Moderating Content-Hosting Platforms*

Robin Ng† and Greg Taylor‡

September 26, 2026

Abstract

We study how content moderation facilitates communication on online platforms. A sender transmits information to a receiver, exerting effort to signal her 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 the engagement value of content on the platform. The socially optimal policy may therefore involve limited moderation. Nevertheless, the platform sometimes chooses to undermoderate to preserve engagement. We study the interaction between moderation and other strategic platform choices such as 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 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. Examples of such misleading content include undisclosed sponsored product promotions, fake online reviews, deceptive health or financial advice, and political misinformation. This has become a significant policy issue. Regulators like the Federal Trade Commission and European Commission have

*We are grateful to Heski Bar-Isaac, Felix Chopra, Ole Jann, Rafael Jiménez-Durán, Ester Manna, 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

imposed new compliance requirements on platforms to prevent deceptive product promotion by creators or incentivized online reviews.[1] Meanwhile, the role of misinformation in hotly debated topics such as politics, health, or climate change is also an active subject 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 information they deem to be false or harmful, 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 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 containing a message about that state. The sender can also exert effort to make that content more *engaging*, meaning it is more enjoyable to consume and better able to hold a receiver's attention. 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, such as buying a sponsor's product or voting a particular way, regardless of whether it is in the receiver's best interests.

We say that *communication* takes place if the sender joins the platform and transmits a (semi-)separating signal that contains at least some information about the state. Conversely, no communication takes place if the sender does not join or if she transmits an uninformative signal. When a signal is uninformative, the receiver may still learn from the (non-) presence of a message.[3] To facilitate communication, a platform can moderate content by deleting a fraction of content whose message is untruthful.

As a benchmark, we show that moderation or sender effort in isolation does little to enable communication. If a sender cannot invest in making content engaging then communication can affect the receiver's action only in the knife-edge case of perfect moderation, where every piece of content is moderated without error. Conversely, without moderation, communication can only be sustained if a sender dissipates her entire surplus through effort, leaving her with

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

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

[3] There are also pooling equilibria where the platform moderates with enough intensity that the receiver will trust messages even though they contain no information. The platform then becomes a pure certifier rather than a communications medium.

no incentive to join the platform in the first place.

The picture changes substantially when moderation and sender effort are paired together. Moderation makes effort a stronger signal of truthfulness because a sender is reluctant to exert effort on crafting engaging but false content that will probably be deleted. This means that a sender who is reporting truthfully finds it less costly to persuade the receiver of her 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, the sender responds by exerting less effort in equilibrium. Therefore, when the receiver values engaging content, moderation creates a trade-off between inducing sender participation and preserving receiver engagement.

A platform driven by ad revenue must get both the sender and the receiver on board. Because the receiver pays more attention to engaging content, such content is able to attract higher ad revenue. In the truthful equilibrium, the platform optimally chooses some, but not perfect, moderation: just enough to encourage sender participation, but not enough to crowd out engagement that generates ad revenue. If a platform can also select among non-truthful equilibria, it optimally chooses to allow some false but engaging content on the platform. With the right level of moderation, the receiver can be made to trust this false content, which can then be monetized with ads.

We then turn to a regulator concerned with user (sender and receiver) surplus. The regulator faces the same participation-effort trade-off as the platform. Consequently, in low-stakes environments where the receiver benefits more from being entertained than from learning the truth, the regulator's preferred level of moderation may be relatively low, even if moderation is completely costless. Nevertheless, we show that the regulator often prefers more moderation than the platform because the platform does not fully internalize the receiver's benefit from learning the truth.

Finally, we consider moderation in the wider context of platform strategy. We start with two modest institutional departures from the baseline model. We show a platform has an incentive to maintain a reputation as a reliable moderator when it faces a commitment problem, and that it has an incentive to limit moderation when it shares part of its revenue with the sender. Next we turn to two more substantive strategic questions. We examine a firm's decision between operating as a content-hosting platform or as a traditional broadcaster. Being a platform means that the firm gets user-generated content for free rather than having to produce it itself. But a platform must manage senders' participation-effort trade-off. When inducing participation is easy, the platform model becomes more attractive, whereas if it is difficult to simultaneously induce the sender to enter and exert high effort, direct first-party content production becomes more appealing. Lastly, we study the implications of

AI-generated content. We model AI as lowering the cost of producing engaging content while also being prone to hallucination. Because false content may already be deleted, the sender has less to lose from hallucination when her message is untruthful and is therefore more willing to use AI when lying. This makes truthful communication more difficult to sustain and can require more moderation. AI can raise platform profit only when it is sufficiently reliable that the sender also uses it to transmit truthful content, allowing its cost-saving benefit to outweigh the potential losses caused by hallucination.

In sum, the paper yields a number of empirical predictions and policy implications. As barriers to entry fall for creators, 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 if such content is posted to the platform. 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 engaging content.

The rest of the paper is structured as follows. After reviewing the literature, Section 2 describes our setting and Section 3 characterizes equilibrium and analyzes platform moderation. Section 4 studies the optimal regulation of a platform's moderation policy. Section 5 examines platform commitment, the role of transfers from the platform to the sender, a firm's choice between platform and broadcaster business models, and the implications of AI-generated content. Section 7 concludes. Omitted proofs are in Appendix A, and the Online Appendix provides additional results.

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) show that limiting content sharing can improve overall information quality. In a dynamic information disclosure setting, 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) describe 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) show that a platform's revenue model can affect its moderation strategy, and discuss how imperfect moderation technology can lead a platform to allow extreme content. Our paper addresses a similar question to Dwork et al. (2024), who study how content moderation can increase user participation on a platform. Mostagir and Siderius (2022) and Acemoglu et al. (2024) show that 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 a sender's effort to decrease, suggesting that imposing high levels of moderation onto platforms could lead to less engaging content. Moreover, in Bedre Defolie et al. (2026), a platform finds it optimal to host some bad sellers so good sellers are incentivized to exert higher effort to become certified. We also find that a platform has an incentive to tolerate misinformation. In our setting, creators providing misinformation exert high effort to mimic truthful signals and this drives advertising revenue. 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 communication between users as a signaling game, showing how moderation can sustain truthful communication, and studying the platform's incentive to tolerate false but engaging content.

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 (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 a third-party platform that conducts moderation and its commitment to a moderation rule at the beginning of the game rather than as a response to the observed signal. A key departure from the literature is that, in our setting, the effort required to induce a truthful equilibrium is endogenous and depends on the platform's moderation strategy.

Because our platform's moderation rule affects which signals reach the receiver, our analysis relates to information design (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2019), with the distinction that the resulting information structure is generated endogenously by sender entry, message choice, and effort rather than directly chosen by the platform.

By considering a benevolent regulator, we also relate to the branch on *mediated communication* following Myerson (1982) and Forges (1985) (see also Arieli et al., 2023; Ganguly and Ray, 2023; Ivanov, 2014; Salamanca, 2021; Myerson, 1991). 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; Krishna and Morgan, 2004; 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 the sender 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) study a communications game where competing senders can either provide signals informative of their own quality or misinform receivers of their opponent's quality, and characterize when higher quality senders choose to employ misinformation. We approach misinformation differently. In our model, false content can arise endogenously in non-truthful equilibria, and a platform is willing to tolerate such content because false but engaging 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; Henry et al., 2022; Horta Ribeiro et al., 2023), there is limited empirical evidence on the behavior of content creators. For example, Beknazar-Yuzbashev et al. (2025) and Andres and Slivko (2021) show that the number of toxic posts on platforms decreases following the use of moderation; Jiménez-Durán (2023) finds a negligible effect on creator participation. We provide one possible mechanism through which content moderation can promote the participation of content creators. Consistent with our findings, Matias (2019) shows in a field experiment on Reddit that raising awareness of

existing moderation rules encouraged newcomer participation. We examine how a platform's moderation and revenue sharing strategies jointly affect creators' participation and effort incentives. This is relevant for platforms such as YouTube, where Abou El-Komboz et al. (2023) find that more stringent partner program rules led creators to produce less content and lowered content quality. Our model also speaks to the accuracy and transparency of moderation, such as Meta's Cross-Check program.[4] We show that accurate moderation can lower the extent of moderation required to sustain truthful communication, and highlight the value of commitment to announced moderation policies for encouraging creator participation and effort. More broadly, our framework helps interpret empirical evidence by distinguishing creator participation from investment in engaging content, and highlights how effort and moderation can work in opposite directions. This helps for understanding the indirect consequences of moderation. Not only do they affect the content being removed, but also the amount and engagement value of remaining content.

2 Model setup

Players and strategies We consider a game of incomplete information with three players: a *sender*, a *receiver*, and a *platform*. At the beginning of the game, the platform publicly and costlessly commits to a moderation rule, $I \in [0,1]$, which can be understood as the probability with which content containing 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, $w \in \{0,1\}$, where $\Pr(w=1) = p$. She then chooses a signal $s = (m, e)$, comprising a message $m \in \{0,1\}$ and an effort $e \in \mathbb{R}_+$ exerted to create engaging content containing the message. 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

[4]See `https://transparency.meta.com/enforcement/detecting-violations/reviewing-high-visibility-content-accurately/` accessed 1 September 2026.

the post-moderation signal, $s \in \mathbb{S} \cup \{\emptyset\}$, and forms posterior belief $\beta(s) = \Pr(w = 1|s)$. The receiver takes an action $r \in \{0, 1\}$. Denote a generic receiver strategy by $\rho(s) = \Pr(r = 1|s)$. If $\beta(s) = 1/2$, we use the tie-breaking rule $\rho(s) = 1$.[5]

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 $\pi_h''(e) \leq 0$.[6] 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. One example satisfying these assumptions is $\pi_h(e) = q(e) + b$ and $\pi_l(e) = q(e) - d$, where $q(e)$ is the utility from consuming content with engagement level e, and $b, d > 0$ are the benefit and damage from taking the correct and incorrect action respectively.

The sender has state-independent preferences: she seeks to persuade the receiver to play $r = 1$ in order to obtain some exogenous benefit v. Hence, her payoff (gross of entry cost, c) is

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

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 content; we assume he does so in increasing relation to how engaging the content is, e. Formally, let $A(e)$ represent the ad revenue generated by the attention of the receiver, with $A(0) = 0$, $A(e) > 0$ for every $e > 0$, $A'(e) \geq 0$, and $A''(e) < 0$.[7]

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

1. The platform publicly announces I.

[5]This tie-breaking convention is immaterial for the truthful equilibrium. In semi-separating and pooling cases, it ensures that the platform can attain, rather than merely approach, its preferred equilibrium payoff by including the relevant boundary in its choice set.

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

[7]A can easily be microfounded. For instance, suppose that the receiver gets a flow utility of $\sqrt{q(e)A}$ from spending A units of time ("attention") consuming content with engagement value $q(e)$, with $q''(e) < 0$ and a unit opportunity cost of time. Then the receiver maximizes $\sqrt{q(e)A} - A$, implying $A = q(e)/4$.

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 signal s.

6. The receiver updates his beliefs based on the signal and takes action r.

We search for equilibria where *communication* occurs, that is, the sender joins the platform with positive probability and transmits a (semi-)separating signal that contains at least some information about the state. Our solution concept is perfect Bayesian equilibrium satisfying the intuitive criterion under the following assumptions.

Assumption 1. $0 < p < 1/2$.

This assumption implies 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 $r = 1$. Under the converse, $p \geq 1/2$, the receiver chooses $r = 1$ at his priors, so there is no incentive for the sender to enter.

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

The assumption that $pv > \underline{c}$ ensures sender entry with some positive probability if there is a way to convince the receiver of the truth without effort. This is a necessary condition to sustain communication in equilibrium. Assuming $\overline{c} > v$ ensures that non-entry occurs with some positive probability, which allows $\beta(\emptyset)$ to be pinned-down by Bayes' rule in any equilibrium.

Applications To help anchor the model, we briefly outline some potential applications.

In a first application, 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 policies 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 a scam and the platform removes some fraudulent content.

In a second application, 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.

Third, in our leading interpretation, moderation is used to remove false or deceptive content. But the underlying mechanism can also apply when moderation targets content deemed intrinsically harmful to the receiver, such as hate speech, abusive content, or disturbing imagery. In that case, the harm to the receiver comes directly from viewing the content rather than from being persuaded by the content to take an inappropriate action. Because it requires a slight reinterpretation of the model, we present the mapping to this application in Part II of the Online Appendix, where we show that the same moderation–effort mechanism studied below arises.

3 Equilibrium

We first show that without either moderation or sender effort, it is difficult to establish communication. Second, we show how moderation and sender effort work together to support truthful communication. We then characterize all active-platform equilibria (i.e., equilibria in which the sender enters with non-zero probability). Lastly, we turn our attention to the problem of equilibrium selection and the platform-optimal moderation strategy.

3.1 Two benchmarks

Before studying how moderation and sender effort work together to support separating communication, we show that in isolation they do relatively little to enable communication. Proofs are in Appendix A.

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

The intuition for the two cases is slightly different. Suppose there is no moderation, $I = 0$. Then, conditional on entry, 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. As the proof shows, any separating equilibrium with positive entry must induce different receiver actions after different messages. Regardless of the state, the sender then prefers a message that induces $r = 1$. 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.[8] In practice, it is likely that moderating every message ($I = 1$) on a large online platform will be prohibitively costly.[9]

3.2 Truthful communication when costly messages are moderated

The picture changes substantially if we allow both effort and moderation, $e \geq 0$ and $I > 0$. To build intuition, it is helpful to first construct a truthful equilibrium.

Lemma 2. *Suppose $I > 0$. For some $e^* \in [0, v)$, there exists an equilibrium 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.[10] In order to sustain truthful communication, we need the sender to

[8] 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.

[9] 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), separating communication is possible only if the platform moderates error-free.

[10] 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.

prefer setting $s = (0, 0)$ when $w = 0$.[11] The incentive compatibility constraint is

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

When $I = 0$, the signal conveyed by effort is weak because the sender's payoff is identical regardless of the state. Communication is possible only if the sender dissipates her entire surplus, as discussed in 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 deleted whereas high-effort truthful messages are never deleted. 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 $v(1 - I)$ and correspondingly increasing the expected sender utility from $-c$ to $p[v - e^*] - c$. The sender enters if her participation constraint is satisfied,

$$e^* \leq v - \frac{c}{p}. \tag{PC}$$

Together, (IC) and (PC) pin down the set of e^* and I that can be sustained in a truthful equilibrium with positive probability of sender entry. For this set to be non-empty we need $v(1 - I) < v - \frac{c}{p}$, or

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

When (3) is satisfied, any $e \in [e^*, v - \frac{c}{p})$ is compatible with a truthful equilibrium with a positive probability of sender entry. The intuitive criterion selects the least-cost equilibrium, $e = e^*$. Moderation and sender effort, though both needed for communication, are substitutes in the truthful equilibrium because 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 a truthful sender to establish her credibility with less effort. In this sense, moderation crowds out sender effort, which, we will see, has important implications for the optimal level of moderation.

3.3 Equilibrium signals

We now provide a more general characterization of equilibria in which the platform is active. The following definition will be useful:

[11] 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 (PC).

Definition 1. *Let $\mathcal{S}(\zeta, e)$ denote the following profile of strategies and beliefs:*

1. *if $w = 1$ the sender transmits $s^* := (1, e)$.*
2. *if $w = 0$ the sender transmits s^* with probability ζ, and $s_0 = (0, 0)$ otherwise.*
3. *the receiver updates using Bayes' rule, plays $r = 1$ following s^* and plays $r = 0$ following s_0 or $\emptyset$.*
4. *off-path beliefs are pessimistic: $\beta(s) < 1/2$ for any $s \notin \{\emptyset, s_0, s^*\}$.*[12]

In words, $\mathcal{S}(\zeta, e)$ describes a profile of strategies where a higher ζ corresponds to a higher probability that the sender lies when $w = 0$. Given $\mathcal{S}(\zeta, e)$, the receiver's posterior satisfies

$$\beta(s^*) = \frac{p}{p + (1-p)\zeta(1-I)} \geq \frac{1}{2} \iff \zeta \leq \bar{\zeta} := \min\left\{1, \frac{p}{(1-p)(1-I)}\right\}. \tag{4}$$

Intuitively, as long as false messages are sufficiently rare (ζ low enough) and are deleted sufficiently often (I high enough), the receiver remains willing to trust equilibrium messages even if he knows the sender sometimes lies. Once the receiver is willing to trust messages, we can make the sender indifferent between s_0 and s^* when $w = 0$ by setting $e = e^* = v(1 - I)$. As in the truthful equilibrium constructed in Section 3.2, the equilibrium effort level is decreasing in I.

Proposition 1. *Fix any $0 \leq \zeta \leq \bar{\zeta}$. The profile $\mathcal{S}(\zeta, e^*)$ describes an equilibrium conditional on sender entry. Sender gross surplus is pvI, and the platform is active if and only if $I > \underline{I}$.*

When $\zeta < 1$, Proposition 1 describes (semi-)separating equilibria in which the sender's signal contains information about the state. Notice that the truthful equilibrium we constructed in Section 3.2 is the special case $\mathcal{S}(0, e^*)$. If, on the other hand,

$$\bar{\zeta} = 1 \iff I \geq \bar{I} := \frac{1-2p}{1-p},$$

we can sustain a pooling equilibrium with $\zeta = 1$. Indeed, once moderation is sufficiently intensive, the mere fact that a message wasn't deleted becomes sufficiently good news to induce receiver trust, even when the sender regularly lies. In such a pooling equilibrium the sender's signal contains no information. The platform therefore ceases to operate as a medium for communication and instead acts as a *pure certifier*. Moreover, because pooling

[12] $\beta(s^*)$ is given in (4). Since s_0 is transmitted only if $w = 0$, $\beta(s_0) = 0$. Lastly, a null signal arises following either non-entry or deletion of a message. Thus, letting N denote the probability of sender entry, $\beta(\emptyset) = \frac{p(1-N)}{(1-N)+N(1-p)\zeta I} \leq p$.

no longer requires separation of types, we can sustain pooling equilibria at effort levels below e^*, as the following result establishes.

Proposition 2. *The profile $\mathcal{S}(1, e)$ describes a (pooling) equilibrium conditional on sender entry for any $I \geq \bar{I}$ and $e \in [0, e^*]$. The sender's gross payoff is $pv + (1-p)(1-I)v - e$ and the platform is active if and only if $I < \frac{v-e-c}{(1-p)v}$.*

We now show that, subject to a mild additional assumption, Propositions 1 and 2 exhaust the set of equilibrium outcomes that involve an active platform and satisfy the intuitive criterion (up to payoff equivalence).[13]

Proposition 3. *Up to payoff equivalence, Propositions 1 and 2 exhaust the set of active-platform equilibria if the sender transmits $m = 1$ whenever $w = 1$. Moreover, the platform is active in every equilibrium if $I > \underline{I}$.*

The condition that $m = 1$ when $w = 1$ requires that the sender tells the truth when it is in her favor. This rules out counter-intuitive reverse-signaling equilibria where the sender plays $m = 0$ when $w = 1$ and the receiver interprets $m = 0$ as meaning $w = 1$. With such reverse signaling, null signals could be treated as good news because the sender lies when $w = 1$ and gets her message deleted.

To summarize, ζ indexes three types of equilibria. Truthful ($\zeta = 0$) and semi-separating ($\zeta \in (0, 1)$) equilibria have $e = e^* = v(1 - I)$. Pooling equilibria ($\zeta = 1$) have $e \in [0, e^*]$.

3.4 Profit-maximizing moderation strategy

We now turn to the question of how the platform will choose its moderation rule. As a first step, we observe the following:

Proposition 4. *Full moderation, $I = 1$, is weakly dominated by every $I < 1$, and strictly dominated by $I \in (\underline{I}, \bar{I})$ if $\underline{I} < \bar{I}$.*

Intuitively, under full moderation, the sender will not exert effort because she is certain that the platform will certify her truthfulness. Thus, $I = 1$ yields zero profit because $A(0) = 0$. Suppose the platform deviates to $I \in (\underline{I}, \bar{I})$. Then pooling equilibria are infeasible but there exists an active-platform separating equilibrium. The selected equilibrium therefore has $e = e^* > 0$ and yields positive profit.

We now consider two specific ways to select equilibrium. First, we select the equilibrium that is optimal from the platform's point of view. Second, we consider the truthful equilibrium, which we show to be uniquely robust.

[13] We say "up to payoff equivalence" because, in any separating 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 these possibilities lead to $r = 0$.

3.4.1 Platform-optimal equilibrium selection

Suppose, for a given I, we select the equilibrium that maximizes the platform's profit.

First, if $I < \bar{I}$ then pooling is infeasible and a (semi-)separating equilibrium of the form described in Proposition 1 is selected, $\mathcal{S}(\zeta, e^*)$. Platform profit can then be written as

$$\begin{aligned}\Pi &= G(p(v-e^*))\Pr(e=e^*)A(e^*) \\ &= G(pvI)\underbrace{[p+(1-p)\zeta(1-I)]}_{Z(I)}A(v(1-I)).\end{aligned} \tag{5}$$

This expression comprises the probability of sender entry (G), the likelihood with which the sender successfully transmits engaging content (Z), and the resulting unit ad revenue from receiver attention (A).

Notice that Π is increasing in ζ, so, among non-pooling equilibria the platform-optimal equilibrium is $\mathcal{S}(\bar{\zeta}, e^*)$, yielding $Z(I) = 2p$.[14] Making this substitution, we have

$$\frac{\partial \Pi}{\partial I} = Z(I)v\left[\underbrace{G'(pvI)pA(v(1-I))}_{\text{participation effect}} - \underbrace{G(pvI)A'(v(1-I))}_{\text{crowding-out effect}}\right]. \tag{6}$$

Equation 6 can be interpreted as reflecting the usual trade-off for a two-sided platform that needs to get both content producers and audiences on board. The (positive) *participation effect* tells us that higher I makes it less costly for the sender to influence the receiver, making her more inclined to enter the platform. Conversely, the *crowding-out* effect arises because moderation acts as a substitute for sender effort. A higher I reduces the equilibrium e^*, and thus the supply of receiver attention. This effect induces the platform to limit the intensity of moderation and force the sender to work harder to establish her credibility.[15] Setting (6) equal to zero yields I^* as the unconstrained optimal moderation strategy conditional on $\mathcal{S}(\zeta, e^*)$. Since G is log-concave and A is strictly concave, Π is quasi-concave and I^* is unique. If $I^* < \bar{I}$ then it becomes a candidate solution to the platform's maximization problem. If $I^* \geq \bar{I}$ then profit is non-decreasing over $I \in [0, \bar{I}]$, which brings us to the second case.

Second, suppose $I \geq \bar{I}$, implying $\bar{\zeta} = 1$. From (5), the platform's profit is increasing in ζ. Thus, its preferred equilibrium is one of pooling, $\mathcal{S}(1, e)$. Holding e fixed, a reduction in I is

[14] If A were dependent on truthfulness, e.g. because advertisers don't like appearing beside false content, then the platform's incentive to increase the amount of false content diminishes. It may no longer prefer the highest possible ζ.

[15] It is easy to find examples in which I^* is very low indeed. For instance, 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 platform's optimal level of moderation satisfies $I^* \to 0$.

profitable because it increases sender entry and allows more content to survive. Thus, the platform-optimal moderation is $I = \bar{I}$, the minimum level needed to guarantee credibility of pooled messages. Recall, however, that this corresponds to pure certification rather than communication.

In sum, the platform chooses between a pooling equilibrium with $I = \bar{I}$ and, if $I^* < \bar{I}$, a separating equilibrium with $I = I^*$. Either option can be optimal, depending on the shape of G and A.

Proposition 5. *Suppose the equilibrium is selected to maximize platform profit for each I. The platform-optimal moderation policy satisfies $I \in \{I^*, \bar{I}\}$ such that $I \leq \bar{I}$ and I^* is the unique solution to*

$$\frac{pG'(pvI^*)}{G(pvI^*)} = \frac{A'(v(1-I^*))}{A(v(1-I^*))}. \tag{7}$$

Moreover, the platform's preferred equilibrium is one where false content is transmitted with positive probability, $\zeta > 0$.

Proposition 5 implies that the platform chooses to allow some false content to survive on the platform, even though we have assumed it could remove such content without cost. Indeed, following entry the receiver sees false content with probability $(1-p)\zeta(1-I) = p > 0$.

3.4.2 Truthful and robust equilibrium

We now determine the platform's optimal moderation strategy under the alternative rule that selects the truthful equilibrium, $\mathcal{S}(0, e^*)$. Besides being potentially focal, the truthful equilibrium is uniquely robust in the following sense. Suppose we perturb the game such that, with probability $\varepsilon > 0$, the receiver is *incredulous*: he will play action $r = 1$ only if he is certain it is correct—i.e., only if $\beta(s) = 1$. With probability $1 - \varepsilon$ the receiver behaves as in the baseline game. The receiver's type is private knowledge. An equilibrium outcome of our main model is *robust* if, as $\varepsilon \to 0$, it can be approached by equilibria of this perturbed game satisfying the intuitive criterion. Proposition 6 shows that only the truthful equilibrium is robust (up to payoff equivalence). Intuitively, starting from a non-truthful equilibrium, there is an arbitrarily low-cost way for the $w = 1$ sender to credibly persuade an incredulous receiver that $m = 1$ is truthful, yielding a profitable deviation even as ε approaches zero.

Proposition 6. *Fix $I \in [0, 1]$. In every robust active-platform equilibrium, the sender transmits $m = 1$ with probability 1 when $w = 1$. All such equilibria are payoff-equivalent to the truthful equilibrium.*

Conditional on selecting the truthful equilibrium, platform profit is given by (5), with $Z(I) = p$ for any I. Thus, the platform trades off the participation and crowding-out effects

in exactly the same way as it did in (6) because the relevant terms in square brackets are unchanged. We therefore obtain the following result immediately:

Corollary 1. *In the truthful equilibrium, $\mathcal{S}(0, e^*)$, the platform's optimal moderation strategy is I^*, given by* (7)*. In particular, the level of moderation is interior, $I^* < 1$.*

In the remainder of the paper we focus on the truthful equilibrium both because of its robustness, and also because it simplifies analysis by eliminating mixed strategies. The Online Appendix shows that qualitatively similar results are obtained using platform-optimal equilibrium selection.

4 Regulating the platform

We now turn to the question of whether the platform's choice of moderation policy is consistent with the interests of its users (content creators and viewers), or whether a regulator may wish to intervene. We consider a regulator that can impose a moderation policy, I, and seeks to maximize either the joint surplus of the sender and receiver, or the receiver surplus alone. Accordingly let $\mathbf{1}_S$ be an indicator function that takes value 1 if the regulator internalizes the sender's surplus. The welfare of the platform's users is

$$W(I) = G(pvI)\left[\mathbf{1}_S pvI + p\pi_h(v(1-I)) + (1-p)\pi_h(0)\right] - \mathbf{1}_S \int_{\underline{c}}^{pvI} c\, dG(c) + \left[1 - G(pvI)\right]\left[p\pi_l(0) + (1-p)\pi_h(0)\right]. \quad (8)$$

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.

Under our maintained assumptions $W(I)$ is quasi-concave, meaning there is a unique optimal $I \in (\underline{I}, 1]$ and the regulator's optimum, when interior, solves

$$W'(I^{**}) = pv\left[\underbrace{G'(pvI^{**})p\left[\pi_h(v(1-I^{**})) - \pi_l(0)\right]}_{\text{participation effect}} - \underbrace{G(pvI^{**})\left[\pi_h'(v(1-I^{**})) - \mathbf{1}_S\right]}_{\text{crowding-out effect}}\right] = 0, \quad (9)$$

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 higher probability of sender entry, while the crowding-out effect arises because moderation crowds out sender 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 ($\pi_h' > \mathbf{1}_S$), 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 7. *In the truthful equilibrium:*

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

$$\frac{G'(pv)}{G(pv)}p < \frac{\pi_h'(0) - \mathbf{1}_S}{\pi_h(0) - \pi_l(0)}. \tag{10}$$

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

$$\frac{A'(v(1-I^*))}{A(v(1-I^*))} > \frac{\pi_h'(v(1-I^*)) - \mathbf{1}_S}{\pi_h(v(1-I^*)) - \pi_l(0)}. \tag{11}$$

The left-hand side of (10) 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 7 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 far below the maximum feasible level when little moderation is sufficient to induce a high probability of sender entry.

The second part of Proposition 7 says the platform undermoderates when ad revenues are more sensitive to engagement or when the receiver cares 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 the receiver would benefit from higher moderation that results in a large supply of informative messages. If the receiver cares *only* about learning the truth (i.e., if $\pi_h'(e) = 0$) then the regulator would choose $I = 1$ and the platform always implements too little moderation. Thus, even in the 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 focused on a regulator that maximizes the surplus of platform users. Whenever (11) is satisfied we have $I^* < I^{**}$. Incorporating the platform's profits into the regulator's objective would then cause the regulator to choose an even lower I. Our result that the regulator may choose an $I < 1$ would then be further strengthened.[16]

5 Platform strategies

The analysis so far has focused on a platform's moderation decision. We now analyze how moderation interacts with other strategic considerations for the platform. We do this in two layers. First, Sections 5.1 and 5.2 establish the robustness of the mechanism in slightly modified institutional environments. We show how the platform responds to a lack of commitment power and how it moderates when sharing revenue with the sender. We then expand the scope and turn to more substantive strategic questions. In Section 5.3 we examine a firm's choice between operating as a content-hosting platform or as a more traditional broadcaster that bears all costs of content production. Finally, in Section 5.4 we study the platform's incentive to enable AI tools for content production.

5.1 Platform 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, a sender transmitting an untruthful message 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 3. *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 demonstrate this in a repeated version of our model: Consider an infinitely-repeated stage game in which a new short-lived

[16]We do not separately include advertiser surplus because advertiser demand is not explicitly modeled.

sender-receiver pair arrives in each period. Each pair plays the communication game from the baseline model. The state w is independently realized for each pair in each period. The platform is long-lived and has a discount factor of 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. At the end of each period, a publicly visible audit verifies that the platform complied with its pre-announced I.

Assuming moderation outcomes are ex post observable is relatively mild. First, large platforms, such as YouTube, are subject to regular (twice annually) transparency reports under the Digital Services Act. These reports include statistics on moderation activities ranging from content reported by authorities to content that is automatically removed. These reports are public. Together, these reports and (quiet) changes to moderation policies are often publicized by the popular media.[17]

Proposition 8. *Fix a pre-announced $I \in (\underline{I}, 1]$. If $\delta \geq \frac{1}{1+pG(pvI)}$, there is an equilibrium of the repeated game in which the platform adheres to its pre-announced I, while senders and receivers play the truthful equilibrium, $\mathcal{S}(0, e^*)$, in every stage game.*

The platform develops a reputation for deleting false messages because this is essential to attract engaging content. Thus, reputation can sustain the announced policy if the platform is sufficiently patient that the value of continuation payoffs to the platform is at least as large as an immediate gain from monetizing false content.

5.2 Transfers to the sender

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.[18] We consider how a platform can simultaneously deploy moderation and transfers to sustain communication.

Consider a setting where the platform shares a fraction $\psi \in [0, 1)$ of its ad revenue with the sender if she enters and her message is not deleted. In our leading application, v captures affiliate income that the sender obtains when she persuades the receiver to buy the promoted product, whereas $\psi A(e)$ captures her share of the platform's advertising revenue.

The game is otherwise as in the baseline. Define $U_\psi := \max_{e \geq 0}\{\psi A(e) - e\}$ as the sender's maximum achievable profit from transfers. We assume the associated maximizer, e_L, exists

[17] See `https://www.nytimes.com/2025/06/09/technology/youtube-videos-content-moderation.html`, `https://www.cbc.ca/news/entertainment/youtube-content-moderation-rules-1.7559931`, `https://mashable.com/article/youtube-content-moderation-changes`, accessed 25 August 2026.

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

and is finite; a sufficient condition is $\lim_{e\to\infty} \frac{A(e)}{e} = 0$. Let

$$e_H = \min\{e \geq e_L : U_\psi \geq (v + \psi A(e))(1 - I) - e\}. \tag{12}$$

In words, e_H is the effort level that is closest to the optimal one (e_L) while still being high enough to credibly signal that $w = 1$. Lastly, for convenience, define

$$I_\psi := \frac{v}{v + \psi A(e_L)}.$$

Therefore, $e_H = e_L$ if $I \geq I_\psi$ and $e_H > e_L$ otherwise.

We can show the following:

Proposition 9. *There exists an equilibrium with truthful communication if and only if*

$$I > \frac{\underline{c} - U_\psi}{p(v + \psi A(e_H))}. \tag{13}$$

Whenever (13) *holds, there exists a truthful equilibrium with effort*

$$e = \begin{cases} e_L & \text{if } w = 0, \\ e_H & \text{if } w = 1. \end{cases}$$

For $I > 0$, this is the least-cost truthful equilibrium satisfying the intuitive criterion.

When $I < I_\psi$, the equilibrium has a familiar structure with $w = 1$ senders exerting higher effort, $e_H > e_L$, to establish their credibility. However, an interesting feature of Proposition 9 is that, once $I \geq I_\psi$, truthful communication can be sustained even though the sender exerts the same state-independent effort, e_L. Since effort is independent of the underlying state, 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 $w = 0$ and $e_H = e_L$. The sender could transmit the untruthful signal $s = (1, e_L)$ to induce $r = 1$ and gain $v(1 - I)$. However, since this message is potentially deleted, there is an expected loss of transfer payments $I\psi A(e_L)$. When $I \geq I_\psi$ 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 threshold I_ψ again embeds an inverse relationship between moderation and effort. For $I < I_\psi$, greater moderation reduces the effort e_H required to distinguish the $w = 1$ sender. Moreover, a more generous revenue share lowers I_ψ, sustaining truth-telling without signaling at a lower moderation intensity. However, the next proposition shows that moderation policies

at or above I_ψ are strictly dominated.

Proposition 10. *Every moderation policy satisfying $I \geq I_\psi$ is strictly dominated by some policy satisfying $I < I_\psi$. In particular, full moderation is never optimal for the platform.*

For intuition, recall that reducing moderation induces the sender 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 the sender's continuation payoff. Near the boundary at which the sender chooses her transfer-maximizing effort in both states, 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 the receiver itself rather than relying on the sender 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 the sender and the receiver, or to provide that content itself like a traditional *broadcaster*. We show that a firm may prefer the platform business model, even if the sender has no informational advantage over the firm.

Start with the platform business model. The platform delegates the task of informing the receiver to the sender. 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 $(1, e^*)$ 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 (PC), 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

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

We assume this problem has an interior solution.[19]

The following proposition describes the optimal firm behavior.

Proposition 11. *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 a weakly 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) If the firm prefers being a platform then following an increase in v it must continue to prefer being a platform.

(iii) If the firm prefers being a platform then, following a first-order stochastic dominance shift in G to $\hat{G}$ such that $\hat{G}(c) \geq G(c)$ for all c, entry costs decrease and it must continue to prefer being a platform.

Proposition 11 (i) shows that platforms induce weakly 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, sender entry becomes only slightly more likely, while the sender reduces their effort. Hence, if the semi-elasticity of sender participation with respect to moderation is too small, the platform prefers low levels of moderation that induce higher effort.

Proposition 11 (ii) and (iii) provide conditions under whick a firm would prefer being a platform. Indeed, a larger v makes the sender more willing to join the platform at any level of moderation. Likewise, when there is a decrease in sender entry cost, such as following a first-order stochastic dominance shift in G, small levels of moderation are sufficient to induce sender entry with high probability. Together, they show that a firm prefers being a platform when it is able to minimize the participation-effort trade-off.

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 1.[20]

[19] A mild sufficient condition is $A'(0) > 1$ and $\lim_{e\to\infty} \frac{A(e)}{e} = 0$.

[20] With these expressions for G and $A(e)$, the platform implements $e = v/(1+2\alpha)$, while the broadcaster

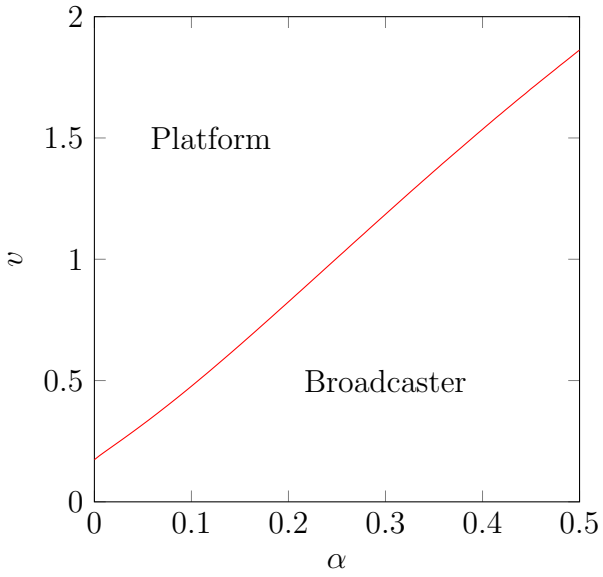

Figure 1: Firm's preferred business model when $G(c) = (c/2)^\alpha$, $A(e) = \frac{2}{3}\sqrt{e}$, and $p = 0.4$. A smaller α corresponds to a reduction in sender entry costs.

One possible reason for a first-order stochastic dominance shift in G is the introduction of cost-reducing technologies such as generative AI. We now turn to examining the effects of a platform allowing, or even facilitating, AI-generated content, before discussing a platform's adoption of AI moderation tools.

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 with engagement e at cost $k(e)$ where $k'(e) \in (0,1)$ $\forall e > 0$, $k(0) = 0$ and $\lim_{e\to\infty} k(e) > v$.[21] However, with probability $h \in (0,1)$ an AI-generated signal is garbled into unintelligible noise. Following this noise, the receiver plays $r = 0$ and does not engage with the content, so the platform makes

chooses $e = 1/9$. Substituting these values into the respective profits shows that a platform is more profitable if $v > \left[\frac{(1+2\alpha)^{\alpha+1/2}}{6p^{\alpha+1}\alpha^\alpha}\right]^{\frac{2}{2\alpha+1}}$.

[21] 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.

no ad revenue.[22] We can interpret h as a simple measure of the technological advancement of AI.

The sender may choose whether to use AI. 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 produces without 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 $\underline{s} = (0,0)$. If $w = 1$ the sender transmits $s^* = (1, e_A^*(I,h))$. 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 4. *Consider a truthful equilibrium satisfying the intuitive criterion in the subgame following a given $I \in (0,1)$. Then there exists $0 < \underline{h}(I) < \overline{h}(I) < 1$ such that*

(i) If $h \geq \overline{h}(I)$, AI is used neither if $w = 1$ nor in the binding off-path deviation by an untruthful sender when $w = 0$. The equilibrium effort is $e_A^*(I,h) = v(1-I)$.

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

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

The most important insight from Lemma 4 is that $\underline{h}(I) < \overline{h}(I)$. In words: the sender is more willing to use AI to produce false content than truthful content. This happens because the sender anticipates that false content may be deleted by the platform, giving her 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 false content off path because it is too unreliable to be used for truthful content. Only when AI's reliability improves sufficiently will it be

[22] One interpretation is that there are many independent topics for which the sender could inform the receiver. The sender's topic is selected at random before the state is observed. This way, garbling conceals the topic, so the receiver plays his priors and chooses $r = 0$.

adopted more widely to produce engaging truthful content. This has immediate implications for platform profit and moderation strategy:

Proposition 12. *Fix a moderation intensity I such that AI is not used when $w = 1$ (cases 1 and 2 of Lemma 4). 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 $\underline{h}(I) \leq h < \overline{h}(I)$).*

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, $e_A^*(I, h)$, must weakly increase to deter such a deviation. Importantly, any increase in $e_A^*(I, h)$ harms the sender when $w = 1$, who does not benefit from the efficiency savings of AI but must still exert more effort to separate herself 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 12 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 a sender producing false content, which makes incentive compatibility more difficult to satisfy.

An immediate implication of Proposition 12 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 2 provides an illustration of these results.

Proposition 13. *Consider the truthful equilibrium satisfying the intuitive criterion, with I set to maximize the platform's profit. There exists an $h^* \in (0, \frac{v-k(v)}{v})$ 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 $h < \underline{h}(I)$.*

Taken together, the results in this section tell a story of AI's emergence. Initially, the technology is so unreliable that the sender does not use it. 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

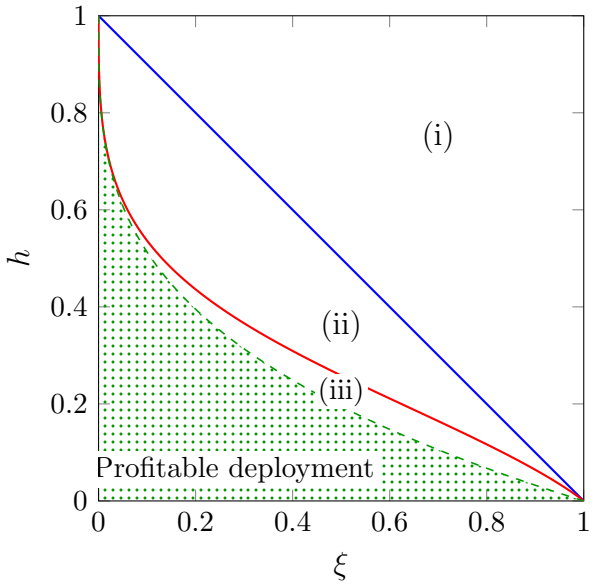

Figure 2: Platform and sender strategies when $G(c) = c^{\alpha}$ fixing $\alpha = 1/10$, $A(e) = \sqrt{e}$, $p = 1/3$, $v = 1$, and setting $k(e) = \xi e$, with $0 < \xi < 1$. The region on or above the top blue line, (i), reflects where the sender never prefers to use AI. The region on or above the middle red line and strictly below the top blue line, (ii), reflects where the sender only prefers to use AI when lying if $w = 0$. The region below the red line, (iii), reflects where the sender prefers to use AI both if $w = 1$ and when lying if $w = 0$. The dotted green area represents the region where AI is profitably deployed.

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.[23] These results also highlight an indirect cost of AI. Creators who do not adopt it may need to invest more effort to establish their credibility.

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 a fruitful area for further investigation.

6 Discussion of robustness

We now discuss how our analysis is robust to relaxation of several modeling decisions. Formal results are provided in Part I of the Online Appendix.

First, effort in our model affects payoffs but does not directly affect the information content

[23]The Online Appendix shows that AI can also raise profit in a pooling equilibrium in which only the $w = 0$ sender uses it, because garbling partly substitutes for moderation. Profit therefore need not be monotone in h when considering equilibria with no communication.

of the message. This cleanly highlights the role of effort as a signaling device, clarifying our main mechanism. We can alternatively consider a model where effort also makes messages more informative; this yields 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.[24] However, our insights are robust when the sender knows w before joining the platform.

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

Fourth, we assume moderation is free from error. We can relax this to consider moderation that is prone to error: a false message can sometimes pass inspection and survive, and a truthful message can similarly fail inspection and be deleted. This means, conditional on entry, more effort must be exerted by the sender when $w = 1$ to sustain truthful communication, but she is also subject to the risk of wasting this effort. This makes entry less attractive. Hence, a higher level of inspection is required to induce sender participation and sustain truthful communication. If the platform makes too many errors, then truthful communication becomes impossible.

Fifth, our model focuses on a one-shot communications game. Our results are robust to a multi-period game where the sender incurs a cost of entry and can be removed from the platform following moderation action. If the sender is removed from the platform, rejoining the platform incurs the entry cost again. This disciplines her on-platform behavior. Hence, communication can be sustained with less moderation, especially so when the sender places a greater emphasis on future payoffs.

We also show that our results are qualitatively robust to non-linear sender effort costs.

Finally, 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.

[24] 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.

7 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 more rigorous moderation than the platform.

We also show how these forces interact with other strategic concerns of the platform. Committing to a moderation policy is necessary to sustain profitable communication, so a platform has an incentive to build a reputation for moderating consistently. Moderation and revenue sharing can complement each other by increasing the opportunity cost of misleading content. More substantively, a firm is more likely to operate as a content-hosting platform rather than a broadcaster of first-party content when the participation-effort trade-off is less severe. This is because content-hosting platforms benefit from shifting the cost of content production to users, but only if the platform can easily attract a critical mass of creators of high-quality content. Lastly, AI-generated content introduces a new tension as it can reduce the cost of producing engaging content, but also makes it easier to fabricate misleading claims. Thus, platforms benefit from sufficiently advanced AI that represents a pure efficiency gain, but may be harmed by mediocre technology that is mostly used to create false content.

References

Abou El-Komboz, Lena, Anna Kerkhof, and Johannes Loh (2023). "Platform partnership programs and content supply: Evidence from the YouTube 'Adpocalypse'".

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

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, Jeffrey 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*.

Bedre Defolie, Özlem, Bjørn Olav Johansen, and Leonardo Madio (2026). "Curation with moral hazard: Why platforms host low quality". *Available at SSRN*.

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.

Bergemann, Dirk and Stephen Morris (2019). "Information design: A unified perspective". *Journal of Economic Literature* 57.1, pp. 44–95.

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.

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.

Forges, Françoise (1985). "Correlated equilibria in a class of repeated games with incomplete information". *International Journal of Game Theory* 14.3, pp. 129–149.

Ganguly, Chirantan and Indrajit Ray (2023). "Simple Mediation in a Cheap-Talk Game". *Games* 14.3, 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.

Kamenica, Emir and Matthew Gentzkow (2011). "Bayesian persuasion". *American Economic Review* 101.6, pp. 2590–2615.

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, 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, and David G. Rand (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*.

Matias, J. Nathan (2019). "Preventing harassment and increasing group participation through social norms in 2,190 online science discussions". *Proceedings of the National Academy of Sciences* 116.20, pp. 9785–9789.

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

Myerson, Roger B. (1982). "Optimal coordination mechanisms in generalized principal–agent problems". *Journal of Mathematical Economics* 10.1, pp. 67–81.

— (1991). *Game Theory: Analysis of Conflict*. Harvard University Press. ISBN: 9780674341166. URL: http://www.jstor.org/stable/j.ctvjsf522 (visited on 08/15/2026).

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 $e = 0$ and $I \in [0, 1)$, and consider a putative separating equilibrium. Because the equilibrium is separating and the message space is binary, both $m = 0$ and $m = 1$ are on path conditional on entry.

We first show that $\rho(0) \neq \rho(1)$. Suppose instead that both messages induce $r = 0$. If $\rho(\emptyset) = 0$, entry yields zero gross surplus and therefore cannot occur with positive probability. Hence $\rho(\emptyset) = 1$. This requires $I > 0$, since when $I = 0$ the null signal arises only from non-entry and has posterior $p < 1/2$. But when $I > 0$, a false message yields $Iv > 0$, whereas a truthful message yields zero. The sender therefore transmits $m \neq w$. Because $I < 1$, a surviving message $m = 0$ then reveals $w = 1$ and must induce $r = 1$, a contradiction.

Now suppose both messages induce $r = 1$. If $\rho(\emptyset) = 1$, every on-path observation induces $r = 1$, contradicting Bayes plausibility because $p < 1/2$. Hence $\rho(\emptyset) = 0$. If $I = 0$, every message sent following entry survives, so Bayes plausibility conditional on entry again requires some on-path message to induce $r = 0$, a contradiction. If $I > 0$, a truthful message yields v, whereas a false message yields $(1 - I)v < v$. The sender therefore reports truthfully in both states, so $m = 0$ reveals $w = 0$ and must induce $r = 0$, again a contradiction.

Thus, after relabeling the messages, $\rho(m) = 1$ and $\rho(m') = 0$. If $\rho(\emptyset) = 1$, message m yields v in either state, whereas m' yields at most $Iv < v$. Hence m strictly dominates m'. We must therefore have $\rho(\emptyset) = 0$. But then m yields either v if truthful or $(1 - I)v > 0$ if false, while m' yields zero in either state. Message m again strictly dominates m', contradicting the requirement that m' is on path.

It remains to establish existence of a separating equilibrium when $I = 1$. Let the sender report truthfully, $m = w$. Then $\beta(m) = m$, and $\beta(\emptyset) = p < 1/2$ because the null signal only arises following non-entry. A false-message deviation is deleted with certainty and therefore induces $r = 0$, so no sender profits from deviating. Gross expected surplus from entry is pv, and hence the sender enters with probability $G(pv) > 0$. □

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 there is no equilibrium in which the sender enters with positive probability. □

A.2 Proof of Proposition 1

Proof. By (4), $\zeta \leq \bar{\zeta}$ implies $\beta(1, e^*) \geq 1/2$. Hence, following $(1, e^*)$, the receiver optimally plays $r = 1$. Deviation is not profitable because the change in his expected payoff from deviating is

$$[1 - 2\beta(1, e^*)][\pi_h(e^*) - \pi_l(e^*)] \leq 0.$$

Similarly, following any other signal $s \neq (1, e^*)$, $\beta(s) < 1/2$ and the receiver optimally plays $r = 0$. Deviation is not profitable because the change in his expected payoff from deviating is

$$[2\beta(s) - 1][\pi_h(e(s)) - \pi_l(e(s))] < 0.$$

We now turn to the sender's optimal strategy. When $w = 1$, a sender playing $(1, e^*)$ obtains the gross payoff $v - e^* = vI \geq 0$, and any other signal yields the gross payoff $-e \leq 0$. When $w = 0$, sending $(1, e^*)$ with probability ζ earns $(1 - I)v - e^* = 0$, sending s_0 yields the gross payoff 0, and any other signal yields a gross payoff $-e \leq 0$. Therefore deviating from either s^* or s_0 is not profitable. The sender's expected gross payoff when $w = 0$ is 0, and her expected gross payoff prior to learning the state is $p(v - e^*) + (1 - p)[\zeta[(1 - I)v - e^*] + [1 - \zeta]0] = pvI$.

The sender's net expected payoff from entry is $pvI - c$, so she prefers to enter whenever $I > \frac{c}{pv}$. If $I > \frac{c}{pv}$, then $G(pvI) > 0$ and entry occurs with positive probability. Conversely, if $I \leq \frac{c}{pv}$, then the sender has weakly negative surplus from entry. By the continuity of G entry occurs with zero probability. An active equilibrium following $\mathcal{S}(\zeta, e^*)$ thus exists if and only if $G(pvI) > 0$, which requires $I > \frac{c}{pv}$. □

A.3 Proof of Proposition 2

Proof. By (4), when $\zeta = 1$ and $I \geq \bar{I}$, we have $\beta(s^*) \geq 1/2$ for any $e \in [0, e^*]$. Therefore, following a signal s^*, the receiver plays $r = 1$. Any deviation toward $r = 0$ changes his expected payoff by

$$[1 - 2\beta(1, e)][\pi_h(e) - \pi_l(e)] \leq 0.$$

Following any other signal $s \neq s^*$, his beliefs satisfy $\beta(s) < 1/2$ and the receiver optimally plays $r = 0$. Deviating toward $r = 1$ would change his expected payoff by

$$[2\beta(s) - 1][\pi_h(e(s)) - \pi_l(e(s))] < 0.$$

We now turn to the sender's optimal strategy. When $w = 1$, transmitting s^* yields the gross payoff $v - e \geq vI \geq 0$. When $w = 0$, this yields the gross payoff $(1 - I)v - e \geq 0$. Any other signal s leads to $r = 0$ and the gross payoff $-e(s) \leq 0$. Thus, the sender has no profitable deviation.

The sender's expected payoff from entry is thus $pv + (1 - p)(1 - I)v - e - c$. The sender therefore enters with probability $G(pv + (1 - p)(1 - I)v - e)$, and this probability is positive if and only if $I < \frac{v-e-c}{(1-p)v}$. □

A.4 Proof of Proposition 3

Proof. Consider any active equilibrium that satisfies the intuitive criterion and in which the sender transmits $m = 1$ with probability 1 whenever $w = 1$. The receiver plays $r = 0$ following the null signal or any on-path signal with $m = 0$. We first show that S_1 is a singleton. Positive entry implies a strictly positive gross payoff when $w = 1$. Moreover, indifference requires any $s \in S_1$ to yield the same payoff $v - e$ when $w = 1$. This implies that the sender exerts the same effort for all signals. Therefore, there is a single s^*.

We next show that signals other than s^* transmitted when $w = 0$ have zero effort. These signals induce $r = 0$, giving the sender gross payoff $-e(s) \leq 0$. Since zero effort guarantees a non-negative payoff, the sender can strictly improve her payoffs by exerting zero effort instead of positive effort. Hence, when $w = 0$, any equilibrium signal other than s^* has $e = 0$ which induces zero payoff and is payoff equivalent to s_0. Thus, when $w = 0$, the sender transmits s^* with some probability ζ and, up to payoff equivalence, s_0 otherwise.

Suppose $\zeta < 1$ so the equilibrium is separating. The sender thus sends s_0 with positive probability when $w = 0$ and obtains payoff 0. Therefore, for signal s^* to be incentive compatible, we require $(1 - I)v - e \leq 0$, so $e \geq e^*$. The intuitive criterion selects the least-cost separating equilibrium in which the constraint binds. To see this, suppose $e > e^*$ and consider

an off-path signal $(1, \bar{e})$, where $\bar{e} \in (e^*, e)$. Even under receiver acceptance, this deviation yields a negative payoff when $w = 0$, whereas it increases the sender's payoff when $w = 1$. The intuitive criterion therefore requires the receiver to infer $w = 1$ and play $r = 1$, making the deviation profitable. Hence $e = e^*$.

We next consider $\zeta = 1$. In the pooling equilibrium, the sender always transmits s^*. When $w = 0$, this signal survives with probability $1 - I$, giving the sender a gross payoff $(1 - I)v - e$. Since she can guarantee a non-negative payoff by choosing zero effort, we have $(1 - I)v - e \geq 0$, or $e \leq e^*$. Following Bayes' rule, the receiver's posterior belief is $\frac{p}{p+(1-p)(1-I)}$. To induce $r = 1$, this posterior must be at least $1/2$, which requires $I \geq \bar{I}$. Thus, the equilibrium satisfies the conditions in Proposition 2.

Next, we show that the off-path beliefs $\beta(s) = 0$ satisfy the intuitive criterion. If the receiver chooses $r = 1$ following an off-path signal $s = (1, e(s))$, the change in the sender's payoff from deviating to s is $e - e(s)$ for all w. Thus, if this deviation is profitable when $w = 1$, it is also profitable when $w = 0$. Consider an off-path signal $s = (0, e(s))$. If the receiver chooses $r = 1$, the change in the sender's payoff is $e - e(s) - Iv$ when $w = 1$ and $e - e(s) + Iv$ when $w = 0$. The gain from this deviation is therefore strictly greater when $w = 0$. Hence, for any $\beta(s)$, there is no off-path signal that is more attractive when $w = 1$ than when $w = 0$. Thus, $\beta(s) = 0$ is compatible with the intuitive criterion.

Finally, we show that if $I > \underline{I}$ then there is no equilibrium in which the platform is inactive. Suppose that this claim was false. Choose a cost type $c > \underline{c}$ and $\varepsilon > 0$ sufficiently small that $c < p(vI - \varepsilon)$. Such a (c, ε) exists because $I > \underline{I} = \underline{c}/pv$. Let V_w denote the sender's gross continuation payoff following entry in state w. Inactivity requires $pV_1 + (1 - p)V_0 \leq c$, while $V_0 \geq 0$ because the sender can guarantee a non-negative payoff by transmitting $(0, 0)$. Hence $V_1 \leq c/p < vI - \varepsilon$. Now consider the following deviation: the sender enters, transmits $\widetilde{s} = (1, e^* + \varepsilon)$ if $w = 1$, where $e^* = v(1 - I)$, and transmits $(0, 0)$ if $w = 0$. If the receiver plays $r = 1$ after $\widetilde{s}$, a sender with $w = 1$ obtains $v - (e^* + \varepsilon) = vI - \varepsilon > V_1$, whereas a sender with $w = 0$ obtains $(1 - I)v - (e^* + \varepsilon) = -\varepsilon < 0 \leq V_0$. The intuitive criterion therefore requires the receiver to assign $\beta(\widetilde{s}) = 1$ and hence to choose $r = 1$. The deviation then gives the sender expected net payoff of at least $p(vI - \varepsilon) - c > 0$ from entering, contradicting equilibrium inactivity. □

A.5 Proof of Proposition 4

Proof. First consider $I = 1$. Any surviving message must be truthful. A surviving message with $m = 0$ induces $r = 0$, so its sender exerts zero effort. A surviving message with $m = 1$ reveals $w = 1$; under the intuitive criterion, its sender cannot exert positive effort, since

she could profitably deviate to a slightly lower effort that could never be observed following $w = 0$. Hence all surviving content has zero effort, and $I = 1$ yields zero platform profit. Since platform profit is always non-negative, every $I < 1$ weakly dominates $I = 1$.

Now suppose $\underline{I} < \bar{I}$ and fix any $I \in (\underline{I}, \bar{I})$. The intuitive criterion rules out an inactive-platform equilibrium (Proposition 3). Suppose that the selected equilibrium gives the platform zero profit. Since $I < 1$, every transmitted signal survives with positive probability; because $A(e) > 0$ for every $e > 0$, zero platform profit therefore requires every on-path signal to have $e = 0$. Lemma 1 rules out a separating equilibrium with zero effort when $I < 1$, so the sender's message distribution must be independent of w. For any pooled message $m = 1$ that survives moderation, the receiver's posterior is

$$\frac{p}{p + (1-p)(1-I)} < \frac{1}{2},$$

where the inequality follows from $I < \bar{I}$. A surviving pooled message $m = 0$ induces an even lower posterior. The receiver therefore plays $r = 0$ following every surviving message. If he also plays $r = 0$ following the null signal, entry gives the sender zero gross surplus and the platform cannot be active. If instead he plays $r = 1$ following the null signal, the sender in each state strictly prefers to transmit the false message, since it is deleted with probability $I > 0$, whereas a truthful message survives and induces $r = 0$. The $w = 1$ sender therefore transmits $m = 0$, while the $w = 0$ sender transmits $m = 1$, contradicting pooling. Thus every equilibrium selected at $I \in (\underline{I}, \bar{I})$ yields strictly positive platform profit, so every such I strictly dominates $I = 1$. □

A.6 Proof of Proposition 6

Proof. Fix $\varepsilon > 0$ and $I \in (0, 1]$, and consider a positive-entry equilibrium of the perturbed game satisfying the intuitive criterion. Let U_w denote the sender's continuation payoff in state w, and let $\rho_\varepsilon(s)$ denote the probability that $r = 1$ following s. Since $v < \overline{c}$, non-entry occurs with positive probability, so $\rho_\varepsilon(\emptyset) < 1$.

We first show that the sender never transmits $m = 0$ when $w = 1$. Suppose that $s = (0, e)$ is used in that state. Then $U_1 = v\big[(1-I)\rho_\varepsilon(s) + I\rho_\varepsilon(\emptyset)\big] - e$, whereas a $w = 0$ sender can transmit s truthfully or transmit $(1, 0)$, implying

$$U_0 \geq \max\big\{v\rho_\varepsilon(s) - e,\ vI\rho_\varepsilon(\emptyset)\big\}.$$

These inequalities imply $U_1 \leq U_0$. Hence

$$v\big[(1-I)+I\rho_\varepsilon(\emptyset)\big]-U_0 < v-U_1.$$

Choose an off-path $e'>0$ strictly between these expressions. If $s'=(1,e')$ induces $r=1$, it strictly benefits the $w=1$ sender but cannot benefit the $w=0$ sender. The intuitive criterion therefore assigns $\beta(s')=1$, making s' a profitable deviation when $w=1$. Thus, the sender transmits $m=1$ whenever $w=1$. It follows that $\beta(\emptyset)\leq p<1/2$ and $\rho_\varepsilon(\emptyset)=0$.

We next rule out any on-path signal $s=(1,e)$ satisfying $1/2\leq\beta(s)<1$. Such a signal is used in both states and is accepted only by a regular receiver, so

$$U_1=(1-\varepsilon)v-e, \qquad U_0=(1-I)(1-\varepsilon)v-e.$$

Choose Δ such that $\varepsilon(1-I)v<\Delta<\varepsilon v$. The off-path signal $s'=(1,e+\Delta)$ would, if assigned posterior one, raise the $w=1$ sender's payoff by $\varepsilon v-\Delta>0$ but lower the $w=0$ sender's payoff even under her most favorable beliefs, since her gain would be $\varepsilon(1-I)v-\Delta<0$. The intuitive criterion therefore assigns $\beta(s')=1$, making the deviation profitable, a contradiction. Hence every on-path signal has $\beta\in[0,1/2)\cup\{1\}$.

A $w=0$ sender therefore never induces $r=1$ and chooses zero effort, obtaining $U_0=0$. Positive entry meanwhile requires the $w=1$ sender to use a signal that is accepted by both receiver types. If her effort is e, deterrence of imitation by $w=0$ requires $e\geq e^*$. If the inequality were strict, an off-path signal with effort $e'\in(e^*,e)$ would benefit only the $w=1$ sender; the intuitive criterion would assign it posterior one, producing a profitable deviation. Thus $e=e^*$. Every positive-entry equilibrium of the perturbed game is payoff-equivalent to the truthful equilibrium.

For $I=0$, no positive-entry equilibrium exists. Finally, any robust positive-entry equilibrium of the original game is the limit of positive-entry equilibria of the perturbed games as $\varepsilon\to 0$. Since each such equilibrium is payoff-equivalent to the truthful equilibrium, so is its limit. □

A.7 Proof of Proposition 7

Proof. For part (i), strict quasi-concavity implies that $W'(1)>0$ entails $I^{**}=1$, whereas $W'(1)<0$ entails $I^{**}<1$. It remains to consider $W'(1)=0$. Since pv lies inside the support of the log-concave distribution G, we have $G'(pv)>0$, so this equality implies $\pi_h'(0)>\mathbf{1}_S$.

For every $I \in (\underline{c}/(pv), 1)$, log-concavity of G and concavity of π_h therefore give

$$\begin{aligned} p\frac{G'(pvI)}{G(pvI)} &\geq p\frac{G'(pv)}{G(pv)} \\ &= \frac{\pi_h'(0) - \mathbf{1}_S}{\pi_h(0) - \pi_l(0)} \\ &> \frac{\pi_h'(v(1-I)) - \mathbf{1}_S}{\pi_h(v(1-I)) - \pi_l(0)}. \end{aligned}$$

The expression for W' then implies $W'(I) > 0$ throughout the positive-entry interval. Since welfare is constant on the no-entry interval, $I^{**} = 1$ is the unique optimum.

Consequently, $I^{**} \in (0, 1)$ if and only if $W'(1) < 0$. Substituting for $W'(1)$ shows that this is equivalent to (10).

Part (ii): Because W is strictly quasi-concave and I^* is interior, $I^* < I^{**} \iff W'(I^*) > 0$. Substituting the platform's first-order condition (7) into $W'(I^*)$ shows that this inequality is equivalent to (11). □

A.8 Proof of Lemma 3

Proof. Consider an equilibrium with a positive probability of sender entry and positive effort. If a false signal $s = (m, e)$ has $e > 0$, the platform obtains zero by deleting it and $A(e) > 0$ by retaining it. By backward induction, every positive-effort signal must be left undeleted, whether truthful or false.

Let U_w denote the sender's equilibrium continuation payoff in state w. Any on-path positive-effort signal must induce $r = 1$, otherwise it gives payoff $-e < 0$. Moreover, such a signal must be transmitted with positive probability when $w = 1$, since otherwise its posterior would be zero. If $s = (m, e)$ is such a signal, then $U_1 = v - e$. A sender with $w = 0$ can imitate s. Because $e > 0$, the platform retains the signal even if it is false, so imitation also yields $v - e$. Hence $U_0 \geq U_1$. In fact, $U_0 = U_1$. If $U_0 > U_1$, choose e' satisfying

$$v - U_0 < e' < v - U_1 = e.$$

If an off-path signal $s' = (m, e')$ induced $r = 1$, it would strictly benefit the sender when $w = 1$ but not when $w = 0$. Because $e' > 0$, the platform would retain s' in either state. The intuitive criterion therefore assigns posterior one to $w = 1$ following s', so the receiver chooses $r = 1$, making the deviation profitable. This is a contradiction. Define the common continuation payoff as $U := U_0 = U_1$. Positive entry requires $U > \underline{c} \geq 0$, while positive effort implies $U = v - e < v$. Thus, $0 < U < v$.

It remains to consider the receiver's action following the null signal.

First suppose $\rho(\emptyset) = 1$. A sender with $w = 1$ cannot use a zero-effort signal. A truthful zero-effort message gives either 0 or v, neither of which equals U. If she instead sends the false zero-effort message $m = 0$, deletion yields v. If the message survives and is accepted, her payoff is again v. If it survives and is rejected, it must also be transmitted when $w = 0$, otherwise a surviving $m = 0$ would reveal $w = 1$. But a sender with $w = 0$ would obtain zero from that truthful rejected message, contradicting $U > 0$. Hence no zero-effort signal is used (and no message is deleted) when $w = 1$. Consequently, $\beta(\emptyset) \leq p < 1/2$, contradicting $\rho(\emptyset) = 1$.

Now suppose $\rho(\emptyset) = 0$. Suppose some zero-effort signal induces $r = 1$. Then the sender of the appropriate type could transmit that message truthfully and get payoff v. If the signal would induce $r = 0$ then sending it guarantees a payoff of zero. Since $U \notin \{0, v\}$, no zero-effort signal can be on-path. Both states must therefore use only positive-effort signals. Every such signal survives and, as established above, induces $r = 1$. Conditional on entry, every observed signal must therefore have posterior at least $1/2$. This contradicts Bayes plausibility because the average posterior conditional on entry is $p < 1/2$.

Thus no active-platform equilibrium can involve positive sender effort. All on-path content consequently has $e = 0$, and the platform earns zero profit because $A(0) = 0$. □

A.9 Proof of Proposition 8

Proof. We first establish a grim trigger equilibrium in which, so long as the platform has never deviated from its pre-announced I, senders and receivers play the equilibrium $\mathcal{S}(\zeta, e)$, and let N denote the probability of sender entry each period. If the platform ever deviates from I then senders and receivers expect a no-entry equilibrium in which the platform implements $I = 0$ in every future period.

Suppose the platform is called upon to delete a false message but chooses not to do so. Then its revenue immediately increases by $A(e)$. However, the grim trigger strategy is triggered and the platform's subsequent continuation profit is zero because future senders never enter or exert effort. The overall deviation profit is thus $A(e)$.

If, on the other hand, the platform deletes the message, its reputation would remain intact and it would expect to earn discounted profits equal to

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

Comparing the current gain in revenues to the future loss, we find that the platform can

sustain its reputation if

$$\frac{\delta}{1-\delta}N[p+(1-p)\zeta(1-I)]A(e) \geq A(e),$$

which is equivalent to the condition in the proposition in the truthful equilibrium, where $\zeta = 0$ and $N = G(pvI)$. □

A.10 Proof of Proposition 9

Proof. Step 1: Construction of a truthful equilibrium. A sender who truthfully reports $w = 0$ knows that $r = 0$. She therefore chooses $e = e_L$ and obtains payoff $U_\psi = \psi A(e_L) - e_L$. If instead she mimics the $w = 1$ signal with effort e, her false message survives with probability $1 - I$, giving expected payoff $[v + \psi A(e)](1 - I) - e$. Deterring this deviation therefore requires $e \geq [v + \psi A(e)](1 - I) - U_\psi$. By definition, e_H is the least effort $e \geq e_L$ satisfying this constraint.

A sender with $w = 1$ obtains $v + \psi A(e_H) - e_H$ from her equilibrium signal. Any deviation that induces $r = 0$ gives at most $\psi A(e) - e \leq U_\psi$. By the definition of e_H, either $e_H = e_L$ or the incentive compatibility constraint binds. In the first case, the sender's equilibrium payoff when $w = 1$ is $v + \psi A(e_L) - e_L = v + U_\psi > U_\psi$. In the second case, $e_H = [v + \psi A(e_H)](1 - I) - U_\psi$, and hence her equilibrium payoff is

$$v + \psi A(e_H) - e_H = I[v + \psi A(e_H)] + U_\psi \geq U_\psi.$$

Thus, the sender when $w = 1$ has no profitable deviation to a signal inducing $r = 0$. Pessimistic beliefs following other off-path signals complete the construction.

Step 2: Characterizing when e_H and e_L coincide. By (12), $e_H = e_L$ is feasible if and only if $e_L \geq [v + \psi A(e_L)](1 - I) - U_\psi$. Using $U_\psi = \psi A(e_L) - e_L$, this is equivalent to $I \geq I_\psi$. Since e_H is the lowest feasible effort weakly above e_L, it follows that $e_H = e_L$ if and only if $I \geq I_\psi$. If $I < I_\psi$, then $e_H = e_L$ violates incentive compatibility, so $e_H > e_L$.

Step 3: Entry in the least-cost truthful equilibrium when $I > 0$. Fix $I > 0$. First suppose $I < I_\psi$. The incentive compatibility constraint binds and $e_H > e_L$, so gross expected sender surplus is

$$p\big[v + \psi A(e_H) - e_H\big] + (1 - p)U_\psi = pI\big[v + \psi A(e_H)\big] + U_\psi.$$

Positive entry therefore occurs if and only if $pI\big[v + \psi A(e_H)\big] + U_\psi > \underline{c}$, which is equivalent to (13).

Now suppose $I \geq I_\psi$. In this case $e_H = e_L$, and gross expected sender surplus is

$$p\big[v + \psi A(e_L) - e_L\big] + (1-p)\big[\psi A(e_L) - e_L\big] = pv + U_\psi > \underline{c}.$$

Thus, positive entry always occurs in this region. Moreover, $I \geq I_\psi$ implies

$$pI\big[v + \psi A(e_L)\big] + U_\psi \geq pv + U_\psi > \underline{c},$$

so (13) also holds automatically. Hence (13) characterizes positive sender entry in the least-cost truthful equilibrium whenever $I > 0$.

Step 4: Other truthful equilibria. The analysis above characterizes least-cost truthful equilibrium effort and entry conditional on a sender with $w = 1$ choosing $e_H \geq e_L$. We now consider equilibria in which a sender with $w = 1$ chooses $e < e_L$.

First suppose $I > 0$. Consider a candidate truthful equilibrium in which the $w = 1$ sender exerts effort $e < e_L$ and obtains payoff $v + \psi A(e) - e$. Since e_L maximizes $\psi A(e) - e$, we can construct an $e' > e_L$ such that $v + \psi A(e') - e' = v + \psi A(e) - e$. Then a sender with $w = 1$ would be indifferent between e' and e if both would lead to $r = 1$. Meanwhile, if $w = 0$ the sender would strictly prefer a deviation to $s = (1, e)$ over deviating to $s = (1, e')$:

$$\big\{(1-I)[v + \psi A(e')] - e'\big\} - \big\{(1-I)[v + \psi A(e)] - e\big\} = -I\psi[A(e') - A(e)] < 0,$$

where the equality exploits $v + \psi A(e') - e' = v + \psi A(e) - e$. Since incentive compatibility requires the $w = 0$ sender to prefer $(0, e_L)$ to $(1, e)$, she must therefore strictly prefer $(0, e_L)$ to $(1, e')$. By continuity and concavity, there exists an $\epsilon > 0$ such that a deviation from e to $e' - \epsilon$, if it induced $r = 1$, would be strictly profitable for a $w = 1$ sender but strictly unprofitable for a $w = 0$ sender. Thus, a truthful equilibrium satisfying the intuitive criterion cannot have the $w = 1$ sender exert effort below e_L.

Since $e_H \geq e_L$ and e_L maximizes $\psi A(e) - e$, the $w = 1$ sender's payoff is decreasing in effort above e_H. Therefore, if there is no entry in the least-cost truthful equilibrium, no other truthful equilibrium satisfying the intuitive criterion has positive entry.

Finally, suppose $I = 0$. Let X denote the payoff generated by the accepted truthful signal. Incentive compatibility for the $w = 1$ sender requires $X \geq U_\psi$, since she could imitate the $w = 0$ signal. Because false messages are never deleted, the $w = 0$ sender evaluates the accepted signal in exactly the same way, so her incentive compatibility constraint requires $X \leq U_\psi$. Hence $X = U_\psi$, and both sender states obtain U_ψ in every truthful equilibrium. Positive entry consequently occurs if and only if $U_\psi > \underline{c}$, which is precisely (13) when $I = 0$. Any equilibrium at $I = 0$ with $e < e_L$ when $w = 1$ therefore does not affect the proposition's

existence condition. □

A.11 Proof of Proposition 10

Proof. First suppose $e_L = 0$. Then $A(e_L) = 0$ and $I_\psi = 1$. At $I = 1$, the sender exerts zero effort in both states and platform profit is zero. For an $I < 1$ sufficiently close to one, (12) gives $e_H > 0$. Moreover, gross sender surplus converges to $pv > \underline{c}$ as I approaches 1 from below. Sender entry therefore occurs with positive probability and, because $A(e_H) > 0$ and $\psi < 1$, platform profit is strictly positive. Thus $I = 1$ cannot be optimal, proving $I < I_\psi$ when $e_L = 0$.

Now suppose $e_L > 0$. Then $\psi A'(e_L) = 1$ and $I_\psi < 1$. By definition, for every $I \geq I_\psi$ we have $e_H = e_L$ and sender gross utility $pv + U_\psi$. Platform profit is therefore constant for all $I \geq I_\psi$.

Consider a moderation policy that sets I below but close to I_ψ. The incentive compatibility constraint binds, $U_\psi = [v + \psi A(e_H)](1 - I) - e_H$, and expected gross sender participation utility is

$$U := p[v + \psi A(e_H) - e_H] + (1 - p)U_\psi.$$

Differentiating with respect to e_H and evaluating at $e_H = e_L$ yields

$$\left.\frac{dU}{de_H}\right|_{e_H = e_L} = p[\psi A'(e_L) - 1] = 0.$$

Now, platform profit is $\Pi_\psi := (1 - \psi)G(U)\left[pA(e_H) + (1 - p)A(e_L)\right]$, meaning

$$\left.\frac{d\Pi_\psi}{de_H}\right|_{e_H = e_L} = (1 - \psi)\Bigg[G(pv + U_\psi)pA'(e_L)$$
$$+G'(U)\left[pA(e_H) + (1 - p)A(e_L)\right]\underbrace{\left.\frac{dU}{de_H}\right|_{e_H = e_L}}_{=0}\Bigg] > 0.$$

Implicit differentiation of the binding constraint $U_\psi = [v + \psi A(e_H)](1 - I) - e_H$ gives

$$\frac{de_H}{dI} = -\frac{v + \psi A(e_H)}{1 - (1 - I)\psi A'(e_H)} < 0.$$

A small reduction in I below I_ψ therefore raises e_H which strictly raises platform profit and no policy satisfying $I \geq I_\psi$ can be optimal. □

A.12 Proof of Proposition 11

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 broadcaster problem, the optimal e^B that maximizes the firm's objective 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) Let $\Phi(v)$ denote maximized platform profit. Fix $v' > v$ and evaluate platform profit at a moderation policy that is optimal given v. At this fixed policy, both $G(pvI)$ and $A(v(1-I))$ are weakly increasing in v. Hence

$$\Phi(v') \geq pG(pv'I)A(v'(1-I)) \geq pG(pvI)A(v(1-I)) = \Phi(v).$$

Broadcaster profit is independent of v. Therefore, if the platform is strictly preferred at v, it remains strictly preferred at v'.

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

A.13 Proof of Lemma 4

Proof. 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 produces it without AI. The non-trivial incentive constraints are therefore the constraints deterring a $w = 0$ sender from mimicking the signal sent when $w = 1$. 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 weakly less attractive than the non-AI mimicking deviation at the no-AI truthful-equilibrium effort $e^N(I) := v(1-I)$. This is the case if $v(1-I)(1-h) - k(e^N(I)) \leq v(1-I) - e^N(I) = 0$, or equivalently

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

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 effort supporting truthful communication 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 $v(1-h) - k(e^N(I)) > v - e^N(I)$, which is equivalent to

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

This cannot hold when $h \geq \overline{h}(I)$. 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 $h < \overline{h}(I)$. At effort $e^N(I)$, the AI-assisted mimicking deviation is strictly profitable. The relevant mimicking condition is therefore the AI-assisted one. The least effort deterring it is such that

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

Because $h < \overline{h}(I)$, 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 $v(1-h) - k(e_A^*(I,h)) > v - e_A^*(I,h)$, or equivalently

$$F_I(h) := \frac{e_A^*(I,h) - k(e_A^*(I,h))}{v} - h > 0.$$

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 $h < \underline{h}(I)$, then $F_I(h) > 0$, so a $w = 1$ sender uses AI to generate the signal s^*. This

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

A.14 Proof of Proposition 12

Proof. 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

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

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 case 2, $\underline{h}(I) \leq h < \overline{h}(I)$.

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

$$\Pi^N(e^*_A(I, h)) = G(p(v - e^*_A(I, h)))pA(e^*_A(I, h)).$$

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

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

It is immediate that $\Pi^N(e^*_A(I, h)) \leq \Phi^N$. □

A.15 Proof of Proposition 13

Proof. Let $\Pi^N(e) := G\big(p(v-e)\big)pA(e)$ and $\Phi^N := \max_e \Pi^N(e)$ denote the platform profit from implementing effort e without AI and maximized no-AI profit, respectively. By Proposition 12, whenever $w = 1$, the sender produces her signal without AI, and platform profit is weakly below Φ^N. Hence AI can strictly increase maximized profit only through case 3 of Lemma 4.

In case 3, effort e and moderation satisfy $k(e) = v(1-I)(1-h) \iff I(e,h) = 1 - \frac{k(e)}{v(1-h)}$. Feasibility requires $k(e) \leq v(1-h)$, while the $w = 1$ sender strictly prefers AI if $h < \frac{e-k(e)}{v}$. Define the closure of the feasible set of efforts by

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

and let

$$\Pi^A(e,h) := G\big(p[v(1-h)-k(e)]\big)p(1-h)A(e), \qquad \Phi^A(h) := \max_{e \in \mathcal{E}_3(h)} \Pi^A(e,h).$$

Because $k(e)$ and $e - k(e)$ are strictly increasing, $\mathcal{E}_3(h)$ is a non-empty compact interval exactly when $h \leq h_v := \frac{v-k(v)}{v}$, and its endpoints vary continuously with h. Moreover, if $h' > h$, then $\mathcal{E}_3(h') \subseteq \mathcal{E}_3(h)$ and $\Pi^A(e,h') \leq \Pi^A(e,h)$ for every common feasible effort. It follows that Φ^A is continuous and weakly decreasing on $[0, h_v]$.

Let $e_N^* \in (0,v)$ maximize Π^N. Since $k(e_N^*) < e_N^*$, $\Phi^A(0) \geq \Pi^A(e_N^*, 0) > \Pi^N(e_N^*) = \Phi^N$. At $h = h_v$, the only effort in $\mathcal{E}_3(h_v)$ is $e = v$, and therefore $\Phi^A(h_v) = G(0)p(1-h_v)A(v) = 0 < \Phi^N$. Continuity and monotonicity therefore imply that there exists $h^* \in (0, h_v)$ such that $\Phi^A(h) > \Phi^N \iff h < h^*$.

$\mathcal{E}_3(h)$ ensures Φ^A is well-defined. Because the boundary of this set includes a point of indifference where the sender would not use AI under the tie-breaking rule, we have to verify that a strict profit gain identified using the closed set is not generated solely by an unattainable boundary outcome. To do so, consider a boundary effort inducing indifference, $h = \frac{e-k(e)}{v}$. The sender obtains the same surplus from creating a signal with and without AI, so the probability of sender entry is the same. AI production nevertheless yields ad revenues only with probability $1-h$. Thus, the platform prefers the no-AI benchmark at this boundary. Whenever $\Phi^A(h) > \Phi^N$, it must therefore be at a feasible interior point of $\mathcal{E}_3(h)$. Conversely, Proposition 12 implies every non-AI production outcome yields profit weakly below Φ^N. Hence AI strictly increases maximized platform profit if and only if $h < h^*$.

Finally, if AI strictly increases maximized profit, an optimal policy cannot induce case 1 or case 2, since Proposition 12 bounds profit in those cases by Φ^N. It must therefore induce case 3, which requires $h < \underline{h}(I)$. □

Online Appendix for
Moderating Content-Hosting Platforms

Robin Ng and Greg Taylor

This appendix is divided into three parts. Part I provides formal derivations for the robustness exercises described in Section 6 in the main paper. Part II shows how our model can be mapped into an alternative application where moderation removes content that is directly harmful rather than merely misleading. Part III examines the robustness of our results when we select the platform-optimal equilibrium instead of the truthful one.

Contents

Part I
Robustness of assumptions

i. 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 a message produced with effort e is 'understood' by the receiver with probability $\phi(e)$ such that $\phi'(e) > 0$ and $\phi''(e) < 0$. With probability $1 - \phi(e)$ the signal is unintelligible, akin to the garbling in Section 5.4, and the receiver plays $r = 0$. Assume $\max_e\{p(\phi(e)v - e)\} > \underline{c}$, which is a necessary condition for the sender to ever want to enter in a truthful equilibrium.

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

$$\phi(e)v(1 - I) - e \leq 0, \tag{OA1}$$

and sender participation,

$$p(\phi(e)v - e) > \underline{c}. \tag{OA2}$$

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

If $I = 0$ then it is impossible to satisfy (OA1) and (OA2), so truthful communication depends on a positive level of moderation.

For I small, we have $\phi(e_s(0))v(1-I) - e_s(0) > 0$. Thus, whenever truthful communication with positive probability of entry can be sustained, the sender when $w = 1$ must raise her effort above her payoff-maximizing level, $e^*_\phi(I) > e_s(0)$, in order to credibly signal the veracity of $m = 1$. The effort level selected by the intuitive criterion then makes (OA1) bind and is decreasing in I. There is again a trade-off between participation and effort. Moreover, a regulator 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 from our baseline analysis arises as I becomes large enough. There exists an $\hat{I} > 0$ such that $\phi(e_s(0))v(1 - \hat{I}) - e_s(0) = 0$. For any $I \geq \hat{I}$, $e^*_\phi(I) = e_s(0)$ satisfies both (OA1) and (OA2). 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.

ii. Sender learning the 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 is as follows: (1) The platform 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 the 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 the sender knows $w = 0$ prior to entry, to maintain incentive compatibility, we require $v(1 - I) - e^* - c \leq 0$ and the sender participation constraint is now $v - e^* - c \geq 0$. In the truthful equilibrium, we require the incentive compatibility constraint to be satisfied after taking into account sender entry cost, hence it becomes $e^* \geq v(1 - I) - \underline{c}$. Meanwhile the sender enters with positive probability when $e^* < v - \underline{c}$.

From the intuitive criterion, we have $e^* = \max\{0, v(1 - I) - \underline{c}\}$. We focus on the case where $e^* > 0$, which is the region where sender effort works in combination with moderation to enable communication. From this we see that if $I = 0$ then $G(v - e^*) = G(\underline{c}) = 0$, so the sender does not enter even when $w = 1$, preserving the result of no entry without moderation.

Applying the sender's equilibrium effort, we recover trade-offs similar to those in the main model: higher levels of moderation reduce sender effort, which (i) increases the probability of sender entry and (ii) can hurt the receiver. To see this, observe that welfare is

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

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

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

This yields a formulation which is qualitatively identical to (9). Hence, we recover the welfare trade-off between sender participation and crowding-out of effort, which is key to our understanding of the optimal moderation policy.

iii. Fact checking

We consider an alternative moderation rule where the platform 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 platform takes on a more active role in information provision. The sender may choose $s = (0, 0)$ when $w = 1$ and rely on the platform to inform the receiver 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 enters if her participation constraint is satisfied, $p(v - e^*) - c \geq 0$. Applying e^*, the constraint becomes $pvI \geq c \iff I \geq \frac{c}{pv}$. Finally, to ensure the sender enters with positive probability, her gross expected payoff must strictly exceed the lowest possible value of c, hence $I > \frac{c}{pv}$.

Under the alternative moderation rule, we recover exactly the truthful equilibrium in the main text without the application of the intuitive criterion to pin down equilibrium effort $e^* = v(1 - I)$ and entry probability $G(pvI)$. Thus, in the truthful equilibrium, the platform and regulator face the same participation-effort trade-off.

iv. Platform moderation error

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

The incentive compatibility condition preventing the sender from transmitting $s = (1, e^*_\mu(I))$ when $w = 0$ is $v(1 - I(1 - \mu)) - e^*_\mu(I) \leq 0$. Indeed, there is now a new way for a false message to reach the receiver: following inspection, the platform may erroneously retain it. Meanwhile, the sender finds entry worthwhile if $p[v(1 - \mu I) - e^*_\mu(I)] - c \geq 0$. Entry has become less attractive because there is now the chance that her effort is wasted if the platform mistakenly deletes her truthful message. Combining these conditions, a truthful equilibrium with positive probability of entry can be supported if $(1 - 2\mu)I > \frac{c}{pv}$. Comparison with (3) shows that a higher level of moderation is needed to sustain communication when moderation is imperfect. Indeed, frequent moderation errors may completely prevent truthful communication: it is possible to find an $I \leq 1$ that sustains truthful communication with entry only if $\mu < \frac{1}{2} - \frac{c}{2pv}$.

As in the proof of Proposition 3, the unique truthful equilibrium effort level satisfying the intuitive criterion is the least-cost effort supporting truthful communication, $e^*_\mu(I) = v(1 - I(1 - \mu))$. Thus, if μ is sufficiently small, the error-prone platform still faces the same trade-offs, and moderation affects sender participation and welfare in the same qualitative manner as the main model.

v. Long-lived sender with reputational concerns

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

Proposition OA1. *In an equilibrium with truthful communication, there is a unique effort level surviving the intuitive criterion, $e^*_\delta(I) = \max\{0, v(1-I) - \delta I\underline{c}\}$, that credibly signals $w = 1$. Such an equilibrium exists if and only if*

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

Proof. Following sender entry, the receiver learns the true state of the world in each period, and $r_t = m_t$ where m_t is the message sent in period t. For the equilibrium to be sustained, the sender must not have a profitable one-shot deviation. When $w = 0$, this means the expected payoff from deviation to $(1, e^*_\delta(I))$ must be no greater than the expected payoff from truthfully playing $(0, 0)$. These payoffs are $\left(v + \frac{\delta p(v - e^*_\delta(I))}{1-\delta}\right)(1-I) - e^*_\delta(I) + I\left(-\delta c + \frac{\delta p(v-e^*_\delta(I))}{1-\delta}\right)$ and $0 + \frac{\delta p(v-e^*_\delta(I))}{1-\delta}$ respectively. Hence, the sender's incentive compatibility constraint is $v(1-I) - e^*_\delta(I) - \delta cI \leq 0$. Since this must hold for every possible entry cost, the incentive compatibility constraint is

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

To induce sender entry, the participation constraint has to be satisfied with positive probability. In other words, the expected payoff from entry $\frac{p(v-e^*_\delta(I))}{1-\delta}$ must exceed the cost of entry with positive probability. Therefore the participation constraint is

$$\frac{p(v - e^*_\delta(I))}{1-\delta} > \underline{c}. \tag{PC2}$$

The rest of the proof follows the same steps as the proof of Proposition 3 if we use (IC2) instead of (IC) and (PC2) instead of (PC). Moreover, since effort is nonnegative, $e^*_\delta(I) = \max\{0, v(1-I) - \delta I\underline{c}\}$. □

Long-lived senders are reluctant to lie because, if caught, they incur the reentry cost, c. The sender most tempted to lie is therefore the one with $c = \underline{c}$ who finds reentry least costly.

[1] One can think of this as the situation where a series of moderation policy violations leads to an account ban. For example, YouTube has a 'Three-strike' policy where 3 warnings in a 90-day period lead to a ban. See `https://support.google.com/youtube/answer/2802032?hl=en-GB`.

For this sender, the anticipated reentry cost increases the effective cost of transmitting a false message by $\delta \underline{c} I$, reducing the effort required to credibly signal by the same amount. If $\underline{c} = 0$ then there exists a sender type who has no concern about reentry costs and the effort needed to deter this type from lying is then the same one we find in the baseline model, $e^* = v(1 - I)$.

vi. Non-linear effort cost

Consider the case where effort costs are non-linear. In other words, exerting an effort e costs the sender $C(e)$, where $C(0) = 0$ and C is continuous with $C'(e) > 0$ and $C''(e) > 0$. Focus on the truthful equilibrium as in Lemma 2. Incentive compatibility requires $v(1-I) - C(e^*) \leq 0$, while sender participation requires $p(v - C(e^*)) - c \geq 0$. Applying the intuitive criterion as in Propositions 1 and 3, the least-cost effort supporting truthful communication is $e^* = C^{-1}(v(1 - I))$. Because the sender's payoff is $p(v - C(e^*)) - c$, substituting the least-cost effort supporting truthful communication gives $pvI - c$, which means positive probability of sender entry requires $I > \frac{c}{pv}$, just as in (3).

As in the main model, a higher level of moderation decreases sender effort. This increases the probability of sender entry but lowers the advertising revenue of the platform. Therefore, under the same concavity conditions, the platform's trade-offs are the same as in the main model, and qualitatively, our results on optimal moderation and welfare continue to hold.

vii. Costly moderation

We now allow moderation to be costly. Suppose that each submitted message is inspected with probability I, at a cost $\tau > 0$ to the platform per inspection. The game is otherwise as in the baseline.

In this setting, the sender's incentive compatibility and participation constraints remain (IC) and (PC). Thus, the least-cost truthful equilibrium still has $e^* = v(1 - I)$, and sender entry occurs with probability $G(pvI)$. Since inspection costs are incurred only following entry, the platform's expected profit is

$$\Pi(I) = G(pvI) \left[pA(v(1 - I)) - \tau I \right].$$

Differentiating with respect to I gives

$$\frac{\partial \Pi}{\partial I} = pv \left[pG'(pvI)A(v(1 - I)) - G(pvI)A'(v(1 - I)) \right] - \tau \left[G(pvI) + pvIG'(pvI) \right].$$

The first two terms capture the same participation and crowding-out effects as in the baseline model. The final term reflects the cost of inspecting messages. Higher moderation increases both the probability of inspection and the probability of sender entry, thereby raising expected inspection expenditure. Hence, inspection costs provide an additional marginal disincentive to moderate.

For sufficiently small τ, the platform can still earn positive profit and optimally chooses an interior level of moderation. Indeed, $I = 0$ yields zero profit, while $I = 1$ yields negative

profit because sender effort, and hence advertising revenue, is zero. Thus, the platform's incentive to limit moderation survives the introduction of inspection costs.

Part II
Moderation of harmful content

Our baseline model focuses on moderation of false or deceptive messages. However, the same economic mechanism may arise when platforms moderate content that objectively violates their policies because it is directly harmful to receivers, such as hate speech or abusive material. In that application, a creator who anticipates that her content is likely to be removed has less incentive to invest in making it engaging. The receiver may consequently interpret high engagement as evidence that content is not harmful. Mapping our model onto this application requires two changes of interpretation. First, we interpret w as describing whether the content is safe or harmful. Second, the receiver chooses whether to consume the content, rather than whether to take some other action after consuming it. We formulate this nearby model below and show that its least-cost separating equilibrium has exactly the same effort, participation and platform-profit expressions as the truthful equilibrium in the baseline model, thus yielding the equivalent participation-effort trade-off that drives our welfare results.

Environment. The timing, entry decision and cost distribution are as in the baseline model. In a slight reinterpretation, the sender's content has a type, either safe ($w = 1$) or harmful ($w = 0$). Harmful content such as hate speech or violent content can be presented in its naked form ($m = 0$) or disguised 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) and chooses whether to continue consuming ($r = 1$) or not ($r = 0$). If the receiver continues consuming the content the sender gets ad or sponsorship 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 moderation technology now inspects the content for an objective policy violation. If $w = 0$, the content is deleted with probability I and replaced by $s = \emptyset$. Compliant content is never deleted. Thus, unlike in the baseline model, moderation depends directly on the content's type rather than on whether $m = w$.

After moderation, the receiver observes the surviving signal and chooses whether to consume the content. Let $r = 1$ denote consumption and $r = 0$ denote avoidance. The receiver's payoff is

$$u_R^H(w, r, e) = \begin{cases} (2w - 1)B(e), & r = 1, \\ 0, & r = 0, \end{cases} \tag{OA3}$$

where $B(e) > 0$, $B'(e) \geq 0$ and $B''(e) \leq 0$. The receiver therefore benefits from consuming compliant content and is harmed by consuming policy-violating content. If the receiver does not consume any content then his payoff is zero. Given posterior belief $\beta(s)$, consumption of

the content gives the receiver expected payoff

$$\big[2\beta(s)-1\big]B(e).$$

The receiver consequently chooses $r=1$ if and only if $\beta(s)\geq 1/2$.

As in the baseline model, the sender's payoff is $u_S^H(r,e)=vr-e$. The platform earns $A(e)$ if the content survives and is consumed, and zero otherwise. Hence the platform benefits from attracting more engaging content that the receiver is willing to consume.

Proposition OA2. *Fix $I\in[0,1]$. There exists a separating equilibrium in which the sender chooses $m=w$ and the receiver consumes content if and only if it is not harmful. The least-cost such equilibrium has*

$$s_1^*=\big(1,v(1-I)\big),\qquad s_0^*=(0,0).$$

Effort, sender payoffs, platform profit, and induced receiver action therefore coincide with those in the truthful equilibrium of the baseline model and there is sender entry if $I>\frac{c}{pv}$.

Proof. Suppose the sender uses $s_1=(1,e_1)$ when $w=1$ and $s_0=(0,0)$ when $w=0$. Let N denote the probability of sender entry. A surviving s_1 reveals $w=1$, while a surviving s_0 reveals $w=0$. Hence, $\beta(s_1)=1$ and $\beta(s_0)=0$. The null signal arises from non-entry in either state and from deletion of policy-violating content. Bayes' rule gives

$$\beta(\emptyset)=\frac{p(1-N)}{p(1-N)+(1-p)(1-N+NI)}\leq p<\frac{1}{2}.$$

The receiver therefore consumes after s_1 and avoids the content after s_0 or the null signal.

A sender with $w=0$ obtains zero from s_0. If she mimics the compliant sender by choosing s_1, her content survives and is consumed with probability $1-I$. The deviation gives payoff $v(1-I)-e_1$. Deterring this deviation requires

$$e_1\geq v(1-I). \tag{OA4}$$

A sender with $w=1$ obtains $v-e_1$ from s_1 and zero from deviating to s_0. Her incentive constraint is therefore $e_1\leq v$. The two incentive constraints are compatible for every $I\in[0,1]$.

It remains to establish the least-cost effort. Suppose $e_1>v(1-I)$. Choose an off-path signal $\widetilde{s}=(1,\widetilde{e})$ satisfying $v(1-I)<\widetilde{e}<e_1$. If the receiver consumes after $\widetilde{s}$, the deviation gives a $w=1$ sender payoff $v-\widetilde{e}>v-e_1$. A $w=0$ sender would obtain

$$v(1-I)-\widetilde{e}<0,$$

which is strictly less than her equilibrium payoff. The intuitive criterion therefore assigns $\beta=1$ to $\widetilde{s}$. The receiver consumes, making the deviation profitable for the $w=1$ sender. Hence $e_1>v(1-I)$ cannot survive the intuitive criterion, and $e_1^*=v(1-I)$. Gross expected entry surplus is therefore $p(v-e_1^*)=pvI$.

Only compliant content is consumed in equilibrium. It occurs with probability p, is never deleted and has engagement $v(1-I)$. Platform profit is therefore $pG(pvI)A\big(v(1-I)\big)$, which is the expression obtained in the truthful equilibrium of the baseline model. □

The proposition shows that the paper's central mechanism does not depend specifically on our focus on the truthfulness of content. In both models, moderation creates a gap between the expected return to effort across states. In the baseline model, effort invested in a false message may be wasted because the message is deleted. In the present application, effort invested in harmful content that violates the platform's rules may be wasted for the same reason. In either case, moderation reduces the effort required for the sender type not exposed to deletion to distinguish herself.

Because platform profit is identical to that under truthful communication in the baseline model, all results that depend only on this objective carry over directly. In particular, the participation-effort trade-off survives.

Part III
Robustness to platform-optimal equilibrium selection

The purpose of this part is to show that the main participation and effort trade-offs of Sections 4 and 5 extend to platform-optimal equilibrium selection.

As in Section 3.4, the resulting problem has two cases: the one where a separating equilibrium containing the greatest sustainable amount of false content is platform-optimal, and the one where the platform prefers a pooling equilibrium.

i. Selection rule and notation

We adopt the platform-optimal equilibrium-selection rule characterized in Section 3.4. Define

$$\bar{I} := \frac{1-2p}{1-p}, \qquad \bar{e} := v(1-\bar{I}) = \frac{pv}{1-p}.$$

Equilibrium effort $e \leq \bar{e}$ is implemented in a pooling equilibrium at $I = \bar{I}$, whereas effort $e > \bar{e}$ is implemented in the maximally misleading separating equilibrium, with

$$I = 1 - \frac{e}{v}, \qquad \zeta = \frac{pv}{(1-p)e}.$$

The platform's profit as a function of the implemented effort is therefore

$$\Pi(e) := \begin{cases} 2pG(2pv-e)A(e), & 0 \leq e \leq \bar{e}, \\ 2pG\big(p(v-e)\big)A(e), & \bar{e} \leq e \leq v. \end{cases} \tag{OA5}$$

If $e = \bar{e}$, the two profits coincide.

ii. Regulating the platform

Suppose that a regulator chooses I, and equilibrium selection is platform-optimal conditional on that policy.

For $I < \bar{I}$, pooling is infeasible and the platform-optimal equilibrium is the maximally misleading separating equilibrium, $\mathcal{S}(\bar{\zeta}, e^*)$. Effort is $e^* = v(1-I)$, the sender's expected payoff gross of entry cost is pvI, and conditional on entry, the probabilities that a truthful or false high-effort message survives are both p. Receiver surplus following sender entry is therefore

$$p\pi_h(e) + p\pi_l(e) + (1-2p)\pi_h(0).$$

For $I \geq \bar{I}$, the platform-optimal equilibrium is pooling. Define $Q(I) := 1 - (1-p)I$ and let $e_{pool}(I)$ denote the unique solution to

$$\max_{e \in [0, v(1-I)]} Q(I)G\big(Q(I)v - e\big)A(e).$$

Gross expected sender surplus and conditional receiver surplus are then $Q(I)v - e_{pool}(I)$ and

$$p\pi_h(e_{pool}(I)) + (1-p)\left[(1-I)\pi_l(e_{pool}(I)) + I\pi_h(0)\right]$$

respectively.

Let $U(I)$ denote the sender's expected payoff gross of entry cost and $R(I)$ the receiver's expected payoff conditional on sender entry under the platform-optimal equilibrium:

$$U(I) := \begin{cases} pvI, & I < \bar{I}, \\ Q(I)v - e_{pool}(I), & I \geq \bar{I}, \end{cases}$$

and

$$R(I) := \begin{cases} p\pi_h(v(1-I)) + p\pi_l(v(1-I)) + (1-2p)\pi_h(0), & I < \bar{I}, \\ p\pi_h(e_{pool}(I)) + (1-p)\left[(1-I)\pi_l(e_{pool}(I)) + I\pi_h(0)\right], & I \geq \bar{I}. \end{cases}$$

If the sender does not enter, the receiver follows the prior and obtains

$$R_0 := p\pi_l(0) + (1-p)\pi_h(0).$$

Using the indicator $\mathbf{1}_S$ defined in Section 4, the two welfare objectives can be written compactly as

$$W(I) := G(U(I))\big[R(I) + \mathbf{1}_S U(I)\big] - \mathbf{1}_S \int_{\underline{c}}^{U(I)} c\, dG(c) + [1 - G(U(I))]R_0.$$

The following result gives sufficient conditions under which the regulator chooses $I < 1$. Define

$$\Delta_0 := \pi_h(0) - \pi_l(0) > 0.$$

Proposition OA3. *Full moderation ($I = 1$) is strictly dominated by some $I < 1$ under the*

receiver-surplus objective if

$$G(pv)\left[pv\pi_h'(0) - (1-p)\Delta_0\right] > p^2 vG'(pv)\Delta_0. \tag{OA6}$$

Full moderation ($I = 1$) is strictly dominated by some $I < 1$ under the user-surplus objective if

$$G(pv)\left[pv\big(\pi_h'(0) - 1\big) - (1-p)\Delta_0\right] > p^2 vG'(pv)\Delta_0. \tag{OA7}$$

Proof. For I sufficiently close to 1, we have $I > \bar{I}$ and the platform-optimal equilibrium is pooling. At $I = 1$, the only feasible pooling effort is zero. Now consider a small reduction in I. Since $A(0) = 0$, $A'(0) > 0$, and $G(pv) > 0$, pooling profit is increasing in effort in a neighborhood below $I = 1$. Hence the platform-optimal pooling effort is $e_{pool}(I) = v(1-I)$ for all I sufficiently close to one. Define $I = 1 - \epsilon$. For sufficiently small $\epsilon > 0$, we therefore have $e_{pool}(I) = v\epsilon$ and $U(I) = pv(1-\epsilon)$. At $I = 1 - \epsilon$, conditional receiver welfare becomes

$$R(I) = p\pi_h(v\epsilon) + (1-p)\left[\epsilon\pi_l(v\epsilon) + (1-\epsilon)\pi_h(0)\right].$$

Differentiating welfare at $\epsilon = 0$ gives

$$\left.\frac{dW}{d\epsilon}\right|_{\epsilon=0} = G(pv)\left[pv\big(\pi_h'(0) - \mathbf{1}_S\big) - (1-p)\Delta_0\right] - p^2 vG'(pv)\Delta_0.$$

Setting $\mathbf{1}_S = 0$ and $\mathbf{1}_S = 1$ yields conditions (OA6) and (OA7), respectively. Hence, under the relevant condition, every sufficiently small reduction in moderation from $I = 1$ strictly raises welfare. □

The result has the same broad interpretation as under truthful equilibrium selection. Full moderation maximizes the accuracy of receiver decisions, but it eliminates sender effort. A small reduction in moderation can therefore raise welfare when the receiver's marginal return to effort is sufficiently large.

Platform-optimal selection introduces an additional cost of reducing moderation. When pooling, some messages sent when $w = 0$ survive and induce the wrong receiver action. The term $(1-p)\Delta_0$ in equations (OA6) and (OA7) captures this misinformation effect. Greater moderation replaces some misleading messages with correct decisions when $w = 0$.

We next compare the platform's preferred moderation policy with the regulator's choice. Define $S(e) := \pi_h(e) + \pi_l(e) - \pi_h(0) - \pi_l(0)$ and assume that $S(e) > 0$, a sufficient condition for which is $\pi_l'(e) > 0$.

Proposition OA4. *Suppose the platform-optimal equilibrium is a separating one with $I = I_P^*$. Suppose, additionally, that $W(I)$ is concave and the moderation policy that maximizes receiver or user surplus also induces a separating equilibrium. Then*

1. *The platform implements too little moderation from the receiver's point of view if*

$$\frac{A'(v(1-I_P^*))}{A(v(1-I_P^*))} > \frac{\pi_h'(v(1-I_P^*)) + \pi_l'(v(1-I_P^*))}{S(v(1-I_P^*))}. \tag{OA8}$$

2. The platform implements too little moderation from the users' point of view if

$$\frac{A'(v(1-I_P^*))}{A(v(1-I_P^*))} > \frac{\pi_h'(v(1-I_P^*)) + \pi_l'(v(1-I_P^*)) - 1}{S(v(1-I_P^*))}. \tag{OA9}$$

Proof. In the platform-optimal separating equilibrium, platform profit is $2pG(pvI)A(v(1-I))$. The platform's interior first-order condition is therefore

$$\frac{pG'(pvI_P^*)}{G(pvI_P^*)} = \frac{A'(v(1-I_P^*))}{A(v(1-I_P^*))}.$$

For either objective, welfare in this equilibrium satisfies

$$\frac{dW(I)}{dI} = pv\Big[pG'(pvI)S(v(1-I)) \\ - G(pvI)\big\{\pi_h'(v(1-I)) + \pi_l'(v(1-I)) - \mathbf{1}_S\big\}\Big].$$

Evaluating this derivative at I_P^*, substituting the platform's first-order condition, and setting $\mathbf{1}_S = 0$ or $\mathbf{1}_S = 1$ yields (OA8) and (OA9), respectively. □

If we include the platform's preference in addition to the receiver's or users', a similar result would hold. Increasing I in the pooling region (above $\bar{I}$) removes additional false content and improves the accuracy of the receiver's action conditional on entry. At the same time, it reduces the probability that engaging content reaches the receiver, and changes equilibrium effort and therefore has an ambiguous effect on sender entry. Users may therefore prefer more moderation than the platform when the harm from misinformation is large, but less moderation when effort and sender participation are especially valuable.

The same participation and effort trade-offs therefore arise under platform-optimal equilibrium selection. A regulator concerned only with the receiver's surplus, or with total user surplus, may optimally choose less than full moderation. Moreover, the platform need not moderate in the direction preferred by its users: it can select either too much or too little moderation.

iii. Platform commitment

The proof of Proposition 8 shows a sufficiently patient platform commits to its moderation policy whenever $\mathcal{S}(\zeta, e)$ is selected following the platform-optimal selection rule.

iv. Transfers to the sender

We now allow the platform to share a fraction $\psi \in [0, 1)$ of its advertising revenue with the sender, as in the main paper. We retain the notation e_L, U_ψ, e_H, and I_ψ defined there. In particular, e_L maximizes the sender's payoff from transfers, U_ψ is the associated payoff, and e_H is the lowest $w = 1$ effort consistent with truthful communication.

We show that the main qualitative result survives platform-optimal equilibrium selection: the platform still chooses less than full moderation.

Let

$$\bar{\zeta}(I) := \min\left\{\frac{p}{(1-p)(1-I)}, 1\right\}.$$

iv-i. Profit in the platform-optimal separating equilibrium

Begin with separating equilibria. Gross expected sender surplus is

$$U_C(I) := p\big[v + \psi A(e_H) - e_H\big] + (1-p)U_\psi.$$

When $I < I_\psi$, e_H is determined by the incentive compatibility constraint. The sender when $w = 0$ is therefore indifferent between the truthful signal $(0, e_L)$ and the signal $(1, e_H)$, and transmits the latter with probability ζ. The receiver is willing to trust the latter signal if

$$\beta(1, e_H) \geq \frac{1}{2} \iff \zeta \leq \min\left\{\frac{p}{(1-p)(1-I)}, 1\right\}.$$

Focus here on the (semi-)separating equilibrium $\zeta < 1$, and consider pooling equilibria $\zeta = 1$ in the next section. Conditional on $w = 0$, replacing the truthful signal for $w = 0$ with the signal sent when $w = 1$ changes expected advertising revenue by $(1-I)A(e_H) - A(e_L)$. The first term is the advertising value of the false high-effort message, accounting for its probability of deletion, while the second is the value of the truthful message it replaces. The platform-optimal separating equilibrium when $I < I_\psi$ therefore uses mixing probability

$$\zeta_C(I) = \begin{cases} \bar{\zeta}(I), & (1-I)A(e_H) > A(e_L), \\ 0, & (1-I)A(e_H) < A(e_L), \\ \text{any } \zeta \in \left[0, \bar{\zeta}(I)\right], & (1-I)A(e_H) = A(e_L). \end{cases}$$

When $I \geq I_\psi$, we have $e_H = e_L$. Thus, replacing a truthful signal $(0, e_L)$ with $(1, e_H)$ causes revenue to fall because $(1-I)A(e_H) - A(e_L) = -IA(e_L)$. Thus the platform-optimal separating equilibrium when $I \geq I_\psi$ uses mixing probability $\zeta_C(I) = 0$.

Profit in the platform-optimal separating equilibrium is then

$$\Pi_C(I) := (1-\psi)G\big(U_C(I)\big)\Big[pA(e_H) + (1-p)\big\{[1-\zeta_C(I)]A(e_L) + \zeta_C(I)(1-I)A(e_H)\big\}\Big],$$

and the best profit attainable in a separating equilibrium is

$$\Phi_C(\psi) := \max_{I \in [0,1]} \Pi_C(I).$$

iv-ii. Profit in the platform-optimal pooling equilibrium

We next consider pooling equilibria. If the sender transmits $(1, e)$ in either state, her payoff when $w = 0$ is

$$(1 - I)\big[v + \psi A(e)\big] - e.$$

She can instead transmit the truthful signal for $w = 0$ and obtain U_ψ.

The intuitive criterion imposes an additional restriction on pooling effort. Define $q(e) := \psi A(e) - e$. Suppose $e_L > 0$ and consider a proposed pooling effort $e < e_L$. Because q is strictly concave and uniquely maximized at e_L, there exists $E > e_L$ such that $q(E) = q(e)$. If the deviation $(1, E)$ were accepted, it would leave the $w = 1$ sender indifferent, whereas it would change the $w = 0$ sender's payoff by $-I\psi[A(E) - A(e)] < 0$. An effort slightly below E therefore strictly benefits the $w = 1$ sender while continuing to make the $w = 0$ sender worse off. The intuitive criterion assigns posterior probability one to $w = 1$ following this deviation, making it profitable. Hence a pooling equilibrium satisfying the intuitive criterion must have $e \geq e_L$.

At the minimum moderation level supporting receiver trust, the set of pooling efforts consistent with sender incentives and the intuitive criterion is therefore

$$\mathcal{E} := \left\{ e \geq e_L : (1 - \bar{I})\big[v + \psi A(e)\big] - e \geq U_\psi \right\}.$$

At a general $I \geq \bar{I}$, the message transmitted when pooling survives with probability $1 - (1 - p)I$. The sender's expected payoff gross of entry cost and platform profit are respectively

$$[1 - (1 - p)I]\big[v + \psi A(e)\big] - e$$

and

$$[1 - (1 - p)I](1 - \psi)G\left([1 - (1 - p)I]\big[v + \psi A(e)\big] - e\right) A(e).$$

Conditional on $\beta(1, e) \geq 1/2$, reducing I increases the probability that a message goes undeleted, raises the probability of sender entry, and relaxes the sender's incentive constraint when $w = 0$. The platform therefore chooses the lowest feasible pooling level of moderation, $I = \bar{I}$. Since $1 - (1 - p)\bar{I} = 2p$, the best pooling profit is

$$\Phi_{pool}(\psi) := \max_{e \in \mathcal{E}} 2p(1 - \psi)G\big(2p[v + \psi A(e)] - e\big)A(e). \tag{OA10}$$

If $\mathcal{E}$ is empty, we adopt the convention that $\Phi_{pool}(\psi) = -\infty$.

iv-iii. Optimal moderation satisfies $I^* < 1$

The platform-optimal equilibrium is whichever of the separating and pooling equilibria yields higher profit. Maximized profit is therefore

$$\max\left\{\Phi_C(\psi), \Phi_{pool}(\psi)\right\}.$$

We are now ready to state the result:

Proposition OA5. *If the platform-optimal equilibrium is separating, every profit-maximizing moderation policy satisfies $I^* < I_\psi \leq 1$. If the platform-optimal equilibrium is pooling, the profit-maximizing moderation policy is $I^* = \bar{I} < 1$. Consequently, every platform-optimal outcome involves less than full moderation.*

Proof. Case 1: We begin with the separating equilibrium case. Suppose first that $e_L > 0$. Since e_L maximizes sender profit from transfers, $\psi A'(e_L) = 1$. For $I \geq I_\psi$, we have $e_H = e_L$ and $\zeta_C(I) = 0$. The platform's profit is

$$(1 - \psi)G(pv + U_\psi)A(e_L),$$

which is independent of I.

As I approaches I_ψ from below, e_H approaches e_L. The change in expected advertising revenue generated by an increase in ζ therefore converges to

$$(1 - I_\psi)A(e_L) - A(e_L) = -I_\psi A(e_L) < 0.$$

It follows that $\zeta_C(I) = 0$ for I close to but below I_ψ. Within this neighborhood, differentiating profit with respect to e_H and evaluating at $e_H = e_L$ gives

$$\begin{aligned}\left.\frac{\partial \Pi_C}{\partial e_H}\right|_{e_H = e_L} &= (1 - \psi)G'(pv + U_\psi)p\big[\psi A'(e_L) - 1\big]A(e_L) + (1 - \psi)G(pv + U_\psi)pA'(e_L) \\ &= (1 - \psi)G(pv + U_\psi)pA'(e_L) > 0,\end{aligned}$$

where the second equality follows from $\psi A'(e_L) = 1$.

For $I < I_\psi$, the sender's incentive constraint when $w = 0$ binds:

$$e_H = (1 - I)\big[v + \psi A(e_H)\big] - U_\psi.$$

Implicit differentiation gives

$$\frac{de_H}{dI} = -\frac{v + \psi A(e_H)}{1 - (1 - I)\psi A'(e_H)} < 0$$

for I sufficiently close to I_ψ. Hence a small reduction in I below I_ψ raises e_H, which we have seen strictly increases platform profit. Every $I \geq I_\psi$ is therefore dominated by some policy below I_ψ.

Case 2: Now continue with the separating equilibrium case but suppose $e_L = 0$. Then $U_\psi = 0$ and $I_\psi = 1$. At $I = 1$, we have $e_H = 0$, so platform profit is zero because $A(0) = 0$. For any I below but sufficiently close to one, $e_H > 0$, the sender's expected payoff gross of entry cost is close to $pv > \underline{c}$, and the separating equilibrium with $\zeta = 0$ generates strictly positive profit. The platform-optimal equilibrium does at least as well. Thus $I = 1$ is again dominated by $I^* < I_\psi$.

Case 3: Finally, consider a pooling equilibrium with $I > \bar{I}$. Holding effort fixed and reducing moderation to $\bar{I}$ yields $\beta = 1/2$ by definition of $\bar{I}$. It also relaxes the sender's incentive constraint when $w = 0$ and strictly reduces the probability of deleting the message. This increases platform profit whenever the original outcome earns positive profit. Since

a separating equilibrium permits strictly positive profit, any pooling equilibrium that is platform-optimal must itself earn positive profit. It follows that every platform-optimal pooling outcome has $I^* = \bar{I}$. □

v. Platform versus broadcast business models

The platform-optimal effort level maximizes $\Pi(e)$, given in (OA5). Let

$$e_P^* \in \arg\max_{e\in[0,v]} \Pi(e)$$

denote effort in a platform-optimal equilibrium. A broadcaster instead chooses

$$e^B = \arg\max_{e\geq 0}\{A(e) - e\},$$

which is interior under the assumptions made in the paper.

Lemma OA1. *If $e_P^* < \bar{e}$, so that the platform-optimal equilibrium is pooling, then*

$$e_P^* \geq e^B \quad \Longleftrightarrow \quad \frac{G'(2pv - e_P^*)}{G(2pv - e_P^*)} \leq \frac{1}{A(e_P^*)}. \tag{OA11}$$

If $e_P^* > \bar{e}$, so that the platform-optimal equilibrium is separating, then

$$e_P^* \geq e^B \quad \Longleftrightarrow \quad p\frac{G'(p(v - e_P^*))}{G(p(v - e_P^*))} \leq \frac{1}{A(e_P^*)}. \tag{OA12}$$

Proof. Suppose first that $e_P^* < \bar{e}$, so the platform-optimal equilibrium is pooling. The platform's first-order condition is

$$\frac{A'(e_P^*)}{A(e_P^*)} = \frac{G'(2pv - e_P^*)}{G(2pv - e_P^*)}.$$

If instead $e_P^* > \bar{e}$, the platform-optimal equilibrium is separating, and its first-order condition is

$$\frac{A'(e_P^*)}{A(e_P^*)} = p\frac{G'(p(v - e_P^*))}{G(p(v - e_P^*))}.$$

The broadcaster's interior first-order condition is

$$A'(e^B) = 1.$$

By the maintained concavity of A, $e_P^* \geq e^B \iff A'(e_P^*) \leq 1$. Substitution of $A'(e_P^*)$ from the relevant platform first-order condition gives (OA11) and (OA12). □

We can now show that the principal insight from the paper survives under platform-optimal equilibrium selection.

Proposition OA6. *Parts (ii) and (iii) of Proposition 11 continue to hold under platform-optimal equilibrium selection.*

Proof. For part (ii), fix an effort e feasible at some v. As v increases, profit from that effort is weakly increasing within either the pooling or separating region. If e moves from the separating region into the pooling region, the two profit expressions coincide at the boundary and pooling profit is thereafter weakly increasing in v. Since the feasible effort set $[0, v]$ also expands with v, maximized platform profit is weakly increasing in v, whereas broadcaster profit is independent of v. Hence, if the platform model is preferred initially, it continues to be preferred after an increase in v.

For part (iii), replacing G with a pointwise higher distribution $\hat{G}$ weakly raises platform profit at every feasible effort, while leaving broadcaster profit unchanged. Evaluating platform profit at an effort that was optimal under G, and then allowing the platform to re-optimize, proves the final claim. □

The participation–effort trade-off governs organizational form whether the platform-optimal equilibrium is pooling or separating. Greater effort raises advertising revenue but discourages sender entry. When pooling, increasing effort by one unit lowers the sender's gross expected payoff by one unit. In the semi-separating equilibrium, that payoff falls by p: when $w = 0$, the sender remains indifferent between transmitting the two equilibrium signals.

vi. AI-generated content

As in the paper, producing effort e without AI costs e, whereas producing it with AI costs $k(e)$, where $k(0) = 0$ and $0 < k'(e) < 1$. With probability $h \in [0, 1)$, AI-generated content is garbled. Garbled content neither persuades the receiver nor generates advertising revenue. We assume that a non-garbled AI output is observationally identical to content produced without AI with the same effort.

Lemma OA2. *Fix a moderation policy I and an observable effort e. Whenever a sender with $w = 1$ prefers AI production to production without AI, a sender with $w = 0$ also weakly prefers AI production. Moreover, the $w = 0$ sender strictly prefers AI while the $w = 1$ sender strictly prefers production without AI whenever $h(1 - I)v < e - k(e) < hv$.*

Proof. A sender with $w = 1$ obtains $v - e$ from production without AI and $(1 - h)v - k(e)$ from AI production. She therefore prefers AI if and only if $e - k(e) \geq hv$. In either a pooling or separating equilibrium, $\rho(s) \in \{0, 1\}$. A sender with $w = 0$ who attempts to induce $r = 1$ obtains $(1-I)v-e$ from production without AI and $(1-I)(1-h)v-k(e)$ from AI production. She prefers AI if and only if $e - k(e) \geq h(1 - I)v$. Because $I \geq 0$, the latter condition is weaker. The final claim follows from the strict versions of the two inequalities. □

The reason for this ordering is that hallucination is less costly to a sender whose content is already at risk of deletion. AI may therefore be used only for attempted deception, whereas the reverse configuration—AI use only when $w = 1$—cannot occur. Given a separating equilibrium, this reproduces the basic mechanism in the main paper. If only the $w = 0$ sender

uses AI, the platform must deter a cheaper AI-assisted imitation. The minimum moderation needed to implement effort e is

$$\max\left\{1-\frac{e}{v}, 1-\frac{k(e)}{v(1-h)}\right\},$$

which is weakly greater than the moderation needed without AI. When the AI-assisted constraint binds, the platform-optimal separating equilibrium gives profit

$$G\big(p(v-e)\big)A(e)\left[p+\min\left\{p, \frac{(1-p)k(e)}{v}\right\}\right].$$

Without AI, platform-optimal separating profit at the same effort is

$$G\big(p(v-e)\big)A(e)\left[p+\min\left\{p, \frac{(1-p)e}{v}\right\}\right].$$

Since $k(e) < e$, using AI only when $w = 0$ cannot raise profit within the separating equilibrium. It makes imitation cheaper, but the additional moderation needed to preserve receiver trust offsets this advantage.

Pooling creates a different possibility.

Proposition OA7. *Suppose $h < \bar{I}$ and there is an effort $e \in (0, \bar{e}]$ satisfying*

$$\frac{h\bar{e}}{1-h} < e - k(e) < hv \qquad \text{and} \qquad 2pv - pe - (1-p)k(e) > \underline{c}.$$

There is a pooling equilibrium at the moderation policy

$$I = 1 - \frac{1-\bar{I}}{1-h} < \bar{I},$$

in which the $w = 1$ sender produces without AI and the $w = 0$ sender uses AI.

Proof. Suppose the $w = 1$ sender produces the pooled signal without AI while the $w = 0$ sender uses AI. Conditional on observing a surviving, non-garbled signal, the receiver's posterior is

$$\beta = \frac{p}{p + (1-p)(1-I)(1-h)}.$$

The minimum moderation needed to make this posterior at least one half therefore satisfies

$$(1-I)(1-h) = \frac{p}{1-p} = 1 - \bar{I}.$$

This gives the stated moderation policy. The condition $h < \bar{I}$ ensures that it lies strictly between zero and $\bar{I}$.

At this policy, a $w = 0$ sender prefers AI to production without AI if and only if

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

A $w = 1$ sender prefers production without AI to using AI if and only if $e - k(e) \leq hv$. The assumed strict inequalities therefore produce state-contingent technology choices. Because $e \leq \bar{e}$ and $k(e) < e$, both states also prefer the pooled signal to obtaining a zero continuation payoff.

The probability that content survives and $w = 1$ is p. The probability that surviving content is non-garbled and $w = 0$ is $(1 - p)(1 - I)(1 - h) = p$. Total monetizable content therefore occurs with probability $2p$, exactly as in the no-AI pool at $I = \bar{I}$.

Gross expected sender surplus in the state-contingent AI pool is

$$p(v - e) + (1 - p)\left[(1 - \bar{I})v - k(e)\right] = 2pv - pe - (1 - p)k(e).$$

This is larger than $\underline{c}$ by hypothesis, so the sender enters with positive probability. In a no-AI pool at the same effort, gross surplus is $2pv - e$. The difference is

$$(1 - p)[e - k(e)] > 0.$$

Platform profit consequently rises because the sender enters with higher probability. □

Relative to a no-AI pool at the same effort, this equilibrium has the same mass of surviving content but strictly greater sender entry. It therefore gives the platform strictly greater profit than the corresponding no-AI pool. In particular, if the platform-optimal no-AI equilibrium is a pool whose effort satisfies the displayed inequalities, AI strictly raises maximized platform profit even though it is used only when $w = 0$. The mechanism is that hallucination partially substitutes for platform moderation. When some false AI content is garbled, the platform can reduce I while leaving the surviving pool equally credible. Some unreliability can therefore be compatible with greater platform profit.

When $h = 0$, both states prefer AI at every positive effort, and the usual cost-saving benefit of reliable AI remains. As h rises, however, the equilibrium may move first to adoption only when $w = 0$ and eventually to no AI adoption. Platform profit need not be monotone in h.