Discussion Paper Series – CRC TR 224

Discussion Paper No. 698
Project B 05

Moderating Content-Hosting Platforms

Robin Ng[1]
Greg Taylor[2]

August 2025

[1] Department of Economics and MaCCI, University of Mannheim. Email: robin@robinng.com
[2] Oxford Internet Institute, University of Oxford. Email: greg.taylor@oii.ox.ac.uk

Support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

Moderating Content-Hosting Platforms*

Robin Ng† and Greg Taylor‡

July 29, 2025

Abstract

We study how content moderation facilitates communication on online platforms. A sender transmits information to a receiver, exerting effort to signal their truthfulness. Communication fails without moderation because the effort required is prohibitive. Moderation resolves this problem by making effort a more powerful signal of veracity. However, moderation crowds-out sender effort, decreasing content quality on the platform. A socially optimal or profit-maximizing policy may therefore involve limited moderation. We study the choice between being a platform or broadcaster, how moderation influences competition for attention, and the effects of misinformation actors, AI-generated content, and moderator errors on the sustainability of communication.

1 Introduction

Many online environments for information exchange are subject to moderation, often conducted by a third party that filters information to remove messages that are misleading or harmful. The advent of 'Web 2.0' introduced a wave of technologies that enabled ordinary (often anonymous) users to create their own online content. Users of platforms such as Wikipedia, Facebook, YouTube, or Amazon Marketplace may produce false or fraudulent content in an attempt to mislead others into acting in their own interests, e.g., to vote in a particular way or to buy their sponsor's product. This has become a significant policy issue. Regulators like the Federal Trade Commission and European Commission have moved to

*We are grateful to Hesi Bar-Isaac, Felix Chopra, Ole Jann, Ellen Muir, Martin Peitz, Francisco Poggi, Jesse Shapiro, Camille Urvoy, Jonas Von Wangenheim, Allen Vong, Jakob Wegmann, Julian Wright, and participants at various seminars. Support by the Deutsche Forschungsgemeinschaft (DFG) through CRC TR 224 (Project B05) is gratefully acknowledged.

†Department of Economics and MaCCI, University of Mannheim; robin@robinng.com

‡Oxford Internet Institute, University of Oxford; greg.taylor@oii.ox.ac.uk

prevent deceptive product promotion by creators or incentivized online reviews.[1] Meanwhile, the role of misinformation in debates around politics, health, or climate change is an active subject of policy attention. In response to these concerns, many online platforms operate moderation policies to remove misleading content. For example, it is policy for YouTube and Facebook to remove content that contains harmful misinformation or undisclosed product placements, while Amazon has a team of thousands working to remove fake and misleading reviews.[2]

This paper studies the role that moderation can play in enabling effective communication when there are strategic incentives to mislead. We consider an environment where a sender can, at some cost, join a content-hosting platform. On the platform the sender is informed about the state of the world and sends a message describing it to a receiver. In addition, the sender may also be able to exert observable effort. An example of a message where effort is costly might be YouTube creators making more entertaining videos with engaging scripts and reducing cognition costs with well-placed infographics. A receiver benefits from both the informative and entertainment aspects of a video. In other words, they benefit from learning the truth, and may benefit more if the sender exerts higher effort. However, the sender is strategic and prefers to persuade the receiver to take a particular action, regardless of whether it is in the receiver's best interests. To facilitate communication between the two parties, a benevolent moderator can inspect a random sample of messages and reject those that are untruthful.

As a benchmark, we show that moderation or sender effort in isolation do little to enable communication. If senders cannot exert effort then communication is possible only if every message is moderated without error, which is unlikely to be feasible on platforms with millions of users. Conversely, without moderation communication can only be sustained if senders dissipate their entire surplus through effort, leaving them with no incentive to join the platform in the first place.

This picture changes substantially when moderation and sender effort are paired together. Moderation reduces senders' willingness to put effort into misleading messages because such messages are sometimes deleted and the effort wasted. This makes it easier to sustain incentive compatibility and means senders can persuade receivers of their truthfulness without dissipating their full surplus in effort. Hence, moderation enables truthful communication

[1] Reviewers paid to mislead are a prevalent problem also monitored by the competition and Markets Authority, https://www.bbc.com/news/technology-65336369, accessed 28 July 2025.

[2] See https://support.google.com/youtube/answer/10834785?sjid=78889283863190386-EU, https://support.google.com/youtube/answer/154235, https://transparency.meta.com/en-gb/policies/community-standards/misinformation/, https://www.facebook.com/business/help/221149188908254 and https://trustworthyshopping.aboutamazon.com/how-amazon-maintains-a-trusted-review-experience, accessed 28 July 2025.

while also increasing sender participation. However, because moderation makes it easier to persuade a receiver, senders respond by exerting less effort in equilibrium. When receivers value high-effort messages, the socially optimal moderation policy therefore involves a trade-off because inducing sender participation crowds-out effort. Consequently, the socially optimal level of moderation may be relatively low even if moderation is completely costless.

A similar trade-off arises when moderation is conducted by a profit-driven content-hosting platform. The platform must get both senders and receivers on board. Moderation attracts senders to the platform by reducing their cost of credible communication. But too much moderation crowds-out effort, which can make the platform less valuable to receivers and advertisers. Thus, platforms will often implement an interior level of moderation. Alternatively, a platform could incur the cost of producing its own content and sidestep the need to manage sender incentives. We show that the choice between being a content-hosting platform or a content-producing 'broadcaster' depends on the severity of the participation-effort trade-off.

Next we consider some factors that may limit the efficacy of moderation. First, we consider sources of misinformation who are sponsored (e.g., by a rogue state). Second, we suppose the moderator or sender sometimes makes a mistake in identifying the true state. Thirdly, we consider situations where senders can use AI to easily generate false messages. All three cases undermine the receiver's willingness to trust a message and call for a higher level of moderation in response. In extreme cases, the friction might be so large that truthful communication can no longer be sustained at all.

We extend the model to a dynamic setting in two ways. When senders arrive sequentially over time, each sender faces a temptation to free-ride by waiting for someone else to incur the effort of communicating with the receiver. Additional moderation is then needed to overcome this free-riding problem and sustain communication. Conversely, if a sender is long-lived and has concerns for future payoffs, truthful communication can occur at lower levels of moderation if moderation involves removing the sender from the platform. This is because costly (future) re-entry is in itself a deterrence for lying.

Next, we study some bigger departures from the baseline setup and illustrate two alternative mechanisms by which moderation can facilitate communication. We consider a situation where two senders with opposing incentives compete in effort for the receiver's attention. Moderation endogenously gives truthful senders a competitive advantage, meaning the market more effectively selects accurate messages. But this softens the competition for attention, resulting in less sender effort. We also consider situations where a platform rewards sender effort (e.g., by sharing ad revenues). In this environment, it is possible to sustain communication even without any signaling because the moderator prevents transfers

to senders caught lying. Although the details of the mechanisms are different, both of these extensions preserve the core participation-effort trade-off.

Lastly, we verify that our results are robust to a change in timing where senders learn the state before joining the platform, a switch from moderation to fact-checking, and where moderation can be conditioned on effort.

The rest of the paper is structured as follows: after reviewing the literature, Section 2 describes our setting and Section 3 provides the analysis of our equilibrium, discussing the key trade-offs. In Section 4 we study the socially optimal level of moderation and a platform's optimal strategy. Some limitations of moderation in sustaining informative communication are considered in Section 5 and Section 6 extends the model to dynamic settings. In Section 7 we discuss some alternative mechanisms and Section 8 explains several ways in which our results are robust. Section 9 concludes. Omitted proofs are in Appendix A, and Appendix B gathers additional robustness analysis.

1.1 Literature review

A limited number of theoretical papers study the *moderation of user-generated content*. When information transmission is noisy, Jackson et al. (2022) shows limiting content sharing can improve overall information quality. Madio and Quinn (2024) describes how profit concerns can result in platforms moderating content beyond the socially optimal level. Our paper addresses a similar question to Dwork et al. (2024), which discusses how content moderation can improve user participation on a platform. Mostagir and Siderius (2022) and Acemoglu et al. (2024) show content moderation can backfire when users are Bayesian, allowing undetected misinformation to become more easily trusted, an effect shown empirically by Pennycook et al. (2020). In Bar-Isaac et al. (2025) a platform selling certification of misleading messages has to attract and certify enough useful messages to make certification credible, which can benefit users. Hence, like in our communications game, they find limited moderation can be optimal.

Most closely related to our work is a nascent literature on *content moderation in communication games*. Kominers and Shapiro (2024) study an environment where moderation policies may be opaque, leading receivers to mistrust the moderator. Even a mistrusted moderator can improve the receiver's payoff by selectively blocking information that enables harmful acts, but the same is not true of content that instead promotes harmful beliefs. Our work differs by endogenizing the incentives of a sender under a known moderation policy, and allowing senders to exert costly effort.

Our paper considers an application of *communication and signaling games* (Crawford and

Sobel, 1982; Spence, 1973) with information asymmetries to an environment of user-generated content on platforms. This includes the classic idea—first introduced by Nelson (1974)—that advertising expenditure can signal product quality. In considering how a receiver may improve his information, many have explored the role of (costly) inspection by the receiver following some action (Cameron and Rosendorff, 1993; Jeffery S Banks, 2013; Bilancini and Boncinelli, 2018; Bester et al., 2021; Rahman, 2012; Figueroa and Guadalupi, 2021). Garfagnini (2017) studies a signaling game where receivers may choose to inspect senders' messages after seeing the message. Our setting differs not only by introducing an independent moderator but also their commitment to an inspection rule at the beginning of the game rather than as a response to the observed signal.

By considering a benevolent moderator, we also relate to the branch on *mediated communication* following Myerson (1986) (see also Arieli et al., 2023; Ganguly and Ray, 2023; Ivanov, 2014; Salamanca, 2021). In these games, a mediator processes a signal, and transforms it into a coarse recommendation to the receiver. Despite introducing noise, which can obscure players' types from each other, the mediator is able to create a more informative equilibrium which cannot be achieved in cheap talk (Ben-Porath, 2003; V. Krishna and Morgan, 2004; R. V. Krishna, 2007). In many environments, such as moderated online fora, moderators cannot arbitrarily transform messages, but rather either accept or delete them. We focus on the possibility of communication under such constrained moderation and when senders can exert effort that is also payoff-relevant for receivers.

Supported by the evidence on the implications of *misinformation and biased information* (DellaVigna and Kaplan, 2007; Alsem et al., 2008; Kartal and Tyran, 2022; Ershov and Morales, 2024), H. Li and W. Li (2013) studies a communications game where competing senders can either provide signals informative of their own quality or misinform receivers of their opponent's quality, characterize when higher quality senders choose to employ misinformation. We approach misinformation differently, focusing on how senders conducting misinformation can distort the entry incentives of other senders and affect the socially optimal moderation level.

While a number of empirical works have studied the role of moderation on user behavior (Chopra et al., 2022; Berger et al., 2025; Lin et al., 2024; Ahmad et al., 2024; Henry et al., 2022; Horta Ribeiro et al., 2023), there is limited empirical evidence on the behavior of content creators. For example, Beknazar-Yuzbashev et al. (2025) and Andres and Slivko (2021) show that the amount of toxic posts on platforms decreases following the use of moderation. When evaluating content creator participation in response to content moderation, the evidence is mixed (Jiménez-Durán, 2023; Beknazar-Yuzbashev et al., 2025; Mattozzi et al., 2022; Andres and Slivko, 2021). We provide one possible mechanism that shows how content moderation

can promote participation of content creators.

2 Setup

The state of the world is $w \in W = \{0, 1\}$, with $\Pr(w = 1) = p$. A *sender* chooses whether to incur idiosyncratic cost c to join an information sharing platform or not, distributed according to differentiable CDF G on support $[\underline{c}, \overline{c}]$, with $\underline{c} \geq 0$. Conditional on joining, the sender observes w and chooses an action $s = (m, e)$, comprising a message ($m \in W$) and an effort level ($e \in \mathbb{R}_+$). If the sender does not join the platform then no message is transmitted and we adopt the convention that $m = \emptyset$ and $e = 0$. In a slight abuse of notation, write $s = \emptyset$ as shorthand for $s = (\emptyset, 0)$.[3] We also use $m(s)$ and $e(s)$ to denote the message and effort components of s.

A *moderator* publicly announces a probability, I, with which it inspects the sender's message. If the message is inspected and $m \neq w$ then the moderator deletes the message (formally, the chosen s is replaced with $s = \emptyset$).[4] Proceeding by backward induction, we initially treat I as a parameter and focus on the communication sub-game that follows. Then, in Section 4, we study the optimal way for the moderator (whether benevolent or profit-driven) to choose I. A natural interpretation of moderation is the removal of content in violation of a platform's policies. This covers situations where content may be harmfully misleading but also the removal of content with undisclosed paid product placements.[5]

To focus on the main trade-off while minimizing unnecessary notation, we assume that any level of moderation can be implemented costlessly. Indeed, one of the main contributions of our paper is to show that the optimal level of moderation is often interior even without any moderation cost. It is easy to add a cost for inspecting messages, which simply provides an extra marginal disincentive to moderate. We provide discussions where a positive moderation cost would play a qualitatively significant role.

After the sender has moved and any moderation has taken place, a *receiver* observes $s \in (W \cup \emptyset) \times \mathbb{R}_+$ and forms posterior belief $\beta(s) = \Pr(w = 1|s)$. The receiver takes an action $r \in W$. In general, both the sender and receiver may use mixed strategies. Denote a generic receiver strategy by $\rho(s) = \Pr(r = 1|s)$. Let S_w denote the set of transmitted signals that are

[3] In principle, a sender could choose to say nothing ($s = \emptyset$) even if they do join the platform. However, $s = \emptyset$ can never be better in equilibrium than the best s with $m \in W$, so it is without loss to require $m \in W$ if a sender enters the platform.

[4] We can obtain the same insights if the moderator 'fact checks' the messages and labels untruthful ones—see our discussion in Section 8.

[5] E.g., YouTube says "If your content violates this policy [on paid product placement], we'll remove the content and send you an email to let you know." (Source: `https://support.google.com/youtube/answer/154235`, 21 July 2025).

on path conditional on sender entry in state w. The receiver's payoff is

$$u_R(w, r, e) = \begin{cases} \pi_h(e) & \text{if } r = w \\ \pi_l(e) & \text{if } r \neq w, \end{cases}$$

such that $\pi_h(e) > \pi_l(e)\ \forall e$ and $\pi_h'(e) \geq 0$.[6] Thus, the receiver would like his action to match the state. We also allow the receiver's payoff to depend on the sender's effort, for example because the receiver derives entertainment value from high-effort messages or because extra effort makes it quicker for the receiver to interpret the message.

The sender has state-independent preferences: she seeks to persuade the receiver to play $r = 1$ and obtains payoff (gross of entry cost, c)

$$u_S(r, e) = \begin{cases} v - e & \text{if } r = 1 \\ -e & \text{if } r = 0. \end{cases}$$

For the purpose of welfare, we assume that $\pi_h(e) - \pi_l(e) > v - e$ for any e. This means it is inefficient to 'trick' receivers into making a mistake.

The timing is as follows:

1. The moderator publicly announces I.
2. The sender chooses whether to enter or not.
3. If the sender entered then she observes w and chooses s.
4. Moderation takes place.
5. The receiver observes s and chooses r.

Entry occurs before the state is realized. Entry is costly and we should interpret this decision as a long-run commitment, such as learning to use a platform and setting up a user account. Once this cost is sunk, the sender can send messages about the random state. For example, an online 'influencer' must spend time building an audience before they are contacted by firms to promote their products, which may be good ($w = 1$) or bad, fraudulent or fake ($w = 0$). Alternatively, we could interpret w as audience-specific (e.g., what kinds of products would be most fitting to recommend) so that w is observed only when the sender is already on the platform and paired with an audience. We can obtain essentially equivalent insights to our baseline analysis in a model where senders know w before joining, as we discuss in Section 8.

[6] If $\pi_h'(e) = 0$, this is akin to money burning in cheap talk games (Austen-Smith and Jeffrey S Banks, 2000; Kartik, 2007).

We search for a Perfect Bayesian Equilibrium under the following parameter assumptions.

Assumption 1. $(1-p)\pi_l(e) + p\pi_h(e) < p\pi_l(e) + (1-p)\pi_h(e) \iff p < 1/2$. *In words, the receiver prefers to play $r = 0$ at his prior beliefs.*

Assumption 2. $\overline{c} > v > \underline{c}/p$. *The assumption that $pv > \underline{c}$ says that at least some senders find it worthwhile to enter if they can convince the receiver of the truth without effort. This is a necessary condition to sustain truth-telling in equilibrium. Assuming $\overline{c} > v$ means not all senders enter, which ensures $\beta(\emptyset)$ is pinned-down by Bayes' rule in any equilibrium.*

The following definition will also be useful.

Definition 1. *(1) There is* truthful *communication if a positive mass of senders enter and all entering senders transmit $m = w$. (2) There is* instrumental *communication if a positive mass of senders enter and $\rho(s) < \rho(s')$ for two on-path signals $s, s' \neq \emptyset$.*

In words, communication is truthful if there are senders on the platform, all of whom honestly report the state. It is instrumental if senders' sometimes choose different signals in a way that influences the receiver's behavior . Every situation with truthful communication is also clearly one of instrumental communication.

3 Equilibrium

3.1 Two benchmarks

Before studying how moderation and sender effort work together to support a truthful equilibrium, we show that in isolation they do relatively little to enable communication.

Lemma 1. *Suppose we constrain either $e = 0$ or $I = 0$. Then no equilibrium supports instrumental communication unless both $e = 0$ and $I = 1$.*

The intuition for the two cases is slightly different. If $e = 0$ then this becomes a game of cheap talk and, given $\rho(m) < \rho(m')$,[7] no sender will transmit m. The one exception is if $I = 1$, in which case a sender is willing to truthfully transmit $m = w$ because $m \neq w$ is deleted with certainty. In practice, it is likely that moderating every message ($I = 1$) on a large online platform will be prohibitively costly,[8] implying that no useful communication is feasible without sender effort.[9]

[7] In a slight abuse of notation, when $e = 0$ we write ρ solely as a function of m.

[8] Even if $I = 1$ is feasible (e.g., algorithmic moderation), instrumental communication is possible only if the moderator makes no mistakes (we return more formally to the issue of moderator mistakes in Section 5.2).

[9] This does not mean moderation is useless. Indeed, the mere fact that a message might have been inspected and hasn't been deleted conveys some information. However, the receiver's information here flows solely from the efforts of the moderator; the messages themselves are not instrumental.

Suppose, instead, $I = 0$ and $\rho(s) < \rho(s')$. Then s and s' can both be on the equilibrium path only if $\rho(s')v - e(s') = \rho(s)v - e(s)$. One can show that at least one on-path signal induces $\rho(s) = 0 \iff \rho(s)v - e(s) \leq 0$. Thus, incentive compatibility would require the sender exerts enough effort to dissipate her entire surplus in any equilibrium with instrumental communication. This makes the expected payoff from joining the platform negative, so the sender never enters and no communication takes place in equilibrium.

3.2 Truthful communication when costly messages are moderated

The picture changes substantially if we allow both effort and moderation, $e \geq 0$ and $I > 0$. Begin with the following result, which establishes an equilibrium with truthful communication in the communication sub-game following sender entry.

Lemma 2. *Suppose $I > 0$. For some $e^* \in (0, v)$, there exists an equilibrium in the continuation game following sender entry as follows:*

- *The sender truthfully reports the state ($m = w$), along with effort*

$$e = \begin{cases} 0 & \text{if } w = 0 \\ e^* & \text{if } w = 1. \end{cases} \tag{1}$$

- *The receiver updates his beliefs, following Bayes' rule where possible:*

$$\beta(s) = \begin{cases} 1 & \text{if } s = (1, e^*) \\ p & \text{if } s = \emptyset \\ 0 & \text{otherwise.} \end{cases}$$

- *$r = 1$ if $\beta(s) \geq 1/2$ and $r = 0$ otherwise.*

It is immediate that the receiver correctly updates his beliefs and acts optimally given the sender's strategy. In order to sustain truthful communication, we need the sender to prefer setting $s = (0, 0)$ when $w = 0$.[10] The incentive compatibility constraint is $v(1 - I) - e^* \leq 0$, implying

$$e^* \geq v(1 - I). \tag{IC}$$

When $I = 0$, the signal conveyed by effort is weak because the sender's payoff is identical regardless of the state. Incentive compatibility thus requires the sender to fully dissipate their

[10] A second incentive compatibility constraint is required to ensure the sender does not transmit $(0, 0)$ when $w = 1$. But satisfaction of this constraint is always implied by (IR).

surplus ($e^* = v$), resulting in no sender entry as already noted in the discussion following Lemma 1. However, setting $I > 0$ drives a wedge between the rewards to exerting effort under different states because high-effort but misleading messages are sometimes wasted whereas high-effort honest messages are never wasted. Inspection therefore makes effort a more powerful signal of veracity, reducing the minimum incentive compatible effort needed to persuade the receiver that $w = 1$ from v to $v(1 - I)$.

In the continuation equilibrium described in Lemma 2, the sender's expected payoff from entering is $p(v - e^*) - c$, which is non-negative if the individual rationality constraint,

$$e^* \leq v - \frac{c}{p}, \tag{IR}$$

is satisfied. Together, (IC) and (IR) constrain the set of e^* and I that can be sustained in an equilibrium with sender entry in the manner characterized by the following result.

Proposition 1. *(i) There exists an equilibrium with truthful communication if and only if there is a positive level of moderation, satisfying*

$$I > \frac{c}{pv}. \tag{2}$$

(ii) In all such equilibria, effort takes the threshold form (1)*. (iii) The unique effort level supporting an equilibrium with truthful communication and satisfying the intuitive criterion is* $e^* = v(1 - I)$*.*

Considering Lemma 1, sender effort and moderation are complements in the sense that (unless the moderator inspects every message) neither alone suffices to sustain instrumental communication. Some sender effort is needed for moderation to bite and, conversely, enough moderation is needed to make sender entry attractive. However, in a truthful equilibrium, an implication of Proposition 1(iii) is that effort and moderation are also substitutes in the sense that more intensive moderation crowds-out sender effort. Intuitively, if more messages are moderated then *every* sender's message becomes more credible, which allows senders to reduce the effort spent persuading the receiver of their truthfulness by slackening (IC). As we will now show, this implies it may be optimal for the moderator to set a positive but very low level of moderation.

4 Optimal moderation and platform strategy

We now turn to the question of how the moderator should optimally choose I, conditional on inducing a truthful equilibrium.[11] Focusing on the truthful equilibrium allows us to reflect on existing and proposed regulations which require platforms to ensure they do not host misleading or fraudulent information.[12]This section proceeds in three steps. First, we consider a benevolent moderator choosing I to maximize either total welfare or receiver surplus. Then we consider a profit-maximizing platform. Lastly, we consider a firm's business model decision between operating as a content-hosting platform or acting as a traditional broadcaster that directly informs receivers through its own content.

4.1 Benevolent moderation and welfare

Suppose the moderator is benevolent and seeks to maximize total (sender plus receiver) welfare. In light of Proposition 1, focus on an equilibrium with $e^* = v(1-I)$. Total welfare is

$$W(e^*) = G(p(v-e^*))\Big[p[v-e^*+\pi_h(e^*)] + (1-p)\pi_h(0)\Big] - \int_{\underline{c}}^{p(v-e^*)} c\, dG(c) + \left[1-G(p(v-e^*))\right]\left[p\pi_l(0)+(1-p)\pi_h(0)\right]. \quad (3)$$

The first line is the welfare when the sender enters the platform ($c < p(v-e)$). The second line is the receiver's expected payoff from following his prior and playing $r=0$ when the sender does not enter.

If π_h or G (or both) are sufficiently concave then W is concave and, because moderation affects W only via e^*, we have

$$\begin{aligned}\frac{\partial W(e^*)}{\partial I} &= \frac{\partial e^*}{\partial I} W'(e^*) \\ &= -\frac{\partial e^*}{\partial I} p\Big\{ \underbrace{pG'(p(v-e^*))\left[\pi_h(e^*)-\pi_l(0)\right]}_{\text{participation effect}} - \underbrace{G(p(v-e^*))\left[\pi_h'(e^*)-1\right]}_{\text{crowding-out effect}} \Big\}, \end{aligned} \quad (4)$$

where $\frac{\partial e^*}{\partial I} = -v$. The first bracketed term in (4) is a *participation effect*: higher I means the sender expects to exert less effort and she is more inclined to enter the platform. Increasing sender entry means the receiver is more likely to receive an informative message, boosting welfare. The second term is a *crowding-out effect*. Once (IR) is satisfied, further increases in I

[11] A common feature of communication games is equilibrium multiplicity. In Section 8 we show that the truthful equilibrium can be made essentially unique with a slight enrichment of the moderator's toolbox.

[12] For example, in the UK, platforms are responsible for preventing unlawful practices such as hidden advertising on their services, and more generally protecting consumers from so-called "false communications".

crowd-out effort because moderation and sender effort are substitutes in inducing instrumental communication. This crowding-out effect is negative if effort is efficient at increasing receiver surplus ($\pi_h'(e^*) > 1$)—for example, because the receiver derives entertainment as well as informational value from a YouTube video. It is clear that if sender participation is relatively inelastic then the crowding-out effect can dominate, causing $\frac{\partial W(e^*)}{\partial I} < 0$. Thus, the optimal level of moderation may be interior, even if moderation is costless, because this forces senders to work hard at establishing their credibility.

We can also consider a moderator that seeks to maximize the receiver's surplus, which is found by stripping sender payoffs out of $W(e^*)$:

$$E(u_R) = W(e^*) - \left[G(p(v-e^*))p(v-e^*) - \int_{\underline{c}}^{p(v-e^*)} c\, dG(c) \right].$$

Again, $E(u_R)$ is concave in I if π_h or G (or both) are sufficiently concave. We have,

$$\frac{\partial W(e^*)}{\partial I} - \frac{\partial E(u_R)}{\partial I} = pvG(p(v-e^*)) > 0$$

and some gains in welfare are driven by improving the sender's payoff.

Summarizing:

Proposition 2. *In an equilibrium with truthful communication that satisfies the intuitive criterion:*

(i) If

$$\frac{G'(pv)}{G(pv)}p < \frac{\pi_h'(0) - 1}{\pi_h(0) - \pi_l(0)}$$

then the socially-optimal level of moderation satisfies $I < 1$, even if moderation is costless.

(ii) If $I < 1$, a moderator that maximizes the receiver's payoff prefers a lower level of moderation than one who maximizes overall welfare.

In fact, even if moderation is costless, the optimal level of moderation may be substantially below the maximum feasible level. As an illustrative example (Figure 1), suppose that effort is efficient in the relevant range ($\pi_h'(e) > 1$). Let G have the constant elasticity form, $G(c) = c^\alpha$. As $\alpha \to 0$, we have $\frac{G'(c)c}{G(c)} \to 0$ and the socially optimal level of moderation satisfies $I \to 0$ by (4). Intuitively, when most senders have very low participation cost, a little moderation suffices to get almost all senders on board and transmitting truthful messages; the main effect of further increases in I is to crowd-out effort.[13]

[13]This trade-off can also affect content variety. Consider two genres—a popular genre 1 and a niche genre 2,

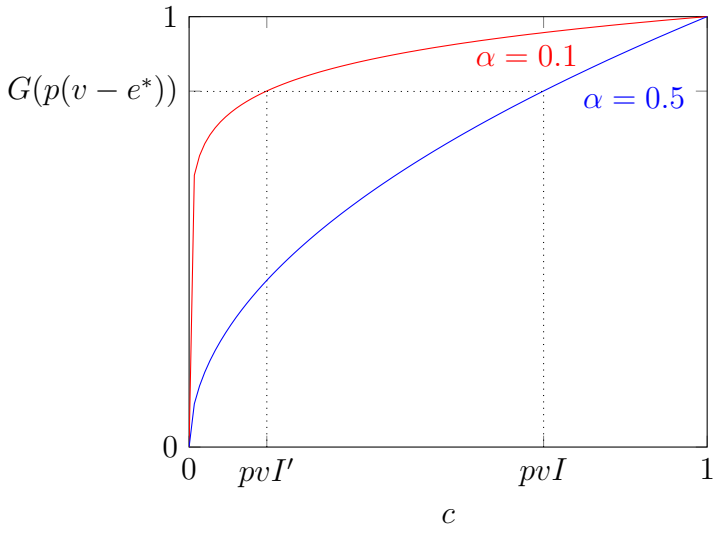

Figure 1: Using $G(c) = c^\alpha$ as an illustration, a decrease in α causes $G(c)$ to become more concave. When $G(c)$ is highly concave a lower level of moderation is sufficient to induce most senders to participate, meaning the crowding-out effect dominates even for small I.

4.2 Profit-maximizing platform

Suppose that I is instead chosen by a platform that, subject to inducing truthful communication, maximizes its profit from two sources of revenue. First, it charges participation fees f_S and f_R for senders and receivers respectively. Secondly, it earns one unit of ad revenue for each unit of receiver attention it supplies to advertisers.[14] In order to supply a unit of attention, two things must happen: First, there must be a message for the receiver to pay attention to (i.e., the sender must enter). Second, the receiver must dedicate attention to consuming the message; we assume he does so in increasing relation to the message's entertainment value, e. In other words, the platform faces the usual problem of getting all sides on board. Formally, we write $B(e)$ for the ad revenue generated by the attention of a receiver, with $B(0) = 0$ and $B'(e) \geq 0$.[15]

In a truthful equilibrium a fee of f_S induces the sender to enter and transmit a message with probability $G[p(v - e^*) - f_S]$. The most a receiver will pay to join the platform

with respective cost distributions $G_1(c)$ and $G_2(c)$ such that G_1 dominates in the left tail ($\frac{d}{dx}[G_1(x) - G_2(x)] < 0$). A lower I induces more effort from participating senders, but also distorts the content mix as senders of the popular genre are more likely to be active.

[14] Advertisers may be willing to pay more to advertise alongside high quality content, which is also consistent with the analysis below. We omitted ad market surplus from our earlier welfare analysis, but could include it without affecting the qualitative conclusions there.

[15] It is easy to microfound B. For instance, suppose that the receiver gets a flow utility of $\sqrt{eB}$ from spending B units of time ("attention") consuming content, and that time has a unit opportunity cost. Then the receiver maximizes $\sqrt{eB} - B$, implying $B = e/4$. The assumption $B(0) = 0$ is made to keep the analytical expressions concise and is not essential.

is $f_R = G(p(v - e^*) - f_S) \times p[\pi_h(e^*) - \pi_l(0)]$. The first term is the supply of messages, while $p[\pi_h(e^*) - \pi_l(0)]$ is the expected improvement in payoff from an informative message. Assembling these ingredients and using $e^* = v(1 - I)$, the platform's profit is therefore

$$\Pi = G\left(pvI - f_S\right)\left\{f_S + pB(v(1-I)) + p[\pi_h(v(1-I)) - \pi_l(0)]\right\}.$$

Provided G is not highly convex, Π is concave in f_S and uniquely maximized at some $f_S = f_S^*(I)$. The platform therefore solves

$$\frac{d\Pi}{dI} = pv\Bigg\{\underbrace{G'(pvI - f_S^*)\left\{f_S^* + pB(e^*) + p[\pi_h(e^*) - \pi_l(0)]\right\}}_{\text{participation effect}} - \underbrace{G(pvI - f_S^*)[B'(e^*) + \pi_h'(e^*)]}_{\text{crowding-out effect}}\Bigg\} + \underbrace{\frac{\partial \Pi}{\partial f_S^*}}_{=0}\frac{\partial f_S^*}{\partial I} = 0. \quad (5)$$

A trade-off similar to the one faced by a benevolent moderator emerges. Increasing the level of moderation induces more senders to participate. This is good news for the platform, allowing it to collect more in participation fees and ad revenue. However, the additional moderation crowds-out sender effort, which reduces both the receiver's willingness to pay and the attention that the platform supplies to advertisers. The optimal I is chosen to balance these two effects.

This can be seen as analogous to the standard multi-sided platform trade-off of shifting surplus from one side to another in order to balance participation. Indeed, when the elasticity of sender participation ($G'(c)c/G(c)$) is high, the participation effect is stronger and the platform caters more to the senders' interests by raising I. Conversely, when receivers' attention ($B(e)$) is highly elastic or effort highly efficient at increasing receiver surplus, the platform will lower I (but still sufficiently high to incentivize sender entry) to induce an equilibrium more favorable to receivers and advertisers.

Alternative revenue models We have allowed the platform to charge sender fees, receiver fees, and earn ad revenues. But the same insights emerge with a reduced set of revenue sources. A platform that did not earn ad revenues would correspond to setting $B(e) = B'(e) = 0$. A platform that did not charge sender fees would correspond to $f_S^* = \frac{\partial f_S^*}{\partial I} = 0$. A platform that did not charge receiver fees would correspond to setting $\pi_h(e^*) - \pi_l(0) = \pi_h'(e^*) = 0$ in (5). It is easily verified that the participation versus crowding-out trade-off remains in tact as long as at least on of ad revenues or receiver fees give the platform a reason to care about sender effort.

4.3 Platform versus broadcast business models

In order to inspect messages the moderator must be able to observe w. So, why doesn't the moderator just directly inform receivers itself rather than relying on senders to do so? In essence, this is a question of business model: a firm can either be a *content-hosting platform* that facilitates communication between senders and receivers, or it can be a traditional *broadcaster* that provides its own content directly to receivers. We show here that a firm may prefer the platform business model, even if senders have no informational advantage over itself. To make the point cleanly, we suppose the firm is purely ad funded, with ad revenues $B(e)$ generated by each message.

Start with the platform business model. The platform delegates the task of informing the receiver to senders, which obliges the sender to exert effort $e^* = v(1-I)$ in an equilibrium with truthful communication. The sender enters with probability $G(p(v-e^*))$. The platform's problem can be written as

$$\max_e \{G(p(v-e))pB(e)\} \qquad \text{such that } e = v(1-I).$$

The key point here is that the underlying communication game, via (IC) and (IR), induces an endogenous trade-off for the platform between encouraging sender participation and incentivizing their effort.[16]

If it acts as a broadcaster, the firm can directly report the truth to the receiver without needing to induce any sender entry or ensure incentive compatibility, so the trade-off observed in the platform model can be sidestepped. However, the firm now incurs the costs associated with content production itself. Its objective is therefore

$$\max_e \{B(e) - e\}.$$

It is immediate that either model can be optimal. If v is large or G is highly concave then senders would be willing to join a platform and exert a lot of effort. The firm can exploit this eagerness by opting for a lightly-moderated platform business model and allowing senders to bear the cost of high-effort content production. If, on the other hand, v is low or many senders have a high c then a platform faces a severe trade-off because it cannot induce high-effort content without substantially reducing sender entry. The firm will then tend to prefer the broadcast business model to escape this trade-off, even if it means bearing the cost of content production.

[16]When considering costly moderation our results hold qualitatively but tips in the favor of firms acting as a broadcaster.

5 Limits of Moderation

We now explore some of the limits of moderation. We show how sources of misinformation, moderator (or sender) mistakes, or sender free-riding incentives requires more moderation to sustain a truthful equilibrium and, where too much moderation is required, can cause communication to break down. We also show how senders with future concerns can relax the need for moderation to sustain communication.

5.1 Misinformation agents

Suppose there are two types of senders: ordinary senders and misinformation senders. The sender is *ordinary* with probability θ and has preferences as in our baseline model. With probability $1-\theta$ the sender is motivated by *misinformation* and aims to mislead receivers. In other words, misinformation senders aim to minimize $\rho(s)$ if $w=1$ and maximize $\rho(s)$ if $w=0$. One can imagine these senders are sponsored by rogue governments or serve the agenda of lobbyists. Hence, misinformation senders face no entry or effort costs.[17]

We can extend Lemma 2 as follows.

Lemma 3. *Suppose*

$$I > \max\left\{\frac{\underline{c}}{pv}, 1-\frac{p\theta G(pvI)}{(1-p)(1-\theta)}\right\}. \tag{6}$$

Then there exists an equilibrium in which: (1) As in Lemma 2, an ordinary sender chooses a truthful message $m=w$ along with effort

$$e=\begin{cases} 0 & \text{if } w=0 \\ e^* & \text{if } w=1. \end{cases}$$

(2) Misinformation senders report $m\neq w$ and

$$e=\begin{cases} e^* & \text{if } w=0 \\ 0 & \text{if } w=1. \end{cases}$$

(3) Communication is instrumental: ordinary senders enter with positive probability, $G(pvI)>0$, and the receiver plays $r=m$ whenever a sender enters.

For the ordinary sender to sustain this equilibrium, we need to satisfy her incentive compatibility constraint. In other words, she should not choose $s=(1,e^*)$ when $w=0$,

[17] Any costs faced by misinformation senders can be seen as compensated through some side payments outside the model.

requiring $e^* \geq v(1-I)$. The ordinary sender is willing to enter the communication game only if $e^* \leq v - \frac{c}{p}$. Combining the IC and IR, we recover that $I \geq \frac{c}{pv}$ is necessary for a sender with cost c to enter. Following Proposition 1 focus on the least-cost separating equilibrium for ordinary senders reporting the truth, $e^* = v(1-I)$.

The presence of misinformation may cause truthful messages from the ordinary sender to be non-instrumental. Intuitively, if both θ and I are small, the receiver may interpret $s = (1, e^*)$ as most likely being misinformation and refuse to act on such a signal. Thus, to sustain the ordinary sender's incentive to exert effort, we need to check that $\beta(1, e^*) \geq \frac{1}{2} \geq \beta(0,0)$. This is the case if and only if

$$I \geq 1 - \frac{p\theta G(p(v-e^*))}{(1-p)(1-\theta)}.$$

To see that such an I exists, starting from $I = 0$ the LHS is strictly increasing in I up to 1 while RHS is weakly decreasing ($\frac{\partial G(p(v-e^*))}{\partial I} = -G'(p(v-e^*))\frac{\partial e^*}{\partial I} \geq 0$) from 1. Hence, truthful and instrumental communication exists if and only if I is sufficiently large, (6). Notice that as misinformation becomes widespread, ($\theta \to 0$), a truthful communication equilibrium exists only if $I \to 1$. The constraint induced by misinformation senders may therefore be binding and force the moderator to inspect more messages than would be socially optimal without misinformation senders.

In summary, the main trade-off from Proposition 1 remains. Higher inspection levels reduce sender effort but stimulate ordinary sender participation. However, when there is a large proportion of misinformation senders a new consideration emerges. Sufficiently stringent moderation is needed to make the ordinary sender's signal instrumental. In such instances inspection is self-reinforcing in this dimension. Higher levels of inspection improves trust in the signal and additionally increase ordinary sender participation which reduces the relative proportion of misinformation senders.

5.2 Imperfect moderation

Suppose that moderation is prone to error, perhaps because it is ill-informed or adopts some imperfect technology. Upon inspecting a message the moderator makes the wrong choice (deleting a truthful message or allowing a false message to go undeleted) with some probability $0 < \mu < 1/2$. We can then construct an equilibrium with truthful communication of an analogous form to that given in Lemma 2.

The incentive compatibility condition preventing a sender transmitting $s = (1, e^*)$ when $w = 0$ is $v(1 - I(1-\mu)) - e^* \leq 0$. Indeed, there is now a new way for false messages to reach the receiver, namely that following inspection the moderator erroneously chooses not to delete it. Meanwhile, senders find entry worthwhile if $p[v(1-\mu I) - e^*] - c \geq 0$. Entry

has become less attractive because there is now the chance that their effort is wasted by a moderator who mistakenly deletes it. Combining these conditions, an equilibrium with truthful communication and entry can be supported if

$$(1-2\mu)I > \frac{\underline{c}}{pv}.$$

Comparison with (2) shows that a higher level of moderation is needed to sustain communication when moderation is imperfect. Indeed, frequent moderation errors may completely prevent truthful communication: it is possible to find an $I \leq 1$ that sustains truthful communication with entry only if

$$\mu < \frac{1}{2} - \frac{\underline{c}}{2pv}. \tag{7}$$

As in Proposition 1, the unique truthful equilibrium effort level satisfying the intuitive criterion is the least-cost separating one, $e^* = v(1-I(1-\mu))$. Thus, the error-prone moderator can provide identical incentives and achieve equivalent (welfare or profit) outcomes as one who doesn't make any mistakes, simply by scaling up the proportion of messages it inspects by a factor $1/(1-\mu)$, provided this is feasible.

The fact that maximized welfare (or profit) is constant in μ for low rates of error depends on the assumption that moderation is costless for two reasons. First, as we have seen, holding welfare (or profit) constant as μ increases requires the moderator to inspect more messages. If there is an increasing cost to inspecting messages then this becomes progressively more costly. Second, if the cost is convex, the moderator will optimally respond to the increasing marginal cost by increasing I less quickly than would be necessary to hold senders' incentives constant (formally, $I^*(\mu) < I^*(0)/(1-\mu)$). This causes senders to exit the platform as the moderator becomes more error-prone.

The following Corollary summarizes:

Corollary 1. *In a truthful equilibrium satisfying the intuitive criterion, suppose welfare (or platform profit) is quasi-concave in I and maximized by $I^*(\mu)$. Then*

- $I^*(\mu) = \min\{I^*(0)/(1-\mu), 1\}$.
- *If moderation is costless the highest welfare (or profit) that can be achieved with truthful communication is constant in μ if $I^*(0) < 1-\mu$, but decreasing in μ if $I^*(0) \geq 1-\mu$.*
- *If the moderator faces a weakly convex cost of inspecting messages then the highest welfare (or profit) that can be achieved with truthful communication is decreasing in μ.*

Moreover, if the rate of moderator error is too high (if (7) fails) then it becomes impossible to sustain an equilibrium with truthful communication.

Moderation by AI One motivation for the analysis in this subsection is that a platform might employ AI to automatic the moderation process. A consequence of deploying algorithmic moderation might be that the moderator makes more or fewer mistakes, which would correspond to an increase or decrease in μ. AI moderation might also reduce the cost of moderating a message. However, our baseline analysis already assumes moderation is costless, so the main incentives and tradeoffs we study would continue to operate even if AI reduces the costs of moderation to zero.

Noisy sender signals In Appendix B we study a similar setting where, instead of imperfect moderators, senders have imperfect information. In this alternative, we consider the situation where senders observe the state of the world with some noise and truthfully report their beliefs about the state. We find results that are qualitatively identical to Corollary 1.

5.3 AI-generated content

The previous subsection included a note about AI moderation. But AI can also be used by senders. Suppose the sender can adopt some generative AI technology to facilitate her effort. Using AI is cheaper as it only costs the sender $K(e) < e$ to produce content of quality e, $K'(e) > 0$.[18] The sender may opt to use AI or not. If she does, the receiver may choose to ignore the message m with some probability κ and updates his beliefs to $\beta(s) = 0$.[19] For example, individuals may dismiss AI generated content as uninformative but are only able to sometimes detect such content. This can occur if it is easier (or harder) to detect AI generated content on different mediums or on different topics or simply some receivers being naive about the existence of AI content.

Hence, gross of entry costs, the payoffs from using AI generated content is

$$u_A(r, e) = \begin{cases} v - K(e) & \text{if } r = 1 \\ -K(e) & \text{if } r = 0. \end{cases}$$

It is immediate to see the sender never uses AI if $\kappa \geq \frac{e-K(e)}{v(1-I)}$ and always uses AI if $\kappa \leq \frac{e-K(e)}{v}$. If the sender never uses AI, results from the base model apply directly. If, instead, the sender always uses AI then higher levels of moderation is required to sustain incentive compatibility. This has a negative effect on the least-cost separating equilibrium level of

[18] We could assume that AI reduces the cost of effort for false messages more than for truthful ones without changing the main insights below.

[19] For example, YouTube policy requires senders to disclose to receivers when content is AI-generated. Ignoring messages flagged as AI-generated is Bayesian consistent whenever the sender only uses AI when $w = 0$, and is a well-specified off-path belief when the sender chooses to either always or never use AI.

effort, $K(e^*) = v(1-I)(1-\kappa)$. Hence, qualitatively similar results to our base model apply to how moderation affects sender participation and effort levels. However, because using generative AI to improve quality is cheaper, the effect on equilibrium content quality is ambiguous.

For the remaining analysis, we consider the more interesting case where sender is only motivated to use AI if $w = 0$. Formally, this occurs when $\kappa \in (\frac{e-K(e)}{v}, \frac{e-K(e)}{v(1-I)})$. Then incentive compatibility for truthful communication requires $K(e^*) \geq v(1-I)(1-\kappa)$ and individual rationality for sender entry requires $e^* \leq v - \frac{c}{p}$. Combining these terms, we have that $I > 1 - \frac{K(v-\frac{c}{p})}{v(1-\kappa)}$ is necessary to ensure at least some senders enter and send truthful signals. Observe the right hand side of this equation is decreasing in κ. Hence, in cases where receivers choose to ignore AI generated messages, the minimum level of moderation which sustains truthful communication *can* be lower than without AI.[20]

From (4), AI does not change the socially optimal e^*: a benevolent moderator adjusts I so as to hold e^* constant as the prevalence of AI changes. It is clear that, for a given I, the least-cost separating e^* (solving $K(e^*) = v(1-I)(1-\kappa)$) is higher with AI than without it. Intuitively, AI makes it easier to fabricate effort, which forces honest senders (who don't use AI in equilibrium) to exert more "real" effort to maintain the credibility of their signal. This is bad for sender participation. The benevolent moderator responds by increasing I to filter out additional (now AI-generated) false messages and reduce the signaling burden on honest senders. For instance, in July 2025 YouTube announced that it was tightening its moderation rules motivated by increased volumes of 'inauthentic' AI-generated content.[21]

6 Dynamic considerations

6.1 Sender arrival

Consider an extension of the main model to an infinite discrete time horizon. The state of the world, w, is drawn at $t = 0$. In each period a sender arrives and chooses whether to enter or not. If a sender enters then she observes w and transmits a signal and the receiver updates his beliefs. The receiver's action, r, is a one-time irreversible decision that he can take at the end of any period. At the moment the receiver acts, he and all senders obtain payoffs as in the main model, discounted at the rate $\delta \in (0, 1)$. We assume that

[20] It is not always true that the minimum inspection required is lower than the base model. Additionally, equilibrium effort levels and welfare depend on the function K.

[21] See `https://support.google.com/youtube/answer/10008196` and `https://techcrunch.com/2025/07/09/youtube-prepares-crackdown-on-mass-produced-and-repetitive-videos-as-concern-over-ai-slop-grows/`, accessed 25 July 2025.

$p\pi_l(0) + (1-p)\pi_h(0) < 0 < \pi_h(0)$, so the receiver would rather wait for more information than act on his prior, but will act if sufficiently well-informed.

Proposition 3. *In an equilibrium with truthful communication, the unique level of effort surviving the intuitive criterion is $e^* = v(1-I)$. Such an equilibrium exists if and only if*

$$I - \frac{G(pvI)\delta}{1-\delta[1-pG(pvI)]} > \frac{\underline{c}}{pv}. \tag{8}$$

Remark 1. *If (8) holds when $I = 1$ and $\frac{(1-\delta)\delta pvG'(pvI)}{(1-\delta(1-pG(pvI)))^2} \leq 1$:*

1. *there exists a unique cutoff, $\bar{I}$ such that any $I > \bar{I}$ sustains the truthful communication equilibrium with entry.*

2. *as δ increases, senders are more patient and a higher $\bar{I}$ is required to induce entry.*

In the one-shot game the condition to sustain truthful communication was $I \geq \frac{\underline{c}}{pv}$, which is less stringent than (8). Intuitively, senders are reluctant to exert effort to persuade the receiver if they can free-ride by waiting for another sender to do it. Thus, the moderator needs to increase I to make signals more credible at lower effort levels.

If we interpret the rate of sender arrival, $\delta G(pvI)$, as the popularity of a topic, (8) implies that popular topics should be more highly moderated—not only in absolute terms but proportionally to the topic's popularity. Lastly, note that if δ is large enough then it is impossible to sustain truthful communication with any level of moderation because the temptation to free-ride is too strong.

6.2 Long-lived sender

Consider an extension of the main model to an infinite discrete time horizon. At the beginning of each period the state of the world w_t is drawn, $w_t = 1$ with probability p and $w_t = 0$ otherwise. A long-lived sender draws her cost c at the beginning of the game and can choose to join the platform. In each period she may choose to send a signal s_t and moderation may occur. If the sender fails inspection, her signals are deleted and she exits the platform.[22] The sender may reenter the platform (e.g., with a new user account) at her cost c. At the time she makes her decisions, the sender discounts her future payoff at a rate $\delta \in (0,1)$. A new receiver arrives each period, observes the resulting signal (if any), and takes the action r_t. Payoffs in each period are as in the main model.

[22] One can think of this as the situation where a series of moderation policy violations leads to an account ban.

Proposition 4. *In an equilibrium with truthful communication, there is a unique effort level surviving the intuitive criterion, $e^* = v(1-I) - \delta I\underline{c}$, that credibly signals $w = 1$. Such an equilibrium exists if and only if*

$$I > \frac{(1-\delta)\underline{c}}{p(v+\delta\underline{c})}.$$

When senders have more concern for future payoffs lower levels of inspection are required to sustain truthful communication. Intuitively, this occurs because lying incurs a future cost of reentry under a new identity. Hence, when more concern is placed on future value, the sender's payoff from lying decreases. As a result, the moderator is able to induce the same effort levels with lower I.[23]

7 Alternative mechanisms

Here we consider two settings that illustrate alternative but related mechanisms by which moderation can help to sustain communication.

7.1 Senders competing for attention

We now study a model where senders compete for attention and show that a similar trade-off to the one in the baseline model can arise through an alternative mechanism.

Suppose there are two senders with respective types $t \in \{0, 1\}$. It is useful to label a sender of type $t = w$ as an (A)ligned sender and when $t \neq w$ an (U)naligned sender. Senders independently draw the cost c from the distribution G and learn w after joining the platform. Their payoffs gross of entry costs are

$$u_S(r, e) = \begin{cases} v - e & \text{if } r = t \\ -e & \text{if } r \neq t. \end{cases}$$

Senders compete for the receiver's attention through *unobservable* effort.[24] Formally, upon entering the communications game, senders' efforts are interpreted as bids in an all-pay auction, with the receiver paying attention only to the highest-effort message. For example, to retain attention, platforms like YouTube may recommend the most entertaining video.

To simplify the analysis, let $p = 1/2$ such that the game is symmetric between the two senders and assume that if $s = \emptyset$ receivers take no action and the game ends. We look for

[23] Notice if $\delta = 0$ the model reverts to the baseline model. Hence, if $I^*(\delta)$ is the socially optimal policy then we have $I^*(\delta) = \frac{I^*(0)v}{v+\delta\underline{c}} \leq I^*(0)$, where $I^*(0)$ is the optimal policy in the baseline model.

[24] Making effort unobservable allows us to shut down it's earlier signaling effect and cleanly highlight the alternative mechanism.

equilibria in which the sender reports $m = t$ and the receiver trusts the message he sees, playing $r = m$.

Lemma 4. *Conditional on entering, in an equilibrium with $m = t$ and $r = m$, the aligned and unaligned senders respectively choose e according to CDFs*

$$F_A(e) = 1 - \frac{\overline{e} - e}{(1-I)vG(u_S^*)} \quad \textit{and} \quad F_U(e) = 1 - \frac{\overline{e} - e}{vG(u_S^*)} \tag{9}$$

on the support $[0, \overline{e}]$, where $\overline{e} = (1-I)vG(u_S^*)$, and u_S^* is a sender's ex ante expected payoff from joining, uniquely defined by $u_S^* = v[\left(1 - \frac{I}{2}\right) - (1-I)G(u_S^*)]$.

Conditional on both aligned and unaligned senders entering and the false message not being deleted, the probability of the unaligned sender winning the receiver's attention is

$$\int_0^{\overline{e}} F_A(e) F_U'(e)\, de = \frac{1-I}{2}. \tag{10}$$

Notice that $(1-I)/2 \leq 1/2$ so the receiver indeed finds it optimal to follow the advice of whichever message he sees. More importantly, the probability that the unaligned sender wins the auction is decreasing in I. Thus, *even when a message is not inspected*, moderation makes the market for attention a more effective device for promoting truthful content. Intuitively, moderation handicaps dishonest senders by increasing the chance that their effort is wasted. As senders endogenously adjust their effort, this translates into a higher equilibrium probability that aligned senders win the competition for attention.

However, similar to the baseline model, this comes at the cost of crowding-out effort. Intuitively moderation makes effort more costly for the unaligned sender because her message is more likely deleted. As the unaligned sender bids less, the aligned sender faces less competition and can also reduce her bid. The following proposition gathers these results.

Proposition 5. *When two opposing senders compete for the receiver's attention as in Lemma 4, higher levels of moderation (i) make it more likely that the aligned sender wins, but (ii) reduce the expected effort of both senders.*

We thus observe a trade-off similar in spirit to the one identified in the baseline model. Higher levels of moderation increase the likelihood that the receiver sees a truthful message, but reduces senders' incentives to invest in the quality of that message. For the optimal level of moderation, which of these two effects dominates depends on the relative importance of the receiver being correct (the size of π_h compared to π_l) and of sender effort (the sizes of $\pi_h'(e)$ and $\pi_l'(e)$).

In the baseline model senders use effort as a signal but would be indifferent between a productive signal (with $\pi_h'(e) > 0$) and a socially wasteful one. This model of competition for attention provides an explanation for why senders might choose productive signals: wasteful signals are unlikely to be as effective in winning receivers' attention.

Competing with uninformative content In Appendix B we study a model where senders of informative content (similar to those in the main model) compete with senders of content with purely entertainment value. In this setting, we show an identical participation-effort trade-off arises. However, because even truthful senders can have their effort wasted, the moderator needs to take on more of the burden of communication by imposing a higher level of inspection.

7.2 Transfers to senders

We now consider environments where the platform makes a transfer, $t(e)$, to any sender whose message is not deleted, with $t(0) \geq 0$, $t'(e) > 0$, and $t''(e) < 0$. For example, one natural formulation would be $t(e) = \psi B(e)$, where B is the platform's ad revenue and ψ is the revenue share paid to content creators. The game is otherwise as in the baseline.

Define $e' = \arg\max_{e \geq 0}\{t(e) - e\}$ as the effort level that maximizes the sender's profit from transfers. Let

$$\hat{e} = \min\{e \geq e' : e \geq (v + t(e))(1 - I) - [t(e') - e']\}. \tag{11}$$

In words, $\hat{e}$ will be the effort level that is closest the optimal one (e') while still being high enough to credibly signal that $w = 1$. We can show the following:

Proposition 6. *Suppose*

$$I > \frac{\underline{c} - [t(e') - e']}{p(v + t(\hat{e}))}. \tag{12}$$

Then there exists an equilibrium with truthful communication in which senders exert effort

$$e = \begin{cases} e' & \text{if } w = 0 \\ \hat{e} & \text{if } w = 1. \end{cases}$$

In particular, $\hat{e} = e'$ if $I \geq \frac{v}{v + t(e')}$ and $\hat{e} > e'$ otherwise.

When $I < \frac{v}{v+t(e')}$, the equilibrium has a familiar structure with $w = 1$ senders exerting higher effort, $\hat{e} > e'$, to establish their credibility. Moreover, applying the implicit function theorem to (11) we find that additional moderation crowds-out effort.

However, an interesting feature of Proposition 6 is that, once $I \geq \frac{v}{v+t(e')}$, truthful communication can be sustained even though all senders exert the same effort, e'. Since effort does not differentiate senders, signaling does not play a role in the receiver's decision—emphasizing that a new (non-signaling) mechanism is at play. For some intuition, suppose $w = 0$. A sender could transmit the untruthful signal $s = (1, e')$ to induce $r = 1$ and gain $v(1 - I)$. However, since this message is potentially deleted there is an expected loss of transfer payments $It(e')$. When $I \geq \frac{v}{v+t(e')}$ this loss is sufficiently large to deter deception. Hence achieving incentive compatibility without signaling. This works through the interaction of transfers and moderation, which jointly make lying costly.

The condition $I \geq \frac{v}{v+t(e')}$ again embeds an inverse relationship between moderation and effort. Indeed, if moderation is costly then a moderator may choose a generous revenue sharing policy (e.g., a high ψ), which not only induces high sender effort, but also reduces $\frac{v}{v+t(e')}$ and makes it easier to satisfy incentive compatibility at low levels of moderation without relying on signaling. Even a small risk of being moderated is enough to deter lying when it would mean losing a very generous transfer.

8 Robustness

Senders learning state before entry Our results are robust to variations in the model timing. We chose to focus on situations where senders learn the state after entering the platform because there are many contexts where this is the most natural timing (e.g., when 'influences' build an audience before receiver a product for review). But we can equally imagine situations where the sender knows the state before joining. We can accommodate this alternative timing while recording the same basic results and trade-offs identified in our baseline analysis.

We find a truthful and instrumental equilibrium qualitatively identical to our original communications game. Without moderation, truthful communication fully dissipates senders' surplus, meaning no senders would enter. Thus, moderation is essential to sustaining truthful communication in equilibrium. But moderation also reduces sender effort, allowing us to recover the main moderation trade-offs of sender participation and crowding-out effort.

Details and further discussion can be found in Appendix B.

Fact-checking We focus on a particular moderation policy where content is removed following an observed violation. In some instances, rather than remove content, moderators instead provide corrections or append additional information to the original content. Examples include YouTube and Facebook's misinformation warning labels which gained prominence

during the COVID-19 pandemic.[25] We call this type of moderation 'fact-checking'.

In an alternative specification where moderators conduct fact-checking instead of deletion, we show that a new incentive compatibility constraint is introduced. In any truthful equilibrium, it must be that senders do not prefer to deceive receivers when the state is $w = 1$ and play $s = (0, 0)$, exerting no effort and relying on fact-checking to provide information. This new constraint means we are able to pin down the same equilibrium effort level without the intuitive criterion refinement. However, the switch from moderation to fact-checking otherwise leaves the analysis unchanged and all our results follow through.

Details and further discussion can be found in Appendix B.

Equilibrium uniqueness and effort-dependent moderation Throughout, we show how truthful communication can be sustained by an interior level of moderation. However, as is common in communication games, this equilibrium is not necessarily unique. In fact, if I is sufficiently large, there is an equilibrium where no senders exert effort because the mere fact their message wasn't deleted by the moderator is already sufficient to induce $r = 1$.

We show that if the moderator can condition it's inspection rule on sender's effort it can ensure that all equilibria are payoff-equivalent to the truthful equilibrium studied above by only inspecting high-effort signals. Intuitively, one might think that a moderator should focus their attention on low-effort messages because they are more easily fabricated. But such a policy can backfire because it has the negative side-effect of increasing the credibility of any low-effort lies that go unnoticed by the moderator. Details and further discussion can be found in Appendix B.

9 Conclusion

Social media allows ordinary users to communicate, perhaps untruthfully, with the world. We have shown that moderation plays a central role in governing information dissemination in these environments. When senders' and receivers' incentives are not aligned, some moderation is essential to support equilibrium meaningful communication. The fact that some messages have been moderated increases the credibility of *all* senders' messages, making it less costly for senders to persuade receivers' of their truthfulness. This induces more senders to participate and transmit truthful messages, but at the cost of crowding-out valuable sender effort. The optimal moderation policy resolves this trade-off and often involves an interior level of moderation, even if moderation is costless. Indeed, if senders are easily persuaded to

[25]Also note Singapore's Protection from Online Falsehoods and Manipulation Act requires that content identified as misinformation by the government are required to carry a similar label.

participate then the optimal level of moderation may be very low.

We discuss how a business serving as an (informative) content-hosting platform chooses it's optimal moderation to maximize profits (depending on ad-revenue and membership fees). A platform chooses it's moderation policy to balance sender participation with providing high-quality content to receivers, depending on which group's participation is most elastic. We additionally show ad revenue sharing models can work in tandem with (but not without) moderation to sustain informative communication. We also discuss how a business decides between being a content-hosting platform or a traditional broadcaster. Both may arise in equilibrium depending on the cost of inducing sender participation and directly providing in formation to consumers.

We explore some limits to moderation. First, sources of misinformation make receivers less inclined to trust messages, calling for higher levels of moderation than would otherwise be optimal. Second, if the moderator and/or sender are prone to mistakes interpreting the true state of the world then the efficacy of moderation is reduced and, in equilibrium, more moderation is required to restore trust. Third, senders may use generative AI to lower the cost of effort required to convince receivers. If receivers do not correctly identify and/or ignore AI generated content, more moderation may be required to sustain communication.

We consider how senders behave in a dynamic environment. If new senders arrive in each period, senders may be tempted to free-ride on others exerting effort to influence the receiver. This calls for the moderator to assume more of the burden of persuading the receiver with a more stringent moderation policy. By contrast, if a long-lived sender faces punishment for deception (such as removal from the platform), an opposing effect arises. The threat of punishment complements moderation in disciplining sender behavior and less moderation is required to sustain truthful communication.

Finally, we cast our research question into different frameworks. First, we study how senders may compete for receiver attention through effort. This recovers the participation-effort trade-off found in our main setting. Studying competition also provides a natural explanation for why senders may choose to exert effort at all. Second, in an environment where platforms may share profits with senders, we show that inspection and profit-sharing can work in tandem to sustain communication.

References

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius (2024). "A model of online misinformation". *The Review of Economic Studies* 91.6, pp. 3117–3150.

Ahmad, Wajeeha, Ananya Sen, Chuck Eesley, and Erik Brynjolfsson (2024). *The role of advertisers and platforms in monetizing misinformation: Descriptive and experimental evidence*. Tech. rep. National Bureau of Economic Research.

Alsem, Karel Jan, Steven Brakman, Lex Hoogduin, and Gerard Kuper (2008). "The impact of newspapers on consumer confidence: does spin bias exist?" *Applied Economics* 40.5, pp. 531–539.

Andres, Raphaela and Olga Slivko (2021). *Combating online hate speech: The impact of legislation on Twitter*. Tech. rep. ZEW Discussion Papers.

Arieli, Itai, Ivan Geffner, and Moshe Tennenholtz (2023). "Mediated cheap talk design". *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 5, pp. 5456–5463.

Austen-Smith, David and Jeffrey S Banks (2000). "Cheap talk and burned money". *Journal of Economic Theory* 91.1, pp. 1–16.

Banks, Jeffery S (2013). *Signaling games in political science*. Routledge.

Bar-Isaac, Heski, Rahul Deb, and Matthew Mitchell (2025). "Selling Certification, Content Moderation, and Attention". *Working Paper*.

Beknazar-Yuzbashev, George, Rafael Jiménez-Durán, Jesse McCrosky, and Mateusz Stalinski (2025). "Toxic content and user engagement on social media: Evidence from a field experiment". *CESifo Working Paper*.

Ben-Porath, Elchanan (2003). "Cheap talk in games with incomplete information". *Journal of Economic Theory* 108.1, pp. 45–71.

Berger, Lara Marie, Anna Kerkhof, Felix Mindl, and Johannes Münster (2025). "Debunking "fake news" on social media: Immediate and short-term effects of fact-checking and media literacy interventions". *Journal of Public Economics* 245, p. 105345.

Bester, Helmut, Matthias Lang, and Jianpei Li (2021). "Signaling versus auditing". *The RAND Journal of Economics* 52.4, pp. 859–883.

Bilancini, Ennio and Leonardo Boncinelli (2018). "Signaling with costly acquisition of signals". *Journal of Economic Behavior & Organization* 145, pp. 141–150.

Cameron, Charles M and B Peter Rosendorff (1993). "A signaling theory of congressional oversight". *Games and Economic Behavior* 5.1, pp. 44–70.

Chopra, Felix, Ingar Haaland, and Christopher Roth (2022). "Do people demand fact-checked news? Evidence from US Democrats". *Journal of Public Economics* 205, p. 104549.

Crawford, Vincent P. and Joel Sobel (1982). "Strategic Information Transmission". *Econometrica* 50.6, pp. 1431–1451.

DellaVigna, Stefano and Ethan Kaplan (2007). "The Fox News effect: Media bias and voting". *The Quarterly Journal of Economics* 122.3, pp. 1187–1234.

Dwork, Cynthia, Chris Hays, Jon Kleinberg, and Manish Raghavan (2024). "Content Moderation and the Formation of Online Communities: A Theoretical Framework". *Proceedings of the ACM on Web Conference 2024*, pp. 1307–1317.

Ershov, Daniel and Juan S Morales (2024). "Sharing News Left and Right: Frictions and Misinformation on Twitter". *The Economic Journal* 134.662, pp. 2391–2417.

Figueroa, Nicolás and Carla Guadalupi (2021). "Testing the sender: When signaling is not enough". *Journal of Economic Theory* 197, p. 105348.

Ganguly, Chirantan and Indrajit Ray (2023). "Simple Mediation in a Cheap-Talk Game". *Games* 14.47, pp. 1–14.

Garfagnini, Umberto (2017). "The Downsides of Managerial Oversight in Signaling Environments". *Working Paper*.

Henry, Emeric, Ekaterina Zhuravskaya, and Sergei Guriev (2022). "Checking and sharing alt-facts". *American Economic Journal: Economic Policy* 14.3, pp. 55–86.

Horta Ribeiro, Manoel, Justin Cheng, and Robert West (2023). "Automated content moderation increases adherence to community guidelines". *Proceedings of the ACM web conference 2023*, pp. 2666–2676.

Ivanov, Maxim (2014). "Beneficial mediated communication in cheap talk". *Journal of Mathematical Economics* 55, pp. 129–135.

Jackson, Matthew O, Suraj Malladi, and David McAdams (2022). "Learning through the grapevine and the impact of the breadth and depth of social networks". *Proceedings of the National Academy of Sciences* 119.34, e2205549119.

Jiménez-Durán, Rafael (2023). "The economics of content moderation: Theory and experimental evidence from hate speech on Twitter". *George J. Stigler Center for the Study of the Economy & the State Working Paper* 324.

Kartal, Melis and Jean-Robert Tyran (2022). "Fake news, voter overconfidence, and the quality of democratic choice". *American Economic Review* 112.10, pp. 3367–3397.

Kartik, Navin (2007). "A note on cheap talk and burned money". *Journal of Economic Theory* 136.1, pp. 749–758.

Kominers, Scott Duke and Jesse M Shapiro (2024). *Content moderation with opaque policies*. Tech. rep. National Bureau of Economic Research.

Krishna, R Vijay (2007). "Communication in games of incomplete information: Two players". *Journal of Economic Theory* 132.1, pp. 584–592.

Krishna, Vijay and John Morgan (2004). "The art of conversation: eliciting information from experts through multi-stage communication". *Journal of Economic Theory* 117.2, pp. 147–179.

Li, Hao and Wei Li (2013). "Misinformation". *International Economic Review* 54.1, pp. 253–277.

Lin, Hause, Haritz Garro, Nils Wernerfelt, Jesse Shore, Adam Hughes, Daniel Deisenroth, Nathaniel Barr, Adam Berinsky, Dean Eckles, Gordon Pennycook, et al. (2024). "Reducing misinformation sharing at scale using digital accuracy prompt ads". *PsyArXiv*.

Madio, Leonardo and Martin Quinn (2024). "Content moderation and advertising in social media platforms". *Journal of Economics & Management Strategy*.

Mattozzi, Andrea, Samuel Nocito, and Francesco Sobbrio (2022). "Fact-checking politicians". *CESifo Working Paper*.

Mostagir, Mohamed and James Siderius (2022). "Naive and bayesian learning with misinformation policies". *Working Paper*.

Myerson, Roger B (1986). "Multistage games with communication". *Econometrica*, pp. 323–358.

Nelson, Phillip (1974). "Advertising as Information". *Journal of Political Economy* 82.4, pp. 729–754. (Visited on 07/21/2025).

Pennycook, Gordon, Adam Bear, Evan T Collins, and David G Rand (2020). "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings". *Management Science* 66.11, pp. 4944–4957.

Rahman, David (2012). "But who will monitor the monitor?" *American Economic Review* 102.6, pp. 2767–2797.

Salamanca, Andrés (2021). "The value of mediated communication". *Journal of Economic Theory* 192, p. 105191.

Spence, Michael (1973). "Job Market Signaling". *The Quarterly Journal of Economics* 87.3, pp. 355–374.

A Proofs

Proof of Lemma 1 (case with $e = 0$). Let $e = 0$, suppose $I \in [0, 1)$, and fix m and $m' \neq m$, with both messages being on the equilibrium path conditional on sender entry. Suppose communication is instrumental, $\rho(m) > \rho(m')$.

As a first step, we establish that $\rho(\emptyset) \leq \rho(m)$. If not, a sender who transmits m' would deviate to m because this would increase ρ regardless of whether their message is deleted or not. Thus, there could be no on-path m, m' such that $\rho(m) > \rho(m')$, contradicting the hypothesis of an instrumental equilibrium.

Now, if $w = m$ the sender strictly prefers to send message m because telling the truth both guarantees that the message passes inspection and yields a higher probability that $r = 1$. This means the probability of obtaining the payoff v is larger when playing m than either m' or $\emptyset$. Thus, m' is sent only when $w = m'$, meaning $\rho(m') = m'$. For m' to be on-path as required, it must therefore be optimal to send that message when $w = m'$:

$$\rho(m')v \geq (1 - I)\rho(m)v + I\rho(\emptyset)v. \tag{13}$$

Suppose $m' = 0$. This causes (13) to fail for any $I \in [0, 1)$ because $\rho(m) > \rho(m') = 0$. Thus, (13) requires that $m' = 1$.

So far, it is established that $\Pr(m = 0|w = 0) = 1$ and, for $m = 1$ to be on-path, $\Pr(m = 0|w = 1) < 1$. Now consider the receiver's beliefs upon observing $m = 0$:

$$\beta(m) = \beta(0) = \frac{p \Pr(m = 0|w = 1)}{p \Pr(m = 0|w = 1) + (1 - p) \Pr(m = 0|w = 0)} \leq p.$$

But $\beta(m) = \beta(0) < p$ implies $\rho(m) = \rho(0) = 0$, contradicting the hypothesis that $\rho(m) > \rho(m')$.

It remains to show that there exists an equilibrium with truthful communication when $I = 1$. Suppose the sender's strategy is to truthfully report the state, $m = w$. Then $\beta(m) = m$ and null signals are observed in equilibrium only if the sender does not enter, yielding $\beta(\emptyset) = p$. Given these beliefs, no sender can profitably deviate because $m \neq w$ is inspected and deleted with probability $I = 1$, leading to $\beta(m) = p < 1/2$ and $\rho(m) = 0$. Lastly, these strategies and beliefs imply that the sender's payoff from entering is pv, so $G(pv) > 0$ senders join the platform as required.

□

Proof of Lemma 1 (case with $I = 0$). Fix $I = 0$ and let $S_{\text{eqm}} = S_0 \cup S_1$ be the set of s on the

equilibrium path. Define $\bar{s}$ such that

$$\bar{s} \in \arg\min_{s \in S_{\text{eqm}}} \beta(s).$$

We must have $\beta(\bar{s}) \leq p$,[26] meaning $\rho(\bar{s}) = 0$ (by Assumption 1). The sender's payoff from $\bar{s}$ is therefore $-e(\bar{s}) \leq 0$.[27] Because the sender can guarantee herself a continuation payoff of at least zero by playing $e = 0$, we must have $e(\bar{s}) = 0$. Moreover, because the sender's payoffs are independent of w, the sender must be indifferent over all $s \in S_{\text{eqm}}$, meaning every equilibrium sender action yields a continuation payoff of zero. The overall payoff from entering is therefore $-c$ and no sender finds it worthwhile to enter, implying that the equilibrium involves no (instrumental) communication.

□

Proof of Proposition 1. To prove part (i): Combining (IC) and (IR), truthful communication requires $e^* \in [v(1-I), v - \frac{c}{p}]$ and a necessary and sufficient condition for $I \in (0, 1]$ to support a fully-informative equilibrium with entry is therefore

$$v(1-I) < v - \frac{c}{p} \iff I > \frac{c}{pv}.$$

Such an I exists by Assumption 2.

For part (ii): in an equilibrium with truthful communication the receiver plays $r = 0$ whenever observing $s \in S_0$, implying sender payoff $-e(s)$. If $s = \emptyset$ then the sender's payoff is $\rho(\emptyset)v - e$. In both cases, the sender would prefer to deviate from any $e > 0$ to $e = 0$. Thus, $e = 0$ whenever $w = 0$. Meanwhile, any $s \in S_1$ leads to $r = 1$ and therefore yields sender payoff $v - e(s)$. Thus, the sender can be indifferent between $s, s' \in S_1$ only if $e(s) = e(s') \equiv e^*$ for some e^*.

For part (iii): If $e^* < v(1-I)$ then the sender would deviate to $s \in S_1$ when $w = 0$ because (IC) is violated. So we must have $e^* \geq v(1-I)$. Now, consider an equilibrium with $e^* > v(1-I)$. Suppose the sender transmits the signal $\tilde{s} = (1, \tilde{e})$, with $\tilde{e} \in (v(1-I), e^*)$. By part (ii), $\tilde{s}$ is off-path. If $w = 0$ then, even under the most favorable beliefs $\beta(\tilde{s}) = 1$, the sender's payoff from the deviation would be

$$v(1-I) + \underbrace{\rho(\emptyset)}_{=0} vI - \tilde{e} < 0,$$

[26] This follows from the law of total probability: $p = \sum_{s \in S_{\text{eqm}}} \beta(s) \Pr(s)$. The right-hand side is a weighted average of βs. For that average to equal p, the smallest of the βs must be no larger than p.

[27] $e(s)$ is the effort component of action s.

while when $w = 1$ the payoff under the same receiver beliefs is $v - \tilde{e}$, which is positive in the candidate $e^* \leq v - \frac{c}{p} \leq v$.[28] Thus, the intuitive criterion calls upon the receiver to hold belief $\beta(\tilde{s}) = 1$ following this deviation and these beliefs make the deviation profitable whenever $w = 1$. □

Proof of Proposition 3. Following similar steps to parts (ii) and (iii) of the proof of Proposition 1, it is easily shown that any truthful communication equilibrium satisfying the intuitive criterion has

$$e = \begin{cases} 0 & \text{if } w = 0 \\ e^* & \text{if } w = 1, \end{cases}$$

with $e^* = v(1 - I)$.

For an equilibrium with truthful communication to be sustained, senders must not have a profitable deviation to $(1, e^*)$ following $w = 0$. The incentive compatibility condition is given by (IC). Indeed, if the sender enters the game and plays $(1, e^*)$ her payoff is $v(1 - I) - e^*$. However, the sender may choose to play $(0, 0)$ or do nothing $s = \emptyset$. Suppose the sender plays $(0, 0)$, the game ends and the sender obtains zero payoff. If instead she plays $s = \emptyset$ then a future sender entering the game is expected to play $(0, 0)$ because the equilibrium is informative. This continuation game still leads to zero payoff for the sender. Therefore, if (IC) is satisfied, when $w = 0$ it is payoff equivalent for the sender to play either $(0, 0)$ or $s = \emptyset$.

To ensure entry at least a sender with cost $\underline{c}$ should be incentivized to enter the communications game. Her payoff from entry is $p(v - e^*) - \underline{c}$. If she does not enter, the next sender will enter with probability $G(p(v - e^*))$. A positive measure of entry takes place if

$$p(v - e^*) - \underline{c} > pv \sum_{t=1}^{\infty} \delta^t [1 - G(p(v - e^*))]^{t-1} G(p(v - e^*)) = \frac{G(p(v - e^*))\delta}{1 - \delta[1 - G(p(v - e^*))]} pv. \quad (14)$$

Substituting $e^* = v(1 - I)$ into (14), we have

$$I - \frac{G(pvI)\delta}{1 - \delta\left[1 - G(pvI)\right]} > \frac{\underline{c}}{pv}$$

and an equilibrium with truthful communication and entry exists if I satisfies this condition. □

Proof of Proposition 4. In the truthful equilibrium the receiver learns the true state of the world in each period, and $r_t = m_t$ where m_t is the message sent in period t. For the equilibrium

[28] $\rho(\emptyset) = 0$ because, on the equilibrium path, no sender transmits a message that is rejected. Thus, $s = \emptyset$ is observed only if no sender entered, in which case $\beta(\emptyset) = p$.

to be sustained, a sender must not have a profitable one-shot deviation. When $w = 0$, this means the expected payoff from deviation to $(1, e^*)$ must be less than the expected payoff from truthfully playing $(0, 0)$. These payoffs are $\left(v + \frac{\delta p(v-e^*)}{1-\delta}\right)(1-I) - e^* + I\left(-\delta c + \frac{\delta p(v-e^*)}{1-\delta}\right)$ and $0 + \frac{\delta p(v-e^*)}{1-\delta}$ respectively. Hence, a sender's incentive compatibility constraint is $v(1 - I) - e^* - \delta cI \leq 0$. Since this must hold for senders of all entry cost, we require the IC,

$$e^* \geq v(1-I) - \delta \underline{c} I \tag{IC2}$$

To induce sender entry, the individual rationality constraint has to be satisfied for at least some senders. In other words, the expected payoff from entry $\frac{p(v-e^*)}{1-\delta}$ must exceed the cost of entry for at least some senders. Therefore the IR is

$$\frac{p(v-e^*)}{1-\delta} \geq \underline{c}. \tag{IR2}$$

The rest of the proof follows the same steps as the proof of Proposition 1 if we use (IC2) instead of (IC) and (IR2) instead of (IR). □

Proof of Lemma 4. A standard result in all-pay auctions is that bidders (senders) mix over an interval support, $[\underline{e}, \overline{e}]$. At least one sender never wins against their rival if $e = \underline{e}$; this implies $\underline{e} = 0$.

Indifference of the aligned sender requires

$$v - \overline{e} = v\left[G(u_S^*)F_U(e) + (1 - G(u_S^*))\right] - e, \tag{15}$$

where $F_U(e)$ is the bid distribution for the unaligned sender. Intuitively, the right-hand side says the aligned sender wins the competition for attention if the unaligned sender doesn't enter or if she does enter but submits a lower bid. The indifference condition for the unaligned sender is analogously

$$v(1-I) - \overline{e} = v(1-I)\left[G(u_S^*)F_A(e) + (1 - G(u_S^*))\right] - e, \tag{16}$$

where $F_A(e)$ is the bid distribution of the aligned sender.

From (15) and (16), we obtain (9). We have $F_U(e) \geq F_A(e)$, so $F_A(0) = 0$. Indeed, if $F_A(0) > 0$ then there would be a positive probability of a tie at $e = 0$ and senders would have an incentive to deviate. Solving $F_A(0) = 0$ for $\overline{e}$ yields $\overline{e} = (1 - I)vG(u_S^*)$.

Lastly, from (15) and (16), the utility of aligned and unaligned senders is $u_A = v - \overline{e}$ and $u_U = v(1 - I) - \overline{e}$. Substituting in the value of $\overline{e}$, we have $u_S^* = \frac{1}{2}(u_A + u_U) = v[\left(1 - \frac{I}{2}\right) - (1 - I)G(u_S^*)]$. □

Proof of Proposition 5. Part (i) is immediate from (10).

To see part (ii): first, by applying the implicit function theorem to $u_S^* = v[\left(1 - \frac{I}{2}\right) - (1 - I)G(u_S^*)]$, we obtain

$$\frac{du_S^*}{dI} = \frac{v(2G(u_S^*) - 1)}{2 + 2(1 - I)vG'(u_S^*)}. \tag{17}$$

The expected effort of an unaligned sender is

$$E(e_U) = \int_0^{\overline{e}} F_U'(e)e\,de = \frac{G(u_S^*)}{2}(1 - I)^2 v.$$

Using (17), we obtain

$$\frac{dE(e_U)}{dI} = -\frac{1}{2}(1 - I)v\left(2G(u_S^*) + \frac{1}{2}\frac{(1 - I)v(1 - 2G(u^*))G'(u_S^*)}{1 + (1 - I)vG'(u_S^*)}\right),$$

which is negative for all $I < 1$. Similarly, the expected effort of the aligned sender is

$$E(e_A) = \int_0^{\overline{e}} F_A'(e)e\,de = \frac{G(u_S^*)}{2}(1 - I)v,$$

which in tandem with (17) yields

$$\frac{dE(e_A)}{dI} = -\frac{1}{2}v\left(G(u^*) + \frac{1}{2}\frac{(1 - I)v(1 - 2G(u_S^*))G'(u_S^*)}{1 + (1 - I)vG'(u_S^*)}\right),$$

again negative when $I < 1$. □

Proof of Proposition 6. Consider a truthful equilibrium. A similar argument to part (ii) of the Proof of Proposition 1 establishes that there is a unique e^* that is played when $w = 1$. To maximize the size of the set of truthful equilibria, suppose that any out of equilibrium strategy is punished with posterior belief $\beta = 0$.

Now, let $w = 0$. Then the payoff to playing $s = (0, \tilde{e})$ is $t(\tilde{e}) - \tilde{e}$, so we must have $\tilde{e} = e'$. The sender's payoff from $(0, e')$ must be strictly preferred to the payoff from the deviation to $(1, e^*)$. Hence, incentive compatibility requires

$$(v + t(e^*))(1 - I) - e^* \leq t(e') - e' \iff e^* \geq (v + t(e^*))(1 - I) - [t(e') - e']. \tag{IC0}$$

Next let $w = 1$. Then to ensure truthful communication the incentive compatibility constraint is

$$t(e'')(1 - I) - e'' \leq v + t(e^*) - e^* \iff e^* \leq v + t(e^*) - [t(e'')(1 - I) - e''], \tag{IC1}$$

where $e'' = \arg\max_{e \geq 0}\{t(e)(1-I) - e\} \leq e'$. It is immediate to see that at least one e^* satisfies both (IC0) and (IC1).

In a truthful equilibrium, the payoff of a sender with $w = 1$ is $[v+t(e^*)]-e^*$, so (momentarily ignoring incentive compatibility) the sender's payoff is maximized by $e^* = e'$. Thus, suppose $e^* = \tilde{e} = e'$. It is easily checked that (IC1) is satisfied. (IC0) is also satisfied if $I \geq \frac{v}{v+t(e')}$. The expected payoff of the lowest type from entering is $p[v + t(e') - e'] + (1-p)[t(e') - e'] - \underline{c} > 0$, so a positive mass of senders enter. Having satisfied both incentive compatibility constraints and individual rationality, we have an equilibrium with truthful communication.

Now suppose $I < \frac{v}{v+t(e')}$. Then $e^* = \tilde{e} = e'$ violates (IC0) and cannot support a truthful equilibrium. We can restore truthful communication by increasing e^* until (IC0) is satisfied. The lowest $e^* \geq e'$ that satisfies (IC0) is, by definition, $\hat{e}$. Since (IC1) also holds whenever (IC0) binds, $e^* = \hat{e}$, $\tilde{e} = e'$ satisfy both incentive compatibility constraints. The sender's payoff from entering is $p[v + t(\hat{e}) - \hat{e}] + (1-p)[t(e') - e'] - c$. It is easily checked that $t(\hat{e}) - \hat{e}$ is increasing in I, so there is some $\bar{I}$ such that a positive mass of entry occurs if $I > \bar{I}$. Substituting $t(\hat{e}) - \hat{e} = t(\hat{e}) - \{(v + t(\hat{e}))(1-I) - [t(e') - e']\}$ from (11), we find that a positive mass of senders enter if (12) is satisfied. □

B Robustness

B.1 Noisy sender signals

Suppose that the sender observes the state with some noise. Formally, with (publicly known) probability λ they incorrectly believe $w = 0$ when the true state is $w = 1$ or vice-versa. We construct an equilibrium in which the sender truthfully reports their belief about w and the receiver plays $r = m$.

The incentive compatibility condition preventing a sender transmitting $s = (1, e^*)$ when she believes $w = 0$ becomes $v(1 - I(1 - \lambda)) - e^* \leq 0$. As previously, the unique effort level satisfying the intuitive criterion causes this inequality to bind. Meanwhile, the IR constraint is constructed by identifying sender beliefs that lead to an $s = (1, e^*)$ signal being sent and reaching the receiver:

$$\big\{(1 - \lambda)p(v - e^*)\big\} + \big\{\lambda(1 - p)[(1 - I)v - e^*]\big\} \geq c.$$

The first two terms respectively cover the cases where the sender correctly reports $m = 1$ and the one where the sender incorrectly reports $m = 1$ but goes unmoderated.

Combining the incentive compatibility and individual rationality conditions, there is an equilibrium with truthful communication if

$$I\left[1 - 2\lambda - \frac{1 - 2p}{p}\lambda^2\right] > \frac{c}{pv}. \tag{18}$$

As with imperfect moderation, the minimum level of moderation needed to sustain truthful communication is increasing in λ and truthful communication is no longer possible if λ is sufficiently large.

Moreover, unlike the case of moderator error, there is now an additional constraint on the set of parameters that support truthful communication, namely that the receiver must be willing to trust a signal of the form $s = (1, e^*)$. Formally, we require

$$\rho(1, e^*) = 1 \implies \beta(1, e^*) = \frac{(1 - \lambda)p}{(1 - I)\lambda(1 - p) + (1 - \lambda)p} \geq \frac{1}{2} \iff I \geq \frac{\lambda - p}{(1 - p)\lambda}.$$

Intuitively, when $\lambda > p$, the sender is more likely to transmit $m = 1$ because she made a mistake than because $w = 1$. The moderator must catch enough of these mistakes to ensure the receiver is willing to act on such a message. When this lower bound on moderation binds, it is another mechanism by which sender errors distort moderation away from its optimal level.

Focusing on the truthful equilibrium satisfying the intuitive criterion such that $e^* = v(1 - I(1 - \lambda))$, we recover welfare analysis that parallels Corollary 1.

Corollary 2. *In a truthful equilibrium satisfying the intuitive criterion, suppose welfare (or platform profit) is quasi-concave in I and maximized by $I^*(\lambda)$. Then*

- $I^*(\lambda) = \min\{\max\{\frac{I^*(0)}{1-\lambda}, \frac{\lambda - p}{(1-p)\lambda}\}, 1\}$.
- *If moderation is costless the highest welfare (or profit) that can be achieved with truthful communication is constant in λ if $\frac{\lambda - p}{(1-p)\lambda} \leq I^*(\lambda) < 1$, but decreasing in λ otherwise.*
- *If the moderator faces an increasing and weakly convex cost of inspecting messages then the highest welfare (or profit) that can be achieved with truthful communication is decreasing in λ.*

Moreover, if the rate of moderator error is too high (such that (18) *fails when $I = 1$) then it is impossible to sustain an equilibrium with truthful communication.*

B.2 Competing with uninformative content

We now consider a model where an ordinary sender (as in our main model) competes with one that provides only entertainment. Suppose there are two senders, an ordinary sender O with payoffs as in our main model and an entertainer L that obtains the payoff v_L from winning the receiver's attention and who sends the uninformative entertainment message m_L. The ordinary sender faces a fixed cost of entry drawn from the distribution G, while the entertainer faces no entry costs. Senders act simultaneously and hence in making their decisions have no information about the decision of the other sender. The timing of the game follows: The ordinary sender decides whether to enter the platform and learns the state. She may then decide on which message to send and both the ordinary sender and entertainer compete in effort for receiver attention. Let $Q(e_O, e_L)$ be a contest success function, where e_O is the ordinary sender's effort and e_L the entertainer's effort, indicating the probability the ordinary sender wins the receiver's attention. We assume $Q(0, e_L) = 0$, $Q'_{e_O} > 0$, $Q'_{e_L} < 0$, and if $e_O = e_L$ let $Q = 1/2$. We focus on finding an equilibrium which sustains truthful communication by the ordinary sender.

Proposition 7. *There exists an equilibrium with truthful communication if and only if $I \geq \min\{1, 1 - Q(e_O^*, e_L^*) + \frac{c}{pv}\}$. In this equilibrium, the ordinary sender sends the signal*

$$s = \begin{cases} (0,0) & \text{if } w = 0 \\ (1, e_O^*) & \text{if } w = 1 \text{ where } e^* = \max\{(Q'_{e_O})^{-1}_{e_O}(1/v), v(1-I)\} \end{cases}$$

and the entertainer chooses an effort level e_L^*, given by (19).[29]

Proposition 7 has two immediate implications. First, higher levels of inspection are required to sustain truthful communication when the ordinary sender faces competition. This is because there is probability that her effort is wasted (to competition). Therefore, to sustain communication the moderator has to bear a larger burden of convincing the receiver, which incentivizes the ordinary sender to enter. This recovers the participation-effort trade-off of the main model. Second, if the optimal inspection level is sufficiently large then equilibrium effort is independent of inspection. This is because the ordinary sender is willing to expand more effort, than necessary to convince the receiver, to improve her chances of winning the contest.

Proof of Proposition 7. Notice the game can be subdivided into two stages: A signaling stage followed by a contest stage. To save on notation, let Q represent $Q(e_O, e_L)$ and Q^* be the equilibrium probability of the ordinary sender winning, $Q^* := Q(e_O^*, e_L^*)$.

First, consider the signaling problem. For incentive compatibility, we require that when $w = 0$ the ordinary sender has no incentive to send the signal $(1, e_O^*)$ and instead sends $(0, 0)$. This means $e_O^* \geq v(1 - I)$. Second, an ordinary sender only enters if $p(vQ^* - e_O^*) - c \geq 0$. Indeed, if she does not enter then she expects the entertainer to win for certain and the receiver to choose $r = 0$ in accordance with his prior. Combining these constraints, and in equilibrium, we have $I \geq 1 - Q^* + \frac{c}{pv}$. This level of moderation is required to ensure ordinary sender enters and send a truthful signal, $(0, 0)$ or $(1, e_O^*)$, in equilibrium.

Second, consider the contest model. Following entry and learning the true state of the world, the ordinary sender chooses $e_O = 0$ if $w = 0$ and the entertainer can win the contest at any minimal effort. However, if $w = 1$, the ordinary sender chooses some e_O^* which maximizes her payoff subject to incentive compatibility,

$$\arg\max_{e_O} \; vQ - e_O \text{ s.t. } e_O \geq v(1 - I) \implies e_O^* = \max\{(Q^{*\prime}_{e_O})^{-1}_{e_O}(1/v), v(1 - I)\}.$$

The entertainer wins the contest at any positive effort with probability $1 - p$ and competes with the ordinary sender when $w = 1$. Therefore it maximizes

$$\arg\max_{e_L} v_L \left[G(p(vQ - e_O)) \left[p(1 - Q) + (1 - p)\right] + (1 - G(p(vQ - e_O)))\right] - e_L$$
$$\implies e_L^* = (Q^{*\prime}_{e_L})^{-1}_{e_L} \left(\frac{-1}{v_L p \left[G(p(vQ^* - e_O^*)) + pvQ^* G'(p(vQ^* - e_O^*))\right]} \right). \tag{19}$$

[29] F'_x means the derivative of F with respect to x. F_x^{-1} means the inverse of function F with respect to the argument x.

Notice that the entertainer's effort depends on inspection only through the ordinary sender's effort. And the ordinary sender's effort depends on inspection only if $v(1-I) > (Q^{*\prime}_{e_O})^{-1}_{e_O}(1/v)$. Hence, if the inspection required to sustain truthful communication in the signaling stage is sufficiently high, inspection plays no role in influencing the outcome of the contest stage. □

B.3 Senders learning state before entry

We consider a slight modification to the timing of the game. Suppose instead the sender learns w prior to entry. The timing of the game follows: (1) The moderator publicly announces I. (2) The sender learns the state of the world w and makes her entry decision. (3) The sender chooses $s = (m, e)$ if she entered, otherwise $s = \emptyset$. (4) Moderation takes place. (5) The receiver observes s and chooses r.

We look for a truthful equilibrium. Notice that for s on the equilibrium path, following $w = 0$, there must be non-entry by the sender. This is because $\rho(0, e) = 0$ in a truthful equilibrium, implying a sender's payoff from entering is $-e - c$.

Suppose that when $w = 1$ the sender reports $(1, e^*)$, and when $w = 0$ the sender does not enter. The receiver's posterior is

$$\beta(s) = \begin{cases} 1 & \text{if } s = (1, e^*) \\ pG(v - e^*) & \text{if } s = \emptyset, \end{cases}$$

and he plays $r = 1$ if $\beta(s) \geq 1/2$ and $r = 0$ otherwise. Note that $pG(v - e^*) \leq p < 1/2$ and $\rho(\emptyset) = 0$.

Because a sender knows $w = 0$ prior to entry, to maintain incentive compatibility, we require $v(1 - I) - e^* - c \leq 0$ and the individual rationality constraint is now $v - e^* - c \geq 0$. In the truthful equilibrium, we require the incentive compatibility constraint to be satisfied for senders of all c, hence it becomes $e^* \geq v(1-I) - \underline{c}$. Meanwhile the condition for a positive mass of senders to enter is $e^* < v - \underline{c}$.

From the intuitive criterion, we have $e^* = v(1 - I) - \underline{c}$. From this we see that if $I = 0$ then $G(v - e^*) = G(\underline{c}) = 0$ senders enter when $w = 1$, so we preserve the result of no entry without moderation.

Applying sender's equilibrium effort we recovery similar trade-offs as the main model: higher levels of moderation reduces sender effort which (i) improves sender participation and (ii) can hurt receivers. To see this, observe welfare is

$$W = p\left[G(v - e^*)\left[v - e^* + \pi_h(e^*)\right] - \int_{\underline{c}}^{v-e^*} cG'(c)\, dc + \left[1 - G(v - e^*)\right]\pi_l(0)\right] + (1-p)\pi_h(0),$$

and $e^* = v(1-I) - \underline{c}$ such that $\frac{\partial e^*}{\partial I} < 0$ and

$$\frac{\partial W}{\partial I} = -p\left[G'(v-e^*)\left(\pi_h(e^*) - \pi_l(0)\right) - G(v-e^*)\left(\pi_h'(e^*) - 1\right)\right]\frac{\partial e^*}{\partial I}.$$

This yields a formulation which is qualitatively identical to (4). Hence, we recover the welfare trade-off of sender participation and crowding-out effort which are key to our understanding of the optimal moderation policy.

B.4 Fact-checking

We consider an alternate moderation rule where the moderator does fact-checking instead of removal. We think of fact-checking as a notice stating the sender's message is misleading. In other words, suppose the sender sends $s = (m, e)$ when $w \neq m$. If the signal is selected for inspection the moderator includes a notice. Hence, following inspection the resulting signal becomes $s = (w, e)$ and the receiver can observe that the message has been fact-checked.

In this setting, Lemma 2 continues to hold. To sustain truthful communication, we must satisfy incentive compatibility such that the sender prefers not to send $s = (1, e^*)$ if $w = 0$. This means, identical to the main model, $v(1-I) - e^* \leq 0 \iff e^* \geq v(1-I)$. However, unlike the main model, the moderator takes on a more active role in information provision. The sender may choose $s = (0, 0)$ when $w = 1$ and rely on the moderator to inform the consumer of the true state following inspection without having to exert effort herself. As a result, a new incentive compatibility constraint is required to sustain truthful communication, $vI \leq v - e^* \iff e^* \leq v(1-I)$. Together, the incentive compatibility constraints imply $e^* = v(1-I)$. Observe that this coincides with the equilibrium effort level that survives the intuitive criterion.

Next, consider the sender's entry decision, she does so if individual rationality is satisfied, $p(v - e^*) - c \geq 0$. Applying e^*, individual rationality becomes $pvI \geq c \iff I \geq \frac{c}{pv}$. Finally, to ensure there exists at least some sender entry, we require individual rationality to be satisfied for at least the lowest c, hence $I > \frac{c}{pv}$.

Under the alternate moderation rule, we recover exactly Proposition 1 without the application of the intuitive criterion to pin down $e^* = v(1-I)$. Since we arrive at the same equilibrium under fact-checking, all results following from Proposition 1 follow through.

B.5 Equilibrium uniqueness and effort-dependent moderation

In the baseline analysis we show truthful communication can be sustained by an intermediate level of moderation. However, this equilibrium is not necessarily unique and, indeed, there

can exist equilibria without instrumental communication despite $I > \underline{c}/pv$. Suppose, for example, the sender always plays $m = 1$, $e = 0$ as part of a putative equilibrium (irrespective of the state) and that an out of equilibrium message is always interpreted as implying $w = 0$. The receiver's posterior when he observes that the signal was not rejected is

$$\beta(s) = \frac{p}{(1-I)(1-p)+p},$$

which is increasing in I. Thus, for $I > 1 - \frac{p}{1-p}$ the receiver will play $r = 1$. For the sender, any alternative strategy yields a non-positive payoff and she faces no profitable deviations. Although the receiver correctly anticipates the state when $w = 1$, this outcome is driven by the moderator's efforts rather than the content of communications from the sender. Formally, communication is not instrumental as on the equilibrium path $\nexists s \neq s'$. The receiver's behavior is affected by the moderation outcome but not at all by the content of the message. Thus, we see a perverse effect of moderation: moderation can make messages so convincing that the sender completely free-rides on the moderator, with no incentive to exert effort or tell the truth.

It is therefore interesting to ask whether the moderator, with a slight enrichment of the space of possible moderation policies, could guarantee that truthful communication takes place. The answer is affirmative if the inspection probability can be conditioned on the sender's effort, $I(e)$. For example, as a first pass for moderation websites such as Wikipedia depend on the use of citations to support claims, while on forums such as Stack Exchange or Facebook communities effort may be proxied by post length. We restrict attention to equilibria that are monotone in the sense that the sender is more likely to lie when the truth is not in her favor:[30]

$$\Pr(m = 1|w = 0) \geq \Pr(m = 0|w = 1). \tag{20}$$

We then have:

Proposition 8. *Suppose the moderator commits to the effort-dependent inspection rule*

$$I(e) = \begin{cases} 0 & \text{if } e < e^* \\ \hat{I} & \text{if } e \geq e^*, \end{cases} \tag{21}$$

where $(e^*, \hat{I})$ *satisfy* $e^* \in (v(1-\hat{I}), v - \frac{c}{p})$. *Then every equilibrium of the communications game that satisfies* (20) *and intuitive criterion is payoff-equivalent to one with truthful communication.*

[30]Practically, (20) guarantees that signals are more often deleted when $w = 0$, meaning a null signal will not be interpreted as evidence that $w = 1$. In other words, it guarantees that $\beta(\emptyset) < 1/2$ and $\rho(\emptyset) = 0$.

Proof of Proposition 8. Suppose moderation policy (21) is in effect. For $i \in \{0,1\}$, partition the non-null signals into the following (exhaustive) types: (i) those with effort $e < e^*$, (ii) those with $m = i$ and effort $e_i \geq e^*$. Denote a generic signal of the first type as $s_- = (m, e_-)$ and one of the second type as $s_i = (i, e_i)$. We will show that high-effort signals are always truthful ($s_i \notin S_{-i}$, 'Step 1') and that the sender never transmits a low-effort signal if $w = 1$ ($s_- \notin S_1$, 'Step 2'). Between, we also prove that the sender gets zero payoff whenever $w = 0$ ('Intermediate step'). Together, these observations imply that the only way for communication to be non-truthful is if $s = (1,0) \in S_0$. But we can then construct a payoff equivalent truthful equilibrium by sending $s = (0,0)$ instead.

STEP 1: To show that high effort signals never contain a false message ($s_i \notin S_{-i}$), notice that s_i would yield sender payoff $\rho(s_i)v(1-\hat{I}) + \rho(\emptyset)v\hat{I} - e_i$ when $w = -i$. Since $e_i \geq e^* > v(1-\hat{I})$, and since $\rho(\emptyset) = 0$ under (20), this payoff is negative.

A consequence of Step 1 is that any on-path s with $e \geq e^*$ must fully reveal the state. The sender therefore never exerts $e \geq e^*$ if $w = 0$.

INTERMEDIATE STEP: Now define $\bar{s} \in \arg\min_{s \in S_0} \beta(s)$. In words, $\bar{s}$ is a signal that minimizes the receiver's posterior among the signals in S_0. By Step 1, $e(\bar{s}) < e^*$, meaning the signal is never deleted by the moderator. We must have $\beta(\bar{s}) \leq p$, meaning $\rho(\bar{s}) = 0$.

The sender's payoff from $\bar{s}$ is therefore $-e(\bar{s})$, implying $e(\bar{s}) = 0$. Since the sender must be indifferent between all $s \in S_0$ if $w = 0$, she therefore gets equilibrium utility of zero whenever $w = 0$.

STEP 2: Signal s_- is inspected with probability $I(e_-) = 0$ and therefore yields sender payoff $v\rho(s_-) - e_-$, regardless of w. This payoff must be weakly negative (otherwise, the sender would deviate to s_- whenever $w = 0$). Thus, to show $s_- \notin S_1$, it suffices to find a signal that delivers a positive sender payoff when $w = 1$. Such a signal is $s^* = (1, e^*)$. From Step 1 we know $s^* \notin S_0$. Thus, if $s^* \in S_1$ it must induce $\rho(s^*) = 1$ and leaves the sender with positive payoff $v - e^* > 0$. If $s^* \notin S_1$ then, under the intuitive criterion, s^* must be interpreted by the receiver as implying $w = 1$ because it yields payoff no higher than $\rho(s^*)v(1-\hat{I}) - e^* < 0$ if $w = 0$. Again, s^* therefore leads to $\rho(s^*) = 1$ and leaves the sender with positive payoff if $w = 1$. □