Discussion Paper Series – CRC TR 224

Discussion Paper No. 698
Project B 05

Moderating Content-Hosting Platforms

Robin Ng[1]
Greg Taylor[2]

June 2026
(Second Version : September 2025)
(First version : 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‡

May 22, 2026

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 engagement value of content on the platform. A socially optimal policy may therefore involve limited moderation, but platforms often have an incentive to allow too much false content. We study the interaction between moderation and other strategic platform choices, including a platform's decision to share revenue with creators, the decision to be a content-hosting platform, and the enablement of AI features.

1 Introduction

Digital platforms enable ordinary (often anonymous) users to create and disseminate their own online content. Users of platforms such as Facebook, YouTube, or Amazon may produce false or fraudulent content in an attempt to mislead others into acting in their own interests, e.g., to buy their sponsor's product or to vote in a particular way. This has become a significant policy issue. Regulators like the Federal Trade Commission and European Commission have imposed new compliance requirements onto platforms to prevent deceptive product promotion by creators or incentivized online reviews.[1] Meanwhile, the role of misinformation in hotly

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

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

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

[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.

debated topics such as politics, health, or climate change, and the prevalence of harmful content are also active subjects of policy attention. In response to these measures, many online platforms operate moderation policies to remove misleading content. For example, YouTube and Facebook have policies to remove content that contains false or harmful information, 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 can generate content describing it to a receiver. The sender can also exert effort to make her content more engaging. A receiver benefits both from learning about the truth and from consuming engaging content. 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 platform can moderate content by inspecting a random sample of messages and deleting those that are untruthful.

As a benchmark, we show that moderation or sender effort in isolation does little to enable communication. If senders cannot invest in engaging content then communication is possible only if every message is moderated without error. 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. Thus, neither moderation nor sender effort alone is enough to sustain meaningful communication.

The picture changes substantially when moderation and sender effort are paired together. Moderation makes effort a more powerful signal of truthfulness because a sender is reluctant to exert effort on crafting an engaging but false message that will probably be deleted. This means that honest senders find it less costly to persuade the receiver of their truthfulness and more attractive to participate on the platform. Hence, moderation enables truthful communication 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 engaging content, the socially optimal moderation policy therefore involves a trade-off because inducing sender participation crowds out user engagement. Consequently, the socially optimal level of moderation may be relatively low, even if moderation is completely costless.

[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.

A platform driven by ad revenue must get both senders and receivers on board. Its incentives depart from those of a planner in two respects. First, the platform tends to prefer equilibria with some false content over the fully informative equilibrium. False content induces higher *quantities* of engaging content to which ads can be attached, without sacrificing receiver engagement. Second, because of the participation-effort trade-off just described, the platform may under-moderate in order to induce senders to create higher *quality* content.

We study several strategic choices of the platform that interact with moderation. First, we discuss if a platform would commit to its moderation policy. Without commitment, the platform is tempted to retain highly engaging misleading content rather than delete it, which undermines the credibility of moderation and impedes informative communication. We show, however, that moderation can be sustained by reputation in a repeated interaction and the platform has a strong incentive to develop a reputation for moderating faithfully.

Second, we consider transfers from the platform to senders, such as revenue sharing. Transfers increase the opportunity cost of producing misleading content, meaning less moderation is sufficient to sustain communication. Therefore, transfers and moderation can complement each other in sustaining truthful communication.

Third, we examine a firm's choice between operating as a content-hosting platform and as a traditional broadcaster that produces content itself. This choice depends on the participation-effort trade-off. When inducing participation is easy, the platform model becomes more attractive, whereas if it is difficult to get senders to join, direct content production becomes more appealing.

We also study the implications of AI-generated content. We model AI as lowering the cost of producing engaging content while also being prone to hallucination. These two forces work in opposite directions. Lower production costs make it easier for dishonest senders to mimic honest ones, while hallucination lowers the expected return from using AI. Dishonest senders are more likely to adopt AI than honest ones, which can make AI tools unattractive to the platform. But the possibility of cheap AI imitations makes honest senders create more engaging content to establish their credibility and this can induce platforms to embrace sufficiently advanced AI.

In sum, the paper yields a number of empirical predictions and policy implications. As the cost of content creation falls, content-hosting platforms become more prominent relative to more traditional publishers or broadcasters. We find that such content-hosting platforms are often half-hearted moderators: although they may write clear content guidelines, they have an incentive not to fully enforce them, turning a blind eye to false but engaging content. Nevertheless, platforms do have an incentive to moderate to some degree and will seek to develop a reputation for doing so. From a policy point of view, these results leave space for

a regulator to mandate tougher moderation standards. However, in some cases, especially those with low stakes, excessively stringent moderation can backfire because it deters the creation of high quality content.

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 equilibrium and the key trade-offs. Section 4 studies optimal moderation by platforms and by a regulator concerned with user welfare. Section 5 examines a platform's commitment to its moderation policy, the role of transfers to senders, a firm's choice between platform and broadcaster business models, and the implications of AI-generated content. Section 6 concludes. Omitted proofs are in Appendix A, and Appendix B provides a formal demonstration that our results are robust in various respects.

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 that limiting content sharing can improve overall information quality. In a dynamic information disclosure setting where followers may prefer low-quality content, Szydlowski (2023) shows how content moderation can induce senders to provide only high-quality content, even if some receivers may prefer low-quality content. Madio and Quinn (2024) describes how profit concerns can result in platforms moderating content beyond the socially optimal level. Likewise, Rendo (2025) shows that excessive moderation may arise because a mainstream platform does not internalize that moderating displaces some users to less safe unmoderated fringe platforms. In a model of horizontally differentiated content, Liu et al. (2022) shows that a platform's revenue model can affect its moderation strategy, and discusses how imperfect moderation technology can lead a platform to allow extreme content. Our paper addresses a similar question to Dwork et al. (2024), which studies 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). Our model uncovers a different backfiring mechanism: moderation can cause senders' efforts to decrease, suggesting that imposing high levels of moderation onto platforms could lead to less engaging content. 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. We contribute to this literature by framing the communication between users as a signaling game

and show how moderation can sustain truthful communication.

More closely related to our work is a nascent literature on *content moderation in communication games*. Kominers and Shapiro (2025) show that the only way to robustly moderate a decentralized communication platform is to destroy some information. Our work differs by endogenizing the incentives of a sender under a known moderation policy and studying equilibrium content production. Rhodes and Wilson (2018) consider a signaling model of false advertising. Like us, they identify a trade-off because fines for false advertisements result in fewer false ads but higher prices. Whereas their model features the effect of fines on truthful advertising and prices, we instead consider how moderation enables communication through an alternative signaling mechanism involving sender effort. This allows us to consider a different set of policy issues relevant to the context of content moderation.

Our paper considers an application of *communication and signaling games* (e.g. 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 by introducing an independent moderator and 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 regulator, 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; Jann and Schottmüller, 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, and characterizes when higher quality senders choose to employ misinformation. We approach misinformation differently. In our model, false content can arise endogenously in partially informative equilibria, and a platform is willing to tolerate such content because engaging false content can still attract advertising revenue.

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), Andres and Slivko (2021), and Mattozzi et al. (2022) show that the amount of toxic posts on platforms decreases following the use of moderation while Jiménez-Durán (2023) finds a negligible effect on creator participation. We provide one possible mechanism that shows how content moderation can promote participation of content creators.

2 Model setup

Players and strategies We consider a game of incomplete information with three players: a *sender*, a *receiver*, and a *moderator platform*. At the beginning of the game, the moderator publicly and costlessly commits to a moderation rule, $I \in [0,1]$, which can be understood as the probability with which an untruthful message is deleted.

The sender moves next. She first decides whether to enter the platform at idiosyncratic cost c, drawn from a differentiable log-concave CDF G with support $[\underline{c}, \overline{c}]$ and $\underline{c} \geq 0$. Conditional on entry, the sender observes the state of the world, $w \in W = \{0,1\}$, where $\Pr(w=1) = p$, $p \in (0,1)$. She then chooses an action $s = (m, e)$, comprising a message ($m \in W$) and an effort to make the content engaging ($e \in \mathbb{R}_+$).[3] Denote by $m(s)$ and $e(s)$ the message and effort components of s. Let $\mathbb{S} = \{0,1\} \times \mathbb{R}_+$ be the set of possible signals. A sender's strategy following entry is $\sigma_w(\cdot) \in \Delta(\mathbb{S})$ with support S_w, where for any $X \subseteq \mathbb{S}$, $\Pr(s \in X | w) = \sigma_w(X)$. If the sender does not enter the platform then nature transmits a null signal, comprising $m = \emptyset$ and $e = 0$. In a slight abuse of notation, we denote this null signal by $s = \emptyset$.

Once the signal is chosen, moderation takes place. If the message is untruthful ($m \neq w$) then, with probability I, the signal is deleted and replaced with the null signal, $s = \emptyset$. Otherwise, the message is passed-on unmodified.

After the sender has moved and any moderation has taken place, the receiver observes the post-moderation signal, $s \in \mathbb{S} \cup \emptyset$, and forms posterior belief $\beta(s) = \Pr(w=1|s)$. The

[3]Effort can be interpreted as an observable investment in content quality. Even if viewers cannot observe the signal ex-ante, they can observe content quality ex-post.

receiver takes an action $r \in W$. Denote a generic receiver strategy by $\rho(s) = \Pr(r = 1|s)$.

Payoffs 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 $\pi_h(e) > \pi_l(e)\ \forall e$, $\pi_h'(e) \geq 0$ and $\pi_h''(e) \leq 0$.[4] Thus, the receiver would like his action to match the state. We also allow the receiver's payoff to depend on how engaging the content is.

The sender has state-independent preferences:[5] she seeks to persuade the receiver to play $r = 1$ and obtains payoff (gross of entry cost, c):

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

The platform is financed by ads and chooses its moderation strategy to maximize ad revenue. It earns a unit ad revenue for each unit of attention it supplies to advertisers. 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 and not have her message deleted). Second, the receiver must dedicate attention to consuming the message; we assume he does so in increasing relation to how engaging the message is, e. Formally, let $A(e)$ represent the ad revenue generated by the attention of the receiver, with $A(0) = 0$, $A'(e) \geq 0$, and $A''(e) \leq 0$. A can easily be microfounded.[6] Thus, when the sender enters with probability N and transmits signal s, the platform's profit is

$$\Pi = N \cdot A(e(s)).$$

Timing, equilibrium, and additional assumptions The timing is as follows:

1. The moderator publicly announces I.

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

[5] We could add the possibility that some fraction of senders have payoffs aligned with the receiver (e.g, because they are altruistic). The below results on when moderation can sustain a truthful equilibrium would not change. However, by aligning the sender's and receiver's payoff, some communication can be achieved even without moderation (e.g., Crawford and Sobel (1982)).

[6] For instance, suppose that the receiver gets a flow utility of $\sqrt{eA}$ from spending A units of time ("attention") consuming content, and that time has a unit opportunity cost of time. Then the receiver maximizes $\sqrt{eA} - A$, implying $A = e/4$.

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 the post-moderation s and chooses r.

The solution concept is Perfect Bayesian Equilibrium under the following parameter assumptions.

Assumption 1. $p < 1/2$.

This assumption says that that the receiver prefers to play $r = 0$ at his prior beliefs. We thus study the interesting situation where the sender has an incentive to convince the receiver to take the action $r = 1$. Consider the converse, $p \geq 1/2$, then the receiver always plays $r = 1$ and the sender has no incentive to send a signal.

Assumption 2. $v \in (\underline{c}/p, \overline{c})$.

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 communication 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 clearly one of instrumental communication.

Commentary on assumptions Our analysis is robust to relaxation of several assumptions. First, effort in our model affects payoffs but does not directly affect the information content of the message. This cleanly highlights the role of effort as a signaling device, clarifying our main mechanism. In Appendix B, we discuss an alternative model where effort also makes messages more informative, yielding essentially the same results.

Second, sender entry occurs before the state is realized. Entry is costly and we should interpret this decision as a long-run commitment.[7] However, as we show in Appendix B, our insights are robust when senders know w before joining the platform.

Third, a receiver is unable to distinguish between a moderated untruthful message and the absence of any signal. In Appendix B, we consider an alternative moderation technology that we call *fact checking*. Instead of deleting false content, the moderator simply adds a public label indicating that $m \neq w$. The switch from moderation to fact checking leaves the analysis substantively unchanged and our results follow through.

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.

Finally, we can expand our state space beyond the binary setting, allow the moderator to make mistakes, allow for sender reputation, and allow for non-linear effort costs without affecting our main insights.

Some applications To help anchor the model, we briefly outline some potential applications:

The sender is an 'influencer' who promotes a sponsor's product. She receives affiliate income if the receiver buys the product. The sender thus has an incentive to persuade the receiver to buy ($r = 1$), even if the product is bad ($w = 0$). The receiver prefers to buy good products. Major platforms have explicit policy to remove undisclosed and deceptive marketing. The sender could alternatively be providing advice on topics like personal health, fitness, or finance. She may try to persuade the receiver to enroll in a paid training program ($r = 1$), but the sender knows the program has no real benefit ($w = 0$). The receiver prefers not to be a victim of scam and the platform removes some fraudulent content.

The sender produces political content but has a hidden agenda to influence the receiver's action or belief, such as inducing him to join a protest, oppose vaccine mandates, or become a climate skeptic ($r = 1$). To achieve this, she may spread false information on topics like migration or climate change ($m \neq w$). The receiver prefers not to be manipulated into inappropriate action, and the platform removes or fact-checks false claims.

With a slight reinterpretation: the sender has a type, either safe ($w = 1$) or harmful

[7]For example, an online 'influencer' must spend time building an audience before they are contacted by firms to promote their products. 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.

($w = 0$). A harmful sender can present content such as hate speech or violent content in its naked form ($m = 0$) or disguise it as safe (e.g., with an innocent title, $m = 1$). The receiver observes m and e (e.g., from the thumbnail and first few seconds of a video)[8] and chooses whether to continue consuming ($r = 1$) or not ($r = 0$). If the receiver continues consuming the content the sender gets ad revenue v, while the receiver obtains either a direct benefit ($\pi_h(e) > 0$) or harm ($\pi_l(e) < 0$) depending on the sender's type. The moderator deletes fraction I of harmful messages. This alternative framing yields the same results as our baseline model.

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 $I = 1$.*

Proofs are in Appendix A. The intuition for the two cases is slightly different. Suppose there is no moderation, $I = 0$. Then conditional on entry, the sender's payoff is $\rho(s)v - e(s)$ and the only way for a $w = 1$ sender to credibly convey information without being imitated by a $w = 0$ sender is to dissipate her entire surplus through effort. But that makes the expected payoff from joining the platform negative, so the sender never enters and no communication takes place in equilibrium.

If, instead, $e = 0$ then this becomes a game of cheap talk and, given $\rho(s) < \rho(s')$, no sender will transmit s. 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. This echoes Rhodes and Wilson (2018)'s result (Proposition 1) that truthful communication can be sustained only if the profits from lying are fully eliminated.[9] In practice, it is likely that moderating every message ($I = 1$) on a large online platform will be prohibitively costly.[10] In our model, this

[8] We could also allow the receiver to observe e only with noise before he starts watching, with little change in results.

[9] In Rhodes and Wilson (2018) profits need not be fully eliminated if advertisers have heterogeneous marginal costs. A contribution of our paper is to show that sender effort along with moderation endogenously generates single-crossing without this cost asymmetry.

[10] Often, a platform's moderation decision is reactive and depends on users reporting the content. Even if $I = 1$ is feasible (e.g., algorithmic moderation), instrumental communication is possible only if the moderator makes no mistakes.

means no useful communication is feasible without sender effort.[11] The rest of this paper is therefore focused on how sender effort in combination with moderation enables a new mechanism for signaling, which can sustain communication even at lower levels of moderation.

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:*

$$\beta(s) = \begin{cases} 1 & \text{if } s = (1, e^*) \\ p & \text{if } s = \emptyset \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

- *$r = 1$ 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.[12] In order to sustain truthful communication, we need the sender to prefer setting $s = (0, 0)$ when $w = 0$.[13] 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

[11] 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.

[12] Equation (2) specifies off-path belief $\beta = 0$. But any $\beta < 1/2$ would lead to $r = 0$, which we use to deter off-path deviations by the sender.

[13] 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. Moderation therefore makes effort an endogenously 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) pin-down the set of e^* and I that can be sustained in an equilibrium with truthful communication.

Proposition 1. *(i) There exists an equilibrium with truthful communication if and only if the level of moderation satisfies*

$$I > \frac{c}{pv}. \tag{3}$$

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

Although both moderation and sender effort are needed to sustain communication (Lemma 1), Proposition 1 shows how they are substitutes in the truthful communication equilibrium (e^* is decreasing in I). Intuitively, higher moderation increases the expected cost of sending a false message: a sender choosing to send a false message risks deletion, which slackens the incentive compatibility constraint. This allows honest senders to establish their credibility with less effort. In this sense, moderation crowds out sender effort, which implies that the optimal level of moderation may be positive but low, even when moderation is costless.

3.3 Partially-informative equilibria

We next extend the analysis to consider equilibria that are partially informative. Our key message that both moderation and sender effort are needed to sustain communication, but become substitutes when communication is instrumental survive. To reduce the number of cases, make the following assumption. In words, Assumption 3 says that the sender tells the truth when the truth in in her favor.

Assumption 3. *If* $w = 1$ *and the sender enters then she transmits* $m = 1$ *with probability 1.*

Up to payoff equivalence,[14] we show that every partially-informative equilibrium is found by taking the truthful equilibrium from Proposition 1 and modifying it so that $w = 0$ senders lie about the state a fraction ζ of the time. More formally:

Proposition 2. *Suppose Assumption 3 holds. Following sender entry, any instrumental equilibrium satisfying the intuitive criterion is payoff equivalent to one with the following structure:*

1. *When $w = 1$, the sender always transmits $s^* = (1, e^*)$, where $e^* = v(1 - I)$.*
2. *When $w = 0$, the sender mixes, transmitting $s^* = (1, e^*)$ with probability $\zeta \leq \frac{p}{(1-p)(1-I)}$ and $\underline{s} = (0, 0)$ with probability $1 - \zeta$.*
3. *The receiver updates his beliefs according to Bayes' rule and plays $\rho(s^*) = 1$ and $\rho(s) = 0$ for any $s \neq s^*$.*

In particular, letting $\zeta = 0$ yields the truthful equilibrium of Proposition 1.

Importantly, ζ must be low enough that the receiver is still willing to trust the message, so a partially informative equilibrium is one in which which senders sometimes trick the receiver into a mistake by creating high-effort but misleading content. Notice that for I sufficiently large we have $\frac{p}{(1-p)(1-I)} \geq 1$, meaning we can sustain an equilibrium in which receivers trust the content they see even if $\zeta = 1$. Intuitively, the moderator needs to remove just enough false content to ensure that any message that remains is credible.

4 Optimal moderation

We next study the platform's optimal moderation strategy. The platform prefers partially informative equilibria with lots of misleading content and has an incentive to limit moderation in order to stimulate the production of additional engaging content. We then consider a regulator that maximizes the surplus of platform users. This involves forcing the platform to implement a truthful equilibrium and, potentially, regulating to adjusting the level of moderation.

[14] We say 'up to payoff equivalence' because, in any instrumental equilibrium, there is an on-path signal of the form $s = (m, 0)$ that fully reveals $w = 0$. It is then payoff irrelevant whether $m = 0$, $m = 1$, or the signal is deleted because all lead to $r = 0$.

4.1 Profit-maximizing moderation strategy

The platform earns zero profit in any non-instrumental equilibrium because senders never join, so we focus here on instrumental equilibria. In any instrumental equilibrium, the platform's profit is

$$\begin{aligned}\Pi &= G(p(v-e^*)) \cdot \Pr(e=e^*) \cdot A(e^*) \\ &= G(pvI) \underbrace{[p+(1-p)\zeta(1-I)]}_{Z(I)} A(v(1-I)).\end{aligned}$$

This expression comprises of the mass of senders who enter (G), the frequency with which they successfully transmit engaging messages (Z), and the resulting unit ad revenue from receiver attention (A). As a first observation, note that Π is increasing in ζ: among all instrumental equilibria, the platform prefers the least-informative one. In less informative equilibria some senders are allowed to deceive the receiver by creating highly engaging (though false messages) that can be monetized through advertising.

Suppose the optimal I is given by the first-order condition (to be verified later). Then, momentarily treating Z as differentiable,[15] the platform solves:

$$\frac{\partial \Pi}{\partial I} = Z(I^*) \left[\underbrace{G'(pvI^*)pA(v(1-I^*))}_{\text{participation effect}} - \underbrace{G(pvI^*)A'(v(1-I^*))}_{\text{crowding-out effect}} \right] v + \underbrace{G(pvI^*)A(v(1-I^*))Z'(I^*)}_{\text{misinformation effect}} = 0, \quad (4)$$

where I^* represents the platforms optimal moderation policy.

First focus on the term in square brackets. This has two parts. The first is a *participation effect*: higher I makes it less costly for the sender to influence the receiver, making her more inclined to enter the platform. Increasing sender entry expands the pool of messages to which ads can be attached, boosting profits. The second term is a *crowding-out effect*. Once (IR) is satisfied, further increases in I reduce the effort of senders. This effect is negative because it reduces receiver engagement with messages and thereby the associated ad revenues.

The final term in (4) is a *misinformation effect*. This represents the fact that a change in I can influence both the frequency with which senders transmit false messages and the likelihood of such messages surviving moderation, both captured by $Z'(I^*)$. However, the misinformation effect disappears under reasonable equilibrium selection criteria. First, it is natural to select the truthful equilibrium ($\zeta = 0$) because any $\zeta > 0$ involves the play of

[15] Under the equilibrium selection criteria we introduce below, Z is indeed differentiable.

weakly dominated strategies and is inherently fragile.[16] Selecting the truthful equilibrium yields $Z(I) = p$ and $Z'(I) = 0$. Second, suppose we select the platform-optimal equilibrium (i.e., the one with the highest ζ). From Proposition 2 we thus have $\zeta = \min\{\frac{p}{(1-p)(1-I)}, 1\}$, which again leads to $Z'(I) = 0$ if $\zeta < 1$. If instead $\zeta = 1$, then $Z'(I) = -(1-p)$, and the misinformation effect is another incentive to reduce moderation.

As a last observation, if $Z'(I) = 0$ then Π is quasi-concave in I and the uniquely optimal I^* is interior and given by (4). Summarizing the platform's optimal strategy:

Proposition 3. *(1) Suppose Π is quasi-concave. In the truthful equilibrium, the optimal moderation strategy is interior, $I^* \in (\frac{c}{pv}, 1)$, and satisfies*

$$\frac{G'(pvI^*)p}{G(pvI^*)} = \frac{A'(v(1-I^*))}{A(v(1-I^*))}. \tag{5}$$

(2) The platform-optimal equilibrium is the instrumental equilibrium that has the largest number of misleading messages, $\zeta = \min\{\frac{p}{(1-p)(1-I)}, 1\}$. (3) In the platform optimal equilibrium, the platform optimal I is weakly lower than the truthful equilibrium, and strictly so if $\zeta = 1$.

Thus, we find that the platform will choose to limit the extent of its moderation, even if it could perfectly and costlessly moderate all messages. The two sides of (5) can be interpreted as reflecting the trade-off for a two-sided platform that needs to get both content producers and audiences on-board. When receivers' attention for adds is inelastic, the main imperative is to attract content producers, which is achieved with a high I. Conversely, the crowding-out effect dominates if sender participation is relatively inelastic, causing moderation intensity to fall. This is because limiting the intensity of moderation forces senders to work hard at establishing their credibility, which results in engaging content that the platform can monetize.

4.2 Regulating the platform

We have seen that the least-informative instrumental equilibrium (i.e., the one that maximizes ζ) is platform-optimal. However, examination of Proposition 2 immediately reveals a potential tension between the platform's and users' interests:[17]

[16] On weak dominance, note that $\underline{s}$ is never worse and sometimes better than s^* for a $w = 0$ sender. On fragility, suppose we perturb the game so that with some (small) probability the receiver is 'incredulous' and will play $r = 1$ only if $\beta = 1$. Any $\zeta > 0$ equilibrium then fails because $w = 1$ senders could strictly separate themselves and convince the credulous receivers with only a small increase in effort. The truthful equilibrium, on the other hand, is robust to such a perturbation.

[17] The Corollary uses the fact that the receiver's expected payoff is $p\pi_h(e^*) + (1-p)(1-\zeta(1-I))\pi_h(0) + (1-p)\zeta(1-I)\pi_l(e^*)$.

Corollary 1. *In any instrumental equilibrium: (i) the sender's payoff is pvI, which is independent of ζ. (ii) the receiver's payoff is decreasing in ζ if and only if $\pi_h(0) > \pi_l(e^*)$.*

Thus the truthful equilibrium ($\zeta = 0$) is Pareto optimal from the point of view of the platform's users as long as the receiver sufficiently values learning the truth ($\pi_h(0) > \pi_l(e^*)$). For brevity, we focus on this case, which seems to be of particular interest to regulators.[18] The platform then has too little interest in truthfulness because it maximizes user engagement (with ads) rather than the informational value of the content. A regulator concerned for users' well-being may therefore find it necessary to compel the platform to implement the truthful equilibrium. Indeed, authorities, including those in Singapore, France, and Germany, have imposed on platforms a duty to ensure truthfulness.[19]

Suppose that the regulator indeed forces the platform to select the truthful equilibrium. We show how the platform can achieve the truthful equilibrium at the end of the section. The welfare of the platform's users is

$$W(I) = G(pvI)\left[pvI + p\pi_h(v(1-I)) + (1-p)\pi_h(0)\right] - \int_{\underline{c}}^{pvI} c\, dG(c) + \left[1 - G(pvI)\right]\left[p\pi_l(0) + (1-p)\pi_h(0)\right]. \quad (6)$$

The first line captures the welfare when the sender enters the platform, and the second line is the receiver's payoff from following his prior when the sender does not enter.

Suppose π_h is sufficiently concave such that $W(I)$ is quasi-concave. Then the regulator problem, when interior, solves

$$\frac{\partial W(I)}{\partial I} = pv\left[\underbrace{G'(pvI^{**})p\left[\pi_h(v(1-I^{**})) - \pi_l(0)\right]}_{\text{participation effect}} - \underbrace{G(pvI^{**})\left[\pi_h'(v(1-I^{**}))-1\right]}_{\text{crowding-out effect}}\right] = 0, \quad (7)$$

where I^{**} represents the regulator's optimal moderation policy.

The regulator faces a qualitatively similar participation-effort trade-off to the platform. The participation effect reflects gains in welfare arising from additional senders entering the platform, while the crowding-out effect arises because moderation crowds out sender

[18]For example, regulators may be more concerned with the informational value of content than with its entertainment value or false content may cause negative externalities for non-users, such as when misinformed receivers forgo vaccination and infect their neighbors. In the reverse case ($\pi_l(e^*) \geq \pi_h(0)$), the platform's preferred ζ would also be the best for its users, but it remains the case that the user-optimal I is often interior and generically different from the platform-optimal one.

[19]Indeed, if we reinterpret the model along the lines of harmful content, many advertisers prioritize brand safety, and do not want their ads placed alongside such content. Hence, a platform caring about ad revenue would want to enforce truthfulness.

effort: in equilibrium, moderation and effort are substitutes: higher moderation relaxes the incentive compatibility constraint, which means less effort is required to establish sender credibility. This reduction in effort is harmful if effort is socially efficient ($\pi_h' > 1$), which can lead the regulator to prefer less than full moderation, even though it does not internalize any moderation cost. The following result formalizes this claim and compares the regulator's choice of I to the one that would be chosen by a platform.

Proposition 4. *Suppose* Π *and* W *are quasi-concave. In the truthful equilibrium:*

(i) The user-optimal level of moderation is interior, $I^{**} < 1$*, if and only if*

$$\frac{G'(pv)}{G(pv)}p < \frac{\pi_h'(0) - 1}{\pi_h(0) - \pi_l(0)}. \tag{8}$$

(ii) The platform prefers less moderation than its users, $I^* < I^{**}$*, if and only if*

$$\frac{A'(v(1-I^*))}{A(v(1-I^*))} > \frac{\pi_h'(v(1-I^*)) - 1}{\pi_h(v(1-I^*)) - \pi_l(0)}. \tag{9}$$

The left-hand side of (8) is proportional to the elasticity of sender participation, while the right-hand side is the ratio of the social value of effort to the social value of information. The first part of Proposition 4 says that a regulator concerned with user well-being may prefer limited moderation to avoid crowding-out sender effort, especially in low-stakes environments where $\pi_h(0) - \pi_l(0)$ is small. In fact, the optimal level of moderation may be very far 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 (7). Intuitively, when most senders have very low participation cost, a little moderation suffices to get almost all senders on board and the main effect of further increases in I is to crowd out effort.[20]

The second part of Proposition 4 says the platform undermoderates when ad revenues are sensitive to content quality and receivers care a lot about learning the truth. In that case, the platform will keep moderation low in order to generate engaging content that boosts ad revenues, whereas receivers would benefit from higher moderation that results in a large supply of informative messages. If receivers care *only* about learning the truth (i.e., if

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

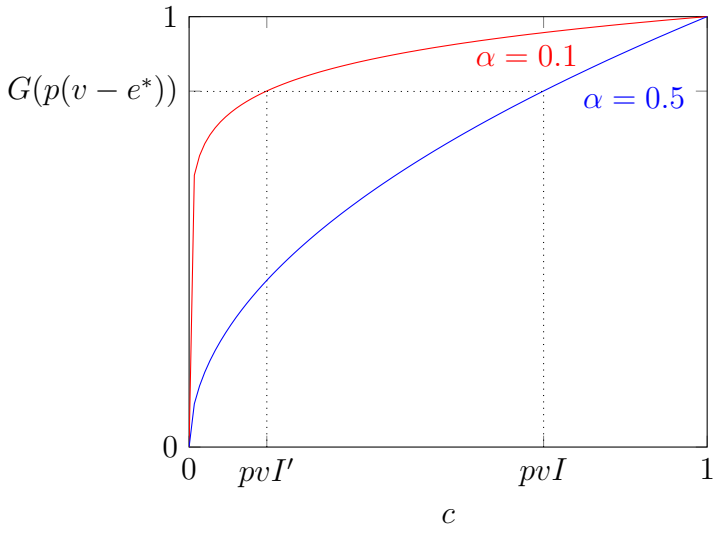

Figure 1: Using $G(c) = c^\alpha$ as an illustration, a decrease in α causes $G(c)$ to become more concave. When $G(c)$ is highly concave a lower level of moderation is sufficient to induce most senders to participate, meaning the crowding-out effect dominates even for small I.

$\pi_h'(e) = 0$) then the regulator would choose $I = 1$ and the platform always implements too little moderation. Thus, even after a regulator forces a truthful equilibrium, there may still be scope to regulate for a more efficient level of moderation.

The proposition therefore suggests that the platform may undermoderate when the marginal returns to engagement in advertising revenue are sufficiently high. Since evidence suggests that more engaged users are more likely to convert to sales (Simonov et al., 2025), advertisers may be willing to pay substantially more when their ads are placed alongside high-effort content. In that case, the platform has a stronger incentive to keep moderation low in order to preserve engagement, even when a higher level of moderation would increase user welfare.

We have considered the effects of a policy forcing the platform to implement a truthful equilibrium. But can platforms comply with such a requirement? The following Lemma addresses this.

Lemma 3. *Suppose the platform can commit to the effort-dependent moderation rule*

$$I(e) = \begin{cases} 0 & \text{if } e < e^* \\ \hat{I} & \text{if } e \geq e^*, \end{cases} \tag{10}$$

where $(e^*, \hat{I})$ *satisfy* $v(1-\hat{I}) < e^* < v - \frac{c}{p}$. *Then, under Assumption 3, every equilibrium of the communications game that satisfies the intuitive criterion is payoff-equivalent to one with truthful communication.*

Thus, with a slight enrichment of the space of possible moderation policies, the answer is affirmative.

5 Platform strategies

The analysis so far has focused on a platform's moderation decision. We now focus on the truthful equilibrium to analyze how moderation interacts with other strategic considerations for the firm.[21] First, we ask if a platform has the platform has an incentive to commit to its moderation policy. Second, we consider a platform that shares ad revenues with senders to complement moderation in sustaining communication. Third, we examine a firm's choice between operating as a content-hosting platform or as a more traditional broadcaster which bears all costs of content production. Finally, we study the platform's incentive to enable AI tools for content production.

5.1 Moderator commitment

A key assumption in our analysis is that the platform can commit to its moderation policy, I. But the platform has a conflict of interest because deleting highly engaging but untruthful content sacrifices the ad revenue, $A(e)$, it could have generated. Anticipating this, untruthful senders may be tempted to craft engaging content in the expectation the platform will let it survive. This ultimately destroys the capacity of effort to signal truthfulness and thus undermines communication. More formally, suppose the platform can choose to deviate from the announced value of I after observing s. We have:

Lemma 4. *Suppose the platform cannot commit to I. Then there is no equilibrium with positive sender effort. The platform earns zero profit.*

Because the temptation to renege drives away all engaging content and eliminates platform profit, the platform has an incentive to commit itself to faithful moderation, for example by developing a reputation as a reliable moderator. We can demonstrate this in a repeated version of our model. Consider an infinitely-repeated stage game in which a unit mass of short-lived sender-receiver pairs arrive and play the communication game from the baseline model (with an independently realized state, w, for each pair). The platform is long-lived and has discount rate δ. It announces I prior to the beginning of the game, but can renege on this promise in any period after signals are sent but before they are seen by receivers. The history of the game can be observed publicly, for example because moderation failures are reported in the media.

[21] Similar results could be obtained under alternative equilibrium selection rules.

Proposition 5. *Fix a pre-announced I such that $I \leq \frac{\delta(p+\zeta(1-p))}{\zeta(1-p)}$. There is an equilibrium in which the platform adheres to its pre-announced I each period, while senders and receivers in each stage game play the instrumental equilibrium indexed by ζ (characterized in Proposition 2).*

The platform develops a reputation for deleting false messages because this is essential to attract engaging content. However, the higher are I and ζ, the more tempted the platform becomes to exploit its reputation by tricking receivers into engaging with false content. Thus, commitment problems limit the level of moderation that can be sustained in equilibrium. An infinitely repeated truthful equilibrium can always be sustained by reputation.

5.2 Transfers to senders

On many content-hosting platforms, content creators are able to monetize their content, receiving a fee from the platform. The ability to do so often depends on their adherence to platform content guidelines. For example, on YouTube, content creators face demonetization if they violate platform guidelines.[22] We consider how a platform can simultaneously deploy moderation and transfers to sustain communication.

Consider an environment where the platform makes a transfer, $t(e)$, to any sender whose message is not deleted, with $t(0) \geq 0$ and $t'(e) > 0$. For example, one natural formulation would be $t(e) = \psi A(e)$, where A 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, $\hat{e}$ is the effort level that is closest to the optimal one (e') while still being high enough to credibly signal that $w = 1$. We can show the following:

Proposition 6. *There exists an equilibrium with truthful communication if and only if*

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

[22] `https://support.google.com/youtube/answer/1311392?hl=en`, accessed February 2026.

holds. In all such equilibria, sender effort takes the following form

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

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.

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 was costly then the 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.

The next proposition shows that if transfers take the common form of a share of ad revenue, then the platform-optimal moderation policy must be interior.

Proposition 7. *Suppose the platform gives senders a share of ad revenues,* $t(e) = \psi A(e)$ *if their message is not deleted. Then the platform-optimal moderation policy has* $I < \frac{v}{v+t(e')}$*. In particular, the platform always chooses* $I < 1$*.*

For intuition, recall that reducing moderation induces senders to exert higher effort. This has two opposing effects. Higher effort raises ad revenue per sender, but it can also reduce sender participation by lowering senders' continuation payoffs. Near the boundary at which all senders choose the transfer maximizing effort, however, a small increase in effort has only a second-order effect on sender participation, by an envelope argument, while it has a first-order effect on ad revenue. Hence, the platform can profitably lower moderation to induce effort. By contrast, inducing the same incentives through a higher revenue share directly reduces the platform's retained ad revenue.

5.3 Platform versus broadcast business models

In order to moderate messages the platform must be able to observe w. So, why doesn't the platform just directly inform receivers itself rather than relying on senders to do so? In essence, this is a question of business model. For any given content, a firm can either choose to act as a *content-hosting platform* that facilitates communication between senders and receivers, or to provide that content itself like a traditional *broadcaster*. We show that a firm may prefer the platform business model, even if senders have no informational advantage over the firm.

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

$$\max_{e^*} \{G(p(v - e^*))pA(e^*)\} \qquad \text{such that } e^* = v(1 - I).$$

The key point 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.

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 function is therefore

$$\max_{e} \{A(e) - e\}.$$

The following proposition describes the optimal firm behavior.

Proposition 8. *Suppose G and A are such that there exists an interior $I^* \in (\frac{c}{pv}, 1)$ that solves the platform problem.*

(i) A platform induces higher level of effort than a broadcaster exerts if and only if $\frac{G'(p(v-e^*))p}{G(p(v-e^*))} \leq \frac{1}{A(e^*)}$.

(ii) There exists some $\hat{v}$ such that for any $v > \hat{v}$ the firm prefers a platform model.

(iii) If the firm prefers being a platform, then when G increases in the first-order stochastic dominance sense, it must continue prefer being a platform.

Proposition 8 (i) shows that platforms induce higher effort if and only if the semi-elasticity of moderation on sender participation is small. Recall from the participation-effort trade-off,

that higher moderation allows a platform to induce more sender participation. However, if this effect is too small, few additional senders join the platform, while all senders reduce their effort. Hence, if the semi-elasticity of sender entry cost to moderation is too small, the platform prefers low levels of moderation that induces higher effort.

Proposition 8 (ii) and (iii) provide conditions for when a firm would prefer being a platform. Indeed, when v is sufficiently large, many senders join the platform even at low levels of moderation. Likewise, when there is a first order stochastic dominant shift in G, small levels of moderation are sufficient to induce many senders to join the platform. Together, they show that a firm prefers being a platform when it is able to minimize the participation-effort trade-off.[23]

Conversely, when the participation-effort trade-off is stark, then the platform 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. This trade-off is illustrated with an example in Figure 2.[24]

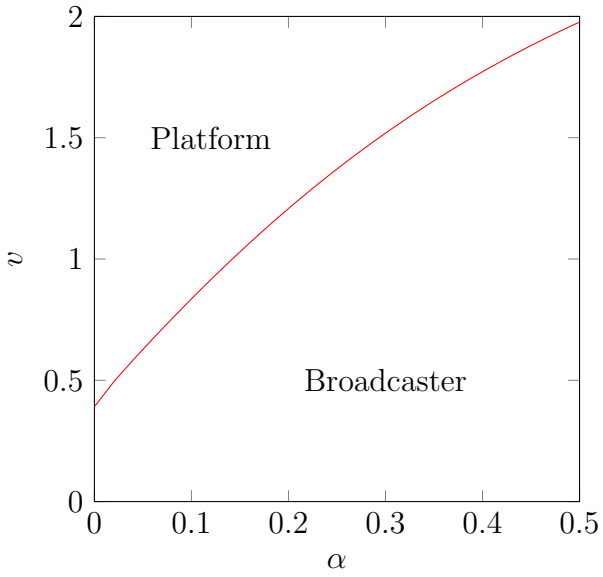

Figure 2: Firm's preferred business model when $G(c) = c^{\alpha}$, $A(e) = \sqrt{e}$, and $p = 0.4$. Note that a smaller α represents a first order stochastic dominant shift in G.

One possible reason for a first order stochastic dominant shift in G is the introduction of

[23] Our analysis focuses on the truthful equilibrium, which is the least profitable for a platform. Considering any other instrumental equilibrium would make the firm even more likely to choose the platform business model.

[24] One can check that with these expressions for G and $A(e)$, the platform would implement $e = v/(1+2\alpha)$, while the broadcaster chooses $e = 1/4$. Substituting these values into the respective profits and comparing the two regimes, a platform is more profitable if $v > \left[\frac{(1+2\alpha)^{\alpha+1/2}}{2^{2+\alpha}p^{\alpha}\alpha^{\alpha}}\right]^{\frac{2}{2\alpha+1}}$.

cost reducing technologies such as generative AI. We now turn to examining the effects of a platform allowing, or even facilitating, AI-generated content.

5.4 AI-generated content

We now consider how the availability of generative AI tools affects content creators' incentives and a platform's moderation strategy. This also allows us to study whether platforms would prefer to embrace or oppose the distribution of AI-generated content. We consider two potential effects of generative AI: (i) it lowers the cost of content creation by automating some processes, but (ii) it is prone to hallucination, which may distort or garble the message.

Accordingly, suppose senders using generative AI can produce content of quality e at cost $k(e)$ where $k'(e) \in (0,1)$ $\forall e > 0$ and $k(0) = 0$.[25] However, with probability $h \in (0,1)$ an AI-generated signal is garbled into unintelligible noise and the receiver learns nothing. We can interpret h as a simple measure of the technological advancement of AI. The sender may choose to use AI or not. And gross of entry costs, the sender's payoff from using AI-generated content is

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

Sender payoffs if she does not use AI are as in the baseline model. We assume that if the sender is indifferent between using AI or not, she uses AI. Besides the sender's choice of whether to use AI or not, everything is as in the baseline model.

The structure of a truthful equilibrium mirrors that in the baseline model. If $w = 0$ then the sender transmits $\underline{s} = (0,0)$ using AI. If $w = 1$ the sender transmits $s^* = (1, e^*)$. The next Lemma addresses when the sender will use AI to generate s^* when lying or telling the truth, and the resulting equilibrium signal s^*.

Lemma 5. *Consider a truthful equilibrium satisfying the intuitive criterion in the subgame following a given I. Then there exists $0 < \underline{h}(I) < \overline{h}(I) < 1$ such that*

1. *If $h > \overline{h}(I)$, AI is not used either if $w = 1$ or in the binding off-path deviation by a dishonest $w = 0$ sender. The equilibrium effort is $e^*(I,h) = v(1-I)$.*

2. *If $\underline{h}(I) < h \leq \overline{h}(I)$, AI is not used if $w = 1$, but a dishonest sender with $w = 0$ would use AI in the relevant off-path deviation. The equilibrium effort is $e^*(I,h) = k^{-1}\big(v(1-I)(1-h)\big)$.*

[25] We model the cost of using AI as increasing in engagement. One interpretation is that producing more engaging AI-generated content requires more sophisticated prompt engineering. Alternatively, AI output may still require additional effort to be turned into engaging content.

3. *If $h \leq \underline{h}(I)$, AI is used in equilibrium by a sender with $w = 1$, and a dishonest sender with $w = 0$ would also use AI in the relevant off-path deviation. The equilibrium effort is $e^*(I, h) = k^{-1}\big(v(1 - I)(1 - h)\big)$.*

The most important insight from Lemma 5 is that $\underline{h}(I) < \overline{h}(I)$. In words: dishonest senders are more willing to use AI than honest senders. This happens because dishonest senders anticipate their false content may be deleted by the moderator, giving them less to lose if the AI hallucinates. Thus, when AI cannot be relied on to produce hallucination-free output, it will be used mostly to create either false content or content of very low quality. Only when AI's reliability improves sufficiently will it be adopted more widely to replace high-quality truthful content generation. This has immediate implications for platform profit and moderation strategy:

Proposition 9. *Fix a moderation intensity I such that AI is not used when $w = 1$ (cases 1 and 2 of Lemma 5). Then:*

1. *I is weakly higher than the level of moderation that would be needed to implement the same level of effort in the absence of AI (strictly so if $\underline{h}(I) < h < \overline{h}(I)$).*

2. *The platform's profit is weakly lower than it could achieve if AI were not available.*

Even if AI is not used to produce payoff-relevant content, it still has an important equilibrium effect by making it less costly to fabricate credible but misleading content. The minimum credible effort level, e^*, must increase to deter such a deviation. Importantly, this increase in e^* harms honest senders who do not benefit from the efficiency savings of AI but must still exert more effort to separate themselves from 'AI slop'. To mitigate this harm, the platform will typically want to offset the effects of AI by moderating more intensively. Notice that if, contrary to our maintained assumption, moderation were costly, this would imply a real economic cost of AI.

Although it may be intuitive that AI leads to higher moderation and lower profit, the mechanism behind Proposition 9 is a bit subtle. The increase in I is not a direct effect of wanting to catch more AI-generated content. Instead, it arises from the fact that AI is disproportionately attractive to dishonest senders, which makes incentive compatibility more difficult to satisfy.

An immediate implication of Proposition 9 is that the platform can benefit from the introduction of AI only if it is used to produce the truthful message s^*. The next result allows the platform to endogenously choose I and verifies that it profits from AI only once the technology is sufficiently reliable. Figure 3 provides an illustration of these results.

Proposition 10. *Consider the truthful equilibrium satisfying the intuitive criterion, with I set to maximize the platform's profit. There exists a $h^* \in (0, \frac{v-k(v)}{v})$ such that AI increases the platform's profit if and only if $h < h^*$. Moreover, if the platform profits from the introduction of AI then it optimally chooses an I such that $h < \underline{h}(I)$.*

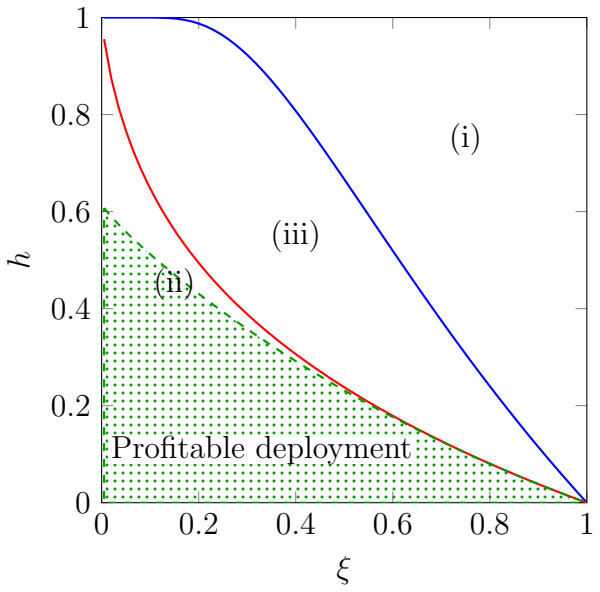

Figure 3: Platform and sender strategies when $G(c) = c^\alpha$ fixing $\alpha = 1$, $A(e) = \sqrt{e}$, $p = 1/3$, $v = 1$, and setting $k(e) = e^{1/\xi}$. The region above the top blue line, (i), reflects where senders never prefer to use AI. The region between the top blue line and middle red line, (iii) reflects where senders only prefer to use AI if $w = 0$. The region below the red line, (ii) reflects where senders always prefer to use AI. The dotted green area represent the region where AI is profitably deployed.

Taken together, the results in this section tell a story of AI's emergence. Initially, the technology is so unreliable that it is used only in payoff-irrelevant zero-effort content. Once the reliability of AI improves enough, it becomes attractive as a way to cheaply generate deceptive content. The platform must respond by moderating more aggressively. Eventually, though, the technology advances to a stage where it can reliably be used to send s^*. The platform then faces a real trade-off between welcoming AI as a cost-saving efficiency, or opposing it as a force that undermines communication. We should expect platforms to switch to actively enabling and encouraging AI as the technology matures.

One challenge of a newly emerging technology like AI is that there are many degrees of freedom in choosing how to model it. We do not intend to claim that the above model is the definitive way to do so, but rather to demonstrate some interesting implications of AI. There are many other interesting variants and extensions of the model and we think the interaction of AI and online content is fruitful area for further investigation.

6 Conclusion

Social media allows ordinary users to communicate, perhaps untruthfully, with the world. We study how content moderation affects communication in environments where content creators have an incentive to persuade rather than inform. Our main result is that creator effort and platform moderation interact in a nontrivial manner. Effort is required to convince users of a creator's truthfulness. Moderation improves the credibility of *all* messages on the platform, thereby encouraging creator participation. But by bearing more of the burden of persuasion, moderation crowds out effort, making content less engaging.

A platform funded by ad revenue values moderation because it attracts creators but also has an incentive to limit moderation to generate engagement. Advertising concerns may also cause a platform to tolerate false but engaging content. A regulator maximizing the surplus of users faces a similar trade-off, but in many cases prefers to enforce a truthfulness and a more rigorous moderation than the platform.

We also show how these forces interact with other strategic concerns of the platform. Moderation and revenue sharing can complement each other by increasing the opportunity cost of misleading content. A firm is more likely to operate as a content-hosting platform rather than a traditional broadcaster when the participation-effort trade-off is less severe. Committing to a moderation policy is necessary to sustain communication. AI-generated content introduces a new tension as it can reduce the cost of producing engaging content, but inadvertently facilitates deception. A platform benefits from facilitating AI only when the technology is sufficiently advanced such that hallucination is limited.

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.

Jann, Ole and Christoph Schottmüller (2024). "Political Debate on Social Media: Theory and Evidence".

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 (2025). *Robust Content Moderation: Theory and Applications*. Tech. rep. Working Paper.

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*.

Liu, Yi, Pinar Yildirim, and Z John Zhang (2022). "Implications of revenue models and technology for content moderation strategies". *Marketing Science* 41.4, pp. 831–847.

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.

Rendo, Iván (2025). "Excessive Content Moderation". *Working Paper*.

Rhodes, Andrew and Chris M Wilson (2018). "False advertising". *The RAND Journal of Economics* 49.2, pp. 348–369.

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

Simonov, Andrey, Tommaso Valletti, and Andre Veiga (2025). "Attention spillovers from news to ads: Evidence from an eye-tracking experiment". *Journal of Marketing Research* 62.2, pp. 294–315.

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

Szydlowski, Martin (2023). *Deprioritizing Content*. 4398140. SSRN.

A Proofs

A.1 Proof of Lemma 1

The proof takes two parts. First we consider the case where $e = 0$, then we look at the case where $I = 0$.

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, $\rho(m) > \rho(m')$.

Suppose to a contradiction that $\rho(\emptyset) \geq \rho(m)$. Then deviating from m' to m causes ρ to increase to $\rho(m)$ if the message is not deleted and to $\rho(\emptyset)$ if it is deleted. This means a deleted m results in the same action as an undeleted m. And there cannot be an equilibrium where m' is on path. Thus, we must have $\rho(\emptyset) \leq \rho(m)$ in any instrumental equilibrium.

Suppose $w = m$. The payoff from m is $\rho(m)v$, whereas the payoff from m' is $\rho(m')v(1-I)$ and the payoff from sending no message is $\rho(\emptyset)v$. Since $\rho(m) > \max\{\rho(\emptyset), \rho(m')\}$, the sender strictly prefers to send message m. Thus, for m' to be on the equilibrium path, it must be sent (only) when $w = m'$, i.e. $\rho(m') = m'$. For m' to be on-path, it must 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, since $m = 1$ is 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 $\beta(\bar{s}) \leq p$ by the law of total probability, meaning, by Assumption 1, $\rho(\bar{s}) = 0$. The sender's payoff from $\bar{s}$ is therefore $-e(\bar{s}) \leq 0$. 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.

□

A.2 Proof of Proposition 1

Proof. 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 truthful equilibrium 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 $-e(s)$. 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

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

while when $w = 1$ the payoff under the same receiver beliefs is $v - \tilde{e}$, which is positive in the candidate $e^* \leq v - \frac{c}{p} \leq v$. Note that $\rho(\emptyset) = 0$ because, on the equilibrium path, $s = \emptyset$ is observed only if no sender entered, in which case $\beta(\emptyset) = p$. 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 $w = 1$. □

A.3 Proof of Proposition 2

We first show a series of Lemmas that help establish the set of equilibria that are instrumental and payoff equivalent to the truthful equilibrium.

The case of $I = 0$ has already been dealt with in Lemma 1, where we showed the equilibrium is necessarily non-instrumental. So focus on cases with $I > 0$.

Lemma 6. *Any instrumental equilibrium satisfying the intuitive criterion and in which senders play pure strategies is payoff equivalent to the truthful equilibrium.*

Proof. Suppose senders use pure strategies, meaning S_i is a singleton. There are two possibilities: First, if $S_0 = S_1$ then the signal is uninformative and we have a non-instrumental equilibrium. Second, suppose $S_0 \neq S_1$ and let $s_0 \in S_0$ and $s_1 \in S_1$ be the strategies of the sender in states $w = 0$ and $w = 1$ respectively. Since $w = 0$ is perfectly revealed by s_0, it must leave the sender with a payoff of zero. Moreover, incentive compatibility requires $v(1-I) - e(s_1) \leq 0 \iff e(s_1) \geq v(1-I) = e^*$. The intuitive criterion selects the least-cost separating effort, $e(s_1) = e^*$. In sum, if $S_0 \neq S_1$ then the state is always perfectly revealed to the receiver and the sender exerts the same effort as in the truthful equilibrium. Thus, the equilibrium is payoff equivalent to the truthful one. □

From now on, allow senders to mix in their choice of signals. First, it is without loss to focus on cases where $I < 1$.

Lemma 7. *When $I = 1$, any equilibrium is payoff equivalent to a truthful equilibrium.*

Proof. Under Assumption 3, the sender always transmits $m = 1$ if $w = 1$. This message alone suffices to reveal the state because untruthful messages are certainly deleted when $I = 1$, so the sender has no reason to exert effort.

A sender with $w = 0$ induces $r = 0$ regardless of which message she transmits, either because the message truthfully reveals $w = 0$, or because the signal is deleted.

Thus, in equilibrium, the sender never exerts effort and never yields a payoff other than $\pi_h(0)$ for the receiver. The sender's payoff is $vw - c$. These are the same payoffs as the truthful equilibrium when $I = 1$. □

Thus, we now focus on cases with $0 < I < 1$. The next Lemma helps to narrow-down the set of equilibria to be considered when $0 < I < 1$.

Lemma 8. *Suppose $I > 0$ and let S_i be the set of signals in the sender's support when $w = i$. Then $|S_0 \cap S_1|$ has at most one element in any instrumental equilibrium.*

Proof. Proceed by contradiction. Suppose there are at least two distinct signals $s, s' \in S_0 \cap S_1$. Under Assumption 3, $m(s) = m(s') = 1$. For s and s' to both be on-path when $w = 1$ the sender must be indifferent: $v\rho(s) - e(s) = v\rho(s') - e(s')$. Similarly, the sender must be

indifferent if $w = 0$: $v(1-I)\rho(s) + vI\rho(\emptyset) - e(s) = v(1-I)\rho(s') + vI\rho(\emptyset) - e(s')$. These two indifference conditions can be rewritten as

$$\frac{1}{v}\left[e(s) - e(s')\right] = \left[\rho(s) - \rho(s')\right] \tag{14}$$

$$\frac{1}{v}\left[e(s) - e(s')\right] = (1-I)\left[\rho(s) - \rho(s')\right]. \tag{15}$$

We know that $e(s) \neq e(s')$ otherwise we would have $s = s'$. Equation (14) therefore requires $\rho(s) \neq \rho(s')$. It follows that both (14) and (15) can be satisfied only if $I = 0$. This establishes that if $I > 0$ then $|S_0 \cap S_1|$ has at most one element. □

Following from Lemma 8, we can now identify the set of potential equilibria of the communication game.

Lemma 9. *Up to payoff equivalence, there are four types of potential equilibria that are partially-informative:*

1. $S_0 = \{s^\dagger, \underline{s}\}, \quad S_1 = \{s^\dagger\},$
2. $S_0 = \{s^\dagger, \underline{s}\}, \quad S_1 = \{s^\dagger, \overline{s}\},$
3. $S_0 = \{s^\dagger\}, \quad S_1 = \{s^\dagger, \overline{s}\},$
4. $S_0 = \{\underline{s}\}, \quad S_1 = \{\overline{s}\},$

where $\underline{s} = (0,0)$ *and* $\overline{s} = (1, e(\overline{s}))$.

Proof. In a partially-informative equilibrium, there is at most one $s^\dagger \in |S_0 \cap S_1|$ by Lemma 8. Any other s on the equilibrium path is therefore transmitted in only one state and fully reveals the state.

It follows that all $s \in S_i \backslash s^\dagger$ has the same effort, $e(s)$. Indeed, if $e(s) < e(s')$ for $s, s' \in S_i \backslash s^\dagger$ then sender i would strictly prefer s. Under Assumption 3, this means $S_1 \subseteq \{s^\dagger, \overline{s}\}$, where $\overline{s} = \{1, e(\overline{s})\}$.

Any $\underline{s} \in S_0 \backslash s^\dagger$ must have $e(\underline{s}) = 0$. Indeed, any s that reveals $w = 0$ leads to $\rho(s) = 0$, so the sender has no incentive to exert effort. Thus, $S_0 \subseteq \{s^\dagger, (0,0), (1,0)\}$. But we can disregard $(1,0)$ without loss because $(0,0)$ and $(1,0)$ both include the same effort and fully reveal w to the receiver and are therefore payoff equivalent. □

Having identified the set of mixed strategies that satisfy payoff equivalence, we can now proceed with the proof of Proposition 2.

Proof of Proposition 2. In any equilibrium, the utility of the sender when $w = 1$ must be at least vI. This is because the sender can guarantee herself vI by choosing $e = v(1-I)+\epsilon$, for $\epsilon > 0$ small. Indeed, by the intuitive criterion the receiver must have belief $\beta(s) = 1$ when $e > v(1-I)$, and therefore plays $\rho(s) = 1$.

Suppose a type 3 equilibrium from Lemma 9 exists. It must be that $\beta(s^\dagger) \leq p < 1/2$ because $s^\dagger$ is sent by all $w = 0$ senders but only some $w = 1$ senders. Thus, $\rho(s^\dagger) = 0$, implying the sender gets a non-positive payoff when $w = 1$, a contradiction.

Suppose the equilibrium is of type 1 or 2. In either case, $\underline{s}$ fully reveals $w = 0$ and therefore yields a sender payoff of zero. This pins down

$$v(1-I)\rho(e(s^\dagger)) + vI\underbrace{\rho(\emptyset)}_{=0} - e(s^\dagger) = 0 \implies e(s^\dagger) = v(1-I)\rho(e(s^\dagger)).$$

In either a type 1 or a type 2 equilibrium, the sender's payoff when $w = 1$ is the payoff she gets from playing $s^\dagger$:

$$v\rho(e(s^\dagger)) - e(s^\dagger) = v\rho(e(s^\dagger)) - v(1-I)\rho(e(s^\dagger)) = v\rho(e(s^\dagger))I.$$

Because the sender's payoff when $w = 1$ must be at least vI, we must have $\rho(e(s^\dagger)) = 1 \implies e(s^\dagger) = v(1-I)$. Since a sender can guarantee $\rho = 1$ with $s^\dagger$, she will never transmit $e > e(s^\dagger)$. This rules out equilibria of type 2 above.

We have established that equilibrium has type 1 or type 4. Letting ζ denote the probability that the sender transmits $s^\dagger$ when $w = 0$, notice that we can treat a type 4 equilibrium as a special case of type 1, with $\zeta = 0$. Thus, we focus on type 1 equilibrium.

We have pinned down the receiver's strategy at $\rho(s^\dagger) = 1$ and $\rho(\underline{s}) = 0$, and have determined the values of $s^\dagger = (1, v(1-I))$ and $\underline{s} \subseteq \{\{0,0\},\{1,0\}\}$ (where $m \in \{0,1\}$ are payoff equivalent). The last component of equilibrium is the value of ζ. This probability must make the receiver willing to play $\rho(s^\dagger) = 1$. We thus require

$$\beta(s^\dagger) = \frac{p}{p + \zeta(1-p)(1-I)} \geq 1/2 \iff \zeta \leq \frac{p}{(1-p)(1-I)}. \qquad \square$$

A.4 Proof of Proposition 4

Proof. Part (i): Given W quasi-concave, $I^{**} = 1$ is optimal if and only if $W'(1) \geq 0$. This immediately yields (8).

Part (ii): At the platform's optimal choice of I, (5) must be satisfied. Evaluating $W'(I^*) > 0$ after making this substitution yields (9). $\square$

A.5 Proof of Lemma 3

Proof. Suppose moderation policy (10) 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 Assumption 3, 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 $w = 1$. □

A.6 Proof of Lemma 4

Proof. Suppose senders and receivers expect untruthful messages to be deleted with probability I and consider an instrumental equilibrium, which must have the form described in Proposition 2. Suppose the platform observes signal $s^* = (1, e^*)$, where $e^* = v(1-I)$. Then the platform

obtains profit zero if it deletes the signal and $A(e^*)$ if it does not delete the signal. There are now two cases. Case (i): If $I < 1$ then $A(e^*) > 0$ and the platform will not delete the signal. Thus, the only $I < 1$ consistent with correct equilibrium expectations is $I = 0$. But $I = 0$ implies $e^* = v$; no senders enter the platform and its profits are zero. Case (ii): $I = 1$. We then have $e^* = 0$ and the result is immediate. □

A.7 Proof of Proposition 5

Proof. We establish grim trigger equilibrium in which, so long as the platform has never deviated from its pre-announced I, senders and receivers play the instrumental equilibrium ζ, with on-path signals $s^* = (1, e^*)$ and $\underline{s} = (0, 0)$. If the platform ever deviates from I then senders and receivers expect all future communication to be non-instrumental, meaning receivers' behavior can no longer be influenced and senders never benefit from entering the platform.

Suppose the platform deviates from its pre-announced I to I' during some period. Then its revenue immediately increases by

$$\underbrace{G(pvI)(1-p)\zeta}_{\#\text{ false messages}} \times (I - I')A(e^*).$$

However, because $I \neq I'$, the grim trigger strategy is triggered. The platform's continuation profit following the deviation is then zero because future senders never enter or exert effort in content creation. It is immediate that the best possible deviation is to $I' = 0$.

If, on the other hand, the platform continues to play the putative equilibrium, its reputation would remain intact and it would expect to earn discounted profits equal to

$$\sum_{t=1}^{\infty} \delta^t G(pvI)[p + (1-p)\zeta(1-I)]A(e^*) = \frac{\delta}{1-\delta} G(pvI)[p + (1-p)\zeta(1-I)]A(e^*).$$

Comparing the current gain in revenues to the future loss, we find that the platform can sustain its reputation if

$$\frac{\delta}{1-\delta} G(pvI)[p + (1-p)\zeta(1-I)]A(e^*) \geq G(pvI)(1-p)\zeta I A(e^*)$$

$$\iff I \leq \frac{\delta(p + \zeta(1-p))}{\zeta(1-p)}.$$

□

A.8 Proof of Proposition 6

Proof. 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 weakly 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.

□

A.9 Proof of Proposition 7

Proof. Define $\bar{I} := \frac{v}{v+t(e')}$. The proof shows that, starting from $I = 1$, the platform has a profitable deviation to $I = \bar{I} - \epsilon$, $\epsilon > 0$ small. We show this for a given ψ. If the platform can also vary ψ then the deviation becomes weakly better.

The platform's profit is

$$\Pi = (1-\psi) \cdot G\Big(p\big(v + t(\hat{e}) - \hat{e}\big) + (1-p)\big(t(e') - e'\big)\Big) \cdot \Big(pA(\hat{e}) + (1-p)A(e')\Big). \tag{16}$$

If $e' = 0$ then $\Pi = 0$ at $I = 1$ and the result is immediate and follows from the base model. Thus, we focus on the case where e' is positive.

For $I \geq \bar{I}$, $\hat{e} = e'$. For $I = \bar{I} - \epsilon$, $\hat{e} > e'$ solves

$$\hat{e} = (v + t(\hat{e}))(1 - I) - [t(e') - e']. \tag{17}$$

By the implicit function theorem, $\frac{d\hat{e}}{dI} < 0$ when (17) holds, so lowering I from 1 to $\bar{I} - \epsilon$ increases effort from e' to $\hat{e}$.

$$\begin{aligned}\frac{d\Pi}{d\hat{e}} = (1-\psi) \cdot G'\Big(p\big(v + t(\hat{e}) - \hat{e}\big) + (1-p)\big(t(e') - e'\big)\Big) \cdot p\big(t'(\hat{e}) - 1\big) \cdot \Big(pA(\hat{e}) + (1-p)A(e')\Big) \\ + (1-\psi) \cdot G\Big(p\big(v + t(\hat{e}) - \hat{e}\big) + (1-p)\big(t(e') - e'\big)\Big) \cdot pA'(\hat{e}).\end{aligned} \tag{18}$$

Evaluating this as $\hat{e} = e'$ means the first term disappears because e' maximizes $t(e) - e$ whenever it is positive, such that $t'(e') - 1 = 0$. What remains is

$$\left.\frac{d\Pi}{d\hat{e}}\right|_{\hat{e}=e'} = (1-\psi) \cdot G\Big(p\big(v + t(e') - e'\big) + (1-p)\big(t(e') - e'\big)\Big) \cdot pA'(e'). \tag{19}$$

This is strictly positive.

Applying the chain rule, $\frac{\partial \Pi}{\partial I} = \frac{\partial \Pi}{\partial \hat{e}} \frac{\partial \hat{e}}{\partial I} < 0$. Hence, the platform strictly gains by setting $I < \bar{I}$.

□

A.10 Proof of Proposition 8

Proof. Under the platform problem, I^* solves

$$G'(pvI^*)pA(v(1 - I^*)) = G(pvI^*)A'(v(1 - I^*)).$$

We can rearrange this to $A'(e^*) = \frac{G'(p(v-e^*))pA(e^*)}{G(p(v-e^*))}$, where $e^* = v(1 - I^*)$.

Under the firm as a broadcaster problem, the optimal e^b that maximizes the firm problem satisfies $A'(e^b) = 1$.

(i) Since $A''(e) \leq 0$, we have

$$e^* \geq e^b \iff A'(e^*) \leq A'(e^b) = 1 \iff \frac{G'(p(v-e^*))p}{G(p(v-e^*))} \leq \frac{1}{A(e^*)}.$$

(ii) There exists some $\hat{v}$ such that for any $v > \hat{v}$ the firm prefers a platform model. To see this, note that the broadcaster profits are independent of v, and in equilibrium the effort exerted by the broadcaster satisfies $A'(e^b) = 1$. We can rewrite the platform profit function as $G(pvI)pA(v(1-I))$. By the envelope theorem, the platform's profit is strictly increasing in v. Therefore, $\hat{v}$ solves $G(p\hat{v}I)pA(\hat{v}(1-I)) = A(e^b) - e^b$.

(iii) Finally, suppose that v is sufficiently large such that the firm prefers a platform business model under G. Consider a first order stochastic dominant shift from G to $\hat{G}$ such that $G(c) < \hat{G}(c)$ for all c. Then $G(pvI^*) < \hat{G}(pvI^*)$. Then for $\hat{I}$ that maximizes the platform profit under $\hat{G}$, it must be that $\hat{G}(pv\hat{I})pA(v(1-\hat{I})) \geq \hat{G}(pvI^*)pA(v(1-I^*)) > G(pvI^*)pA(v(1-I^*))$, which is itself preferred to the broadcaster model by construction. So the firm continues to prefer being a platform. □

A.11 Proofs from Section 5.4

Proof of Lemma 5. Case 1 Fix $I \in (0,1)$. In a truthful equilibrium, a sender with $w = 0$ obtains payoff zero, while a sender with $w = 1$ exerts effort $e > 0$. Because $k(0) = 0$, a sender with $w = 0$ is indifferent between using AI and not using AI to send the zero-effort truthful signal; by assumption she uses AI. The non-trivial incentive constraints are therefore the constraints deterring a $w = 0$ sender from mimicking the high signal. If the deviation is produced without AI, her payoff is $v(1-I) - e$. If the same deviation is produced using AI, her payoff is $v(1-I)(1-h) - k(e)$. Truthful communication requires both expressions to be weakly negative. The intuitive criterion selects the least effort satisfying the relevant binding constraint.

First suppose the AI-assisted mimicking deviation is strictly less attractive than the non-AI mimicking deviation at the no-AI separating effort $e^N(I) := v(1-I)$. This is the case if $v(1-I)(1-h) - k(e^N(I)) < v(1-I) - e^N(I) = 0$, or equivalently

$$h > \overline{h}(I) := 1 - \frac{k(e^N(I))}{e^N(I)} = 1 - \frac{k(v(1-I))}{v(1-I)}.$$

Since $k(0) = 0$ and $0 < k'(e) < 1$, we have $0 < k(e^N(I)) < e^N(I)$, so $\overline{h}(I) \in (0,1)$. In this

region, the least separating effort is $e^N(I) = v(1-I)$ and AI is not used by the mimicking sender. A type-$w = 1$ sender uses AI if and only if $v(1-h) - k(e^N(I)) \geq v - e^N(I)$, which is equivalent to

$$h \leq \frac{e^N(I) - k(e^N(I))}{v} = (1-I)\overline{h}(I).$$

This cannot hold when $h > \overline{h}(I)$. Thus, AI is not used if $w = 1$, and is not used in the binding off-path deviation when $w = 0$, proving case 1.

Cases 2 and 3 Now suppose $h \leq \overline{h}(I)$. At effort $e^N(I)$, the AI-assisted mimicking deviation is weakly profitable. Since an indifferent sender uses AI, the relevant mimicking constraint is the AI-assisted one. The least effort deterring it is such that

$$v(1-I)(1-h) - k(e^A) = 0 \iff e^A(I,h) := k^{-1}\big(v(1-I)(1-h)\big).$$

Because $h \leq \overline{h}(I)$, the AI-assisted mimicking constraint is the binding one. Thus a mimicking $w = 0$ sender would use AI in the relevant off-path deviation.

It remains to determine whether a $w = 1$ sender uses AI on the equilibrium path. Given effort $e^A(I,h)$, on-path AI use when $w = 1$ occurs if and only if $v(1-h) - k(e^A(I,h)) \geq v - e^A(I,h)$, or equivalently

$$F_I(h) := \frac{e^A(I,h) - k(e^A(I,h))}{v} - h \geq 0.$$

At $h = 0$, $F_I(0) > 0$ because $e^A(I,0) > 0$ and $k(e) < e$ for every $e > 0$. At $h = \overline{h}(I)$, we have $e^A(I, \overline{h}(I)) = e^N(I)$, so

$$F_I(\overline{h}(I)) = \frac{e^N(I) - k(e^N(I))}{v} - \overline{h}(I) = -I\overline{h}(I) < 0.$$

Moreover, $F_I'(h) < 0$. Therefore there is a unique $\underline{h}(I) \in (0, \overline{h}(I))$ such that $F_I(\underline{h}(I)) = 0$.

If $h \leq \underline{h}(I)$, then $F_I(h) \geq 0$, so a $w = 1$ sender uses AI to generate the signal s^*. This gives case 3. If $\underline{h}(I) < h \leq \overline{h}(I)$, then $F_I(h) < 0$, so a $w = 1$ sender does not use AI to generate the signal s^*. This gives case 2. □

Proof of Proposition 9. Part 1 When AI is unavailable we have $e = v(1-I) \iff I = I^N(e) := 1 - e/v$. This effort level is the one needed to deter a $w = 0$ sender from mimicking a $w = 1$ sender. Now suppose AI is available but is not used to generate s^* by the sender with $w = 1$. The same deviation to mimicking without using AI must still be deterred, so the condition $I \geq 1 - e/v$ remains necessary. In addition, the type-$w = 0$ sender can mimic the

on-path signal s^* using AI. This deviation is deterred if and only if

$$v(1-I)(1-h)-k(e)\leq 0 \quad \Longleftrightarrow \quad I\geq 1-\frac{k(e)}{v(1-h)}.$$

The least moderation level that implements effort e when AI is available but unused by the sender generating s^* is therefore the maximum of these two lower bounds:

$$I^A(e,h):=\max\left\{I^N(e),1-\frac{k(e)}{v(1-h)}\right\}.$$

This immediately implies $I^A(e,h)\geq I^N(e)$.

The inequality is strict precisely when the AI-assisted mimicking constraint is tighter than the no-AI mimicking constraint, which is exactly when $h<\overline{h}(I)$. Since cases 1 and 2 are the cases in which AI is not used by the sender generating s^*, strictness obtains exactly in the interior of case 2, $\underline{h}(I)<h<\overline{h}(I)$.

Part 2 If AI is not used by the sender generating s^* then the ex ante expected payoff of a sender is $p(v-e^*(I,h))$ and the platform's profit is

$$\Pi^A(e^*)=G(p(v-e^*))pA(e^*).$$

In the absence of AI, the platform's maximised profit is

$$\Phi^N=\max_{e\in[0,v]}G(p(v-e))pA(e).$$

It is immediate that $\Pi^A(e^*)\leq\Phi^N$. □

Proof of Proposition 10. Let $\Pi^N(e):=G(p(v-e))pA(e)$ denote the platform's profit from implementing effort e when AI is unavailable, and let $\Phi^N:=\max_e\Pi^N(e)$ be the corresponding maximised no-AI profit.

By Proposition 9, if the platform chooses an I such that AI is not used when $w=1$, its profit is weakly below Φ^N. Hence AI can strictly raise the platform's maximised profit only if the platform chooses I so that case 3 of Lemma 5 applies. It remains to characterise the best profit attainable in that case.

In case 3, the on-path effort is $e=e^A(I,h)$, so $k(e)=v(1-I)(1-h)$. Equivalently, for a given pair (e,h) the moderation level that implements e in case 3 is

$$I(e,h)=1-\frac{k(e)}{v(1-h)}.$$

$I(e,h)$ is feasible if and only if $k(e) \leq v(1-h)$. Moreover, case 3 requires a $w=1$ sender to weakly prefer using AI:

$$v(1-h)-k(e) \geq v-e \iff h \leq \frac{e-k(e)}{v}.$$

Thus, when the platform chooses I optimally within case 3, it is equivalently choosing effort from the feasible set

$$\mathcal{E}_3(h) := \left\{ e>0 : k(e) \leq v(1-h) \text{ and } h \leq \frac{e-k(e)}{v} \right\}.$$

The set $\mathcal{E}_3(h)$ is non-empty exactly when $h \leq h_v := \frac{v-k(v)}{v}$. For $e \in \mathcal{E}_3(h)$, the platform's case-3 profit is $\Pi^A(e,h) := G\big[p(v(1-h)-k(e))\big]pA(e)$, and the best profit attainable in case 3 is $\Phi^A(h) := \sup_{e \in \mathcal{E}_3(h)} \Pi^A(e,h)$.

Now, we have $\Phi^A(0) > \Phi^N$ (AI increases profit when hallucinations never happen). To see this, let $e^N \in \arg\max_e \Pi^N(e)$. Note that $e^N \in \mathcal{E}_3(0)$. Since $k(e^N) < e^N$, choosing $e = e^N$ when $h=0$ gives

$$\Pi^A(e^N,0) = G[p(v-k(e^N))]pA(e^N) > G[p(v-e^N)]pA(e^N) = \Phi^N.$$

Thus, when hallucination is absent, AI is a pure cost reduction and strictly raises the platform's attainable profit.

Second, at $h=h_v$ the only $e \in \mathcal{E}_3(h_v)$ is $e=v$, so

$$\Phi^A(h_v) = G\big[p(v(1-h_v)-k(v))\big]pA(v) = G(0)pA(v) = 0 < \Phi^N.$$

Finally, $\Phi^A(h)$ is weakly decreasing in h. To see this, note that (1) if $h' > h$, then $\mathcal{E}_3(h') \subseteq \mathcal{E}_3(h)$. And (2) for any fixed feasible effort e,

$$\Pi^A(e,h') = G\big[p(v(1-h')-k(e))\big]pA(e) \leq G\big[p(v(1-h)-k(e))\big]pA(e) = \Pi^A(e,h).$$

Therefore the best case-3 profit is decreasing as hallucination becomes more likely.

Define $h^* := \sup\{h \in [0,h_v] : \Phi^A(h) > \Phi^N\}$. The preceding paragraphs imply that this set is non-empty and bounded away from h_v. Hence $h^* \in (0,h_v)$ and $\Phi^A(h) > \Phi^N \iff h < h^*$. Since the first two cases of Lemma 5 yield profit weakly below Φ^N (Proposition 9), the platform's maximised profit with AI exceeds its no-AI profit if and only if the best case-3 profit exceeds Φ^N. This is exactly the condition $h < h^*$.

It remains to prove the final claim about the platform's choice of moderation. Suppose

the platform profits from the introduction of AI, so that its maximised profit with AI is strictly greater than Φ^N. If a profit-maximising choice of I induced case 1 or case 2 of Lemma 5, then Proposition 9 would then imply that the platform's profit is weakly below Φ^N, a contradiction. Hence any profit-maximising choice of I that strictly raises profit must induce case 3 of Lemma 5, so $h \leq \underline{h}(I)$. If $h = \underline{h}(I)$ a $w = 1$ sender is indifferent between using AI and not using AI, so $v(1-h) - k(e) = v - e$. The platform's profit at this I is therefore

$$G\big[p(v(1-h) - k(e))\big]pA(e) = G[p(v-e)]pA(e) = \Pi^N(e) \leq \Phi^N.$$

This contradicts the supposition that the platform strictly profits from the introduction of AI. □

B Robustness of assumptions

B.1 Informative effort

We now allow effort to directly affect how much information is conveyed by the message. A simple approach to this is to suppose that effort helps ensure the message is properly transmitted and understood, rather than being lost or garbled. Accordingly, suppose that an effort e message is 'understood' by the receiver with probability $\phi(e)$ such that $\phi'(e) > 0$ and $\phi''(e) < 0$. With probability $1 - \phi(e)$ the message is unintelligible and the receiver sticks to his prior. Assume $\max_e\{p(\phi(e)v - e)\} > \underline{c}$, which is a necessary condition for senders to ever want to enter in a truthful equilibrium.

In a truthful equilibrium a sender obtains payoff $-e$ when $w = 0$ and must therefore sets $e = 0$. In order to sustain a truthful equilibrium with some effort, e, when $w = 1$, we must satisfy incentive compatibility,

$$\phi(e)v(1-I) - e \leq 0, \tag{20}$$

and individual rationality,

$$p(\phi(e)v - e) > \underline{c}. \tag{21}$$

Let $e^P(I) := \arg\max_e \phi(e)v(1-I) - e$, and note that $e^P(0) = \arg\max_e \phi(e)v - e$ is the payoff-maximizing effort in a truthful equilibrium.

If $I = 0$ then it impossible to satisfy (20) and (21), so truthful communication depends on a positive level of moderation.

For I small, we have $\phi(e^P(0))v(1-I) - e^P(0) > 0$. Thus, in a truthful equilibrium, senders facing $w = 1$ must distort their effort above their ideal level, $e^* > e^P(0)$, in order to credibly signal the veracity of $m = 1$. The least-cost separating equilibrium is then such that

(20) binds and is decreasing in I. There is again a trade-off between participation and effort. Moreover, a social planner now has an additional incentive to induce high effort because low effort may result in the receiver inefficiently failing to interpret and act on useful information.

One difference to our baseline analysis arises as I becomes large enough. There exists an $\hat{I} > 0$ such that $\phi(e^P(0))v(1-\hat{I}) - e^P(0) = 0$. For any $I \geq \hat{I}$, $e^* = e^P(0)$ satisfies both (20) and (21). Thus, for high levels of moderation, sender effort is motivated by informing rather than persuading the receiver and further moderation no longer crowds out effort.

B.2 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

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

and he plays $r = 1$ if $\beta(s) \geq 1/2$ and $r = 0$ otherwise. Note that $\frac{p(1-G(v-e^*))}{1-pG(v-e^*)} \leq p < 1/2$ 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 (7). 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.3 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 moderation, a correction notice is attached to the message. Hence, following moderation 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 moderation 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 $I > \frac{\underline{c}}{pv}$.

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.