Robust Content Moderation: Theory and Applications

Scott Duke Kominers
Harvard University and a16z crypto

Jesse M. Shapiro
*Harvard University and NBER**

December 2025

Abstract

A sender sends a signal about a state to a receiver who takes an action that determines a payoff. A moderator can block, flag, or even modify the sender's signal before it reaches the receiver—but the moderator cannot control how the receiver will act on what they see, or what the sender will try to send. We find that the only way that the moderator can robustly change the outcome is by removing information. We discuss applications to social media content moderation, state censorship, "flooding the zone," and AI filtering.

JEL Classification: D47, D82, D83, L82, L86

Keywords: Blackwell, misinformation, platforms

*This work subsumes and replaces an earlier paper by the authors entitled "Content Moderation with Opaque Policies." We thank Shai Bernstein, Henrique De Oliveira, Robin Greenwood, Rafael Jiménez-Durán, Miles Jennings, Navin Kartik, Daniel Kornbluth, Sriram Krishnan, Mohamed Mostagir, Alex Nichifor, Leah Plunkett, Marco Reuter, Tim Roughgarden, Suproteem Sarkar, Ludvig Sinander, Greg Taylor, Alex Teytelboym, Liang Wu, Refine.ink, and seminar audiences at Harvard University, Carnegie Mellon University, the University of Pittsburgh, the Marketplace Innovation Workshop, and the University of Chicago for helpful comments. Kominers gratefully acknowledges support from the Washington Center for Equitable Growth, as well as the Ng Fund and the Mathematics in Economics Research Fund of the Harvard Center of Mathematical Sciences and Applications. Part of this work was conducted during the Simons Laufer Mathematical Sciences Institute Fall 2023 program on the Mathematics and Computer Science of Market and Mechanism Design, which was supported by the National Science Foundation under Grant No. DMS-1928930 and by the Alfred P. Sloan Foundation under grant G-2021-16778. Shapiro thanks his dedicated research assistants for their contributions to this project, and acknowledges funding from the Semester Undergraduate Program for Economics Research at Harvard. Kominers is a Research Partner at a16z crypto, which reviewed a draft of this article for compliance prior to publication and is an investor in various online platforms, including social media platforms (for general a16z disclosures, see https://www.a16z.com/disclosures/). Notwithstanding, the ideas and opinions expressed herein are those of the authors, rather than of a16z or its affiliates. Kominers also holds digital assets, including both fungible and non-fungible tokens, and advises a number of companies on marketplace and incentive design, including koodos and Quora. Any errors or omissions remain the sole responsibility of the authors. *E-mails: kominers@fas.harvard.edu · jesse_shapiro@fas.harvard.edu.*

1 Introduction

Modern social media platforms can block, flag, mute, or throttle content on a massive scale. Yet the impact of these seemingly vast powers depends on how users respond. Many users do not adhere to the platforms' policies; some deliberately try to subvert them.[1] Many users say they do not trust the platforms (Kemp and Ekins, 2021; Cato Institute, 2021) or their owners (Edelman Trust Institute, 2024, p. 45); a majority say that algorithms are censoring viewpoints (Rainie et al., 2022; Pew Research Center, 2022) and that fact-checkers are influenced by their own political opinions (Kemp and Ekins, 2021; Cato Institute, 2021). Maybe the platforms are not so powerful, after all.

Already an important problem, content moderation is likely to become even more so as ever larger shares of human communication move to digital platforms. Moreover, content moderation arises in settings beyond social media platforms, for example in authoritarian regimes where censors attempt to moderate the entire digital realm. In these and other settings, the moderator can directly influence the content that people see, but cannot directly control what they believe, and cannot preclude attempts to subvert the moderator's aims.

In this paper, we study the problem of a moderator who can block, flag, or even modify any signal sent by a sender to a receiver, but who cannot directly control what the sender tries to send, or how the receiver acts on the messages they do (or do not) receive. We ask what the moderator can achieve robustly, regardless of what the sender tries to send, or how the receiver chooses to act. We show that the moderator can robustly change the distribution of the receiver's actions only by removing information from the sender's signal. We show why robustness is a sensible criterion, and we characterize when the moderator would want to remove information.

Our setup is simple: There is a state of the world, and a sender who may (or may not) know the state attempts to send a signal to a receiver. A moderator, who may also know the state, chooses between a default policy—say, one that lets the original signal pass through—and a moderation policy—say, one that blocks certain signals. A receiver who does not know the state receives the moderated message and takes a payoff-relevant action. We compare the outcome—i.e., the distribution of the receiver's action conditional on the state—under a given moderation policy to that which would arise

[1] For example, in the second quarter of 2023, Meta reports taking action on 13.6 million pieces of content related to terrorism and 1.1 million pieces related to organized hate on Facebook (Meta, 2023). For examples of users subverting content moderation policies, see Busby (2024), Oversight Board (2024, p. 12), and Institute for Strategic Dialogue (2025).

under the default policy.

We allow for a wide range of moderation techniques, including blocking signals, erasing them, flagging them as false, and even secretly modifying them. We say that a moderation policy *removes information* from the sender's signal if the information about the state contained in the unmoderated message cannot be reconstructed from the moderated message.

We begin by fixing the sender's strategy and examining whether, and when, a moderation policy can change the outcome robustly, i.e., in a way that cannot be undone by a different strategy on the part of the receiver. The most basic answer to this question follows directly from well-known ideas in the literature on comparisons of experiments (Blackwell, 1951, 1953). Specifically: *a moderation policy robustly changes the outcome only if the moderation policy removes information from the sender's signal*, i.e., if the unmoderated message distribution is not a garbling of the moderated message distribution.

Although immediate from classic results, we show that this basic finding has surprisingly rich implications for content moderation. Removing information requires removing (some of the) information about the state contained in the unmoderated message, because only by removing such information is it possible to robustly change the mapping from states to actions. Removing information about the state is not, in general, equivalent to removing the message itself. As a result, even heavy-handed moderation policies such as completely blocking the sender's signal can fail to be robust; for example if the sender's signal is uninformative about the true state.

We then go beyond this basic conclusion to develop foundations and implications that are not immediate from classic results.

Our next set of findings show that receivers' intentions motivate a concern with robustness. Specifically: Unless the receiver believes that the moderation policy introduces information about the state, then the receiver should weakly prefer the unmoderated outcome over any outcome available under the moderation policy. Therefore, the receiver should (at least weakly) prefer to restore the unmoderated outcome when possible. The only way for the moderator to prevent this behavior is to make it impossible for the receiver to restore the unmoderated outcome, which in turn requires removing information. Importantly, this argument concerns not whether the moderation policy introduces information about the state, but whether the receiver *believes* that it does. We show a sense in which any moderation policy, even one that does introduce information about the state, can be interpreted by an uninformed receiver as if it does not introduce infor-

mation about the state.

The findings just described establish a sense in which robustness is important for preventing undesirable outcomes. We next study when it is possible to prevent undesirable outcomes without also preventing desirable ones. We endow the moderator with preferences over outcomes summarized in a payoff function, and consider the set of moderator payoffs that are possible under a given moderation policy and any receiver strategy. We ask when the moderator can eliminate (otherwise-possible) payoffs below some level while still preserving payoffs above that level. We show that this is possible only if the moderator does not prefer more informative outcomes, in the sense that the high-payoff outcomes do not contain more information about the state than low-payoff outcomes. Naturally, this property requires that the receiver is willing to take actions that are not optimal with respect to the moderator's payoff, which could either be because the moderator and receiver do not share the same preferences, or because they do not share the same beliefs.

In practice, not only how the receiver acts, but also what the sender sends, may change in response to the moderation policy. We therefore extend the concepts from the case of a fixed sender strategy to study when the moderation policy can change the outcome in a way that cannot be undone by different strategies on the part of the receiver *or* the sender. We find that this form of robust moderation is possible only if the moderation policy removes information *from any possible signal the sender may send*, in the sense that the information about the state contained in the unmoderated message cannot be reconstructed from the moderated form of any available signal.

In addition to a given sender directly changing their strategy in response to the moderation policy, another possibility is that the receiver changes which sender they choose to listen to. We find that, if the receiver believes that the moderation policy does not introduce information about the state, then the receiver should weakly prefer to restore the unmoderated outcome, and should be willing to switch senders in order to do so. Because switching senders amounts to changing the sender's strategy, and because we again find that any moderation policy can be interpreted by an uninformed receiver as if it does not introduce information about the state, our analysis motivates a concern with robustness to changing sender strategies whenever the receiver does not have a basis for trusting the moderator. We moreover show that our characterization of when the moderator can prevent undesirable outcomes without preventing desirable ones carries over to the many-sender setting.

We illustrate throughout with two stylized running examples.

Example 1 (Instructions for building a bomb)**.** In our first running example, the sender tries to convey instructions for building a bomb, and the receiver tries to build a bomb. If the receiver truly wishes to build a bomb, then the only way to robustly prevent that outcome is to make it impossible, by ensuring that the receiver does not learn how to build one. Doing so, in turn, requires removing the information about how to build a bomb from the sender's signal, and from any other signal that the sender could send (or that the receiver could choose to receive). This requires, for example, removing information not only from instructions transmitted in a literal way, but also from instructions transmitted in disguise, as with a cipher. If knowing how to build a bomb can only cause harm, then policies that remove information, for example by erasing any signals that contain bomb-making instructions, can only eliminate bad outcomes. On the other hand, if the information that can be used to build a bomb can also be used to, say, make a helpful fertilizer, then any policy that eliminates bad outcomes (i.e., bomb-building) must also eliminate some good ones (i.e., fertilizer-making). △

Example 2 (Vaccine safety claim)**.** In our second running example, the sender conveys a vaccine safety claim, and the receiver decides whether to get vaccinated. If the sender's claim contains no information about vaccine safety, then there is (by definition) no way for the moderator to remove information from the sender's signal, and therefore no way to prevent the receiver from acting (as if based) on the sender's claim. Moreover, if the receiver believes that the sender has valuable information, and that the moderator does not, then the receiver prefers to act on the sender's claim when possible. Under these conditions, then, robust moderation is infeasible; i.e., there is no way for the moderator to robustly change the receiver's action. If, on the other hand, the sender's claim does contain information about vaccine safety—for example about the findings of legitimate scientists whom the sender regards as biased—then the moderator may be able to prevent bad outcomes by removing that information, but only at the expense of also preventing any good outcomes that use the same information. △

In further extensions, we consider two additional goals a moderator might have: First, we examine the possibility that a moderator may want to enable an outcome that is not possible under the unmoderated message. We show that doing so requires adding information that is not present in the unmoderated message. We argue that this criterion makes sense in situations where the receiver trusts that the moderator has special information that cannot be obtained from any sender. Second, we examine the possibility that a moderator may care not only about the information content of a

message but also the form in which it is conveyed. We show that this concern can be motivated by the use of a well-understood (and trusted) language, and leads to different, non-informational considerations for the moderator.

In an application to social media, we introduce an original timeline of major platforms' content moderation policies. We note that some of the earliest and most durable policies work precisely by removing information. By contrast, some more recent, and more short-lived, policies do not remove information. We relate our theoretical findings to the fates and effectiveness of these policies, and we contrast the situation of social media with that of mass media.

In an application to state censorship, we argue that many of the favored tactics of authoritarian regimes can be understood as removing information about a payoff-relevant state. We also note that policy subversion of exactly the sort that arises in our model (and running examples) is a prominent phenomenon in censored realms. Last, we note that the modern public relations strategy of "flooding the zone" (Bannon, quoted by Lewis (2018)) can be cast as a form of robust content moderation when the receiver has finite attention.

We conclude with an application to AI filtering, in which we argue that, while in some respects it is more difficult to moderate content from an algorithmic sender, in other respects it is easier.

Our paper contributes, most broadly, to the vast research work on strategic communication (Farrell and Rabin, 1996; Crawford, 1998; Sobel, 2013) and information design (Bergemann and Morris, 2019; Kamenica, 2019). Our setup differs from canonical frameworks in the literature in three main ways:

1. Rather than communicating directly, our moderator must choose whether, and how, to interfere with a sender's signal.

2. Rather than judging the outcome in absolute terms, we judge the moderator's impact relative to the default outcome that would have occurred had the moderator left the signal alone.

3. Rather than fully specifying payoffs and adopting a notion of equilibrium, we characterize properties that hold robustly across a range of strategies for both the receiver and the sender.

Although these elements appear, respectively, in existing research on mediated communication (Ivanov, 2010; Ambrus, Azevedo, and Kamada, 2013; Salamanca, 2021),

regret-minimizing persuasion (Castiglioni et al., 2020; Babichenko et al., 2022), and robust persuasion (Hu and Weng, 2021; Kosterina, 2022; Dworczak and Pavan, 2022), we are not aware of prior work that combines them as we do here. We believe that the combination of modeling elements we pursue here is an especially good fit to the important content moderation settings that we consider as applications.[2]

Our paper also contributes to a smaller, but growing, theory literature specifically addressing social media content moderation. One focus of this literature is on the effect of moderation on other business considerations, such as advertising revenue (Liu, Yildirim, and Zhang, 2022; Madio and Quinn, 2025), content quality (Chang, Segura, and Zhang, 2024), and/or user participation and engagement (Candogan and Drakopoulos, 2020; Papanastasiou, 2020; Dwork et al., 2024; Bar-Isaac, Deb, and Mitchell, 2025; Hojati and Nault, forthcoming). Another focus, more closely related to ours, is on the potential for a platform to limit the spread of falsehoods. Candogan and Drakopoulos (2020), Yang, Li, and Zhu (2023), Hossain et al. (2024), and Ng and Taylor (2025), for example, consider the situation in which a platform can discourage the spread of false signals by credibly revealing information about whether a particular signal is true or by removing signals that are false (see also Papanastasiou (2020)). Jackson, Malladi, and McAdams (2022) study the problem of limiting depth and breadth of content-sharing in a setting where content can mutate as it circulates. Mostagir and Siderius (2022) and Acemoglu, Ozdaglar, and Siderius (2024) highlight potential downsides from interventions to reduce the spread of false information due to users' Bayesian inferences (see also Mostagir and Siderius (2023)). Though our work is broadly related to these other works, our setup, questions, and findings are different. In particular, we are not aware of prior work that studies the robustness of moderation policies in the sense that we consider here.

Our work has connections to other literatures as well, including theoretical research on the comparison of experiments, empirical research on the effectiveness of content moderation, and theoretical and empirical research on state censorship; we highlight these connections throughout the body of the paper.

The remainder of the paper proceeds as follows. Section 2 lays out our notation, establishes some preliminary definitions and facts about information orderings, and initializes our running examples. Section 3 studies robust content moderation when the signal structure is fixed. Section 4 studies robust content moderation when the signal

[2]As an instructive example of why these modeling elements matter: Whereas we find that a moderator can change the outcome robustly only by removing information, Gentzkow and Kamenica (2017) find that a sender in a multi-sender persuasion game can change the outcome unilaterally only by *adding* information.

structure may change in response to the moderation policy. Section 5 discusses applications, including our novel analysis of the timeline of social media content moderation policies. Proofs omitted from the main text are presented in Appendix A.

2 Setup and Preliminaries

In this section we lay out our setup, notation, and main definitions. We then recall some preliminaries on information orderings that are useful in what follows. Along the way, we introduce our running examples.

2.1 Notation and Timing

Nature determines a state $\theta \in \Theta$. A sender may or may not observe the state, and then chooses a signal $s \in \mathcal{S}$ to transmit. A moderator observes the state θ and the signal s, and then chooses a message $m \in \mathcal{M}$ to transmit to a receiver. The receiver observes only the message m, and then chooses an action $a \in \mathcal{A}$.

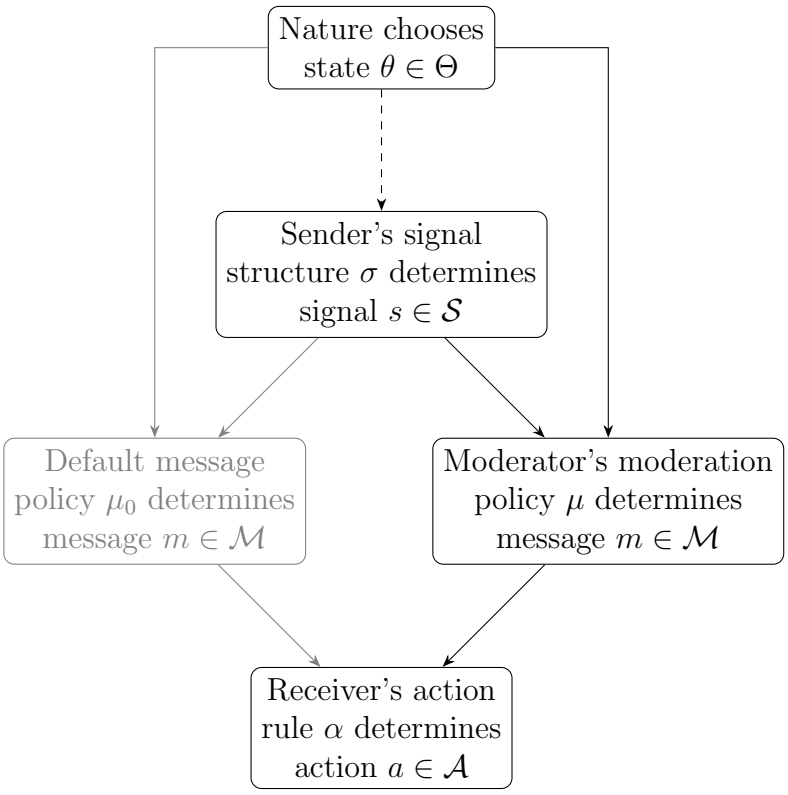
Sets of possible behaviors by the sender and receiver play an important role in our analysis. Accordingly, let a **signal structure** be a mapping $\sigma : \Theta \to \Delta(\mathcal{S})$ from states to (distributions of) signals, with $\overline{\mathscr{S}}$ denoting the set of all signal structures, and $\mathscr{S} \subseteq \overline{\mathscr{S}}$ denoting a generic set of signal structures. Likewise, let an **action rule** be a mapping $\alpha : \mathcal{M} \to \Delta(\mathcal{A})$ from messages to (distributions of) actions, with $\overline{\mathscr{A}}$ denoting the set of all action rules.

To define the tool available to the moderator, we say that a **moderation policy** is a mapping $\mu : \left[\Theta \times \mathcal{S}\right] \to \Delta(\mathcal{M})$ from (state, signal) pairs to (distributions of) messages. We assume there is a well-defined **default messaging policy** $\mu_0 : \left[\Theta \times \mathcal{S}\right] \to \Delta(\mathcal{M})$ that we can loosely think of as reflecting the case of an "unmoderated message." A leading case is where $\mathcal{S} \subseteq \mathcal{M}$ and μ_0 simply transmits the signal. Another case is where μ_0 represents a default algorithm for filtering content, which may transmit the signal (with some probability) or may instead transmit some other piece of content.

Figure 1 summarizes the timing and main definitions.

Example 1, Part i (Instructions for building a bomb, setup)**.** The state $\theta_{💣} \in \Theta_{⚒}$ is the correct way to build a bomb, with $|\Theta_{⚒}| \geq 2$. The sender is a social media user who knows how to build a bomb. The sender can transmit some (possibly incorrect) instructions, which we can think of as a signal s in a space $\mathcal{T} \cong \Theta_{⚒}$ that is isomorphic

Figure 1: Summary of Timing and Main Definitions



to the state space. Or the sender can remain silent, which we can think of as a signal $s = \bigcirc$, so that the signal space is $\mathcal{S} \coloneqq \mathcal{T} \cup \{\bigcirc\}$.

The receiver is another user who can try (and perhaps fail) to build a bomb, which we can think of as an action a in a space $\mathcal{B} \cong \Theta_{\bigcirc}$ that is isomorphic to the state space. The receiver can alternatively do nothing, which we can think of as an action $\emptyset$, so that the action space is $\mathcal{A} \coloneqq \mathcal{B} \cup \{\emptyset\}$.

The moderator is the social media platform. The default messaging policy μ_0 passes the signal through, sending $m = s$. The moderator can alternatively erase the signal, sending $m = \bigcirc$. Or the moderator can explicitly block the signal, sending $m = \bullet$, so that the receiver will know that the moderator intervened. Therefore, the message space is $\mathcal{M} \coloneqq \mathcal{S} \cup \{\bullet\}$.

One available moderation policy, which does not depend directly on the true state, is to erase instructions, $\mu(\theta_{\bullet}, s) = \bigcirc$ for all $s \in \mathcal{S}$, or to explicitly block them, $\mu(\theta_{\bullet}, s) = s$ if $s = \bigcirc$ and $\mu(\theta_{\bullet}, s) = \bullet$ otherwise. Another available moderation policy, which depends directly on the true state, is to erase only correct instructions, $\mu(\theta_{\bullet}, s) = \bigcirc$ when $s \cong \theta_{\bullet}$ and $\mu(\theta_{\bullet}, s) = s$ otherwise. $\triangle$

Example 2, Part i (Vaccine safety claim, setup)**.** The state $\theta_{\text{safety}} \in \Theta_{\text{safety}}$ is the extent of the scientific evidence that a particular vaccine is safe. The sender is a social media user who may or may not be informed of θ_{safety} and can make a claim about it, say $s \in \mathcal{S} \cong \Theta_{\text{safety}}$. The receiver is another user who chooses whether or not to get vaccinated, denoted by $a \in \{\mathsf{no}, \mathsf{yes}\} =: \mathcal{A}$.

The moderator (platform) can choose to pass the sender's signal through, sending $m = s$, or to block it, sending $m = \bullet$. Or the moderator can choose to flag the signal as false, sending $m = (s, \mathfrak{false})$, so that $\mathcal{M} := \mathcal{S} \cup \{\bullet\} \cup \big[\mathcal{S} \times \{\mathfrak{false}\}\big]$. Moderation policies can depend directly on the true state, for example blocking or flagging false claims and leaving true claims alone. $\triangle$

2.2 Preliminaries on Information Orderings

Our results rely on standard information orderings (Blackwell, 1951, 1953). As notation, for stochastic maps $\nu : \mathcal{X} \to \Delta(\mathcal{Y})$ and $\zeta : \mathcal{Y} \to \Delta(\mathcal{Z})$, and any $x \in \mathcal{X}$, we let $\big[\zeta \circ \nu\big](x) \in \Delta(\mathcal{Z})$ denote the distribution on $\mathcal{Z}$ found by passing draws from $\nu(x)$ through ζ, i.e., $\big[\zeta \circ \nu\big](x) \equiv \mathrm{E}_{y \sim \nu(x)}[\zeta(y)]$.

We say that:

- $\nu : \mathcal{X} \to \Delta(\mathcal{Y})$ **garbles** $\nu' : \mathcal{X} \to \Delta(\mathcal{Y})$ if there is a stochastic map $\Gamma : \mathcal{Y} \to \Delta(\mathcal{Y})$ such that $\nu = \Gamma \circ \nu'$.

If ν garbles ν', we may alternatively say that ν' is (weakly) more informative than ν, in the sense that the information about $x \in \mathcal{X}$ contained in the output of ν' can alternatively be obtained from the output of ν. Accordingly, we say that:

- $\nu : \mathcal{X} \to \Delta(\mathcal{Y})$ is **uninformative** if ν garbles all $\nu' : \mathcal{X} \to \Delta(\mathcal{Y})$.
- $\nu' : \mathcal{X} \to \Delta(\mathcal{Y})$ is **maximally informative** if all $\nu : \mathcal{X} \to \Delta(\mathcal{Y})$ garble ν'.
- $\nu : \mathcal{X} \to \Delta(\mathcal{Y})$ and $\nu' : \mathcal{X} \to \Delta(\mathcal{Y})$ are **informationally equivalent** if ν garbles ν' *and* ν' garbles ν.
- $\nu : \mathcal{X} \to \Delta(\mathcal{Y})$ and $\nu' : \mathcal{X} \to \Delta(\mathcal{Y})$ are **informationally equivalent under** $\zeta : \mathcal{Y} \to \Delta(\mathcal{Z})$ if $\zeta \circ \nu$ and $\zeta \circ \nu'$ are informationally equivalent.

We make use of a form of the Blackwell Informativeness Theorem, which we reproduce here as a lemma.

Lemma 1 (Blackwell Equivalence)**.** *If ν garbles ν' then we have*

$$\left\{\zeta \circ \nu : \zeta \in \Delta(\mathcal{Z})^{\mathcal{Y}}\right\} \subseteq \left\{\zeta' \circ \nu' : \zeta' \in \Delta(\mathcal{Z})^{\mathcal{Y}}\right\} \tag{1}$$

for any $\mathcal{Z}$.

Moreover, if the spaces $\mathcal{X}$ and $\mathcal{Y}$ are Polish, (1) holds for any Polish space $\mathcal{Z}$, and a common measure dominates $\nu'(x)$ for all $x \in \mathcal{X}$, then ν garbles ν'.

Proof. For the first part, suppose that ν garbles ν'. Then (definitionally), there exists $\Gamma : \mathcal{Y} \to \Delta(\mathcal{Y})$ such that $\nu = \Gamma \circ \nu'$. Now, for any $\zeta \in \Delta(\mathcal{Z})^{\mathcal{Y}}$, we have $\zeta' := \left[\zeta \circ \Gamma\right] \in \Delta(\mathcal{Z})^{\mathcal{Y}}$ (because Γ is a stochastic map on $\mathcal{Y}$); and moreover (because $\nu = \Gamma \circ \nu'$), we have

$$\zeta \circ \nu = \zeta \circ (\Gamma \circ \nu') = (\zeta \circ \Gamma) \circ \nu' = \zeta' \circ \nu'.$$

The preceding observations together show that for any $\zeta \in \Delta(\mathcal{Z})^{\mathcal{Y}}$, we must have

$$\left[\zeta \circ \nu\right] \in \left\{\zeta' \circ \nu' : \zeta' \in \Delta(\mathcal{Z})^{\mathcal{Y}}\right\};$$

this proves (1).

For the second part, see Appendix A. □

Remark 1. Any countable space is Polish under the discrete metric. Additionally, if $\mathcal{Y}$ is countable, then the condition in Lemma 1 requiring a dominating measure holds automatically (see Khan, Yu, and Zhang (2024, p. 4)).

The first part of the proof of Lemma 1 is elementary (and essentially immediate). For the second part, we require regularity conditions. The proof in Appendix A leverages the sufficient conditions of Khan, Yu, and Zhang (2024); for a self-contained proof in the case of finite spaces, see, for example, de Oliveira (2018).

3 Robust Content Moderation with a Fixed Sender

In this section we study robust content moderation with a fixed signal structure σ. First, we define robustly effective moderation and characterize when and how it can be achieved. Next, we formalize a simple justification, in terms of the intentions of the receiver, for being concerned with robustly effective moderation. Then, we introduce a payoff function for the moderator and characterize when the moderator can eliminate un-

desirable outcomes while preserving desirable ones. Lastly, we discuss why a moderator concerned with modes of expression, rather than actions, need not remove information.

3.1 Robustly Effective Moderation Requires Removing Information

Our first question is whether, and when, the moderator can guarantee an outcome different from the one that would result from the unmoderated message. In our setting, an outcome is a distribution of actions conditional on the state. Specifically, any given signal structure–moderation policy–action rule tuple (σ, μ, α) induces an **outcome** $\Xi(\sigma, \mu, \alpha) \in \left[\Delta(\mathcal{A})\right]^{\Theta}$ given by the stochastic map $\Xi(\sigma, \mu, \alpha) \equiv \alpha \circ \mu \circ \sigma$, where, here and throughout, we ease notation by suppressing the dependence of μ on θ. The preceding observation leads naturally to the following definitions.

Definition. A moderation policy μ is **effective for signal structure** σ **and action rule** α if $\Xi(\sigma, \mu, \alpha) \neq \Xi(\sigma, \mu_0, \alpha)$.

If, furthermore, there is no $\alpha' \in \overline{\mathscr{A}}$ such that $\Xi(\sigma, \mu, \alpha') = \Xi(\sigma, \mu_0, \alpha)$, then we say that μ is **robustly effective for signal structure** σ **and action rule** α. If this is the case for *some* action rule α, we say that μ is **robustly effective for signal structure** σ.

For a given signal structure and action rule, a moderation policy is effective if it changes the outcome relative to the default messaging policy. If a moderation policy is effective for some action rule, and no other action rule recovers the outcome that would have arisen under the default, then the policy is robustly effective.

Our first result is that robustly effective moderation requires removing information about the state. To express this result, we must define what it means to remove information.

Definition. A moderation policy μ **removes information from signal structure** $\sigma \in \overline{\mathscr{S}}$ if $\mu_0 \circ \sigma : \Theta \to \Delta(\mathcal{M})$ does not garble $\mu \circ \sigma : \Theta \to \Delta(\mathcal{M})$; i.e., if $\mu \circ \sigma$ is not (weakly) more informative than $\mu_0 \circ \sigma$.

A moderation policy removes information from a given signal structure if the message distribution under the policy cannot be transformed, without direct knowledge of the state, to recover the default message distribution. In such a case, the default messaging policy reveals some information about the state that is not recoverable under the given moderation policy.

We can now state our first result.

Proposition 1 (Robustly effective moderation requires removing information)**.** *A moderation policy μ is robustly effective for signal structure σ only if it removes information from signal structure σ.*

Proof. Towards proof of the contrapositive, suppose that μ does not remove information from σ. Then $\mu_0 \circ \sigma$ garbles $\mu \circ \sigma$, which by the first part of Lemma 1—taking $\nu = \mu_0 \circ \sigma$ and $\nu' = \mu \circ \sigma$—implies that for any $\alpha \in \overline{\mathscr{A}}$, there is some $\alpha' \in \overline{\mathscr{A}}$ such that $\alpha' \circ \big[\mu \circ \sigma\big] = \alpha \circ \big[\mu_0 \circ \sigma\big]$. Hence, we see that μ is not robustly effective for σ. □

Proposition 1 states that a robustly effective policy must remove information about the state. Intuitively, if the policy does not do so, then the receiver can always recover the distribution of actions they had available under the unmoderated message.

Example 1, Part ii (Instructions for building a bomb, robustly effective moderation)**.** Consider the signal structure $\sigma_\dagger$ that always transmits the correct instructions, i.e., $\sigma_\dagger(\theta_{💣}) \cong \theta_{💣}$ for all $\theta_{💣} \in \Theta_{⚒}$. Imagine the action rule $\alpha_\dagger$ that attempts to build a bomb if the message contains instructions, $\alpha_\dagger(m) \cong m$ if $m \in \mathcal{T}$, and does not attempt to build a bomb otherwise, $\alpha_\dagger(m) = \emptyset$ if $m \notin \mathcal{T}$. Under the policy μ_0 that simply passes along the sender's signal "as is," if the receiver uses $\alpha_\dagger$, then the receiver always (i.e., in all states) builds the bomb.

Now imagine the alternative policy $\mu_{\bigcirc}$ that erases all instructions, $\mu_{\bigcirc}(\theta_{💣}, s) = \bigcirc$ for all $\theta_{💣} \in \Theta_{⚒}, s \in \mathcal{S}$. Under the policy $\mu_{\bigcirc}$, if the receiver uses $\alpha_\dagger$, then the receiver never builds the bomb. Moreover, under the policy $\mu_{\bigcirc}$, there is *no* action rule $\alpha \in \overline{\mathscr{A}}$ under which the receiver always builds the bomb. Therefore, $\mu_{\bigcirc}$ is robustly effective for $\sigma_\dagger$.

The policy $\mu_{\bigcirc}$ removes information from $\sigma_\dagger$, as there is no way to extract the correct instructions from the messages moderated under $\mu_{\bigcirc}$. Via Proposition 1, this property explains why the policy $\mu_{\bigcirc}$ can be robustly effective. The policy removes information about the state, making it impossible for a receiver to build the bomb with probability 1 in all states.

Not all policies that are effective for $\sigma_\dagger$ and $\alpha_\dagger$ are robustly effective for $\sigma_\dagger$. For example, consider the state-dependent policy that blocks the signal, transmitting $m = \bullet$, if and only if the state $\theta_{💣}$ takes on a particular value $\hat{\theta_{💣}}$, and otherwise passes the signal through. That policy is effective for $\sigma_\dagger$ and $\alpha_\dagger$, but is not robustly effective for $\sigma_\dagger$. The reason is that there is another action rule, one that differs from $\alpha_\dagger$ only in choosing $a = \hat{a} \cong \hat{\theta_{💣}}$ whenever the signal is blocked, under which the receiver always builds the bomb. Notice that the policy does not remove information from $\sigma_\dagger$. △

A moderation policy that is robustly effective for a given signal structure and some action rule need not be robustly effective for that same signal structure and another action rule. Indeed, for certain action rules, no moderation policy is robustly effective, regardless of the signal structure.

Remark 2 (No robustly effective moderation of a constant action rule)**.** Consider an action rule α that is **constant** in the sense that it prescribes the same distribution over actions regardless of the message, i.e., $\alpha(m) = \alpha(m')$ for all $m, m' \in \mathcal{M}$. Then, regardless of the signal structure, no moderation policy is robustly effective for α, because (definitionally) the outcome associated with the given action rule does not depend on the message distribution.

Example 1, Part iii (Instructions for building a bomb, constant action rule)**.** Consider the action rule $\hat{\alpha}$ that chooses $a = \hat{a} \cong \hat{\theta_{\bullet}}$ regardless of the message. If the receiver uses $\hat{\alpha}$, then regardless of the signal structure and moderation policy, the outcome is that the bomb is built if and only if $\theta_{\bullet} \cong \hat{\theta_{\bullet}}$. As a result, no moderation policy is robustly effective for $\hat{\alpha}$ under any signal structure. △

On the other hand, if a given moderation policy removes information from a given signal structure, then the policy is robustly effective for that signal structure and some action rule in some action space.

Remark 3 (Removing information guarantees robustly effective moderation of some action rule on some action space)**.** Under suitable regularity conditions, a moderation policy μ that removes information from a signal structure σ is robustly effective for σ and some action rule α' in some action space $\mathcal{A}'$. That this is true follows from the second part of Lemma 1.

Proposition 1 shows that removing information is a necessary condition for robustly effective content moderation. Removing information is only possible when the unmoderated message contains some information about the true state. It follows that, when the sender's signal is uninformative about the true state, robustly effective content moderation is impossible. The following corollary of Proposition 1 formalizes this intuition.

Corollary 1 (No robustly effective moderation of an uninformative signal)**.** *If $\mu_0 \circ \sigma$ is uninformative, then no moderation policy is robustly effective for σ.*

Corollary 1 applies, for example, when the signal does not depend on the state (e.g., because the sender does not know the state), and the default messaging policy passes the signal through.

Example 2, Part ii (Vaccine safety claim, robustly effective moderation)**.** The sender is a skeptic who, regardless of the true state, will claim that there is no evidence supporting vaccine safety. The receiver is credulous and will act according to what they are told. So, if the moderator passes the signal through as is, the receiver will not get vaccinated.

Suppose that the moderator instead adopts the policy of flagging the signal as false whenever it contradicts the true state, and passing it through otherwise. In that case, the credulous receiver will get vaccinated when it is safe to do so. So, the moderator's flagging policy is effective for the skeptical sender and the credulous receiver.

But the moderator's flagging policy is not robustly effective for the skeptical sender. Consider, for example, the action rule that never gets vaccinated regardless of the message. If the receiver uses that action rule, then the outcome under the flagging policy is the same as in the case of the credulous receiver under the default messaging policy.

More generally, since the skeptical sender's signal contains no information about the true state, the combination of the skeptical sender's signal structure and the default messaging policy garbles any state-dependent distribution of messages. Therefore, by Corollary 1, no moderation policy is robustly effective for the skeptical sender. △

It is important that we define information removal with respect to the message distributions $\mu \circ \sigma, \mu_0 \circ \sigma$ rather than the policies themselves μ, μ_0.

Remark 4 (Removing information about the signal vs. about the state)**.** That $\mu_0 : \left[\Theta \times \mathcal{S}\right] \to \Delta(\mathcal{M})$ does not garble $\mu : \left[\Theta \times \mathcal{S}\right] \to \Delta(\mathcal{M})$ is not sufficient to conclude that $\mu_0 \circ \sigma : \Theta \to \Delta(\mathcal{M})$ does not garble $\mu \circ \sigma : \Theta \to \Delta(\mathcal{M})$.

Example 2, Part iii (Vaccine safety claim, state vs. signal)**.** Suppose that the exact form of the skeptic's claim is random (but still unrelated to the true state). A policy that blocks the signal removes information about the signal in the sense that there is no garbling that transforms the moderated message distribution (a mapping from states and signals to distributions on messages) into the default message distribution. But, it remains true that such a policy does not remove information about the state and is therefore not robustly effective. △

A robustly effective moderation policy guarantees a change in the outcome no matter how the receiver adapts their behavior to the policy. Such a guarantee seems intuitively appealing in settings where the moderator cannot directly control the receiver's behavior. In the next subsection, we provide a direct motivation for a concern with robustness.

3.2 Receiver Intentions Motivate a Concern with Robustness

Here, we motivate a concern with robustness by showing how the receiver's intentions can undermine a moderation policy that is not robustly effective. To do this, we endow the receiver with preferences and beliefs.

We endow the receiver with a complete preference ordering over outcomes. We summarize this preference ordering by a real-valued payoff function $\mathbf{U} : \left[\Delta(\mathcal{A})\right]^{\Theta} \to \mathbb{R}$. The payoff function could, but need not, be the expectation of a von Neumann-Morgenstern utility function with respect to the receiver's prior on the state.[3]

We also endow the receiver with a belief about the message distribution. We say that a given belief $\beta : \Theta \to \Delta(\mathcal{M})$ justifies an action rule α if no other action rule leads to a more preferred outcome given that belief. Formally, belief β **justifies** action rule $\alpha \in \overline{\mathscr{A}}$ if $\mathbf{U}(\alpha \circ \beta) \geq \mathbf{U}(\alpha' \circ \beta)$ for all $\alpha' \in \overline{\mathscr{A}}$.

Now suppose that the receiver has beliefs $\beta(\mu_0)$ and $\beta(\mu)$ about the default and moderated message distributions, respectively; these beliefs need not be correct. If $\beta(\mu)$ justifies actions that yield the same outcomes as those justified by $\beta(\mu_0)$, then we say that the default outcomes are sticky.

Definition (Stickiness)**.** The **default outcomes for σ under $\beta(\mu_0)$ are sticky under $\beta(\mu)$** if, for every action rule $\alpha \in \overline{\mathscr{A}}$ that is justified by $\beta(\mu_0)$, there is an action rule $\alpha' \in \overline{\mathscr{A}}$ that is justified by $\beta(\mu)$ for which $\Xi(\sigma, \mu, \alpha') = \Xi(\sigma, \mu_0, \alpha)$.

Notice that, while the beliefs $\beta(\mu_0), \beta(\mu)$ need not be correct, the map $\Xi(\cdot, \cdot, \cdot)$ is (by definition) correct. Therefore, stickiness requires that the receiver is willing and able to reconstitute the default outcome under the moderated message, regardless of their potentially incorrect beliefs.

When the default outcomes are sticky, the receiver's intentions undermine the moderation policy, in the sense that the receiver is (at least weakly) motivated (and able) to maintain the same outcome under the policy as would have arisen under the default. Robustly effective moderation prevents this undermining.

[3] Say that the receiver receives utility $U(\theta, a)$ that depends on the state and action via the function $U : \left[\Theta \times \mathcal{A}\right] \to \mathbb{R}$, and that the receiver holds a prior $\pi \in \Delta(\Theta)$ on the state. Then we may define $\mathbf{U}(X) \equiv \mathrm{E}_{(\theta,a)\sim(\pi,X)}[U(\theta, a)]$ where the expectation is taken with respect to the joint distribution (π, X) over the state and action. Alternatively, we may define $\mathbf{U}(X) \equiv \inf_{\theta\in\Theta} \left\{\mathrm{E}_{a\sim X(\theta)}[U(\theta, a)]\right\}$, where the receiver may now hold no prior on the state, and instead evaluates each action distribution under the worst possible state for that action distribution.

Claim 1 (Robustly effective moderation prevents stickiness with multiple senders)**.** *If moderation policy μ is robustly effective for σ and some action rule α that is justified by $\beta(\mu_0)$, then the default outcomes for σ under $\beta(\mu_0)$ are not sticky under any $\beta(\mu)$.*

Proof. We consider some action α that is justified by $\beta(\mu_0)$. If μ is robustly effective for σ and some action α, then—by the definition of robust effectiveness—there is *no* $\alpha' \in \overline{\mathscr{A}}$ such that $\Xi(\sigma, \mu, \alpha') = \Xi(\sigma, \mu_0, \alpha)$. Thus, there is no $\alpha' \in \overline{\mathscr{A}}$ justified by *any* beliefs $\beta(\mu)$ such that $\Xi(\sigma, \mu, \alpha') = \Xi(\sigma, \mu_0, \alpha)$. □

Claim 1 establishes a sense in which robustly effective moderation is sufficient to prevent stickiness. We now discuss important cases in which robustly effective moderation is also necessary to prevent stickiness.

3.2.1 When the Moderator Does Not Have More Information Than the Sender

Suppose that the moderated message is no more informative than the default message, in the sense that $\mu \circ \sigma$ garbles $\mu_0 \circ \sigma$. We can think of this as a situation in which the default policy passes the sender's signal through, and the moderation policy does not incorporate additional information beyond what is in the signal, for example because the moderator does not have information that the sender does not have.

Suppose also that the receiver's beliefs are **accurate** in the sense that $\beta(\mu_0) = \mu_0 \circ \sigma$ and $\beta(\mu) = \mu \circ \sigma$. Our next result states that, in this case, default outcomes are sticky whenever μ is not robustly effective for σ.

Proposition 2 (Default outcomes are sticky with an uninformed moderator)**.** *Suppose that moderation policy μ is not robustly effective for σ. Then if $\mu \circ \sigma$ garbles $\mu_0 \circ \sigma$, default outcomes for σ under accurate beliefs $\beta(\mu_0) = \mu_0 \circ \sigma$ are sticky under accurate beliefs $\beta(\mu) = \mu \circ \sigma$.*

Proof. We consider some μ and μ_0, along with an action rule $\alpha \in \overline{\mathscr{A}}$ that is justified by $\beta(\mu_0) = \mu_0 \circ \sigma$. Now, if μ is not robustly effective for σ, then the outcome achieved by α under (σ and) μ_0 can be achieved by some action rule $\alpha' \in \overline{\mathscr{A}}$ under (σ and) μ. Then, because $\mu \circ \sigma$ garbles $\mu_0 \circ \sigma$ by hypothesis, we know from the first part of Lemma 1 that the action rule α' must be justified under $\beta(\mu) = \mu \circ \sigma$. □

Remark 5. An immediate consequence of Proposition 2 is that, if μ is not robustly effective for σ, then the default outcomes are sticky whenever the receiver's beliefs are accurate and the default message $\mu_0 \circ \sigma$ is maximally informative. We may think of

the case where the default message is maximally informative as one in which the sender transmits the true state.

Example 1, Part iv (Instructions for building a bomb, sender transmits the true state)**.** Suppose that the receiver's payoff is strictly increasing in the probability that the bomb is built in each state $\theta_{\bullet}$. Recall that the signal structure $\sigma_\dagger$ transmits the correct instructions and the default messaging policy μ_0 passes the signal through. Therefore, the default message $\mu_0 \circ \sigma_\dagger$ is maximally informative.

Because the default message reveals the true instructions, any action rule justified by $\mu_0 \circ \sigma_\dagger$ ensures that the bomb is built with probability 1. If this outcome remains feasible under policy μ, then the outcome will be attained by an action rule justified under $\mu \circ \sigma_\dagger$, and no other outcome will be. Any moderation policy that is not robustly effective is therefore undermined by the receiver's intentions. △

3.2.2 When the Receiver Thinks the Moderator Does Not Have More Information Than the Sender

Suppose now that the moderated message may contain information that is not in the default message, in the sense that $\mu \circ \sigma$ need not garble $\mu_0 \circ \sigma$. Here we show that, if the moderation policy does not remove information from σ, then there are always beliefs the receiver may hold under which default outcomes are sticky. We do this by constructing beliefs such that $\beta(\mu)$ garbles $\beta(\mu_0)$. The construction makes use of a regularity condition on the set of possible messages $\mathcal{M}$.

Assumption (Fixed-pointiness)**.** We say that the set of messages $\mathcal{M}$ is **fixed-pointy** if for any map $\Lambda : \mathcal{M} \to \Delta(\mathcal{M})$, a fixed point exists for the corresponding map $\Lambda_\Delta : \Delta(\mathcal{M}) \to \Delta(\mathcal{M})$ that takes value $\Lambda_\Delta(\phi) \equiv \mathrm{E}_{m\sim\phi}[\Lambda(m)]$ for generic element $\phi \in \Delta(\mathcal{M})$.

Remark 6 (Sufficient conditions for fixed-pointiness)**.** Fixed-pointiness holds by Brouwer's fixed point theorem whenever $|\mathcal{M}| < \infty$, because in this case $\mathrm{E}_{m\sim\phi}$ is continuous in ϕ and $\Delta(\mathcal{M})$ is compact and convex. Fixed-pointiness also holds under a variety of alternative sufficient conditions including regularity conditions on the maps Λ.[4]

Under fixed-pointiness, default outcomes are sticky for some beliefs of the receiver whenever the moderation policy does not remove information from σ.

[4]For example, if $\mathcal{M}$ is compact and $\Delta(\mathcal{M})$ is endowed with the weak-* topology, then whenever Λ can be guaranteed weakly continuous, Λ_Δ is continuous (this is a consequence of the "Portmanteau Theorem" [Billingsley, 2013, Theorem 2.1]; see also Kallenberg (2021, pp. 367–368), as well as the proof of Theorem 4.17 of Hairer (2018)); we then have the existence of the desired fixed point by Schauder's fixed-point theorem.

Proposition 3 (Default outcomes are sticky with a skeptical receiver)**.** *Suppose that $\mathcal{M}$ is fixed-pointy, and that moderation policy μ does not remove information from σ and is therefore not robustly effective for σ. Then there exist beliefs $\beta(\mu_0)$ and $\beta(\mu)$ under which the moderator does not have more information than the sender, i.e., $\beta(\mu)$ garbles $\beta(\mu_0)$, and default outcomes for σ under $\beta(\mu_0)$ are sticky under $\beta(\mu)$.*

Proof. If μ does not remove information from σ, then (definitionally) we can write $\mu_0 \circ \sigma = \Gamma \circ \mu \circ \sigma$ for some $\Gamma : \mathcal{M} \to \Delta(\mathcal{M})$. And we can always define a map $\upsilon : \big[\mathcal{M} \times \big[\Theta \times \mathcal{S}\big]\big] \to \Delta(\mathcal{M})$ such that $\upsilon \circ (\mu_0 \circ \sigma, \sigma) = \mu \circ \sigma$, where we continue to ease notation by suppressing the dependence on θ. For any full-support distribution $\pi \in \Delta(\Theta)$, we can furthermore define the average map $\Upsilon : \mathcal{M} \to \Delta(\mathcal{M})$ where $\Upsilon(m) = \mathrm{E}_{\theta\sim\pi}\big[\mathrm{E}_{s\sim\sigma(\theta)}[\upsilon(m;\theta,s)]\big]$ for all $m \in \mathcal{M}$.

We define $\beta(\mu)$ so that, for each $\theta \in \Theta$, $\beta(\mu)[\theta]$ is a fixed point of the map that takes value $\mathrm{E}_{m\sim\phi}\big[\big[\Upsilon \circ \Gamma\big](m)\big]$ for generic element $\phi \in \Delta(\mathcal{M})$; such a fixed point exists by fixed-pointiness. Furthermore, we define $\beta(\mu_0) = \Gamma \circ \beta(\mu)$.

Now, we have $\beta(\mu) = \big[\Upsilon \circ \Gamma\big] \circ \beta(\mu) = \Upsilon \circ \big[\Gamma \circ \beta(\mu)\big] = \Upsilon \circ \beta(\mu_0)$, where the first equality follows from our fixed point construction of $\beta(\cdot)$.

Thus, we see that $\beta(\mu)$ garbles $\beta(\mu_0)$, as desired. Moreover, for any action rule α that is justified under $\beta(\mu_0)$, it follows by the first part of Lemma 1 that action rule $\alpha \circ \Gamma$ is justified under $\beta(\mu)$. Because $\Xi(\sigma, \mu, \alpha \circ \Gamma) = \Xi(\sigma, \Gamma \circ \mu, \alpha) = \Xi(\sigma, \mu_0, \alpha)$, this completes the proof. □

Remark 7 (How a skeptical receiver reasons)**.** The proof of Proposition 3 constructs a set of beliefs for the receiver under which the moderated message does not contain any information beyond the default message. These beliefs have an intuitive economic interpretation. Specifically, whenever the moderated message depends directly on the state, the receiver believes that the moderated message instead depends on a random variable unrelated to the state. The receiver is therefore motivated to restore the default outcome under the moderated message. Moreover, while the receiver may have incorrect beliefs about the default and moderated message distributions, the receiver has a *correct* belief about how to recover the unmoderated message distribution from the moderated message. The receiver is therefore able (as well as motivated) to restore the default outcome under the moderated message.

Example 2, Part iv (Vaccine safety claim, receiver is skeptical of the moderator)**.** The sender always (i.e., in all states) claims that there is no evidence supporting vaccine safety. But, the receiver believes that the sender always transmits the true state.

The moderator flags the signal as false whenever it contradicts the true state, and passes it through otherwise. The receiver believes that the moderator flags the signal as false whenever it contradicts what the moderator perceives to be the true state. But, the receiver believes that the moderator's perception of the state is unrelated to the true state, for example because the moderator's perception is driven entirely by confusion or bias.

Under this belief of the receiver, the moderator's flag does not add any additional information about the state. The receiver is therefore motivated to act just as they would if the moderator had instead allowed the sender's signal to pass through. To act in this way, the receiver simply ignores the moderator's flag. Importantly, despite the receiver's incorrect beliefs, ignoring the moderator's flag is the correct way to restore the unmoderated message. △

Remark 8 (Non-falsifiability of the skeptical receiver's beliefs)**.** One way the skeptical receiver might learn that their beliefs are incorrect is if they received a message outside the support of their beliefs. The proof of Proposition B.1 in Section B.1 shows how to extend the construction in the proof of Proposition 3 to ensure that all messages in the support of the true (moderated and default) message distributions are in the support of the distributions the receiver perceives, thus eliminating the possibility of such falsification.

3.3 Moderator Payoff Determines Scope for Robust Improvement

We have seen that robustly effective moderation requires removing information. Removing information makes some outcomes impossible. Whether this is desirable depends on the goals of the moderator. Here, we introduce a payoff for the moderator and study the conditions under which the moderator can make undesirable outcomes impossible, while still preserving the possibility of desirable outcomes.

Formally, let the real-valued payoff function $\mathbf{V} : \left[\Delta(\mathcal{A})\right]^{\Theta} \to \mathbb{R}$ summarize the moderator's preferences. The payoff might, but need not, be the expectation of a von Neumann-Morgenstern utility function with respect to the moderator's prior over the state. The payoff might coincide with that of the receiver. Even in such a case, there may be scope for helpful moderation, because the receiver may have mistaken beliefs about the message distribution. The payoff might also differ from that of the receiver, for example because of effects (externalities) of the receiver's action on others, including on the moderator.

We will compare the set of possible payoffs $\boldsymbol{\mathcal{V}}(\sigma,\mu) \equiv \bigcup_{\alpha\in\overline{\mathscr{A}}}\{\mathbf{V}(\Xi(\sigma,\mu,\alpha))\}$ under a given moderation policy μ to the set $\boldsymbol{\mathcal{V}}(\sigma,\mu_0)$ under the default messaging policy. We can motivate a concern with the set of possible payoffs by imagining either that the moderator does not know the receiver's intentions or that the moderator is concerned with the actions of a large set of receivers each with different intentions. Either way, the moderator entertains the possibility that the receiver may choose any feasible outcome.

Ideally, the moderator would like to eliminate undesirable outcomes while preserving desirable ones.

Definition. A moderation policy μ **improves the worst payoff (to $\underline{V}$) under signal structure** σ if $\inf \boldsymbol{\mathcal{V}}(\sigma,\mu) \equiv \underline{V} > \inf \boldsymbol{\mathcal{V}}(\sigma,\mu_0)$.

If, furthermore, we have $V \in \boldsymbol{\mathcal{V}}(\sigma,\mu)$ for any $V \in \boldsymbol{\mathcal{V}}(\sigma,\mu_0)$ with $V > \underline{V}$, then we say that μ **preserves better payoffs than $\underline{V}$ under signal structure** σ; this holds **nontrivially** whenever there is at least one $V \in \boldsymbol{\mathcal{V}}(\sigma,\mu_0)$ with $V > \underline{V}$.

A moderation policy that improves the worst payoff eliminates the least desirable outcomes according to the moderator's payoff. A moderation policy that furthermore preserves better payoffs does so without eliminating more desirable outcomes.

It follows from Proposition 1 that improving the worst payoff requires removing information. If removing information preserves better payoffs, then it must be that these payoffs can be attained without relying on the information that was removed. This, in turn, requires that the moderator not (strictly) prefer outcomes that use more information about the state. To describe this aspect of the moderator's preferences, we use the fact that, as outcomes are stochastic maps, two outcomes can be related by garbling.

Definition. We say that **information is helpful at** V' if any outcome that achieves a payoff strictly above V' can be garbled to some outcome with payoff strictly below V'.

That is, information is helpful at V' if any $X \in \left[\Delta(\mathcal{A})\right]^{\Theta}$ with $\mathbf{V}(X) > V'$ is garbled by some $X' \in \left[\Delta(\mathcal{A})\right]^{\Theta}$ with $\mathbf{V}(X') < V'$.

We can now state our result.

Proposition 4 (Necessary conditions for improving the worst payoff and preserving better payoffs)**.** *If the moderation policy μ improves the worst payoff under signal structure σ, then μ is robustly effective for σ, and hence removes information from σ.*

If, furthermore, the moderation policy μ improves the worst payoff to $\underline{V}$ and nontrivially preserves better payoffs than $\underline{V}$ under signal structure σ, then information is not helpful at $\underline{V}$.

The first part of Proposition 4 follows closely from Proposition 1. Proposition B.2 in Appendix B gives example sufficient conditions for the existence of a moderation policy μ that improves the worst payoff. The proof of the second part of Proposition 4 proceeds via Lemma A.1 in Appendix A, which states that if some outcome is possible under a given moderation policy, then so is any outcome that garbles that outcome.

Example 1, Part v (Instructions for building a bomb, moderator payoffs)**.** The moderator's utility takes value $-L < 0$ if the bomb is built and 0 otherwise. The moderator's prior puts measure zero on each $\theta_{💣} \in \Theta_{⚒}$; so, implicitly, $\Theta_{⚒}$ is uncountable. The moderator's payoff is the expectation of the moderator's utility. When the sender sends the correct instructions, the policy $\mu_{◯}$ that erases the instructions improves the worst payoff to 0. Correspondingly, $\mu_{◯}$ removes information from the sender's signal. The policy $\mu_{◯}$ also (trivially) preserves better payoffs than 0. A payoff of 0 can be obtained by a constant outcome (never build the bomb); this outcome cannot be garbled to any outcome with strictly worse payoff (which would require information on the correct instructions).

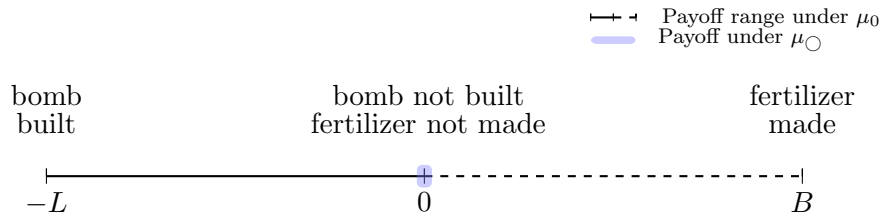
Suppose that the correct instructions for building a bomb are also the correct instructions for making a harmless fertilizer. The moderator's utility takes value $-L < 0$ if the bomb is built, $B > 0$ if the fertilizer is made, and 0 otherwise. The policy $\mu_{◯}$ that erases the instructions improves the worst payoff to 0 by eliminating the possibility that the bomb is built. But, this policy does not preserve better payoffs at 0 because it also eliminates the possibility that fertilizer is made. Preventing the bad outcome (bomb-making) requires eliminating the information about $\theta_{💣}$, which also prevents the good outcome (fertilizer-making). Correspondingly, information is helpful at 0, from which it follows that any policy that improves the worst payoff to 0 does not preserve better payoffs than 0. Figure 2a illustrates the range of moderator payoffs in this case.

△

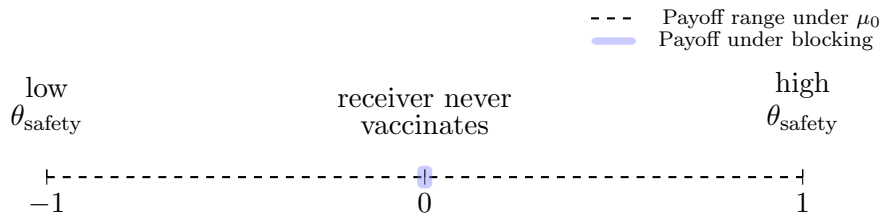
Example 2, Part v (Vaccine safety claim, moderator payoffs)**.** The moderator's utility takes value $\theta_{\text{safety}} \in \Theta_{\text{safety}} = [-1, 1]$ if the receiver gets vaccinated and 0 otherwise. The moderator's prior implies that θ_{safety} has mean 0. The moderator's payoff is the expectation of the moderator's utility. If the sender always claims there is no evidence in support of vaccine safety (i.e., if $s \cong -1$ always) and the receiver acts accordingly, then the receiver never gets vaccinated and the moderator's payoff is 0. No moderation policy improves the worst payoff under this signal structure, because no policy eliminates the option to never get vaccinated. Correspondingly, no policy removes information from this signal structure, as the signal contains no information about the state.

Figure 2: Range of Moderator Payoffs in Running Examples

(a) Instructions for building a bomb



(b) Vaccine safety claim



Notes: Figure 2 illustrates the range of possible moderator payoffs under different moderation policies in the two running examples. In Panel a, the dashed line adds the possible moderator payoffs under the extension in which the correct instructions for building a bomb also enable making a harmless fertilizer. In Panel b, the dashed line adds the possible moderator payoffs under the extension in which the sender conveys the true scientific evidence but also claims that science is a scam.

Suppose instead that the sender always reveals the true scientific evidence θ_{safety} but claims that science is a scam. In this case, a credulous receiver (who trusts the sender) might only get vaccinated when θ_{safety} is low, leading to a negative payoff for the moderator. The moderation policy that always blocks the sender's signal improves the worst payoff to 0, because it eliminates this bad possibility. But, this policy does not preserve better payoffs at 0, because it eliminates the possibility that a discerning receiver (who ignores the sender's skepticism) gets vaccinated only when θ_{safety} is high. These two properties of the policy are linked: to prevent the bad outcome requires eliminating the information about θ_{safety}, which also prevents the good outcome. Correspondingly, information *is* helpful at 0, from which it follows that any policy that improves the worst payoff to 0 does not preserve better payoffs than 0. Figure 2b illustrates the range of moderator payoffs in this case. △

3.4 Moderation of Modes of Expression Does Not Require Removing Information

We have defined the outcome to be the distribution of actions given the state. Yet in some settings, the message itself may be consequential. We could represent such settings as those in which the moderator's payoff depends on the distribution of messages rather than on the distribution of actions.[5] We could consider the goal of adopting a moderation policy that changes the message distribution, relative to the default messaging policy, without the possibility that the change can be undone by the receiver's actions. Such a goal is trivially achievable in our setup because the moderator directly controls the message distribution. In particular, because the goal does not concern the relationship between the action and the state, removing information is no longer necessary.

This observation highlights an important contrast between moderating to change the mode of expression—i.e., *to change the actual message sent*—versus moderating to change the outcome.

Example 1, Part vi (Instructions for building a bomb, modes of expression)**.** The sender is an extremist who, in addition to conveying the instructions, expresses hate towards a group of people (against whom a bomb could be used). The receiver is an extremist who will attempt to build the bomb using the conveyed instructions only if the expression of hate is included.

[5]To cast such settings into our notation, suppose that $\mathcal{M} \subseteq \mathcal{A}$ and that the only available action rule is the identity map.

One goal of the moderator could be to prevent the sender from expressing hate on the platform. The moderator can achieve this goal using a policy that redacts the expression of hate. Because the goal concerns the message itself, rather than the receiver's action, it does not matter that the policy does not remove information about the state.

Another goal of the moderator could be to prevent the receiver from building the bomb. If the receiver builds the bomb only when the sender expresses hate, then the policy that redacts the expression of hate is effective. But, because this policy does not remove information from the default message, it is not robustly effective. The receiver can, for example, undermine the policy by adopting an action rule that builds the bomb whenever any part of the sender's signal is redacted. △

4 Robust Content Moderation with Many Senders

In this section we examine when content moderation is robust to the possibility that the signal structure σ may change in response to the moderation policy. First, we extend our definition and characterization of robustly effective moderation to the case of multiple signal structures. Then, we show that both our justification for being concerned with robustly effective moderation, and our characterization of the role of the moderator's payoff, extend to this setting. Lastly, we show that particular forms of receiver trust can motivate different considerations for the moderator.

4.1 Robustly Effective Moderation Requires Removing Information

We now return to the question of whether, and when, the moderator can guarantee an outcome different from the one that would result from the unmoderated message. Here, in addition to allowing the receiver to change their action rule in response to the moderation policy, we furthermore allow the sender to change their signal structure σ in response to the moderation policy, in principle choosing any signal structure in some set $\mathscr{S}$. This perspective leads naturally to the following definition.

Definition. A moderation policy μ is **robustly effective for action rule α and signal structure σ across the set of signal structures** $\mathscr{S}$ if there is no $(\sigma', \alpha') \in \left[\mathscr{S} \times \overline{\mathscr{A}}\right]$ such that $\Xi(\sigma', \mu, \alpha') = \Xi(\sigma, \mu_0, \alpha)$; if this is the case for *some* action rule α and signal structure $\sigma \in \mathscr{S}$, then we say that μ is **robustly effective across signal structures** $\mathscr{S}$.

For a given signal structure and action rule, a moderation policy is robustly effective across a given set of available signal structures if no combination of action rule and

available signal structure recovers the outcome that would have arisen under the default messaging policy; if this is true for some signal structure and action rule, then the policy is robustly effective across the available set of signal structures. When the available set of moderation policies is a singleton (i.e., when $\mathscr{S} = \{\sigma\}$), robust effectiveness across the set is equivalent to our definition of robustly effective moderation with a fixed signal structure. Otherwise, the definition is more demanding. Notice that the setup here still invests the moderator with substantial power; for example, the moderator could in principle choose to block any given signal, or even all signals.

Our main result in this subsection is that robustly effective content moderation requires removing information about the state *across all available signal structures*; to express this, we extend the definition of removing information to account for the possibility of multiple signal structures.

Definition. A moderation policy μ **removes information from a signal structure** $\sigma \in \overline{\mathscr{S}}$ **across a set of signal structures** $\mathscr{S} \subseteq \overline{\mathscr{S}}$ if for any $\sigma' \in \mathscr{S}$, $\mu_0 \circ \sigma$ does not garble $\mu \circ \sigma'$.

We can now state our result.

Proposition 5 (Robustly effective moderation requires removing information)**.** *A moderation policy μ is robustly effective across signal structures $\mathscr{S}$ only if it removes information from some signal structure $\sigma \in \mathscr{S}$ across $\mathscr{S}$.*

Proposition 5 states that a policy that is robustly effective regardless of the sender's signal structure must remove information from some signal across *all* available signal structures. Intuitively, if the policy does not do so, then the sender can always convey the information available under the unmoderated message by changing the form in which the information is sent. The following corollary makes this clear using the definition of information equivalence from Section 2.2.

Corollary 2 (Robustly effective moderation requires removing information thoroughly)**.** *A moderation policy μ is robustly effective across signal structures $\mathscr{S}$ only if it removes information across $\mathscr{S}$ from all signal structures $\sigma' \in \mathscr{S}$ that are informationally equivalent to some $\sigma \in \mathscr{S}$ under μ_0.*

Example 1, Part vii (Instructions for building a bomb, removing information thoroughly)**.** Recall the policy $\mu_{\bigcirc}$ that erases all instructions, $\mu_{\bigcirc}(\theta_{\text{💣}}, s) = \bigcirc$ for all $\theta_{\text{💣}} \in \Theta_{\text{⚒}}$ and $s \in \mathcal{S}$. Under the policy $\mu_{\bigcirc}$, there is *no* signal structure $\sigma \in \overline{\mathscr{S}}$ and action rule $\alpha \in \overline{\mathscr{A}}$ under which the receiver always builds the bomb. Therefore, $\mu_{\bigcirc}$ is robustly effective across $\overline{\mathscr{S}}$.

The policy $\mu_{\bigcirc}$ also removes information from $\sigma_\dagger$ across $\overline{\mathscr{S}}$, as there is no way to generate the correct instructions from the messages generated by $\mu_{\bigcirc}$ regardless of which signal structure the sender uses. The policy $\mu_{\bigcirc}$, in particular, removes information from any signal structure that is informationally equivalent to $\sigma_\dagger$ under the default messaging policy, for example a signal structure σ_{encrypt} that encrypts the correct instructions by reversing the order of the steps to build a bomb. Via Proposition 5, these properties explain why the policy $\mu_{\bigcirc}$ can be robustly effective.

Not all policies that are robustly effective for $\sigma_\dagger$ are robustly effective across $\overline{\mathscr{S}}$. For example, consider the state-dependent policy μ_{nobomb} that erases the instructions, transmitting $m = \bigcirc$, if the instructions are correct, $s \cong \theta_{💣}$, and otherwise passes the signal through. The policy μ_{nobomb} is robustly effective for $\sigma_\dagger$, but it is not robustly effective across $\overline{\mathscr{S}}$. To see this, observe that the signal structure σ_{encrypt} transmits instructions that are technically incorrect (i.e., reversed). Under moderation policy μ_{nobomb} and signal structure σ_{encrypt}, the action rule α_{decrypt} that attempts to build the bomb based on decrypting (i.e., reversing) the transmitted instructions results in the same outcome as would arise under the signal structure $\sigma_\dagger$, default messaging policy μ_0, and action rule $\alpha_\dagger$. Notice that the policy μ_{nobomb} removes information from $\sigma_\dagger$, but not from all signal structures that are informationally equivalent to $\sigma_\dagger$ under the default messaging policy. △

4.2 Sender and Receiver Intentions Motivate a Concern with Robustness

Here, we motivate a concern with robustness by showing how the receiver's and sender's intentions can undermine a policy that is not robustly effective. Unlike before, we now allow both the receiver's action rule and the sender's signal structure to change in response to the moderation policy.

The possibility that the signal structure adapts to the moderation policy admits multiple interpretations. One is that the sender has a choice of signal structures and may change that choice in response to the moderation policy. Another is that the receiver has a choice of senders and may change that choice in response to the moderation policy. Notice that, provided that there is always *some* sender willing to help the receiver achieve their desired outcome—perhaps because the sender shares the receiver's goals—these two interpretations both imply that the signal structure aligns with the receiver's goals. We therefore continue to focus on the receiver's preferences over outcomes, which we

continue to summarize with the payoff function $\mathbf{U} : \left[\Delta(\mathcal{A})\right]^{\Theta} \to \mathbb{R}$.

We envision a set of senders $i \in \mathcal{I}$, each of whom adopts some signal structure $\sigma_i(\mu) \in \mathscr{S}$ that is an element of the set $\mathscr{S}$ and that may depend on the moderation policy μ. We index the senders explicitly because when receiver beliefs are incorrect, it no longer suffices to identify a sender with an associated (policy-dependent) signal structure. For any policy μ, we assume that the set $\mathcal{I}$ is rich enough so that, for every $\sigma \in \mathscr{S}$, there are arbitrarily many $i \in \mathcal{I}$ such that $\sigma_i(\mu) = \sigma$.

For each moderation policy and each sender, we endow the receiver with a belief about the resulting message distribution. Specifically, for each moderation policy μ and each sender $i \in \mathcal{I}$, we endow the receiver with a belief $\beta(\mu; i) : \Theta \to \Delta(\mathcal{M})$; as before, this belief need not be correct.

We say that a given belief $\beta(\mu; \cdot)$ justifies an action rule–sender tuple $(\alpha; i)$ under policy μ if no other action rule and sender lead to a more preferred outcome given that belief. Formally, belief $\beta(\mu; \cdot)$ **justifies** $(\alpha; i) \in \left[\overline{\mathscr{A}} \times \mathcal{I}\right]$ **under** μ if $\mathbf{U}(\alpha \circ \beta(\mu; i)) \geq \mathbf{U}(\alpha' \circ \beta(\mu; i'))$ for all $(\alpha'; i') \in \left[\overline{\mathscr{A}} \times \mathcal{I}\right]$.

Definition (Stickiness with multiple senders)**.** The **default outcomes for** $\mathscr{S}$ **under** $\beta(\mu_0; \cdot)$ **are sticky under** $\beta(\mu; \cdot)$ if, for every tuple $(\alpha; i) \in \left[\overline{\mathscr{A}} \times \mathcal{I}\right]$ that is justified by $\beta(\mu_0; \cdot)$, there is a tuple $(\alpha'; i')$ that is justified by $\beta(\mu; \cdot)$ for which $\Xi(\sigma_{i'}(\mu), \mu, \alpha') = \Xi(\sigma_i(\mu_0), \mu_0, \alpha)$.

Claim 2 (Robustly effective moderation prevents stickiness)**.** *If moderation policy μ is robustly effective across $\mathscr{S}$ for some $(\alpha, \sigma) = (\alpha, \sigma_i(\mu_0))$ with $(\alpha; i)$ justified by $\beta(\mu_0; \cdot)$, then the default outcomes for $\mathscr{S}$ under $\beta(\mu_0; \cdot)$ are not sticky under any $\beta(\mu; \cdot)$.*

Claim 2 establishes a sense in which robustly effective moderation is sufficient to prevent stickiness. We next discuss important cases in which robustly effective moderation is also necessary to prevent stickiness.

4.2.1 When the Moderator Does Not Have Special Information

Suppose that the moderated message for any sender is no more informative than the default message for *some* sender, in the sense that, for any $\sigma \in \mathscr{S}$, $\mu \circ \sigma$ garbles $\mu_0 \circ \sigma'$ for some $\sigma' \in \mathscr{S}$. We can think of this as a situation in which the default policy always passes the sender's signal through, and the moderation policy does not incorporate additional information beyond what is in *some* sender's signal, for example because the moderator does not have special information unavailable to any individual sender.

Suppose also that the receiver's beliefs are **accurate** in the sense that $\beta(\mu_0; i) = \mu_0 \circ \sigma_i(\mu_0)$ and $\beta(\mu; i) = \mu \circ \sigma_i(\mu)$ for all $i \in \mathcal{I}$. Our next result states that, in this case, default outcomes are sticky when μ is not robustly effective across $\mathscr{S}$.

Proposition 6 (Default outcomes are sticky without special information)**.** *Suppose that moderation policy μ is not robustly effective across $\mathscr{S}$. Then if, for any $\sigma \in \mathscr{S}$, $\mu \circ \sigma$ garbles $\mu_0 \circ \sigma'$ for some $\sigma' \in \mathscr{S}$, default outcomes for $\mathscr{S}$ and accurate beliefs $\beta(\mu; i) = \mu \circ \sigma_i$ are sticky under accurate beliefs $\beta(\mu_0; i) = \mu_0 \circ \sigma_i$ (for all $i \in \mathcal{I}$).*

Remark 9 (Stickiness with accurate beliefs when the default message is maximally informative)**.** An immediate consequence of Proposition 6 is that, if μ is not robustly effective, then the default outcomes are sticky whenever the receiver's beliefs are accurate and the default message $\mu_0 \circ \sigma_i$ for some sender i is maximally informative.

Example 1, Part viii (Instructions for building a bomb, a sender transmits the true state)**.** Suppose that under μ_0 the sender i uses the signal structure $\sigma_\dagger$ that transmits the true instructions. Under the default messaging policy, a receiver who receives from i and uses action rule $\alpha_\dagger$ that attempts to build the bomb using the transmitted instructions will always build the bomb. If the receiver has correct beliefs, then this is the outcome they intend. If the moderator adopts the policy μ_{nobomb} that only erases correct instructions, then the receiver can switch to a different sender who sends σ_{encrypt}, or alternatively the original sender i can switch to σ_{encrypt}. Either way, the action rule α_{decrypt} fulfills the receiver's intention. △

Remark 10 (A sender who cannot be moderated)**.** Our setup directly nests situations in which, for technological or other reasons, the moderation policy cannot affect the message associated with a given sender, i.e., in which for some particular sender $i \in \mathcal{I}$ the moderator must choose a policy μ such that $\mu \circ \sigma_i = \mu_0 \circ \sigma_i$.

Example 1, Part ix (Instructions for building a bomb, receiver can switch platforms)**.** Suppose that under μ_0 the sender i uses the signal structure $\sigma_\dagger$ that transmits the true instructions. Being a social media platform, the moderator can block this signal. Suppose that there is another sender i' who uses this same signal structure but sends their signal on a different platform. The moderator cannot set policy on another platform, and so cannot block this other sender's signal; therefore $\mu \circ \sigma_{i'} = \mu_0 \circ \sigma_{i'}$. The moderator cannot prevent the receiver from building the bomb if that is the receiver's intention. △

4.2.2 When the Receiver Thinks the Moderator Does Not Have Special Information

Suppose now that the moderator may have special information, such that there may be $\mu \circ \sigma, \sigma \in \mathscr{S}$ that does not garble any $\mu_0 \circ \sigma', \sigma' \in \mathscr{S}$. Here we show that, if the moderation policy does not remove information from some σ across $\mathscr{S}$, there is always some belief the receiver may hold under which default outcomes are sticky.

Proposition 7 (Default outcomes are sticky with a skeptical receiver)**.** *Suppose that moderation policy μ does not remove information from any $\sigma \in \mathscr{S}$ across $\mathscr{S}$ and is therefore not robustly effective across $\mathscr{S}$. Then there exist beliefs $\beta(\mu_0; \cdot)$ and $\beta(\mu; \cdot)$ under which the moderator does not have special information—i.e., for any $i \in \mathcal{I}$, $\beta(\mu; i)$ garbles $\beta(\mu_0; i')$ for some $i' \in \mathcal{I}$—and default outcomes for $\mathscr{S}$ under $\beta(\mu_0; \cdot)$ are sticky under $\beta(\mu; \cdot)$.*

The proof of Proposition 7 extends the construction in the proof of Proposition 3.

Example 2, Part vi (Vaccine safety claim, receiver trusts their chosen sender)**.** One sender is the skeptic who, regardless of the state, always claims there is no evidence supporting vaccine safety. Another sender is a medical journal that always transmits the true state. Under the default messaging policy, which passes the signal through, the receiver chooses to receive from the skeptic.

The moderator knows that the journal always transmits the true state. Consider, then, the moderation policy that transmits the signal of the journal alongside that of any sender whose signal it contradicts.

The receiver believes that the journal is in the pocket of the vaccine industry and that the moderator has no special information about which sender tells the truth. If the receiver wanted to hear from the journal, they could have done so under the default. So, the receiver ignores the additional signal from the journal that the moderator insists on transmitting. △

4.3 Moderator Payoff Determines Scope for Robust Improvement

We now compare the set of possible payoffs $\boldsymbol{\mathscr{V}}(\mathscr{S}, \mu) \equiv \bigcup_{\alpha \in \overline{\mathscr{A}}, \sigma \in \mathscr{S}} \{\mathbf{V}(\Xi(\sigma, \mu, \alpha))\}$ for the moderator under a given moderation policy μ to the set $\boldsymbol{\mathscr{V}}(\mathscr{S}, \mu_0)$ under the default messaging policy. Our earlier definitions and results extend directly.

Definition. A moderation policy μ **improves the worst payoff (to $\underline{V}$) under signal structures** $\mathscr{S}$ if $\inf \boldsymbol{\mathscr{V}}(\mathscr{S}, \mu) \equiv \underline{V} > \inf \boldsymbol{\mathscr{V}}(\mathscr{S}, \mu_0)$.

If, furthermore, $V \in \boldsymbol{\mathcal{V}}(\mathscr{S}, \mu)$ for any $V \in \boldsymbol{\mathcal{V}}(\mathscr{S}, \mu_0)$ with $V > \underline{V}$, then we say that μ **preserves better payoffs than** $\underline{V}$ **under signal structures** $\mathscr{S}$; this holds **nontrivially** whenever there is at least one $V \in \boldsymbol{\mathcal{V}}(\mathscr{S}, \mu_0)$ with $V > \underline{V}$.

Proposition 8 (Necessary conditions for improving the worst payoff and preserving better payoffs)**.** *If the moderation policy μ improves the worst payoff under signal structures $\mathscr{S}$, then μ is robustly effective across $\mathscr{S}$, and hence removes information from some σ across $\mathscr{S}$.*

If, furthermore, the moderation policy μ improves the worst payoff to $\underline{V}$ and (nontrivially) preserves better payoffs than $\underline{V}$ under signal structures $\mathscr{S}$, then information is not helpful at $\underline{V}$.

4.4 Additive Moderation Requires an Expert, Trusted Moderator

Our analysis has focused on changing the outcome, relative to the default, in a way that cannot be undone by the sender or the receiver. An alternative goal of moderation could be to enable an outcome that was impossible under the default. We can use the analysis in this section to understand two important requirements for such additive moderation, and thus to better understand its limitations.

First, additive moderation requires that the moderator has special information, in the sense that, for some $\sigma \in \mathscr{S}$, $\mu \circ \sigma$ does not garble $\mu_0 \circ \sigma'$ for any $\sigma' \in \mathscr{S}$. This follows closely from the analysis in Section 4.1 and is formalized in Appendix C.

Second, additive moderation requires that the receiver trusts the moderator, in the sense that the receiver believes the moderator has special information. Absent such trust, the analysis in Section 4.2 implies that the receiver's (and sender's) intentions can undermine any policy that does not remove information.

Example 2, Part vii (Vaccine safety claim, moderator selects the sender)**.** The state comprises both the extent of the scientific evidence that recommended vaccines are safe *and* the identity of the sender who reports this information accurately. Each sender makes a claim about vaccine safety (that they believe is true), but only the moderator knows which sender is accurate. The receiver wishes to be vaccinated only when it is safe to do so.

The default messaging policy passes through the signal of whichever sender the receiver chooses. We may think of such a policy as an algorithm that responds to the receiver's engagement. Under this policy, the receiver's choice of sender may fail to enable the receiver's preferred outcome, because any given sender may be inaccurate.

An alternative moderation policy passes through the signal of whichever sender is accurate. We may think of such a policy as an algorithm that boosts the signal of whichever sender is accurate. Such a policy enables the receiver's preferred outcome.

If the receiver doubts the expertise or intentions of the moderator, however, the receiver may instead try to recover the signal of the default sender, and act on that recovered signal, just as in Example 2, Part vi. △

4.5 Receiver Trust Motivates Moderating Modes of Expression

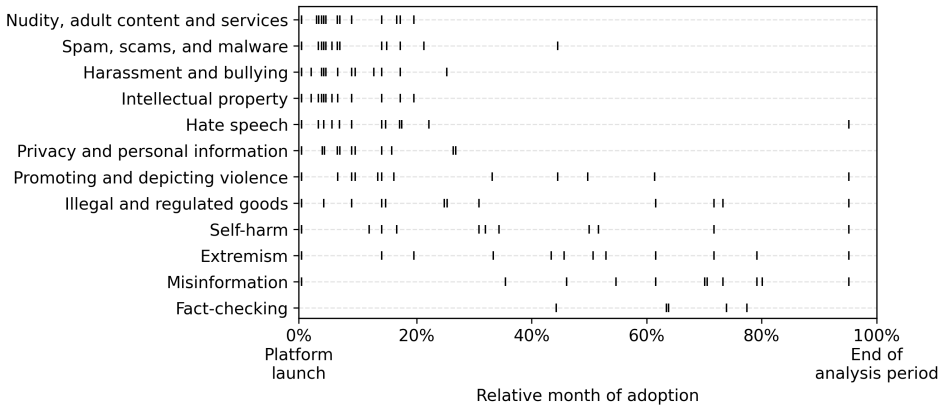
By the first part of Lemma 1, two message distributions that are informationally equivalent are also equivalent in terms of the outcomes they enable. That result requires that the receiver be able to adapt their action rule to the message distribution. One prominent case where such adaptation is absent is where the receiver trusts that the message is delivered in a literal, natural language (Farrell, 1993; Kartik, Ottaviani, and Squintani, 2007; Reny, 2025). In such a case, the receiver's action rule may result in different outcomes under informationally equivalent message distributions; receiver trust may then motivate the moderator to be concerned with modes of expression.

Example 2, Part viii (Vaccine safety claim, trusted language)**.** Suppose now that the moderator is a platform that hosts articles summarizing vaccine safety information for use by medical professionals. The state is the safety of each vaccine for each age group, young and old. The receiver is a medical professional. The default messaging policy passes the signal through.

One sender is a writer who writes in good faith, transmitting $s \cong \theta$. Another sender is a writer who writes in bad faith, transmitting the safety for the old as if it is for the young, and vice versa. These two signal structures are informationally equivalent under the default messaging policy. However, they need not be equivalent if the receiver trusts the language in which information is conveyed.

Suppose that the receiver prescribes, for each age group, whatever vaccines the message indicates is safe, and does so regardless of the true signal structure. Then the good-faith writer's signal results in accurate prescribing and the bad-faith writer's signal results in inaccurate prescribing. To avoid the possibility of inaccurate prescribing, the moderation policy needs to "unscramble" the bad-faith writer's signal, for example by editing the article. Enforcing such a policy requires that the moderator know the state. △

Figure 3: Relative Timing of First Moderation Policy by Category



Notes: Data are from the Social Media Content Moderation Policy Archive (Appendix D). For each category, each marker denotes the relative month in which a platform began moderating the given category. The relative month is defined relative to the platform's launch date (0%) and the end of our analysis period (April 2024, 100%). Points are jittered to reduce overlap. Policy categories are listed in descending order according to the share of platform-months in which the given category of content is moderated.

5 Applications

5.1 Social Media

To facilitate the discussion in this section, we created the Social Media Content Moderation Policy Archive (henceforth, *the Archive*). The Archive consists of monthly snapshots of major platforms' historical moderation policies, drawn from web snapshots and existing policy collections, and grouped into substantive categories. Appendix D describes the construction and contents of the Archive in detail; moderation policy quotes in this section are from the Archive except where stated. Figure 3 shows, for each policy category, the distribution of months when the platforms first moderated content of the given category, measured relative to each platform's launch month. Figure 4 shows a more detailed view of the timeline of each policy category in each platform in the Archive.[6]

5.1.1 Policies that Remove Information

Our theoretical analysis shows that moderation is robust only when it removes information, viz., when it prevents the sender from transmitting information about an otherwise

[6]For previous systematic descriptions of social media content moderation policies, see, e.g., Singhal et al. (2023); Dubois and Reepschlager (2024); Nahrgang et al. (2025).

unknown state to the receiver. Many policy categories aim to remove information. Examples include:

- Privacy and personal information: "Reddit is a pretty open and free speech place, but it is not ok to post someone's personal information." Here, the state is *the person's personal information*.
- Illegal and regulated goods: Pinterest will "take down instructions for making a bomb or evading the police when street racing." Here, the state is *how to make a bomb or to evade the police*.
- Intellectual property: "Don't upload [to YouTube] videos you didn't make, or use content in your videos that someone else owns the copyright to." Here, the state is *the exact content of the copyrighted video*.
- Spam, scams, and malware: "Do not post, upload, stream, or share [on TikTok]: Content that promotes investment schemes with promise of high returns, fixed betting, or any other types of scams." Here, the state is *the way to participate in the (fraudulent) investment scheme*.

Figures 3 and 4 show that most platforms have had policies in several of these categories—notably, "Privacy and personal information," "Intellectual property," and "Spam, scams, and malware"—consistently since their launch (or soon after).

Our theoretical analysis shows that moderating modes of expression does not require removing information. Some policy categories include restrictions on content that arguably do not depend on the information conveyed. Examples include Self-harm ("[Snapchat doesn't] allow the glorification of self-harm"), Extremism ("[Instagram doesn't] allow content that[...] represents events that Meta designates as terrorist attacks"), Adult content (Weixin prohibits "erotic texts, videos, photographs, and cartoons"), and Harassment and bullying ("Pinterest isn't a place to insult, hurt or antagonize individuals or groups of people"). Figures 3 and 4 show that most platforms have had policies regarding "Adult content" and "Harassment and bullying" consistently since their launch.[7]

[7] Ribeiro, Cheng, and West (2023) study the automated enforcement of such policies. Within economics, much of the (still relatively small) empirical literature studying content moderation by online platforms focuses on measuring the effects of moderation on user welfare (Jiménez-Durán, 2023), user engagement (Beknazar-Yuzbashev et al., 2025), platform content (Andres and Slivko, 2023; Müller and Schwarz, 2023), and offline behaviors (Jiménez-Durán, Müller, and Schwarz, 2023), particularly in the context of hate speech and toxicity.

Even within categories that address modes of expression, many policies aim to remove information. Examples include:

- Self-harm: "Do not post, upload, stream, or share [on TikTok]: Content that provides instructions for suicide or how to engage in self-harm." Here, the state is *the way to engage in self-harm*.
- Extremism: Weixin prohibits "organizing or inducing others to engage in terror or violent activities or activities that disrupt the social order." Here, the state could be *the time and location of an attack*.
- Adult content: "Users [of 𝕏] under 18 or viewers who do not include a birth date on their profile cannot click to view marked content." Here, the state could be *the nature of adult sexual acts*.
- Harassment and bullying: Instagram's policy explicitly prohibits posting "non-public phone numbers or non-public email addresses." Here, the state is *the person's personal information*.

In addition to showing that only policies that remove information can be robust, our theoretical analysis clarifies when and how to implement such policies. Information-removing policies are a robust improvement when the moderator prefers that the receiver's actions be based on less information. For some forms of information, such as the personal details of public figures, the case for such a preference is clear. For other forms of information, such as scientific knowledge that can be used in multiple ways (some good and some bad), the case is more complex. Precisely these tradeoffs arise in setting, adjudicating, and enforcing social media content moderation policies.[8]

Likewise, our theoretical analysis shows that, to achieve the strongest guarantees, the moderator must remove information not only from one signal structure but from any equivalent signal structure. This may involve, for example, identifying situations in which a piece of illicit content is hidden among licit content. Precisely such challenges arise in implementing social media content moderation policies (Zhu et al., 2021; Drug Enforcement Administration, 2021; Levine, 2022).

[8]See, for example, van der Boon et al. (2024) regarding sharing information about challenging medical cases and Oversight Board (2023) regarding sharing details of alleged human rights abuses. Even the platforms' relatively strong prohibitions on posting personal information routinely exempt situations with the potential for social benefit, such as charitable solicitations (Meta, 2024) or feedback to elected officials (Reddit, 2024).

Our theoretical analysis also highlights, via Remark 10 and Example 1, Part ix, that the option for the receiver to switch platforms can make a moderator's task more difficult. And indeed, precisely such challenges arise in enforcing social media content moderation policies.[9]

5.1.2 Policies that Do Not Remove Information

Meanwhile, our theoretical analysis shows that policies that do not remove information are not robust. Some policy categories do not aim to remove information, or even aim to add information. Specifically:

- Misinformation: "We aim to ensure that content and behavior on WeChat remains authentic, by attempting to eliminate false news, disinformation, misinformation..."
- Fact-checking: "If the [third-party] fact checking organizations identify a story as fake, it will get flagged [on Facebook] as disputed and there will be a link to the corresponding article explaining why."

If users cannot infer something true from the false content, then policies against misinformation do not remove information according to the definition in our analysis. If fact-checking overlays the platform's or third party's view of the facts on an original post, then it does not remove information, and possibly adds information.

Figures 3 and 4 show that platforms tended to adopt misinformation and fact-checking policies more recently (and more sparingly) than other policy categories. In recent years, at least two major platform owners, 𝕏 and Meta, have discontinued or greatly curtailed third-party fact-checking efforts (Fu, 2024; Kaplan, 2025).

There are likely many reasons why third-party fact-checking has been shorter-lived than other policy categories. For example, whereas third-party fact-checking is typically done using human intervention and is therefore expensive and difficult to scale (Allen et al., 2021; Corse, Bobrowsky, and Horwitz, 2025), some other policies can be enforced algorithmically (Karimi, Squicciarini, and Wilson, 2022).

Our theoretical analysis provides an additional reason for the short lifespan of third-party fact-checking: it is not robust. Section 3.2 shows that if users do not trust that fact-checking adds useful information, users have an incentive to undermine the intent of the policy. Yet precisely because fact-checking is invoked on controversial matters, it is

[9] Rauchfleisch and Kaiser (2024), Mekacher, Falkenberg, and Baronchelli (2023), and Klinenberg (Forthcoming) study substitution of engagement across platforms following removal of content from a platform.

difficult to ensure such trust (Stewart, 2021).[10] Platform executives have noted that fact-checkers are, or are perceived to be, biased, and have not garnered the trust needed to be effective.[11] Other prominent voices have made similar claims.[12] Our theoretical analysis shows that, precisely because users do not know who is telling the truth, users may take actions contrary to, or even adversarial towards, the intent of the fact-checking policy.[13] Notice that simply exposing the mechanics of fact-checking, for example by clarifying the process by which it is done, is not sufficient to resolve this limitation, because what matters is whether users believe that fact-checkers tell the truth—and if users knew the truth, there would be no need for fact-checkers![14]

Our theoretical analysis also clarifies why the problem of moderating social media platforms differs from that of editing mass media platforms such as CNN Primetime or *The New York Times*. For a mass media platform, the moderator is the platform (or its designated editor), the sender is a journalist, and the receiver is a consumer who has specifically chosen to consume media from the platform. Such a sender and receiver seem more likely to trust the information and intentions of the moderator (Gentzkow and Shapiro, 2006) than in the case of social media, thus avoiding the undermining that our theoretical analysis highlights.

[10]"People are highly resistant to fact-checks if they don't like the fact checker." (DiResta, quoted in The Economist, 2024); see also the work of Appel, Pan, and Roberts (2023).

[11]In 2023, 𝕏's Musk tweeted that "the so-called fact-checkers are huge liars and incredibly biased" (Musk, 2023). In 2025, Zuckerberg said that "fact-checkers have just been too politically biased and have destroyed more trust than they created" (Zuckerberg, 2025).

[12]A *New York Post* editorial column, for example, asserts that Facebook's "'fact checks' are really just (lefty) opinion" (Post Editorial Board, 2021). See also Lomborg (2022).

[13]We are not aware of any direct evidence from real-world social media platforms on the effectiveness of third-party fact-checking in improving outcomes. However, there is a large literature studying fact-checking in online survey experiments. A number of studies find that participants in such experiments attach less credence to statements flagged as false (Fridkin, Kenney, and Wintersieck, 2015; Vraga and Bode, 2017; Pennycook et al., 2020; Zhang et al., 2021). Martel and Rand (2024) find this pattern even for participants expressing relatively little trust in fact-checkers. Bachmann and Valenzuela (2023) find that fact-checking is less effective and more likely to lead the participant to conclude that media are hostile when fact-checking contradicts a participant's pre-existing views. In filings under the EU's Code of Conduct on Disinformation, platforms have indicated that fact-checking warnings inhibit resharing of content (see, e.g., TikTok (2024, SLI 18.1.1); Facebook (2025, QRE 18.1.3)). Guriev et al. (2025) evaluate different approaches to reducing sharing of false content using online survey experiments and a structural model.

[14]Sarkar Diba et al. (2025) discuss the possibility of using blockchain technology to improve the transparency of social media content moderation. Drolsbach, Solovev, and Pröllochs (2024) find that adding more details to community notes can improve trust in fact-checking.

5.2 State Censorship and Flooding the Zone

In addition to privately run platforms, state actors also engage in activities that can be cast as content moderation. Here we discuss two prominent forms of state intervention in the information environment and relate them to our setup and findings.

5.2.1 State Censorship

State censorship has elements in common with social media content moderation. The moderator (censor) cannot fully control what senders (citizens) will try to transmit or how receivers (other citizens) will act on what they learn. Citizens may be suspicious of the regime, or even actively opposed to it. Citizens may therefore seek to circumvent or even undermine the regime's aims.

Our analysis implies that censorship that removes information about an unknown state will be more robust than censorship that does not. Many forms of censorship seek to remove such information. Examples include preventing users from posting information that coordinates dissident activity, such as the time and place of a protest (Soo and Press, 2022);[15] King, Pan, and Roberts (2013) find that such content is among the most rigorously censored in the People's Republic of China (see also Corduneanu-Huci and Hamilton (2022) and Lei (2018)).

Other forms of censorship do not seek to remove information. Examples include censorship of opinion. King, Pan, and Roberts (2013, 2014) find evidence that censorship in the People's Republic of China is more tolerant of dissident opinion than of information that coordinates dissident activity (though see the caveats of Gueorguiev and Malesky (2019) and the findings from a different context of Beazer et al. (2022)). Our analysis suggests a possible explanation: removing information that can be used to coordinate dissent is more robust than removing opinions.

Our analysis also clarifies some of the specific challenges of policies that remove information. Our analysis shows that censorship is a robust improvement only when the regime prefers that its citizens have less information. Information about the time and location of a protest is plausibly useful only to achieve outcomes the regime would like to avoid, and is therefore an attractive target for censorship. Information about, say, foreign business opportunities is more complicated: such information may enable defection but may also enable useful economic activity. Censors routinely confront just

[15] Other examples might include evidence of regime weakness (Reuters in Bangkok, 2015) or corruption (BBC, 2016).

this sort of tradeoff (Engle, 2015; West, 2016; Zheng and Wang, 2020; Xu, Xuan, and Zheng, 2021; Kong et al., 2022; Tagat, Phokeer, and Kreitem, 2024; Bernard et al., 2025).[16]

At a more technical level, our analysis shows that effective censorship requires removing a given piece of information in all of its informationally equivalent forms. Modern digital censors devote substantial attention to detecting the oblique or hidden messages, such as visual references to protest, that dissidents/disobedients routinely use (Ser, 2016; Schraer, 2022; Amadon, n.d.).

5.2.2 Flooding the Zone

State censorship often seeks to remove undesirable messages. Another, superficially very different, approach is to overwhelm public discussion with alternative messages. In a now-famous interview, Steve Bannon referred to this approach as "flood[ing] the zone with [a particular expletive]" (Lewis, 2018); see also Pfister (2011).

We can readily cast flooding the zone into our setup. In particular, suppose that citizens' attention is limited, and let the default moderation policy describe the (distribution over) messages that citizens will encounter absent intervention. Now, imagine that the moderator (say, a leader or other public figure) can produce many messages (say, via social or mass media) that attract citizens' attention. We can represent this situation as one in which the moderation policy (flooding the zone) changes the distribution of messages by obscuring (or crowding out) the messages that would have been seen by default.

Our analysis suggests that such tactics will be most effective when they remove information. Suppose, for example, that experts find evidence of an economic downturn which may be blamed on a leader. The unmoderated message, which citizens will receive absent intervention, simply describes experts' findings, as in "experts find evidence of recession." The leader can intervene (moderate) by propagating messages of their own. Messages of the form "experts will tell you that the economy is doing badly, but it is not" do not remove information, because such messages allow receivers to infer the information in the unmoderated message. By contrast, messages that divert attention to entirely different topics do remove information, provided such messages crowd out the unmoderated message, and thus prevent citizens from learning experts' findings.

Such diversions are a common tool of zone-flooders. For example, Brennan (2025)

[16] Censorship may also remove information useful to the state's own bureaucrats, reducing the effectiveness of the state (Egorov, Guriev, and Sonin, 2009; Lorentzen, 2014).

argues that "Trump's offensive AI pope picture [was] a distraction from his failing economy" (see also Tavernise and Gardiner (2019); Illing (2020); Lewandowsky, Jetter, and Ecker (2020)).[17] See also discussions of related tactics used by authorities in Russia (Paul and Matthews, 2016; Field et al., 2018; Jasser, Eilen, and Garibay, 2022), Turkey (Butler and Spicer, 2021; Szyma'nski and Cihangiroğlu, 2025), and China (Roberts, 2018, Chapter 1).[18] While our framework does not explain precisely what kind of content attracts attention, our framework does imply that attracting the receiver's attention away from information the leader finds unhelpful can be a robust way to influence the outcome.

5.3 AI Filtering

There is increasing attention to the problem of moderating spaces, such as social media or chat platforms, in which a human receiver interacts with an algorithmic sender such as a large language model.

In some respects, the problem of moderating an algorithmic sender is more difficult than that of moderating a human sender. The requirement of removing information across a set of signal structures (Proposition 5) becomes more demanding as the set of signal structures grows. Algorithmic agents such as large language models can readily cipher and decipher text (e.g., Wei, Haghtalab, and Steinhardt, 2023; Lemkin, 2024) or otherwise optimize output (e.g., Lu et al., 2024) to evade detection. A growing literature studies such vulnerabilities and how to combat them (Shi et al., 2024; Yi et al., 2024; Lin et al., 2025). Glukhov et al. (2023) discuss inherent limits on attempts to censor large language models.

In other respects, the problem of moderating an algorithmic sender is easier than that of moderating a human sender. While it may be difficult to control what a human sender knows or sends, if the moderator controls the technology underlying the algorithm, the moderator may be able to constrain the set of signal structures directly. For example,

[17] Brennan (2025) writes, "Shrugging off a recession was a bad look. Trump knew it and needed a plot twist. If he couldn't win the day, maybe he could distract us from how he was losing. A picture as pope, posted for his 9.7 million followers on Truth Social, would get that done[...]. Attention is the fuel for distraction."

[18] Magicians employ similar tactics. Tamariz (2020, pp. 11–12), for example, explains (emphasis in original):

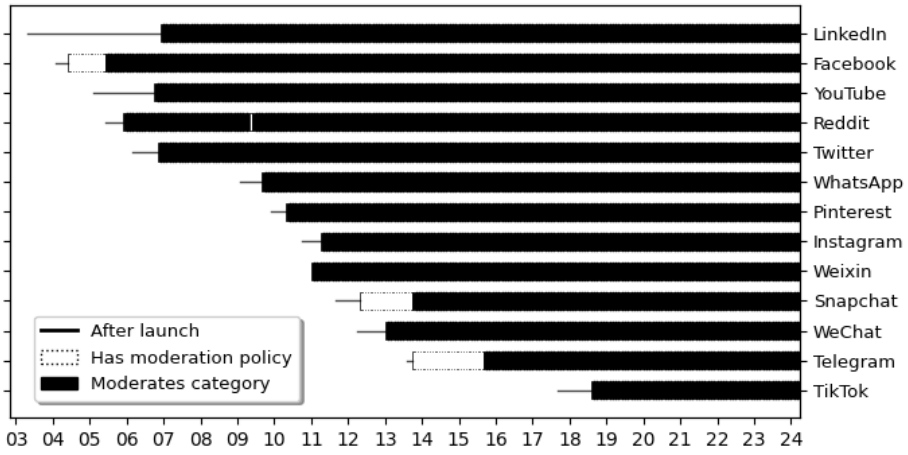
> Gaze at your hands or at an object; then suddenly shift your gaze toward the audience or another object and immediately execute the move. The spectators who gazed at your hands and the object feel forced to shift to the new source of information [...]. THE RESULT IS [t]he spectators *continue to look* at the hands and object but they actually do not *see* what they think they are looking at the whole time.

if a developer of a large language model does not want the model to explain how to build a bomb, it is possible in principle to remove that information from the corpus on which the model is trained, so that the model does not "know" how to build a bomb, and therefore cannot transmit that information to a receiver.[19]

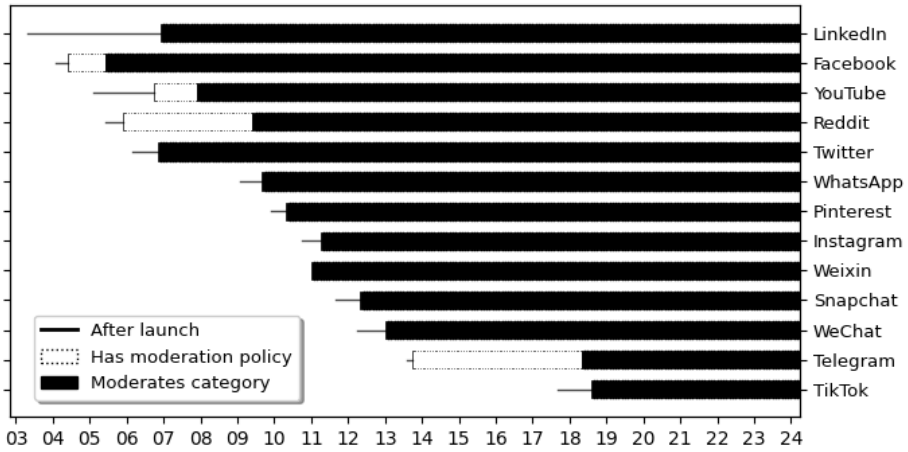
[19]In practice, removing information from a training corpus can be difficult and can reduce model output quality (Pal et al., 2024), leading to efforts to remove information from the model itself after training (Ishibashi and Shimodaira, 2024; Li et al., 2025); see also (Shumailov et al., 2024).

Figure 4: Timeline of Content Moderation Adoption by Policy Category

(a) Nudity, adult content and services



(b) Spam, scams, and malware



(c) Harassment and bullying

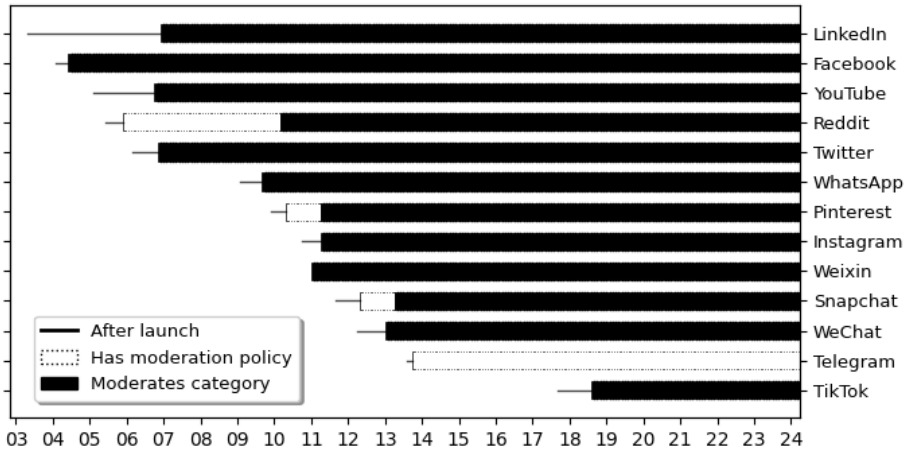
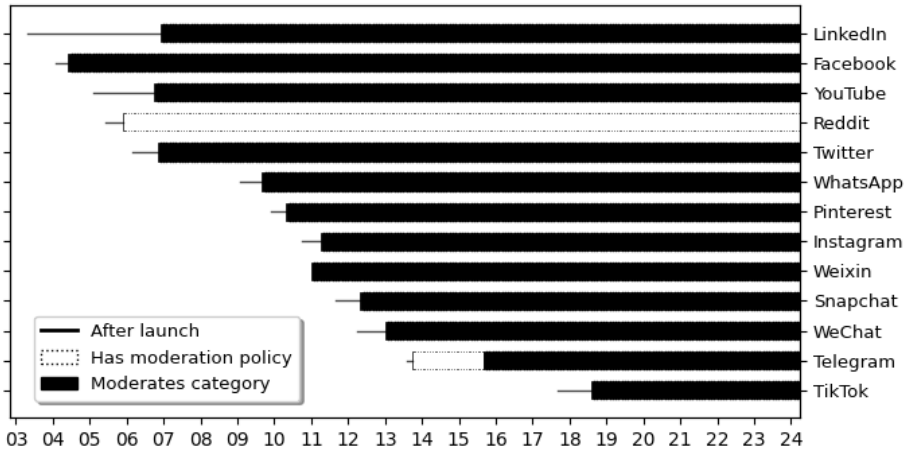
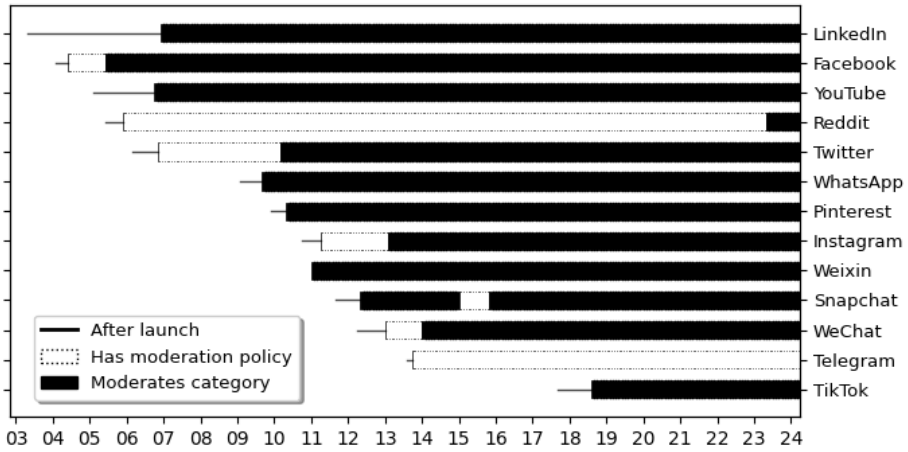


Figure 4: Timeline of Content Moderation Adoption by Policy Category (cont.)

(d) Intellectual property



(e) Hate speech



(f) Privacy and personal information

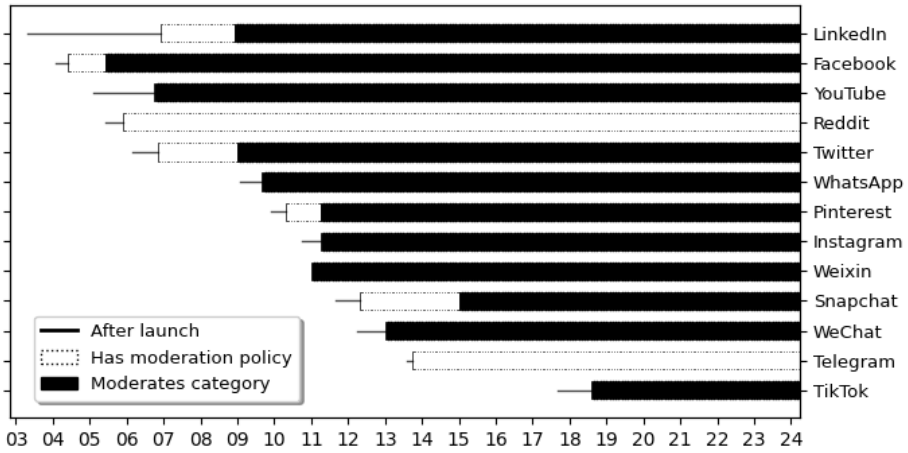
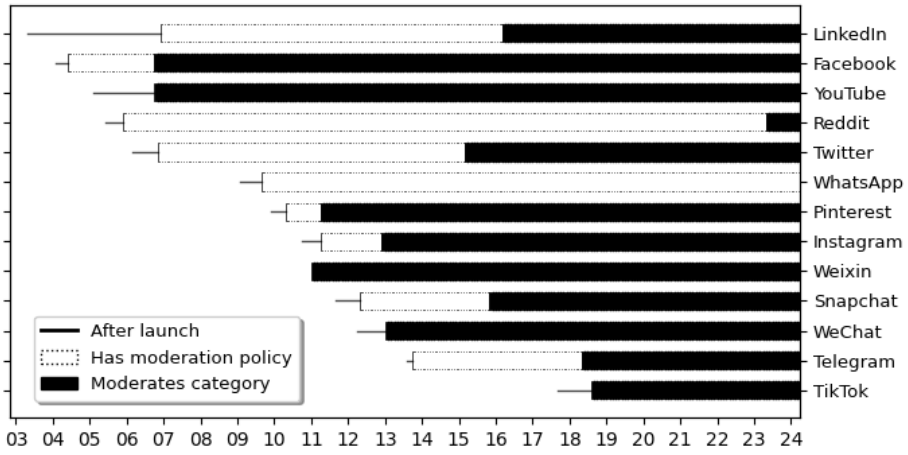
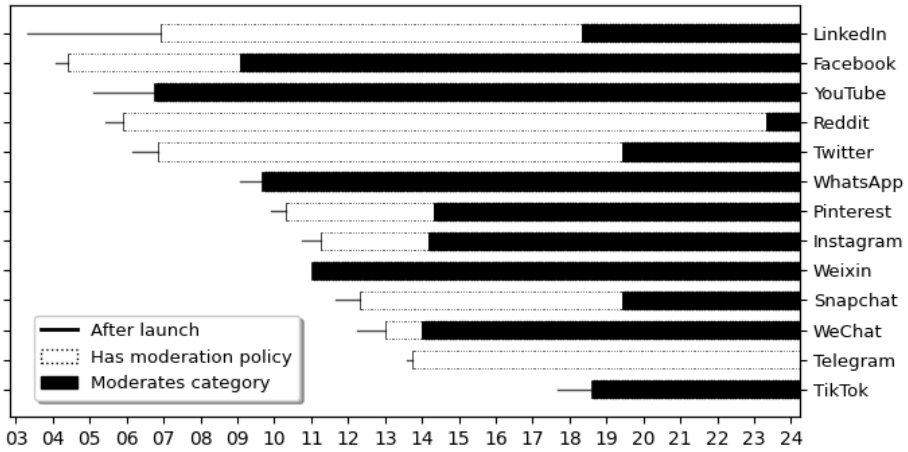


Figure 4: Timeline of Content Moderation Adoption by Policy Category (cont.)

(g) Promoting and depicting violence



(h) Illegal and regulated goods



(i) Self-harm

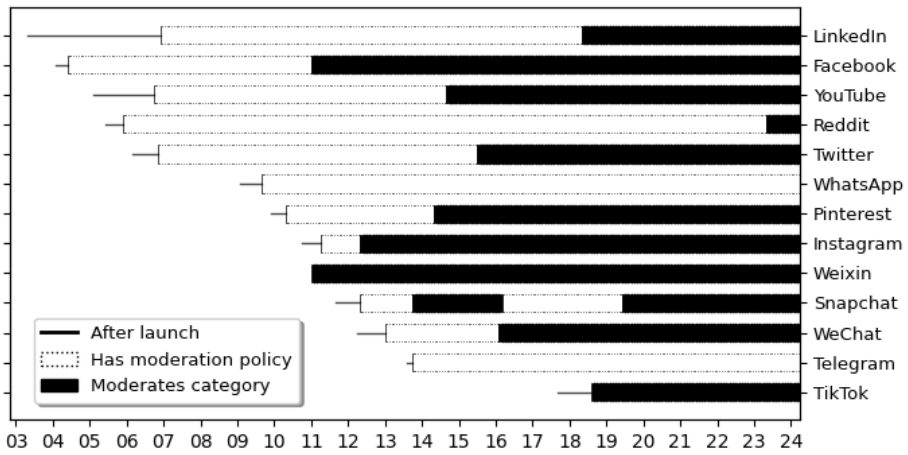
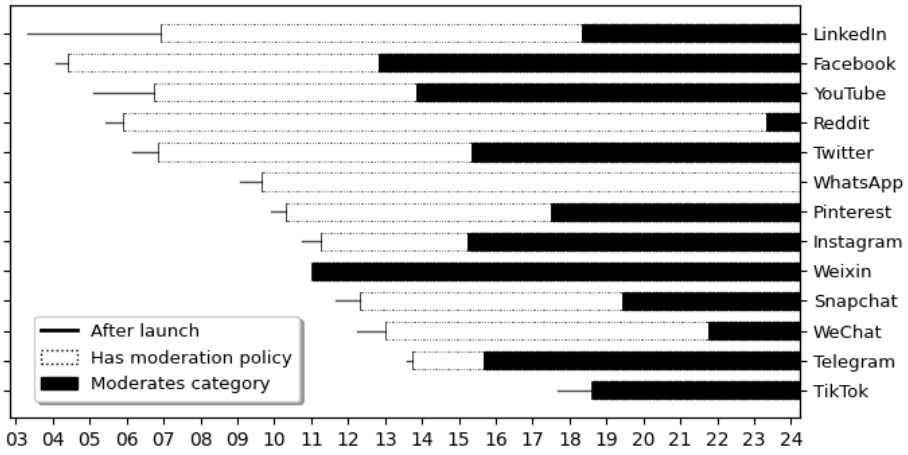
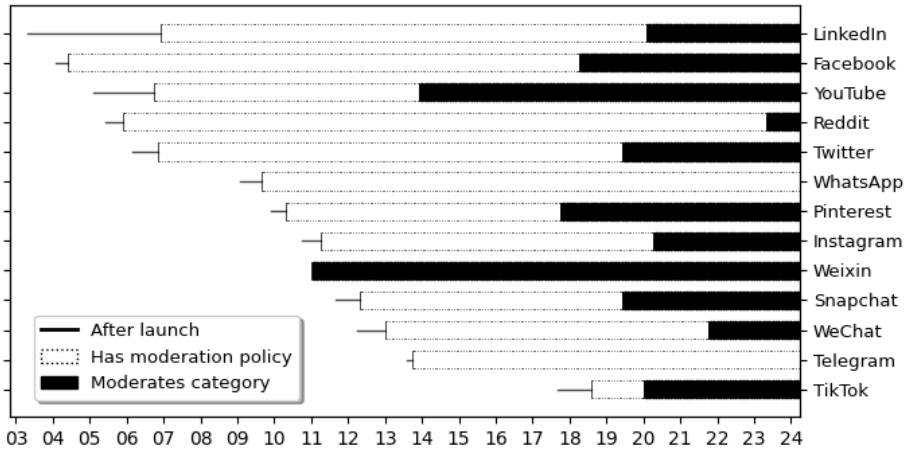


Figure 4: Timeline of Content Moderation Adoption by Policy Category (cont.)

(j) Extremism



(k) Misinformation



(l) Fact-checking

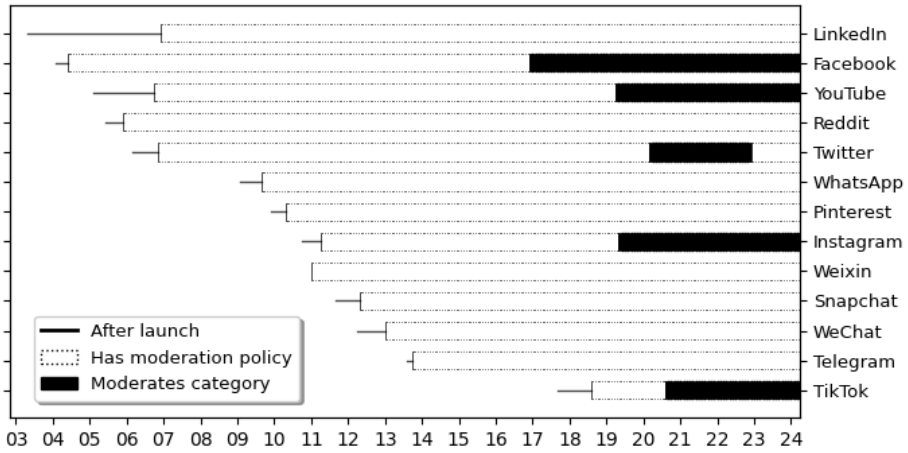


Figure 4: Timeline of Content Moderation Adoption by Policy Category (cont.)

Notes: Data are from the Social Media Content Moderation Policy Archive (Appendix D). Platforms are listed in ascending order by their launch date. Each plot shows the timeline of each platform's moderation of a given category of content, with time measured in months and x-axis labels corresponding to years. The plots are in descending order according to the share of platform-months in which the given category of content is moderated. Black horizontal lines ("After launch") show the period between the platform's launch and the first available policy in the Archive; white bars ("Has policy") show the period when the platform has some content moderation policy but none moderating content in the given category; black bars ("Moderates category") show the period when the platform moderates content in the given category (or, in the case of third-party fact-checking, when such fact-checking is in place).

A Proofs Omitted from the Main Text

Proof of the second part of Lemma 1

For the second part of Lemma 1, we suppose that (1) holds for any Polish space $\mathcal{Z}$. Consider any Polish space $\mathcal{Z}$ and any function $f : [\mathcal{X} \times \mathcal{Z}] \to \mathbb{R}$ for which $\mathrm{E}_{z\sim[\zeta'\circ\nu']_{(x)}}[f(x,z)]$ exists for all $\zeta' \in \Delta(\mathcal{Z})^{\mathcal{Y}}$ and $x \in \mathcal{X}$. It is immediate from the hypothesis (1) that for any $\zeta \in \Delta(\mathcal{Z})^{\mathcal{Y}}$,

$$\text{there is } \zeta' \in \Delta(\mathcal{Z})^{\mathcal{Y}} \text{ s.t. } \mathrm{E}_{z\sim[\zeta\circ\nu]_{(x)}}[f(x,z)] \leq \mathrm{E}_{z\sim[\zeta'\circ\nu']_{(x)}}[f(x,z)] \text{ for all } x \in \mathcal{X}. \quad (2)$$

But now, interpreting f as a(ny) utility function contingent on "state" x and "action" z, (2) implies that for any randomized strategy ζ under ν, there is a randomized strategy ζ' under ν' with weakly higher expected utility; i.e., ν' is *more informative in randomized strategies than* ν in the sense of Khan, Yu, and Zhang (2024, Definition 3), from which the result follows from Remark 1 of Khan, Yu, and Zhang (2024) for Polish $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$, recognizing that the hypothesis of a common dominating measure implies Assumption 1 of Khan, Yu, and Zhang (2024). □

Proof of Corollary 1

If $\mu_0 \circ \sigma$ is uninformative, then (by definition) any $\mu \circ \sigma$ can be garbled to $\mu_0 \circ \sigma$. It follows immediately that no moderation policy removes information from σ, and so (by Proposition 1) no moderation policy is robustly effective for signal structure σ and action rules $\overline{\mathscr{A}}$. □

Proof of Proposition 4

We define the set of **possible outcomes under signal structure σ and moderation policy** μ as

$$\Xi(\sigma, \mu, \overline{\mathscr{A}}) \equiv \bigcup_{\alpha \in \overline{\mathscr{A}}} \{\Xi(\sigma, \mu, \alpha)\}.$$

For the first part, if μ does improve the worst payoff under σ, then (definitionally) we have

$$\inf \boldsymbol{\mathscr{V}}(\sigma, \mu_0) < \inf \boldsymbol{\mathscr{V}}(\sigma, \mu),$$

which means that we must have

$$\Xi(\sigma, \mu_0, \overline{\mathscr{A}}) \not\subseteq \Xi(\sigma, \mu, \overline{\mathscr{A}}). \tag{3}$$

Now, (3) implies that μ must be robustly effective for σ and some α; hence, by Proposition 1, we know it must remove information from σ.

For the second part, we introduce the following lemma, which shows that if an outcome is not achievable under a given signal structure and messaging policy, then no garbling of that outcome is achievable either.

Lemma A.1. *The set $\Xi(\sigma, \mu, \overline{\mathscr{A}})$ is closed under garbling, i.e., if X' garbles some $X \in \Xi(\sigma, \mu, \overline{\mathscr{A}})$, then $X' \in \Xi(\sigma, \mu, \overline{\mathscr{A}})$.*

Proof. Pick any X' that garbles some $X \in \Xi(\sigma, \mu, \overline{\mathscr{A}})$. Then, definitionally, there is some stochastic map $\Gamma : \overline{\mathscr{A}} \to \Delta(\overline{\mathscr{A}})$ such that $X' = \Gamma \circ X$. Because $X \in \Xi(\sigma, \mu, \overline{\mathscr{A}})$, there is some $\alpha \in \overline{\mathscr{A}}$ such that $X = \Xi(\sigma, \mu, \alpha)$. We therefore have

$$\begin{aligned} X' &= \Gamma \circ X \\ &= \Gamma \circ [\Xi(\sigma, \mu, \alpha)] \\ &= \Xi(\sigma, \mu, \Gamma \circ \alpha). \end{aligned}$$

Because $[\Gamma \circ \alpha] \in \overline{\mathscr{A}}$, we have that $\Xi(\sigma, \mu, \overline{\mathscr{A}}) \ni \left[\Xi(\sigma, \mu, \Gamma \circ \alpha)\right] = X'$. Because we picked X' arbitrarily, this completes the proof. □

Now, to prove the second part of the proposition: If μ (nontrivially) preserves better payoffs than $\underline{V}$ under σ, it follows (again definitionally) that there is some $V > \underline{V}$ such that $\mathbf{V}(X) = V$ for some $X \in \Xi(\sigma, \mu, \overline{\mathscr{A}})$. If X' with $\mathbf{V}(X') < \underline{V}$ were to garble X, then, by Lemma A.1, we would have $X' \in \Xi(\sigma, \mu, \overline{\mathscr{A}})$. But by hypothesis, μ improves the worst payoff under σ to $\underline{V}$—hence (definitionally), $X' \notin \Xi(\sigma, \mu, \overline{\mathscr{A}})$ for any X' with $\mathbf{V}(X') < \underline{V}$. Thus, we see that no X' with $\mathbf{V}(X') < \underline{V}$ can garble X.

It follows from the preceding observations that there is some $V > \underline{V}$ and some X with $\mathbf{V}(X) = V$ that is not garbled by any X' with $\mathbf{V}(X') < \underline{V}$. Therefore (definitionally) information is not helpful at $\underline{V}$. □

Proof of Proposition 5

Towards a proof of the contrapositive, we suppose that μ does not remove information from any $\sigma \in \mathscr{S}$ across $\mathscr{S}$. Then (definitionally), for any $\sigma \in \mathscr{S}$, there is some $\sigma' \in \mathscr{S}$

such that $\mu_0 \circ \sigma$ garbles $\mu \circ \sigma'$. Then, for any $\alpha \in \overline{\mathscr{A}}$, the first part of Lemma 1—taking $\nu = \mu_0 \circ \sigma$ and $\nu' = \mu \circ \sigma'$—implies that there is some $\alpha' \in \overline{\mathscr{A}}$ such that $\Xi(\sigma', \mu, \alpha') = \Xi(\sigma, \mu_0, \alpha)$. As our initial choice of $\sigma \in \mathscr{S}$ was arbitrary, we thus see that μ is not robustly effective across $\mathscr{S}$. □

Proof of Corollary 2

Suppose that for each $\sigma \in \mathscr{S}$ there is at least one $\sigma' \in \mathscr{S}$ that is informationally equivalent to σ under μ_0 and from which μ does not remove information across $\mathscr{S}$. Then (definitionally) there is $\sigma'' \in \mathscr{S}$ such that $\mu_0 \circ \sigma'$ garbles $\mu \circ \sigma''$. But because σ' is informationally equivalent to σ under μ_0, we know that $\mu_0 \circ \sigma$ garbles $\mu_0 \circ \sigma'$; it then follows that $\mu_0 \circ \sigma$ garbles $\mu \circ \sigma''$. Thus, we see that μ does not remove information from σ across $\mathscr{S}$. Because the preceding argument applies for each $\sigma \in \mathscr{S}$, it follows that μ does not remove information from *any* $\sigma \in \mathscr{S}$ across $\mathscr{S}$; the corollary then follows from Proposition 5. □

Proof of Claim 2

Suppose μ is robustly effective across $\mathscr{S}$ for $(\alpha, \sigma) = (\alpha, \sigma_i(\mu_0))$ with $(\alpha; i)$ justified by $\beta(\mu_0; \cdot)$. Then (definitionally), there is no $(\sigma', \alpha') \in \left[\mathscr{S} \times \overline{\mathscr{A}}\right]$ such that $\Xi(\sigma', \mu, \alpha') = \Xi(\sigma, \mu_0, \alpha)$. It follows that there is no $\alpha' \in \overline{\mathscr{A}}, i' \in \mathcal{I}$ such that $\Xi(\sigma_{i'}(\mu), \mu, \alpha') = \Xi(\sigma_i(\mu_0), \mu_0, \alpha)$. □

Proof of Proposition 6

We consider some μ and μ_0, along with a tuple $(\alpha, \sigma) = (\alpha, \sigma_i(\mu_0))$ with $(\alpha; i)$ justified by $\beta(\mu_0; \cdot)$. Now, if μ is not robustly effective across $\mathscr{S}$, then the outcome achieved by $(\alpha, \sigma) = (\alpha, \sigma_i(\mu_0))$ under $\beta(\mu_0; i) = \mu_0 \circ \sigma_i(\mu_0)$ can be achieved by some tuple $(\alpha', \sigma') = (\alpha', \sigma_{i'}(\mu)) \in \left[\overline{\mathscr{A}} \times \mathscr{S}\right]$ under $\beta(\sigma_{i'}(\mu); i') = \mu \circ \sigma_{i'}(\mu)$, where the existence of such an $i' \in \mathcal{I}$ follows from our assumption that the set of senders is rich. Then, because each $\mu \circ \sigma$ garbles some $\mu_0 \circ \sigma'$ by hypothesis, we know from the first part of Lemma 1 that the tuple $(\alpha'; i')$ must be justified under accurate beliefs. □

Proof of Proposition 7

Consider any $i' \in \mathcal{I}$. Because μ does not remove information from $\sigma_{i'}(\mu_0)$ across $\mathscr{S}$, there is some $i \in \mathcal{I}$ such that $\mu_0 \circ \sigma_{i'}(\mu_0) = \Gamma_{i \rightsquigarrow i'} \circ \mu \circ \sigma_i(\mu)$ for some $\Gamma_{i \rightsquigarrow i'} : \mathcal{M} \to \Delta(\mathcal{M})$.

Note that we can always define a map $\upsilon : \big[\mathcal{M} \times \big[\Theta \times \mathcal{S}\big]\big] \to \Delta(\mathcal{M})$ such that $\upsilon \circ (\mu_0 \circ \sigma, \sigma) = \mu \circ \sigma$ (where we ease notation by suppressing the dependence on θ). For any full-support distribution $\pi \in \Delta(\Theta)$, we can furthermore define the average map $\Upsilon : \mathcal{M} \to \Delta(\mathcal{M})$ where $\Upsilon(m) = \mathrm{E}_{\theta\sim\pi}\big[\mathrm{E}_{s\sim\sigma(\theta)}[\upsilon(m;\theta,s)]\big]$ for all $m \in \mathcal{M}$.

Now, define $\beta(\mu; i)$ so that, for each $\theta \in \Theta$, $\beta(\mu; i)[\theta]$ is a fixed point of the map that takes value $\mathrm{E}_{m\sim\phi}\big[\big[\Upsilon \circ \Gamma_{i\rightsquigarrow i'}\big](m)\big]$ for generic element $\phi \in \Delta(\mathcal{M})$; such a fixed point exists by fixed-pointiness. Furthermore define $\beta(\mu_0; i') = \Gamma_{i\rightsquigarrow i'} \circ \beta(\mu; i)$. Now, we have $\beta(\mu; i) = \big[\Upsilon \circ \Gamma_{i\rightsquigarrow i'}\big] \circ \beta(\mu; i) = \Upsilon \circ \big[\Gamma_{i\rightsquigarrow i'} \circ \beta(\mu; i)\big] = \Upsilon \circ \beta(\mu_0; i')$, where the first equality follows from our fixed-point construction of the belief $\beta(\mu; i)$.

We can repeat the construction just described for all $i' \in \mathcal{I}$, where by our richness hypothesis on $\mathcal{I}$, we can always choose a different i to correspond with each individual i'. For any $\tilde{i} \in \mathcal{I}$ not thus assigned, suppose that $\beta(\mu; \tilde{i})$ is a constant (which therefore garbles all message distributions). We have ensured that for any $i \in \mathcal{I}$, $\beta(\mu; i)$ garbles $\beta(\mu_0; i')$ for some $i' \in \mathcal{I}$.

Moreover, for any tuple $(\alpha; i')$ that is justified under $\beta(\mu_0; \cdot)$, it follows by the first part of Lemma 1 that some $(\alpha \circ \Gamma_{i\rightsquigarrow i'}; i)$ with $\mu_0 \circ \sigma_{i'}(\mu_0) = \Gamma_{i\rightsquigarrow i'} \circ \mu \circ \sigma_i(\mu)$ is justified under $\beta(\mu; \cdot)$. Because we have $\Xi(\sigma_i(\mu), \mu, \alpha \circ \Gamma_{i\rightsquigarrow i'}) = \Xi(\sigma_{i'}(\mu_0), \mu_0, \alpha)$, this completes the proof. □

Proof of Proposition 8

For the first part, we note that improving the worst payoff requires that $\boldsymbol{\mathscr{V}}(\mathscr{S}, \mu_0)$ not be a subset of $\boldsymbol{\mathscr{V}}(\mathscr{S}, \mu)$; this in turn requires that μ be robustly effective across $\mathscr{S}$, and hence, by Proposition 5, requires that μ removes information from some σ across $\mathscr{S}$.

For the second part, the hypothesis that μ (nontrivially) preserves better outcomes at $\underline{V}$ under signal structures $\mathscr{S}$ means that there is some X with $\mathbf{V}(X) > \underline{V}$ such that $X = \Xi(\sigma, \mu, \alpha)$ for some $(\sigma, \alpha) \in \big[\mathscr{S} \times \overline{\mathscr{A}}\big]$. If information were helpful at $\underline{V}$, then there would be some X' with $\mathbf{V}(X') < \underline{V}$ that garbles X, i.e., with $X' = \Gamma \circ X$ for some $\Gamma : \mathcal{A} \to \Delta(\mathcal{A})$. But then, following the proof of Lemma A.1, we would have $X' = \Xi(\sigma, \mu, \Gamma \circ \alpha)$ with $(\sigma, \Gamma \circ \alpha) \in \big[\mathscr{S} \times \overline{\mathscr{A}}\big]$, so $\mathbf{V}(X') \in \boldsymbol{\mathscr{V}}(\mathscr{S}, \mu)$—contradicting the hypothesis that μ improves the worst payoff to $\underline{V} > \mathbf{V}(X')$. □

B Auxiliary Results

B.1 Receiver Beliefs with Rich Support

Proposition B.1. *Suppose that $\mathcal{M}$ is fixed-pointy, and that moderation policy μ does not remove information from σ and is therefore not robustly effective for σ. Then there exist beliefs $\beta(\mu_0), \beta(\mu)$ that satisfy the conclusions of Proposition 3 such that, for any $\theta \in \Theta$, the support of $\beta(\mu)[\theta]$ contains that of $\mu \circ \sigma(\theta)$ and the support of $\beta(\mu_0)[\theta]$ contains that of $\mu_0 \circ \sigma(\theta)$.*

Proof. The first part of the argument repeats its counterpart from the proof of Proposition 3. If μ does not remove information from σ, then (definitionally) we can write $\mu_0 \circ \sigma = \Gamma \circ \mu \circ \sigma$ for some $\Gamma : \mathcal{M} \to \Delta(\mathcal{M})$. And we can always define a map $\upsilon : \big[\mathcal{M} \times \big[\Theta \times \mathcal{S}\big]\big] \to \Delta(\mathcal{M})$ such that $\upsilon \circ (\mu_0 \circ \sigma, \sigma) = \mu \circ \sigma$, where we continue to ease notation by suppressing the dependence on θ. For any full support distribution $\pi \in \Delta(\Theta)$, we can furthermore define the average map $\Upsilon : \mathcal{M} \to \Delta(\mathcal{M})$ where $\Upsilon(m) = \mathrm{E}_{\theta \sim \pi, s \sim \sigma(\theta)}[\upsilon(m; \theta, s)]$ for all $m \in \mathcal{M}$.

To begin the new part of the argument, we let $\Psi : \mathcal{M} \to \Delta(\mathcal{M})$ be any map such that, for any $m \in \mathcal{M}$, $\Psi(m)$ has full support on $\mathcal{M}$. Let $\hat{\Upsilon} : \mathcal{M} \to \Delta(\mathcal{M})$ be a map such that, for any $m \in \mathcal{M}$, $\hat{\Upsilon}(m)$ mixes with some fixed interior probability between $\Upsilon(m)$ and $\Psi(m)$, and therefore has full support on $\mathcal{M}$. We define $\beta(\mu)$ so that, for each $\theta \in \Theta$, $\beta(\mu)[\theta]$ is a fixed point of the map that takes value $\mathrm{E}_{m \sim \phi}\Big[\big[\hat{\Upsilon} \circ \Gamma\big](m)\Big]$ for generic element $\phi \in \Delta(\mathcal{M})$; such a fixed point exists by fixed-pointiness, and by construction $\beta(\mu)[\theta]$ has full support on $\mathcal{M}$. Furthermore, we define $\beta(\mu_0) = \Gamma \circ \beta(\mu)$.

Now, to show the desired full-support properties of the beliefs we just constructed, we consider some $\theta \in \Theta$. Because $\beta(\mu)[\theta]$ has full support on $\mathcal{M}$ by construction, the support of $\beta(\mu)[\theta]$ contains the support of $\mu \circ \sigma(\theta)$. Moreover, because $\beta(\mu_0) = \Gamma \circ \beta(\mu)$ by construction, the support of $\beta(\mu_0)[\theta]$ is the same as the support of $\big[\Gamma \circ \beta(\mu)\big][\theta]$; hence, the support of $\beta(\mu_0)[\theta]$ contains the support of $\mu_0 \circ \sigma(\theta)$. Because the preceding argument applies to any $\theta \in \Theta$, we have established the desired properties of the supports of $\beta(\mu_0)$ and $\beta(\mu)$.

What remains is to establish the conclusions of Proposition 3. Now, we have $\beta(\mu) = \hat{\Upsilon} \circ \Gamma \circ \beta(\mu) = \hat{\Upsilon} \circ \beta(\mu_0)$, where the first equality follows from our fixed-point construction of $\beta(\cdot)$. Thus, we see that $\beta(\mu)$ garbles $\beta(\mu_0)$, as desired. Moreover, for any action rule α that is justified under $\beta(\mu_0)$, it follows by the first part of Lemma 1 that action rule $\alpha \circ \Gamma$ is justified under $\beta(\mu)$. Because $\Xi(\sigma, \mu, \alpha \circ \Gamma) = \Xi(\sigma, \Gamma \circ \mu, \alpha) = \Xi(\sigma, \mu_0, \alpha)$, this

completes the proof. □

B.2 Sufficient Conditions for Improving the Worst Payoff

Here we establish sufficient conditions for the existence of a moderation policy that improves the worst payoff.

Proposition B.2. *If*

- $\min\{\mathbf{V}(X) : X(\theta) = X(\theta') \text{ for all } \theta, \theta' \in \Theta\}$ *exists, i.e., there is a worst outcome among the set of constant outcomes;*
- $\min \boldsymbol{\mathscr{V}}(\sigma, \mu_0)$ *exists, i.e., there is a worst outcome under signal structure* σ *and default messaging policy* μ_0*;*
- *for any* X' *with* $X' = \Xi(\sigma, \mu_0, \alpha')$ *for some* $\alpha' \in \overline{\mathscr{A}}$ *and* $\mathbf{V}(X') = \min \boldsymbol{\mathscr{V}}(\sigma, \mu_0)$*, there is some* X *such that* X *garbles* X' *but* X' *does not garble* X*; and*
- *there exists a moderation policy that removes information from* σ*;*

then there exists a moderation policy μ *such that* $\inf \boldsymbol{\mathscr{V}}(\sigma, \mu) > \inf \boldsymbol{\mathscr{V}}(\sigma, \mu_0)$*, i.e., such that* μ *improves the worst payoff under* σ *relative to* μ_0*.*

Proof. Let μ be any **constant policy**, i.e., any policy that always reports the same message, irrespective of the signal. If there exists any moderation policy μ' that removes information from σ, then μ must remove information from σ as well.[20]

Now, we have assumed that for any for any X' with $X' = \Xi(\sigma, \mu_0, \alpha')$ for some $\alpha' \in \overline{\mathscr{A}}$ and $\mathbf{V}(X') = \min \boldsymbol{\mathscr{V}}(\sigma, \mu_0)$, there is some X such that X garbles X' but X' does not garble X. We show that this implies that any such X'—of which at least one exists by our assumption that there is a worst outcome under σ and μ_0—must not be feasible under μ, i.e., must not be equal to $\Xi(\sigma, \mu, \alpha)$ for any $\alpha \in \overline{\mathscr{A}}$. To see this, note that if X' were feasible under μ, then X' would have to have the same distribution on $\mathcal{A}$ for all messages $m \in \mathcal{M}$ (since it factors through a constant messaging policy μ). And, then, X' would garble any X; a contradiction to our third hypothesis.

The argument logic shows that when μ is a constant policy, none of the outcomes that can arise under $\mu \circ \sigma$ can be a worst outcome under $\mu_0 \circ \sigma$. Moreover, since $\mu \circ \sigma$

[20] Otherwise, if μ did not remove information from σ, then $\mu_0 \circ \sigma$ would garble a constant message function, in which case it would also have to garble *any* $\mu' \circ \sigma$.

garbles $\mu_0 \circ \sigma$, any outcome that can arise under $\mu \circ \sigma$ can also arise under $\mu_0 \circ \sigma$, by Lemma 1. It follows that if $\min \boldsymbol{\mathscr{V}}(\sigma, \mu)$ exists, then

$$\inf \boldsymbol{\mathscr{V}}(\sigma, \mu) = \min \boldsymbol{\mathscr{V}}(\sigma, \mu) > \min \boldsymbol{\mathscr{V}}(\sigma, \mu_0) = \inf \boldsymbol{\mathscr{V}}(\sigma, \mu_0),$$

as desired. And moreover, we know that $\min \boldsymbol{\mathscr{V}}(\sigma, \mu)$ exists by our hypothesis that there is a worst outcome among the set of constant outcomes—because all such outcomes can arise under $\mu \circ \sigma$, and so $\min \boldsymbol{\mathscr{V}}(\sigma, \mu) = \min \{\mathbf{V}(X) : X(\theta) = X(\theta') \text{ for all } \theta, \theta' \in \Theta\}$. □

C Additive Moderation

The main text focuses on when the moderator can robustly prevent an outcome that is possible under the default messaging policy. Here we consider instead when the moderator can (robustly) enable an outcome that is not possible under the default messaging policy.

Recall that a moderation policy μ is effective for signal structure σ and action rule α if $\Xi(\sigma, \mu, \alpha) \neq \Xi(\sigma, \mu_0, \alpha)$. If, furthermore, there is no $\alpha' \in \overline{\mathscr{A}}$ such that $\Xi(\sigma, \mu, \alpha) = \Xi(\sigma, \mu_0, \alpha')$, then we say that μ is **additively effective for signal structure σ and action rule α**. If this is the case for *some* action rule α, then we say that μ is **additively effective for signal structure σ**.

We then have the following result.

Proposition C.1 (Additively effective moderation with a fixed sender requires adding information)**.** *A moderation policy μ is additively effective for signal structure σ only if $\mu \circ \sigma$ does not garble $\mu_0 \circ \sigma$.*

Proposition C.1 is a consequence of a more general result, which we introduce and prove next as Proposition C.2.

To address the case of many senders (signal structures), say that a moderation policy μ is **additively effective for action rule α and signal structure σ across the set of signal structures** $\mathscr{S}$ if there is no $(\sigma', \alpha') \in \left[\mathscr{S} \times \overline{\mathscr{A}}\right]$ such that $\Xi(\sigma, \mu, \alpha) = \Xi(\sigma', \mu_0, \alpha')$. If this is the case for *some* action rule α and signal structure $\sigma \in \mathscr{S}$, then we say that μ is **additively effective across the set of signal structures** $\mathscr{S}$.

We then have the following result.

Proposition C.2 (Additively effective moderation with many senders requires special information)**.** *A moderation policy μ is additively effective across the set of signal struc-*

tures $\mathscr{S}$ only if there exists $\sigma \in \mathscr{S}$ such that $\mu \circ \sigma$ does not garble $\mu_0 \circ \sigma'$ for any $\sigma' \in \mathscr{S}$.

Proof. We prove the contrapositive: Suppose that for every $\sigma \in \mathscr{S}$, there is some $\sigma' \in \mathscr{S}$ such that $\mu \circ \sigma$ *does* garble $\mu_0 \circ \sigma'$, i.e., such that $\mu \circ \sigma = \Gamma \circ \mu_0 \circ \sigma'$ for some $\Gamma : \mathcal{M} \to \Delta(\mathcal{M})$. We consider any given signal structure $\sigma \in \mathscr{S}$ and action rule $\alpha \in \overline{\mathscr{A}}$, and observe that if we take $\alpha' = \left[\alpha \circ \Gamma\right] \in \overline{\mathscr{A}}$ (where here $\Gamma : \mathcal{M} \to \Delta(\mathcal{M})$ is such that $\mu \circ \sigma = \Gamma \circ \mu_0 \circ \sigma'$, as just described), then we have $\Xi(\sigma, \mu, \alpha) = \Xi(\sigma', \mu_0, \alpha')$. Because the construction just described works for any $\sigma \in \mathscr{S}$ and $\alpha \in \overline{\mathscr{A}}$, we see that μ is not additively effective for any action rule α and signal structure σ across the set of signal structures $\mathscr{S}$. □

D Construction of Social Media Content Moderation Policy Archive

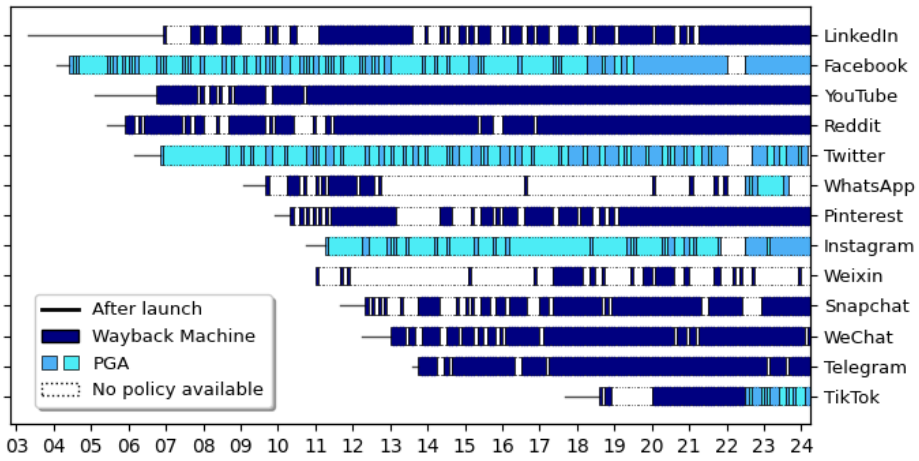
D.1 Sources for Content Moderation Policies

We began with the list of social media platforms with over 400 million monthly active users by May 2024 according to Wikipedia contributors (2024). We excluded Douyin, Kuaishou, Weibo, QQ, and Qzone because they do not regularly publish policies in English. We excluded Quora because we were not able to find historical snapshots of its policies. We excluded Messenger because it shares the same policies as Facebook. We included Weixin because it does not share the same policies as WeChat.

For each of the remaining 13 platforms, we attempted to obtain historical snapshots of that platform's English-language content moderation policies in each month since the platform's inception from the Wayback Machine of the Internet Archive (2003–2024) and the Platform Governance Archive (PGA, Katzenbach et al. (2023*a*,*b*,*c*)), prioritizing the Wayback Machine for platforms and periods in which its coverage is sufficiently complete. In cases where a platform described its policies in multiple documents (e.g., in both its "terms of use" and its "community guidelines"), we used all such documents. In cases where the source included multiple versions of a given document for a given platform in a given month, we used the one with the largest file size. We refer to the resulting collection of documents for a given platform in a given month as the platform's *policy* in that month.

For each platform, we define an analysis period beginning with the platform's first

Appendix Figure A1: Coverage and Sources for Content Moderation Policies



Notes: The unit of analysis is platform-month. Platforms are listed in ascending order by launch date. The black horizontal line ("After launch") represents the months following the platform's launch and before the first available policy in our archive. The shaded bar represents the months following the first available policy for which a policy is available; its shade denotes the data source. The unshaded bar ("No policy available") represents the months following the first available policy for which a policy is not available. For PGA, lighter shading reflects months when PGA considers the policy to be unchanged from the previous month.

observed policy and ending in April 2024. For each platform in each month of the analysis period, we use the most recently available policy if no policy is available for that month. Figure A1 visualizes the coverage of the archive and the data sources.

D.2 Classification of Content Moderation Policies

We define a set of policy categories that represent important recurring themes in content moderation. We classify policies into categories by segmenting them into sentences and then applying rules based on regular expression patterns. We consider a platform to have a policy in a given category if at least one sentence in the policy is classified as belonging to that category.

To segment policies into sentences, we first clean each policy document by fixing malformed or stray punctuation, collapsing whitespace, and unwrapping link text. We then segment text into sentences using punctuation (e.g., period, question mark, bullet characters), while avoiding splits for common abbreviations (e.g., "e.g.," "i.e.," "etc."). We begin a new sentence after section-like numbering (e.g., "1.2.3"). We remove leading and trailing punctuation. We exclude from classification very short sentences or those

that are formatted as titles.

We define each policy category by a set of necessary groups of regular expressions and a set of inadmissible regular expressions, which we determined by reading the policies. We consider a sentence to belong to a category if the sentence has at least one regular expression in each of the necessary groups and no inadmissible regular expressions. For example, a sentence concerns privacy and personal information if and only if it contains an action (e.g., "share", "disclose"), a subject (e.g., "others", "third parties"), and a type of private information (e.g., "personal information", "email address"), and does not contain references suggesting it is about the terms between the platform and the user (e.g., "third-party service providers", "advertisers", "legal requirements"). Likewise, a sentence concerns intellectual property if and only if it contains an action (e.g., "infringing", "posting"), an intellectual property asset (e.g., "copyright", "trademark"), and a subject (e.g., "others"), and again does not contain references suggesting it is about the terms between the platform and the user (e.g., "settlement", "dispute").

We allow two exceptions to this approach. First, in the case of misinformation, we allow richer boolean logic across groups. Specifically, a sentence concerns misinformation if and only if it explicitly mentions "misinformation" or "disinformation", or contains one reference to false or misleading claims (e.g., "false", "misleading") together with one reference to civic/health topics (e.g., "election", "vaccine"), and again does not include language that suggests a different intent. Second, in the case of fact-checking, we coded these policies by hand as they often appear in separate documents and use varying language.[21]

Table A1 lists three representative sentences in each category.

Appendix Table A1: Representative Sentences by Policy Category

Platform	Sentence
Nudity, adult content and services	
Facebook	[Do not post] Digitally-created depictions of nude minors, unless the image is for health or educational purposes / Imagery that depicts non-sexual child abuse regardless of sharing intent / Content that praises, supports, promotes, advocates for, provides instructions for or encourages participation in non-sexual child abuse

[21]The source documents are Mosseri (2016); Katz (2019); Instagram (2019); Gadde and Derella (2020); Pappas (2020); O'Sullivan (2022).

Snapchat	We prohibit promoting, distributing, or sharing pornographic content
Instagram	[Do not post] Content that offers or asks for pornographic material including, but not limited to, sharing of links to external pornographic websites
Spam, scams, and malware	
Snapchat	We prohibit spam, including undisclosed paid or sponsored content, pay-for-follower promotions or other follower-growth schemes, the promotion of spam applications, or the promotion of multilevel marketing or pyramid schemes
LinkedIn	Do not engage in spam or scams: We don't allow untargeted, irrelevant, obviously unwanted, unauthorized, inappropriately commercial or promotional, or gratuitously repetitive messages or other similar content
TikTok	[Do not post] Content or activity that seeks to artificially inflate popularity on the platform is prohibited
Harassment and bullying	
Facebook	[Do not post] Content that violates the bullying and harassment policies for private individuals or
Instagram	[Do not] Post content calling for or stating an intent to engage in behavior that would qualify as bullying and harassment under our policies
LinkedIn	Do not post harassing content: We don't allow bullying or harassment
Intellectual property	
TikTok	Do not post, upload, stream, or share: Content that violates or infringes someone else's copyrights, trademarks, or other intellectual property rights
Instagram	If you violate these Terms of Use or our policies, if you repeatedly infringe other people's intellectual property rights, or where we are required to do so by law, we may –

Facebook	We may also suspend or disable your account if you repeatedly infringe other people's intellectual property rights or where we are required to do so for legal reasons
Hate speech	
TikTok	We do not tolerate content that attacks or incites violence against an individual or a group of individuals on the basis of protected attributes
YouTube	Content that promotes violence or hatred against individuals or groups based on certain attributes isn't allowed under our Hate speech policies
Twitter	You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease
Privacy and personal information	
Twitter	You may not publish or post other people's private and confidential information, such as credit card numbers, street address or Social Security/National Identity numbers, without their express authorization and permission
YouTube	Other types of content that violate this policy Revealing someone's private information, such as their home address, email addresses, sign-in credentials, phone numbers, passport number, or bank account information
WeChat	We are committed to protecting our users' privacy and private/confidential information and prohibit the sharing or publication of any such information without the relevant user's express consent or other legal basis
Promoting and depicting violence	
Twitter	[Do not post] Graphic violence, adult content, and hateful imagery • live video and profile images –

Pinterest	We limit the distribution of or remove such content, including: Content that shows the use of violence / Disturbing scenes from before or after violent events / Threats or language that glorifies violence
Facebook	[Do not post] Content that contains sadistic remarks and any visual or written depiction of real people experiencing premature death, serious physical injury, physical violence or domestic violence
Illegal and regulated goods	
Facebook	[Do not post] Content that attempts to buy, sell, trade, donate or gift or asks for hazardous goods and materials
TikTok	Do not post, upload, stream, or share: Content that displays firearms, firearm accessories, ammunition, or explosive weapons / Content that offers the purchase, sale, trade, or solicitation of firearms, accessories, ammunition, explosive weapons, or instructions on how to manufacture them / Drugs, controlled substances, alcohol, and tobacco
Instagram	[Do not post] Attempts to buy, sell, trade, co-ordinate the trade of, donate, gift or ask for high-risk drugs
Self-harm	
YouTube	When you create content that contains suicide or self-harm related topics, take into account the possible negative impact of your content on other users, especially minors and users who may be sensitive to this content
Facebook	We prohibit content that promotes or encourages suicide or any other type of self-injury, including self
TikTok	Do not post, upload, stream, or share: Content that depicts, promotes, normalizes, or glorifies suicide or self-harm / Content that provides instructions for suicide or how to engage in self-harm / Suicide or self harm games, dares, challenges, pacts, or hoaxes
Extremism	

Snapchat	Terrorist organizations and hate groups are prohibited from using our platform and we have no tolerance for content that advocates or advances violent extremism or terrorism
TikTok	We do not allow dangerous individuals or organizations to use our platform to promote terrorism, crime, or other types of behavior that could cause harm
Twitter	There is no place on X for violent and hateful entities, including (but not limited to) terrorist organizations, violent extremist groups, perpetrators of violent attacks, or individuals who affiliate with and promote their illicit activities
Misinformation	
Snapchat	We prohibit spreading false information that causes harm or is malicious, such as denying the existence of tragic events, unsubstantiated medical claims, undermining the integrity of civic processes, or manipulating content for false or misleading purposes
TikTok	We do not permit misinformation that could cause harm to our community or the larger public
Pinterest	We don't allow false or misleading content that impedes the integrity of an election or an individual's or group's civic participation, including registering to vote, voting, and being counted in a census
Fact-checking	
Instagram	In May of this year, we began working with third-party fact-checkers in the US to help identify, review, and label false information.
TikTok	Fact-checking helps confirm that we remove verified misinformation and reduce mistakes in the content moderation process.
YouTube	By following this process, any eligible publisher can contribute fact check articles that could show in search results on Google Search, Google News, and now, YouTube.

Notes: Data are from the Social Media Content Moderation Policy Archive (Appendix D). Each panel lists the three most representative sentences in each category, ranked by descending representativeness. The panels are in descending order according to the share of platform-months in which the given category of content is moderated. To select representative sentences, we first sample up to

1000 sentences in each category, without replacement and with sampling probabilities that guarantee equal representativeness across platform-months. (For the fact-checking category, we instead use all sentences about fact-checking or trusted partners appearing in the platform's first official announcement of its third-party fact-checking program.) We then define a sentence's representativeness as its average pairwise cosine similarity with all other sampled sentences, using embeddings from the *sentence-transformers/all-mpnet-base-v2* transformer model (Song et al. 2020). Within each category we consider for display only the most representative sentence for each platform. We display sentences after trimming some extraneous text and using slashes to separate items in unpunctuated lists.

References

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius. 2024. "A model of online misinformation." *Review of Economic Studies*, 91(6): 3117–3150. 7

Allen, Jennifer, Antonio A. Arechar, Gordon Pennycook, and David G. Rand. 2021. "Scaling up fact-checking using the wisdom of crowds." *Science Advances*, 7(36). 36

Amadon, Patrick. n.d.. "VeilPNG Alpha." Accessed 2025-09. https://www.transient.xyz/mint/veilpng-alpha. 39

Ambrus, Attila, Eduardo M. Azevedo, and Yuichiro Kamada. 2013. "Hierarchical cheap talk." *Theoretical Economics*, 8(1): 233–261. 6

Andres, Raphaela and Olga Slivko. 2023. "Combating online hate speech: The impact of legislation on Twitter." *ZEW Discussion Paper No. 21-103*. 34

Appel, Ruth E., Jennifer Pan, and Margaret E. Roberts. 2023. "Partisan conflict over content moderation is more than disagreement about facts." *Science Advances*, 9(44): eadg6799. 37

Babichenko, Yakov, Inbal Talgam-Cohen, Haifeng Xu, and Konstantin Zabarnyi. 2022. "Regret-minimizing Bayesian persuasion." *Games and Economic Behavior*, 136: 226–248. 7

Bachmann, Ingrid and Sebastián Valenzuela. 2023. "Studying the downstream effects of fact-checking on social media: Experiments on correction formats, belief accuracy, and media trust." *Social Media + Society*, 9(2): 20563051231179694. 37

Bar-Isaac, Heski, Rahul Deb, and Matthew Mitchell. 2025. "Selling certification, content moderation, and attention." *arXiv preprint arXiv:2506.12604*. 7

BBC. 2016. "Panama papers: China censors online discussion." *BBC News*. Accessed 2025-03. https://www.bbc.com/news/world-asia-china-35957235. 38

Beazer, Quintin H., Charles D. Crabtree, Christopher J. Fariss, and Holger L. Kern. 2022. "When do private actors engage in censorship? Evidence from a correspondence experiment with Russian private media firms." *British Journal of Political Science*, 52(4): 1790–1809. 38

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 No. 11644*. 34

Bergemann, Dirk and Stephen Morris. 2019. "Information design: A unified perspective." *Journal of Economic Literature*, 57(1): 44–95. 6

Bernard, Andrew B., Esther Bøler, Davin Chor, Sirig Gurung, and Wei Lu. 2025. "The Great Firewall and knowledge diffusion." *Dartmouth College Working Paper*. 39

Billingsley, Patrick. 2013. *Convergence of Probability Measures*. John Wiley & Sons. 18

Blackwell, David. 1951. "Comparison of experiments." In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. Jerzy Neyman, pp. 93–102. University of California Press. 3, 10

Blackwell, David. 1953. "Equivalent comparisons of experiments." *Annals of Mathematical Statistics*, 24(2): 265–272. 3, 10

Brennan, Chris. 2025. "Trump's offensive AI pope picture is a distraction from his failing economy." *USA Today*. Accessed 2025-05. https://www.usatoday.com/story/opinion/columnist/2025/05/05/trump-pope-ai-image-economy-recession/83445449007/. 39, 40

Busby, Mattha. 2024. "Drug dealers have moved on to social media." *Wired*. Accessed 2025-09. https://www.wired.com/story/drug-dealers-have-moved-onto-social-media/. 2

Butler, Daren and Jonathan Spicer. 2021. "Erdogan's critics say expulsion call is diversion from economic woe." *Reuters*. Accessed 2025-05. https://www.reuters.com/world/middle-east/

erdogans-critics-say-demand-expulsions-is-distraction-economy-woes-2021-10-24/. 40

Candogan, Ozan and Kimon Drakopoulos. 2020. "Optimal signaling of content accuracy: Engagement vs. misinformation." *Operations Research*, 68(2): 497–515. 7

Castiglioni, Matteo, Andrea Celli, Alberto Marchesi, and Nicola Gatti. 2020. "Online Bayesian persuasion." In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 16188–16198. 7

Cato Institute. 2021. *Cato 2021 Speech and Social Media National Survey: Topline Results.* Cato Institute. Accessed 2025-09. https://www.cato.org/sites/cato.org/files/2021-12/cato-social-media-survey-report-toplines.pdf. 2

Chang, Dongkyu, Adrian Segura, and Pengfei Zhang. 2024. "Decentralizing content moderation." *Available at SSRN 4709599*. 7

Corduneanu-Huci, Cristina and Alexander Hamilton. 2022. "Selective control: The political economy of censorship." *Political Communication*, 39(4): 517–538. 38

Corse, Alexa, Meghan Bobrowsky, and Jeff Horwitz. 2025. "Social-media companies decide content moderation is trending down." *The Wall Street Journal*. Accessed 2025-05. https://www.wsj.com/tech/social-media-companies-decide-content-moderation-is-trending-down-25380d25. 36

Crawford, Vincent. 1998. "A survey of experiments on communication via cheap talk." *Journal of Economic Theory*, 78(2): 286–298. 6

de Oliveira, Henrique. 2018. "Blackwell's informativeness theorem using diagrams." *Games and Economic Behavior*, 109: 126–131. 11

Drolsbach, Chiara Patricia, Kirill Solovev, and Nicolas Pröllochs. 2024. "Community notes increase trust in fact-checking on social media." *PNAS Nexus*, 3(7): pgae217. 37

Drug Enforcement Administration. 2021. "Emoji Drug Code." Accessed 2025-06. https://www.dea.gov/sites/default/files/2021-12/Emoji%20Decoded.pdf. 35

Dubois, Elizabeth and Anna Reepschlager. 2024. "How harassment and hate speech policies have changed over time: Comparing Facebook, Twitter and Reddit (2005-2020)." *Policy & Internet*, 16(3): 523–542. 33

Dworczak, Piotr and Alessandro Pavan. 2022. "Preparing for the worst but hoping for the best: Robust (Bayesian) persuasion." *Econometrica*, 90(5): 2017–2051. 7

Dwork, Cynthia, Chris Hays, Jon Kleinberg, and Manish Raghavan. 2024. "Content moderation and the formation of online communities: A theoretical framework." In *Proceedings of the ACM Web Conference 2024*, pp. 1307–1317. Association for Computing Machinery. 7

Edelman Trust Institute. 2024. *2024 Edelman Trust Barometer Global Report.* Edelman Trust Institute. Accessed 2025-09. https://www.edelman.com/sites/g/files/aatuss191/files/2024-02/2024%20Edelman%20Trust%20Barometer%20Global%20Report_FINAL.pdf. 2

Egorov, Georgy, Sergei Guriev, and Konstantin Sonin. 2009. "Why resource-poor dictators allow freer media: A theory and evidence from panel data." *American Political Science Review*, 103(4): 645–668. 39

Engle, Stephen. 2015. "Tiananmen anniversary date makes money transfers tricky." *Bloomberg*. Accessed 2025-05. https://www.bloomberg.com/news/videos/2015-06-04/tiananmen-anniversary-date-makes-money-transfers-tricky. 39

Facebook. 2025. "Transparency Report March 2025." Facebook Transparency Center. Accessed 2025-08. https://web.archive.org/web/20250621151215/https://disinfocode.eu/reports/facebook/5/text. 37

Farrell, Joseph. 1993. "Meaning and credibility in cheap-talk games." *Games and Economic Behavior*, 5(4): 514–531. 32

Farrell, Joseph and Matthew Rabin. 1996. "Cheap talk." *Journal of Economic Perspectives*, 10(3): 103–118. 6

Field, Anjalie, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. "Framing and agenda-setting in Russian news: A computational analysis of intricate political strategies." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*Association for Computational Linguistics. 40

Fridkin, Kim, Patrick J. Kenney, and Amanda Wintersieck. 2015. "Liar, liar, pants on fire: How fact-checking influences citizens' reactions to negative advertising." *Political Communication*, 32(1): 127–151. 37

Fu, Angela. 2024. "It's easy to find misinformation on social media. It's even easier on X." *Poynter*. Accessed 2025-08. https://www.poynter.org/fact-checking/2024/how-elon-musk-twitter-takeover-accelerated-misinformation/. 36

Gadde, Vijaya and Matt Derella. 2020. "An update on our continuity strategy during COVID-19." *Twitter Blog*. Accessed in 2025-08. https://web.archive.org/web/20250824095751/https://blog.x.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19. 56

Gentzkow, Matthew and Emir Kamenica. 2017. "Bayesian persuasion with multiple senders and rich signal spaces." *Games and Economic Behavior*, 104: 411–429. 7

Gentzkow, Matthew and Jesse M. Shapiro. 2006. "Media bias and reputation." *Journal of Political Economy*, 114(2): 280–316. 37

Glukhov, David, Ilia Shumailov, Yarin Gal, Nicolas Papernot, and Vardan Papyan. 2023. "LLM censorship: A machine learning challenge or a computer security problem?" *arXiv preprint arXiv:2307.10719*. 40

Gueorguiev, Dimitar D. and Edmund J. Malesky. 2019. "Consultation and selective censorship in China." *Journal of Politics*, 81(4): 1539–1545. 38

Guriev, Sergei, Emeric Henry, Théo Marquis, and Ekaterina Zhuravskaya. 2025. "Curtailing false news, amplifying truth." *Available at SSRN 4616553*. 37

Hairer, Martin. 2018. "Ergodic Properties of Markov Processes." Notes from lecture given at the University of Warwick, Spring 2006. Accessed 2025-10. https://hairer.org/notes/Markov.pdf. 18

Hojati, Afrouz and Barrie R. Nault. forthcoming. "Content moderation with shadowbanning." *Information Systems Research*. 7

Hossain, Safwan, Andjela Mladenovic, Yiling Chen, and Gauthier Gidel. 2024. "A persuasive approach to combating misinformation." In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, pp. 18926–18943. 7

Hu, Ju and Xi Weng. 2021. "Robust persuasion of a privately informed receiver." *Economic Theory*, 72(3): 909–953. 7

Illing, Sean. 2020. ""Flood the zone with ****": How misinformation overwhelmed our democracy." *Vox*. Accessed 2025-05. https://www.vox.com/policy-and-politics/2020/1/16/20991816/impeachment-trial-trump-bannon-misinformation. 40

Instagram. 2019. "Combatting misinformation on Instagram." *Instagram Blog*. Accessed in 2025-08. https://web.archive.org/web/20250529000944/https://about.instagram.com/blog/announcements/combatting-misinformation-on-instagram. 56

Institute for Strategic Dialogue. 2025. "Hidden hate: How Amharic is being used to evade hate speech detection on TikTok." *Institute for Strategic Dialogue*. Accessed 2025-09. https://www.isdglobal.org/digital_dispatches/hidden-hate-how-amharic-is-being-used-to-evade-hate-speech-detection-on-tiktok/. 2

Internet Archive. 2003–2024. "Wayback Machine, various entries." *San Francisco, CA: Internet Archive*, Accessed 2023–2024 https://web.archive.org/. 54

Ishibashi, Yoichi and Hidetoshi Shimodaira. 2024. "Knowledge sanitization of large language models." *arXiv preprint arXiv:2309.11852*. 41

Ivanov, Maxim. 2010. "Communication via a strategic mediator." *Journal of Economic Theory*, 145(2): 869–884. 6

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. 7

Jasser, Jasser, Dan Eilen, and Ivan Garibay. 2022. "Flooding the zone: A censorship and disinformation strategy that needs attention." In *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media*AAAI. 40

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*. 34

Jiménez-Durán, Rafael, Karsten Müller, and Carlo Schwarz. 2023. "The effect of content moderation on online and offline hate: Evidence from Germany's NetzDG." *Available at SSRN 4230296*. 34

Kallenberg, Olav. 2021. *Foundations of Modern Probability*. Springer. 18

Kamenica, Emir. 2019. "Bayesian persuasion and information design." *Annual Review of Economics*, 11(1): 249–272. 6

Kaplan, Joel. 2025. "More speech and fewer mistakes." *Meta Newsroom*. Accessed 2025-08. https://web.archive.org/web/20250819070022/https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/. 36

Karimi, Younes, Anna Squicciarini, and Shomir Wilson. 2022. "Automated detection of doxing on Twitter." *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–24. 36

Kartik, Navin, Marco Ottaviani, and Francesco Squintani. 2007. "Credulity, lies, and costly talk." *Journal of Economic Theory*, 134(1): 93–116. 32

Katzenbach, C., D. Dergachava, A. Fischer, A. Kopps, S. Kolesnikov, D. Redeker, and P. Viejo Otero. 2023*a*. "Platform Governance Archive (PGA) v2. [data set]." Accessed 2025-07. https://www.platformgovernancearchive.org/data/dataset-pga-v2-ongoing-collection/. 54

Katzenbach, Christian, Adrian Kopps, Joao C. Magalhaes, Dennis Redeker, and Tom Suhr. 2023*b*. "Platform Governance Archive (PGA) v1. [data set]." Accessed 2025-07. https://www.platformgovernancearchive.org/data/dataset-pga-v1-historical-dataset/. 54

Katzenbach, Christian, Joao C. Magalhaes, Adrian Kopps, Tom Suhr, and Larissa Wunderlich. 2023*c*. "The Platform Governance Archive (PGA)." Accessed 2025-07. https://platformgovernancearchive.org. 54

Katz, Tim. 2019. "Bringing greater transparency and context for news content on YouTube in India." *Google India Blog*. Accessed in 2025-08. https://web.archive.org/web/20250525004416/https://india.googleblog.com/2019/04/bringing-greater-transparency-and.html. 56

Kemp, David and Emily Ekins. 2021. "Poll: 75% don't trust social media to make fair content moderation decisions, 60% want more control over posts they see." *Cato Institute – Survey Reports*. Accessed 2025-09. https://www.cato.org/survey-reports/poll-75-dont-trust-social-media-make-fair-content-moderation-decisions-60-want-more. 2

Khan, M. Ali, Haomiao Yu, and Zhixiang Zhang. 2024. "On comparisons of information structures with infinite states." *Journal of Economic Theory*, 218: 105841. 11, 47

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. "How censorship in China allows government criticism but silences collective expression." *American Political Science Review*, 107(2): 326–343. 38

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2014. "Reverse-engineering censorship in China: Randomized experimentation and participant observation." *Science*, 345(6199): 1251722. 38

Klinenberg, Danny. Forthcoming. "Does deplatforming work?" *Journal of Conflict Resolution*. 36

Kong, Dongmin, Chen Lin, Lai Wei, and Jian Zhang. 2022. "Information accessibility and corporate innovation." *Management Science*, 68(11): 7837–7860. 39

Kosterina, Svetlana. 2022. "Persuasion with unknown beliefs." *Theoretical Economics*, 17(3): 1075–1107. 7

Lei, Ya-Wen. 2018. *The Contentious Public Sphere: Law, Media, and Authoritarian Rule in China*. Princeton, NJ:Princeton University Press. 38

Lemkin, Benjamin. 2024. "Using hallucinations to bypass GPT4's filter." *arXiv preprint arXiv:2403.04769*. 40

Levine, Alexandra S. 2022. "From camping to cheese pizza, 'algospeak' is taking over social media." *Forbes*. Accessed 2025-06. https://www.forbes.com/sites/alexandralevine/2022/09/16/algospeak-social-media-survey/?sh=5bcf6c55e109. 35

Lewandowsky, Stephan, Michael Jetter, and Ullrich K. H. Ecker. 2020. "Using the president's tweets to understand political diversion in the age of social media." *Nature Communications*, 11(1): 1–12. 40

Lewis, Michael. 2018. "Has anyone seen the President?" *Bloomberg*. Accessed 2025-05. https://www.bloomberg.com/view/articles/2018-02-09/has-anyone-seen-the-president. 6, 39

Li, Na, Chunyi Zhou, Yansong Gao, Hui Chen, Zhi Zhang, Boyu Kuang, and Anmin Fu. 2025. "Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects." *IEEE Transactions on Neural Networks and Learning Systems*, 36(8): 13709–13729. 41

Lin, Lizhi, Honglin Mu, Zenan Zhai, Minghan Wang, Yuxia Wang, Renxi Wang, Junjie Gao, Yixuan Zhang, Wanxiang Che, Timothy Baldwin, et al. 2025. "Against the Achilles' heel: A survey on red teaming for generative models." *Journal of Artificial Intelligence Research*, 82: 687–775. 40

Liu, Yi, Pinar Yildirim, and Z. John Zhang. 2022. "Implications of revenue nodels and technology for content moderation strategies." *Marketing Science*, 41(4): 831–847. 7

Lomborg, Bjorn. 2022. "Facebook, other tech giants censor inconvenient facts about climate change." *New York Post*. Accessed 2023-10. https://web.archive.org/web/20231026004136/https://nypost.com/2022/02/07/facebook-other-tech-giants-censor-facts-about-climate-change/. 37

Lorentzen, Peter. 2014. "China's strategic censorship." *American Journal of Political Science*, 58(2): 402–414. 39

Lu, Ning, Shengcai Liu, Rui He, Yew-Soon Ong, Qi Wang, and Ke Tang. 2024. "Large language models can be guided to evade AI-generated text detection." *arXiv preprint arXiv:2305.10847*. 40

Madio, Leonardo and Martin Quinn. 2025. "Content moderation and advertising in social media platforms." *Journal of Economics & Management Strategy*, 34(2): 342–369. 7

Martel, Cameron and David G. Rand. 2024. "Fact-checker warning labels are effective even for those who distrust fact-checkers." *Nature Human Behaviour*, 8(10): 1957–1967. 37

Mekacher, Amin, Max Falkenberg, and Andrea Baronchelli. 2023. "The systemic impact of deplatforming on social media." *PNAS Nexus*, 2(11): pgad346. 36

Meta. 2023. "Facebook Community Standards Enforcement Report for Dangerous Organizations." Accessed 2023-10. https://transparency.fb.com/reports/community-standards-enforcement/dangerous-organizations/facebook/. 2

Meta. 2024. "Privacy Violations." Accessed 2025-06. https://transparency.meta.com/policies/community-standards/privacy-violations/. 35

Mosseri, Adam. 2016. "Addressing hoaxes and fake news." *Meta Newsroom*. Accessed in 2025-08. https://web.archive.org/web/20250808223815/https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/. 56

Mostagir, Mohamed and James Siderius. 2022. "Naive and Bayesian learning with misinformation policies." *Working paper, University of Michigan and Massachusetts Institute of Technology*. 7

Mostagir, Mohamed and James Siderius. 2023. "When should platforms break echo chambers?" *Working paper, University of Michigan and Massachusetts Institute of Technology*. 7

Müller, Karsten and Carlo Schwarz. 2023. "The effects of online content moderation: Evidence from President Trump's account deletion." *Available at SSRN 4296306*. 34

Musk, Elon. 2023. "The so-called fact-checkers are huge liars and incredibly biased." *Accessed 2025-05. https://x.com/elonmusk/status/1669017475659251713*. 37

Nahrgang, Mia, Nils B. Weidmann, Friederike Quint, Sebastian Nagel, Yannis Theocharis, and Margaret E. Roberts. 2025. "Written for lawyers or users? Mapping the complexity of community guidelines." In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 19, pp. 1295–1314. 33

Ng, Robin and Greg Taylor. 2025. "Moderating content-hosting platforms." *Collaborative Research Center Transregio 224 Discussion Paper No. 698*. 7

O'Sullivan, Donnie. 2022. "Twitter is no longer enforcing its Covid misinformation policy." *CNN*. Accessed in 2025-08. https://web.archive.org/web/20250612112655/https://edition.cnn.com/2022/11/29/tech/twitter-covid-misinformation-policy/index.html. 56

Oversight Board. 2023. "Armenian Prisoners of War Video – Upheld." Accessed 2025-06. https://www.oversightboard.com/decision/fb-ylrv35wd/. 35

Oversight Board. 2024. "Content Moderation in a New Era for AI and Automation." *Oversight Board*. Accessed 2025-09. https://www.oversightboard.com/wp-content/uploads/2024/09/

Oversight-Board-Content-Moderation-in-a-New-Era-for-AI-and-Automation-September-2024.pdf. 2

Pal, Anwesan, Radhika Bhargava, Kyle Hinsz, Jacques Esterhuizen, and Sudipta Bhattacharya. 2024. "The empirical impact of data sanitization on language models." *Presented at the Safe Generative AI Workshop at NeurIPS 2024*. 41

Papanastasiou, Yiangos. 2020. "Fake news propagation and detection: A sequential model." *Management Science*, 66(5): 1826–1846. 7

Pappas, Vanessa. 2020. "Combating misinformation and election interference on TikTok." *TikTok Newsroom*. Accessed in 2025-08. https://web.archive.org/web/20250815014057/https://newsroom.tiktok.com/en-us/combating-misinformation-and-election-interference-on-tiktok. 56

Paul, Christopher and Miriam Matthews. 2016. "The "Russian firehose of falsehood" propaganda model: Why it might work and options to counter it." RAND. 40

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): 4944–4957. 37

Pew Research Center. 2022. "2021 Pew Research Center's American Trends Panel Wave 99 – Internet & Science Topline." *Pew Research Center*. Accessed 2025-09. https://www.pewresearch.org/wp-content/uploads/sites/20/2022/03/PI_2022.03.17_ai-he-TOPLINE.pdf. 2

Pfister, Damien Smith. 2011. "The logos of the blogosphere: Flooding the zone, invention, and attention in the Lott imbroglio." *Argumentation and Advocacy*, 47(3): 141–162. 39

Post Editorial Board. 2021. "Facebook admits the truth: 'Fact checks' are really just (lefty) opinion." *New York Post*. Accessed 2023-10. https://web.archive.org/web/20230919155331/https://nypost.com/2021/12/14/facebook-admits-the-truth-fact-checks-are-really-just-lefty-opinion/. 37

Rainie, Lee, Cary Funk, Monica Anderson, and Alec Tyson. 2022. "Mixed views about social media companies using algorithms to find false information." *Pew Research Center*. Accessed 2025-09. https://www.pewresearch.org/internet/2022/03/17/mixed-views-about-social-media-companies-using-algorithms-to-find-false-information/ 2

Rauchfleisch, Adrian and Jonas Kaiser. 2024. "The impact of deplatforming the far right: An analysis of YouTube and BitChute." *Information, Communication & Society*, 27(7): 1478–1496. 36

Reddit. 2024. "Is posting someone's private or personal information okay?" Accessed 2025-06. https://support.reddithelp.com/hc/en-us/articles/360043066452-Is-posting-someone-s-private-or-personal-information-okay. 35

Reny, Philip J. 2025. "Natural language equilibrium I: Off-path conventions." *University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2025-64*. 32

Reuters in Bangkok. 2015. "Thai printer blanks out another International New York Times article." *The Guardian*. Accessed 2023-05. https://www.theguardian.com/world/2015/dec/04/thai-printer-blanks-another-international-new-york-times. 38

Ribeiro, Manoel Horta, Justin Cheng, and Robert West. 2023. "Automated content moderation increases adherence to community guidelines." In *Proceedings of the ACM Web Conference 2023*, pp. 2666–2676. Association for Computing Machinery. 34

Roberts, Margaret. 2018. *Censored: Distraction and Diversion Inside China's Great Firewall*. Princeton University Press. 40

Salamanca, Andrés. 2021. "The value of mediated communication." *Journal of Economic Theory*, 192: 105191. 6

Sarkar Diba, Bidita, Jayonto Dutta Plabon, Nishat Tasnim, Mehjabin Hossain, Durjoy Mistry, Sourav Sarker, M. F. Mridha, Yuichi Okuyama, and Jungpil Shin. 2025. "From centralization to decentralization: Blockchain's role in transforming social media platforms." *IEEE Access*, 13: 80478–80507. 37

Schraer, Rachel. 2022. "The Russians using emojis to evade censors." *BBC*. Accessed 2025-05. https://www.bbc.com/news/60649725. 39

Ser, Kuang Keng Kuek. 2016. "How China has censored words relating to the Tiananmen Square anniversary." Accessed 2025-05. https://theworld.org/stories/2016/06/03/tiananmen-square-anniversary-all-words-worth-censored. 39

Shi, Dan, Tianhao Shen, Yufei Huang, Zhigen Li, Yongqi Leng, Renren Jin, Chuang Liu, Xinwei Wu, Zishan Guo, Linhao Yu, Ling Shi, Bojian Jiang, and Deyi Xiong. 2024. "Large language model safety: A holistic survey." *arXiv preprint arXiv:2412.17686*, abs/2412.17686. 40

Shumailov, Ilia, Jamie Hayes, Eleni Triantafillou, Guillermo Ortiz-Jiménez, Nicolas Papernot, Matthew Jagielski, Itay Yona, Heidi Howard, and Eugene Bagdasaryan. 2024. "UnUnlearning: Unlearning is not sufficient for content regulation in advanced generative AI." *arXiv preprint arXiv:2407.00106*. 41

Singhal, Mohit, Chen Ling, Pujan Paudel, Poojitha Thota, Nihal Kumarswamy, Gianluca Stringhini, and Shirin Nilizadeh. 2023. "SoK: Content moderation in social media, from guidelines to enforcement, and research to practice." In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pp. 868–895. IEEE. 33

Sobel, Joel. 2013. "Giving and receiving advice." In *Advances in Economics and Econometrics – Economic Theory – Tenth World Congress*, Vol. 49 of *Econometric Society Monographs*, ed. Daron Acemoglu, Manuel Arellano, and Eddie Dekel, pp. 305–341. Cambridge University Press. 6

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. "MPNet: Masked and permuted pre-training for language understanding." *Advances in Neural Information Processing Systems*, 33: 16857–16867. 61

Soo, Zen and Associated Press. 2022. "Chinese users work to save protest content against massive censorship." *PBS NewsHour*. Accessed 2025-05. https://www.pbs.org/newshour/world/chinese-users-work-to-save-protest-content-against-massive-censorship. 38

Stewart, Elizabeth. 2021. "Detecting fake news: Two problems for content moderation." *Philosophy & Technology*, 34(4): 923–940. 37

Szyma'nski, Adam and Ahmet Furkan Cihangiroğlu. 2025. "Deliberate polarization as a distractive political strategy in economic downturns: the case of Turkey." *British Journal of Middle Eastern Studies*. First published online 2023. 40

Tagat, Anirudh, Amreesh Phokeer, and Hanna M. Kreitem. 2024. "Net loss: An econometric method to measure the impact of Internet shutdowns." *ACM Journal on Computing and Sustainable Societies*, 2(2). 39

Tamariz, Juan. 2020. *The Five Points in Magic: A Treatise on the Body's Role in Deception.* Hermetic Press Incorporated. Revised from a translation by Donald B. Lehn. 40

Tavernise, Sabrina and Aidan Gardiner. 2019. "'No one believes anything': Voters worn out by a fog of political news." *The New York Times*. Accessed 2025-05. https://www.nytimes.com/2019/11/18/us/polls-media-fake-news.html. 40

The Economist. 2024. "Disinformation is on the rise: How does it work?" *The Economist*. Accessed 2025-09. https://www.economist.com/science-and-technology/2024/05/01/disinformation-is-on-the-rise-how-does-it-work. 37

TikTok. 2024. "Code of Practice on Disinformation – Report of TikTok for the period 1 January 2024 – 30 June 2024." TikTok Transparency Center. Accessed 2025-08 https://web.archive.org/web/20250529041247/https://disinfocode.eu/reports/download/64. 37

van der Boon, Robert M. A., A. John Camm, C. Aguiar, E. Biasin, G. Breithardt, H. Bueno, I. Drossart, N. Hoppe, E. Kamenjasevic, R. Ladeiras-Lopes, Paul McGreavy, P. Lanzer, R. Vidal-Perez, and Nico Bruining. 2024. "Risks and benefits of sharing patient information on social media: A digital dilemma." *European Heart Journal – Digital Health*, 5(3): 199–207. 35

Vraga, Emily K. and Leticia Bode. 2017. "Using expert sources to correct health misinformation in social media." *Science Communication*, 39(5): 621–645. 37

Wei, Alexander, Nika Haghtalab, and Jacob Steinhardt. 2023. "Jailbroken: How does LLM safety training fail?" In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 80079–80110. 40

West, Darrell M. 2016. "Internet shutdowns cost countries $2.4 billion last year." Center for Technological Innovation at Brookings, Washington, DC. 39

Wikipedia contributors. 2024. "List of social platforms with at least 100 million active users." Accessed 2024-05. https://en.wikipedia.org/wiki/List_of_social_platforms_with_at_least_100_million_active_users. 54

Xu, Yongxin, Yuhao Xuan, and Gaoping Zheng. 2021. "Internet searching and stock price crash risk: Evidence from a quasi-natural experiment." *Journal of Financial Economics*, 141(1): 255–275. 39

Yang, Ya-Ting, Tao Li, and Quanyan Zhu. 2023. "Designing policies for truth: Combating misinformation with transparency and information design." In *2023 21st International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pp. 127–134. IEEE. 7

Yi, Sibo, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. 2024. "Jailbreak attacks and defenses against large language models: A survey." *arXiv preprint arXiv:2407.04295*, abs/2407.04295. 40

Zhang, Jingwen, Jieyu Ding Featherstone, Christopher Calabrese, and Magdalena Wojcieszak. 2021. "Effects of fact-checking social media vaccine misinformation on attitudes toward vaccines." *Preventive Medicine*, 145: 106408. 37

Zheng, Yanfeng and Qinyu (Ryan) Wang. 2020. "Shadow of the Great Firewall: The impact of Google blockade on innovation in China." *Strategic Management Journal*, 41(12): 2234–2260. 39

Zhu, Wanzheng, Hongyu Gong, Rohan Bansal, Zachary Weinberg, Nicolas Christin, Giulia Fanti, and Suma Bhat. 2021. "Self-supervised euphemism detection and identification for content moderation." In *Proceedings of the 42nd IEEE Symposium on Security and Privacy (SP)*, pp. 229–246. 35

Zuckerberg, Mark. 2025. "Zuckerberg Facebook video announcing end of fact-checking program." Accessed 2025-05. https://epublications.marquette.edu/zuckerberg_files_videos/431. 37