new $e_h > e_c$ will mandate excessive and therefore suboptimal effort as the platform at least follows the public standard's specified effort (by assumption). Thus, if the public standard is set to require effort e_1^*, we are done.

However, now suppose that the actual socially-optimal effort is e_2^* and the public standard currently specifies e_1^* as the required effort. Then, an increase in effort to e_2^* is necessary to incentivize the platform to exert the socially-optimal effort (contradiction).

Similarly, a decrease in the effort required by the public standard does not guarantee that platform is always induced to exert the socially-optimal effort level: If e_2^* is socially-optimal, then a decrease in the required effort to some $e_c < e_2^*$ does not guarantee that the platform will continue to exert effort at e_2^*; the platform may exert less and therefore socially suboptimal effort at e_c, since without increasing the required effort, there is no way for a static public standard to induce the platform to exert more than the prescribed effort as shown in Proposition 1. Therefore, if the public standard is set to require e_1^*, the platform is no longer guaranteed to exert the socially-optimal e_2^*.

Thus, because no adjustment to the public standard's required effort works, the regulator needs to know whether the platform's true cost function is c_1 or c_2 to incentivize the socially-optimal effort at all times. And since there exist more than just two possible choices for the platform's actual cost function, there is no way for a regulator to guarantee that the platform exerts socially optimal effort for mitigating disinformation with any increasing or decreasing adjustments to the public standard's required effort. □

REFERENCES

[1] Stephen Boyd and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press. https://doi.org/10.1017/CBO9780511804441

[2] Karen Hao. 2021. He got Facebook hooked on AI. Now he can't fix its misinformation addiction. Available at https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/.

[3] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (second ed.). The MIT Press.

[4] Tax Foundation 2020. Pigouvian Tax Definition. https://taxfoundation.org/tax-basics/pigouvian-tax/. Accessed: 2022-01-20.