Content Moderation with Opaque Policies

Scott Duke Kominers
Harvard University and a16z crypto

Jesse M. Shapiro
Harvard University and NBER[*]

February 2024

Abstract

A sender sends a signal about a state to a receiver who takes an action that determines a payoff. A moderator can block some or all of the sender's signal before it reaches the receiver. When the moderator's policy is transparent to the receiver, the moderator can improve the payoff by blocking false or harmful signals. When the moderator's policy is opaque, however, the receiver may not trust the moderator. In that case, the moderator can guarantee an improved outcome only by blocking signals that enable harmful acts. Blocking signals that encourage false beliefs can be counterproductive.

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

Keywords: social media, strategic communication, platform design

[*]The authors thank Shai Bernstein, Rafael Jiménez-Durán, Daniel Kornbluth, Sriram Krishnan, Mohamed Mostagir, Alex Nichifor, Leah Plunkett, Marco Reuter, Tim Roughgarden, Suproteem Sarkar, Alex Teytelboym, Liang Wu, and seminar audiences at Carnegie Mellon University, the University of Pittsburgh, and Harvard University for helpful comments. Kominers gratefully acknowledges support from the Washington Center for Equitable Growth and 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. 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

Social media platforms can block, flag, mute, or otherwise influence the circulation of content on a massive scale.[1] Such content moderation has become a thorny social problem. Moderation decisions and policies by major platforms have been criticized from both the political left (see, e.g., Thompson, 2020) and the political right (see, e.g., Soave, 2022), and are the subject of evolving legal and regulatory challenges (see, e.g., Fung, 2023; Guynn, 2023; Vanian, 2023).

A theme of much of the criticism of content moderation policies is that they are not transparent—and so may be subject to bias or pressure. Yet transparency is difficult to achieve. Social media policies are embodied in hundreds of thousands of lines of code and in the practices of thousands of human moderators and fact checkers.[2] Even if it were technically feasible to expose the full content moderation policies of a platform, doing so could allow bad actors to find and exploit vulnerabilities.[3] Moreover, some policies, such as fact-checking, require a judgment of truth or falsehood. It is difficult to transparently and decisively convey the basis of such judgments as, by nature, they often concern matters on which the truth is contested (Stewart, 2021).[4]

In this paper, we study the possibilities and limits of content moderation when the moderation policy is opaque. We cast our analysis in a model with an unknown state. A sender observes the state and can send a signal about it. A moderator observes the sender's signal and can block all or part of it in the message the moderator shows to a receiver. The receiver sees the message passed through by the moderator and takes an action which, along with the state, determines a payoff. Consistent with some common practices (e.g., X, 2023) we assume in our main analysis that the receiver sees when a signal has been partially or fully blocked; but we also extend our analysis to the case in which the receiver may not know whether blocking has occurred.

[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. These actions led to 662,000 and 164,000 appeals, respectively (Meta, 2023*a*).

[2] A public release of only a part of Twitter's recommendation algorithm in 2023 included 460,239 lines of code (Twitter, 2023). Meanwhile, Meta reported in 2023 that it had 15,000 reviewers involved in detecting violations of its Community Standards (Meta, 2023*b*).

[3] The public release of a part of Twitter's algorithm omitted "code that would compromise [...] the ability to protect our platform from bad actors" (Wiggers, 2023).

[4] 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).

As a benchmark, we begin the analysis by considering the case where content moderation is transparent, in the sense that the receiver knows the moderator's policy and trusts that it is applied faithfully. In this case, a policy of blocking false or harmful signals can improve the receiver's well-being.

We turn next to the case of opaque moderation, in which the receiver does not know the moderator's policy and need not trust the moderator's motives or information. We ask which content moderation policies are *benevolent*, in the sense that they improve the receiver's payoff regardless of the true state and regardless of the receiver's belief about the moderator's policy. We find that the potential for benevolent moderation depends on the nature of the unknown state about which the sender is communicating.

Formally, we divide the state into two components. One part is a binary *state of the world* that might encode, for example, whether an election has been carried out fraudulently. The other part is a continuous *key* that might encode, say, the time and place of a riot or insurrection. The receiver's payoff is such that the receiver wants to take an action (e.g., riot) under a particular state of the world (e.g., a stolen election), but the receiver's action may be inconsequential if the receiver does not know the key (e.g., the time and place).

Our first finding is that, if the receiver's action is payoff-relevant only when the receiver knows the key, then there is a benevolent content moderation policy that sometimes blocks the key. If, for example, the sender asserts that a public official is corrupt (the state of the world), and shares the official's home address (the key), the moderator might block the address. A receiver skeptical of the moderator's motives might interpret the moderator's decision to block the address as evidence that the public official is indeed corrupt—but without the address, the potential for harm is eliminated.

Our second finding is that, if the receiver's action is payoff-relevant regardless of whether the receiver knows the key, then there is no nontrivial benevolent content moderation policy, even if the moderator knows the true state. The reason is that the receiver can interpret a decision to block content as evidence of an incorrect state, and act accordingly. For example, suppose that the sender signals that the state of the world is such that genetically modified foods are intrinsically dangerous, in which case the receiver should avoid eating such foods. If the moderator concludes the signal is false, the moderator can block the signal. But, if the receiver thinks that moderator may be beholden to agribusiness interests (Uscinski and Parent, 2014, pp. 146–47), the decision to

block the signal could reinforce the belief that genetically modified foods are intrinsically dangerous, thus leading to more (rather than less) avoidance of such foods.

More generally, we show that the scope for benevolent content moderation depends on the role of the key in the receiver's payoffs. Specifically, we show that, the more important is the key, the less harm the moderator risks by sometimes blocking the sender's signal about the key. We furthermore show that, if the sender transmits the true key, any nearly benevolent policy that blocks the sender's signal about the state of the world must also block the sender's signal about the key. In these respects, our analysis reveals an important distinction between content that may directly enable harmful acts (the key) and content that may cause harm only through the beliefs it induces (the state of the world), with opaque moderation much more likely to succeed for the first type of content than for the second.

Within economics, much of the (still relatively small) literature studying content moderation by online platforms focuses on measuring its effects on user welfare (Jiménez-Durán, 2023), user engagement (Beknazar-Yuzbashev et al., 2023), platform content (Andres and Slivko, 2023; Müller and Schwarz, 2023), and offline behaviors (Jiménez-Durán, Müller and Schwarz, 2023).[5] Some prior theory research focuses on how a platform can achieve its goals, studying for example the incentive to moderate content when revenue comes from advertising (Liu, Yildirim and Zhang, 2022; Madio and Quinn, 2023), the effect of content moderation on user participation (Dwork et al., 2023), and the potential for tradeoffs between engagement and information (Candogan and Drakopoulos, 2020; Papanastasiou, 2020).

In the spirit of the literature on market design, our focus is instead on when and how content moderation can improve platform users' welfare. Previous work in this vein has examined the optimal design of moderation policies when those policies can be made transparent to users. Candogan and Drakopoulos (2020), Hossain et al. (2023), and Yang, Li and Zhu (2023), 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.[6] We focus instead on the situation in which moderation policies

[5]See also Ribeiro, Cheng and West (2023). Agarwal, Ananthakrishnan and Tucker (2023) and Vu, Hutchings and Anderson (2023), among others, study the effect of removing infrastructure support from objectionable platforms. Rauchfleisch and Kaiser (2021) and Klinenberg (Forthcoming) study substitution of engagement across platforms following removal of content from a platform, and Zhang, Moon and Veeraraghavan (2022) study effects of platform policies on off-platform news consumption.

[6]See also Papanastasiou (2020). Jackson, Malladi and McAdams (2022) study the problem of limiting

are opaque, and users may not trust the platform to know (or tell) the truth.

Like us, Mostagir and Siderius (2022) and Acemoglu, Ozdaglar and Siderius (Forthcoming) highlight potential downsides from interventions to reduce the spread of false information. These papers study a model in which signals are shared along a network. We instead set aside re-sharing to focus on the role of the moderator, though we note in the text how re-sharing can be cast into our setting. Mostagir and Siderius (2022) show that a platform with arbitrarily good (but imperfect) information about an unknown state may not wish to censor misinformation due to the presence of agents with extreme priors whose inferences lead to social mislearning. Acemoglu, Ozdaglar and Siderius (Forthcoming), meanwhile, show that under Bayesian learning, censoring misinformation can lead a platform's users to assign more credence to uncensored misinformation.[7] These papers focus on the case in which the platform's policies are transparent, rather than opaque, and in which misinformation is harmful because of the beliefs it induces, not because of the actions it enables—both distinctions that our analysis reveals to be important.

Like social media platforms, mass media outlets also filter information (Gentzkow, Shapiro and Stone, 2015), with sources analogous to the sender in our framework, editors and journalists analogous to the moderator, and news consumers analogous to the receiver. In that context, a media outlet's reputation can serve as a means of achieving what we call transparency. Shapiro (2016) finds that reputational incentives can lead to excessive balance in news reporting, akin to what might be considered under-moderation in social media. Gentzkow, Wong and Zhang (2023) study the evolution of trust in information sources in a model with biased reasoning. We show that absent strong reputational mechanisms or other sources of trust, content moderation can be effective only in a specific class of situations.

depth and breadth of content-sharing in a setting where content can mutate as it circulates.

[7] See also Pennycook et al. (2020) and Mostagir and Siderius (2023).

2 Model and Definitions

2.1 Timing and Notation

Nature determines a **state** $\theta = (\theta_\omega, \theta_\kappa) \in \left[\{\mathsf{Down}, \mathsf{Up}\} \times [0,1]\right] \equiv \Theta$, choosing both θ_ω and θ_κ according to non-degenerate distributions, where the distribution of $\theta_\kappa|\theta_\omega$ is continuous. We refer to θ_ω as the **state of the world** and θ_κ as the **(state of the) key**.

A **sender** observes θ and can send any **signal**

$$s = (s_\omega, s_\kappa) \in \left[\{\mathsf{Down}, \mathsf{Up}\} \times [0,1]\right] \equiv \mathcal{S}.$$

We will speak of the sender as a unitary agent for concreteness, but we may alternatively think of the sender as representing, for example, the final member (or aggregate) of a sequence of agents passing content along a graph.

A **moderator** observes (θ, s) and can transmit any **message**

$$m = (m_\omega, m_\kappa) \in \left[\{s_\omega, \mathfrak{B}\} \times \{s_\kappa, \mathfrak{B}\}\right] \equiv \mathcal{M}(s),$$

where we write $\mathcal{M} = \cup_{s \in \mathcal{S}} \mathcal{M}(s)$ for the set of all possible messages. If $m(s)$ contains $\mathfrak{B}$, then we say that the moderator has **blocked** the signal; we will likewise speak of blocking a specific *part* of the signal.

A **receiver** observes m and can take any action $a = (a_\omega, a_\kappa) \in \mathcal{A} \cong \Theta$. The payoff to the receiver is given by $U(\theta, a)$, where

$$\begin{cases} U((\theta_\omega, \theta_\kappa), (\mathsf{Down}, a_\kappa)) = 0 \\ U((\theta_\omega, \theta_\kappa), (\mathsf{Up}, a_\kappa)) = (-1)^{\mathbf{1}_{\theta_\omega = \mathsf{Down}}} \cdot ((1-\gamma) + \gamma \mathbf{1}_{a_\kappa = \theta_\kappa}), \end{cases}$$

for $\gamma \in [0,1]$ a parameter. With this payoff, the receiver aims to match their action a_ω to the state of the world θ_ω, and aims to match their action a_κ to the key θ_κ if and only if the true state of the world is Up.

The parameter γ controls the importance of the key. When $\gamma = 1$, the receiver's action a_ω is consequential only when the receiver chooses the correct key $a_\kappa = \theta_\kappa$.[8]

[8] Here and elsewhere, we abuse notation slightly by explicitly identifying the action, signal, and state spaces, so that a statement like "$a_\kappa = \theta_\kappa$" is meaningful even though technically the action and state spaces are distinct.

When $\gamma = 0$, the receiver's action a_ω is always consequential and the receiver's choice of key a_κ is irrelevant.

2.2 Moderation Policies

We write Σ and M, respectively, for the sets of (mixed) strategies for the sender and moderator.[9] We let $\mathcal{P} \subseteq \left[\Delta(\Sigma) \times \Delta(\mathrm{M})\right]$ denote a set of possible receiver beliefs about the sender's and moderator's strategies.

The space $\mathcal{P}$ parameterizes the receiver's knowledge of the moderator's policy. When all beliefs in $\mathcal{P}$ put probability 1 on the moderator using a particular policy $\mu \in \mathrm{M}$, we will say that the policy μ is **transparent to the receiver (over $\mathcal{P}$)**. When $\mathcal{P} = \left[\Delta(\Sigma) \times \Delta(\mathrm{M})\right]$, we will say that the strategies are **fully opaque to the receiver**, or **opaque** for short. Because the sender's strategy space Σ includes strategies that do not depend on the true state θ (and may even randomize independently of the state), the assumption that the sender knows the true state does not affect the analysis that follows. Moreover, although we suppose that blocking is made explicit to the receiver, we explain in Remark 1 that our results extend directly to a case in which the receiver cannot tell whether content has been blocked.

To formalize our criteria for evaluating the moderator's policy, observe that any beliefs $p \in \mathcal{P}$ induce a function $U_p(\theta, m)$ that describes the receiver's payoff given state θ and received message m under optimal receiver behavior given their beliefs.[10] Because $U_p(\theta, m)$ is contingent on the state, it depends on the beliefs p only through the receiver's action.

Definition 1. For $0 \leq \epsilon < 1$, a messaging policy $\mu \in \mathrm{M}$ is **ϵ-benevolent (over $\mathcal{P}$)** if, for all sender strategies $\sigma \in \Sigma$ and receiver beliefs $p \in \mathcal{P}$ we have that

$$\mathrm{Prob}_{\theta,\sigma,\mu}\left[U_p\left(\theta, m(\theta, s(\theta))\right) + \epsilon \geq U_p\left(\theta, s(\theta)\right)\right] = 1. \tag{1}$$

[9] That is, $\Sigma = \Delta(\mathcal{S})^\Theta$ and $\mathrm{M} = \left\{\mu \in \Delta(\mathcal{M})^{\Theta \times \mathcal{S}} : \mathrm{Prob}_{m \sim \mu(s)}[m \in \mathcal{M}(s)] = 1\right\}$.

[10] That is, we take

$$U_p(\cdot, m) \in \left\{U(\cdot, a) : a \in \arg\max_{a' \in \mathcal{A}} \left\{\int_\Theta U(\theta, a') p(\theta \mid m)\, d\theta\right\}\right\},$$

where we may select arbitrarily among the elements of the set when it is not a singleton without affecting the analysis that follows.

If a messaging policy is 0-benevolent (over $\mathcal{P}$) we say that it is **benevolent (over $\mathcal{P}$)**.

That is, a policy is **ϵ-benevolent** if, relative to simply allowing the signal to reach the receiver unblocked, the policy is guaranteed to lead to a loss to the receiver of no more than ϵ regardless of the sender's strategy or the receiver's beliefs about the sender's and moderator's strategies. An ϵ-benevolent policy is in this sense robust to malicious actions by the sender and/or adverse beliefs on the part of the receiver.

Here and throughout, we take probabilities with respect to the true distribution of the state θ, and make no assumption about whether the sender, moderator, or receiver knows this distribution or instead believes in a different one. We simplify notation in (1) by writing $\text{Prob}_{\theta,\sigma,\mu}$ for $\text{Prob}_{\theta,\, s\sim\sigma(\theta),m\sim\mu(s)}$, and we simplify going forward by omitting measures from probability statements when the measures are irrelevant or are clear from context.

3 Benchmark: Transparent Moderation

When the moderation policy can be made transparent to the receiver, it is possible to implement a benevolent moderation policy that blocks inaccurate signals.

Proposition 1. *The following (pure-strategy) messaging policy is benevolent when it is transparent to the receiver:*

$$m^{\oplus}(\theta, s) \equiv \begin{cases} s & s_\omega = \theta_\omega \\ (\mathfrak{B}, \mathfrak{B}) & s_\omega \neq \theta_\omega. \end{cases}$$

That is, when the moderation policy can be made transparent to the receiver, one benevolent policy is for the moderator to block the sender's full signal if the signal misrepresents the state of the world, and not to block otherwise.

Proof. For all sender signal realizations s under which $s_\omega = \theta_\omega$, we have $m^{\oplus}(s) = s$ and so we have $U_p(\theta, m^{\oplus}(\theta, s)) = U_p(\theta, s)$ *a priori* for all $p \in \mathcal{P}$.

For signal realizations s with $s_\omega \neq \theta_\omega$, meanwhile, because the messaging policy is transparent, the receiver's beliefs are such that (with probability 1) upon receiving message s, the receiver believes that the state of the world is s_ω, and so will take

$a_\omega = s_\omega \neq \theta_\omega$. Consequently, the receiver's payoff upon receiving message s is upper-bounded by 0 in the case that $\theta_\omega = \mathsf{Up}$, and upper-bounded by $-(1-\gamma)$ in the case that $\theta_\omega = \mathsf{Down}$—but with probability 1, these are *lower* bounds on the receiver's payoff when they receive the message $(\mathfrak{B}, \mathfrak{B})$ in states of the world Up and Down, respectively. To see this, note first that when $\theta_\omega = \mathsf{Up}$, 0 is an absolute lower bound on the receiver's payoff. And when $\theta_\omega = \mathsf{Down}$, the receiver's payoff can only be less than $-(1-\gamma)$ if the receiver chooses $a_\omega = \mathsf{Up}$ and correctly matches the key, i.e., $a_\kappa = \theta_\kappa$. But because the distribution of θ_κ is continuous, when the signal about the key is blocked, the probability that $a_\kappa = \theta_\kappa$ under any belief-induced distribution of a_κ is 0.

Thus we see that the inequality in (1) holds point-wise for all realizations of the signal s as a function of θ, which proves the result. □

The policy $m^{\oplus}(\cdot)$ characterized in Proposition 1 has other desirable properties in addition to benevolence. For example, relative to the case of no content moderation (i.e., the policy $m(\theta, s) = s$), for a receiver who knows the sender's strategy, transparently adopting the policy $m^{\oplus}(\cdot)$ does not reduce the greatest *ex ante* payoff (which is attained when the sender reports the state), and improves upon the lowest *ex ante* payoff (which is attained when the sender garbles). Moreover, a range of benevolent policies beyond $m^{\oplus}(\cdot)$ are available when moderation is transparent.[11]

Unfortunately, transparency is a strong requirement that, as we have argued, can be difficult to achieve in practice. We therefore devote the rest of the analysis to the case where the strategies are opaque to the receiver.

4 Opaque Moderation

When the moderator's policy need not be transparent, the scope for benevolent moderation depends on the parameter γ, which controls the importance of the key.

4.1 Blocking Signals that Enable Harmful Actions

We begin with the special case of $\gamma = 1$, in which the receiver's action is consequential only when the receiver chooses the correct key. In this case, the following result char-

[11] In particular, the proof of Proposition 1 immediately implies that any messaging policy μ' that takes $m'(\theta, s) = s$ when $s_\omega = \theta_\omega$ and $m'(\theta, s) = (\mathfrak{B}, s_\kappa)$ otherwise is benevolent when it is transparent to the receiver, even if it never blocks the key.

acterizes a policy that is benevolent regardless of the space $\mathcal{P}$ of the receiver's possible beliefs about the sender's and moderator's strategies.

Proposition 2. *If $\gamma = 1$ then the following (pure-strategy) messaging policy is benevolent over any $\mathcal{P} \subseteq \left[\Delta(\Sigma) \times \Delta(\mathrm{M})\right]$:*

$$m^{\ominus}(\theta, s) = \begin{cases} s & \theta_{\omega} = \mathsf{Up} \\ (s_{\omega}, \mathfrak{B}) & \theta_{\omega} = \mathsf{Down}. \end{cases}$$

That is, when $\gamma = 1$, it is benevolent for the moderator to block the sender's signal about the key if the state of the world is Down*, and not to block otherwise.*

Proof. When $\theta_{\omega} = \mathsf{Up}$, then $m^{\ominus}(s) = s$ (for any realization of s) and so we have $U_p(\theta, m^{\ominus}(\theta, s)) = U_p(\theta, s)$ *a priori* for all $p \in \mathcal{P}$, so in particular we have

$$\mathrm{Prob}_{\theta,\sigma}\left[U_p\left(\theta, m^{\ominus}(\theta, s)\right) = U_p\left(\theta, s\right)\right] = 1$$

for any σ.

When $\theta_{\omega} = \mathsf{Down}$, meanwhile, we have $U_p(\theta, s) \leq 0$, so we can only have

$$U_p(\theta, m^{\ominus}(\theta, s)) < U_p(\theta, s)$$

(again, for any realization of s) if $U_p(\theta, m^{\ominus}(\theta, s)) < 0$, but for this to happen, we must have both $a_{\omega} = \mathsf{Up}$ and $a_{\kappa} = \theta_{\kappa}$ with strictly positive probability over θ in response to the message $m^{\ominus}(\theta, s) = (s_{\omega}, \mathfrak{B})$.

Now, under any beliefs p and received message $(s_{\omega}, \mathfrak{B})$ such that $a_{\omega} = \mathsf{Up}$, optimal receiver behavior induces some (possibly atomistic) distribution of a_{κ}. But because the distribution of θ_{κ} is continuous and the support of s_{ω} is discrete, the probability that $a_{\kappa} = \theta_{\kappa}$ under the induced distribution of a_{κ} is 0. We thus have

$$\mathrm{Prob}_{\theta_{\kappa}}[U_p(\theta, m^{\ominus}(\theta, s)) < 0] = 0$$

even when $a_{\omega} = \mathsf{Up}$; the result then follows. □

The fact that the policy $m^{\ominus}(\cdot)$ is benevolent over any $\mathcal{P}$ implies that it is benevolent in the opaque case, which means it is benevolent even when the receiver entertains the

possibility that the moderator is malicious (e.g., intentionally blocking the truth). The reason this works is that when $\gamma = 1$, blocking the key when the state of the world is Down removes the potential for a harmful action—which improves outcomes regardless of the receiver's inferences about the state of the world.

To return to an example from the introduction, blocking the sender from sharing a public official's home address (the key) may convince the receiver that the moderator is protecting the official from scrutiny and so lead the receiver to conclude that the official is corrupt (the state of the world). But even so, absent the address, the potential for direct harm is eliminated.

Observe that, because we have not assumed that any agents in the model know the true distribution of θ, our analysis—and, hence, Proposition 2—applies directly to cases where the sender and receiver mistakenly believe the true state of the world is always Up, and so will always act if in possession of the key. An example is "flash robbing," in which a sender coordinates a group property crime by publicly sharing a time and place. In such cases, the sender's or receiver's belief about the correct action a_ω—for example, whether it is appropriate to commit property crimes—may differ systematically from those of the platform (or of society as a whole). Nevertheless, the results in Proposition 2 continue to imply that it is effective to block such content. A similar argument extends our analysis to content, such as hate speech, whose mere consumption is intrinsically harmful to the receiver.

4.2 Blocking Signals that Encourage False Beliefs

We turn next to the opposite case, $\gamma = 0$, in which the receiver's action is consequential regardless of whether the receiver chooses the correct key. We may think of this case as one in which the potential harm from a signal comes from encouraging a false belief, rather than from enabling a harmful act. In this case, when the strategies are opaque any benevolent moderation policy must be trivial, in the sense that it never blocks any part of the signal.

Proposition 3. *If $\gamma = 0$ and the strategies are opaque, then under any benevolent messaging policy μ, we must have $\text{Prob}_{\theta,\, s\sim\sigma(\theta),\, m\sim\mu(s)}[m(\theta, s) = s] = 1$ under any sender strategy $\sigma \in \Sigma$. That is, when $\gamma = 0$ and the strategies are opaque, no benevolent moderation policy can block any part of the sender's signal with positive probability.*

Proof. We consider some messaging policy μ such that $\text{Prob}_{\theta,\sigma,\mu}[m(\theta,s)=s]<1$; for such a policy, there is a positive-measure set of states $\Theta'\subseteq\Theta$ such that for all $\theta\in\Theta'$, we have

$$m(\theta,s(\theta))\in\{(s_\omega(\theta),\mathfrak{B}),(\mathfrak{B},s_\kappa(\theta)),(\mathfrak{B},\mathfrak{B})\}$$

with strictly positive probability over $s\sim\sigma(\theta)$ and $m\sim\mu(s)$. Because θ_ω takes only two possible values, there must be at least one $\widehat{\theta_\omega}\in\{\mathsf{Down},\mathsf{Up}\}$ such that a strictly positive measure of the states $\theta\in\Theta'$ have $\theta_\omega=\widehat{\theta_\omega}$.

First, we consider the case in which $\widehat{\theta_\omega}=\mathsf{Down}$, i.e., signals are (at least partially) blocked with strictly positive probability in a strictly positive measure of Down-states. We let $\hat{p}_{\mathfrak{B}\to\mathsf{Up}}$ be the belief for the receiver such that whenever the receiver sees an unblocked signal, they believe the true state of the world is Down, and whenever the receiver sees a partially or fully blocked signal, they believe the true state of the world is Up (with some key). Under belief $\hat{p}_{\mathfrak{B}\to\mathsf{Up}}$, the receiver's payoff when they see a blocked signal in states $\theta\in\Theta'$ with $\theta_\omega=\widehat{\theta_\omega}=\mathsf{Down}$ is given by

$$(-1)^{\mathbf{1}_{\theta_\omega=\mathsf{Down}}}=-1$$

because when the receiver sees a blocked signal, they believe the true state of the world is actually Up (with some key), and so they take $a_\omega=\mathsf{Up}$. But this means that

$$U_{\hat{p}_{\mathfrak{B}\to\mathsf{Up}}}(\theta,m(\theta,s(\theta)))=-1<0=U_{\hat{p}_{\mathfrak{B}\to\mathsf{Up}}}(\theta,s(\theta)) \tag{2}$$

with strictly positive probability over $s\sim\sigma(\theta)$ and $m\sim\mu(s)$ for all states $\theta\in\Theta'$ with $\theta_\omega=\widehat{\theta_\omega}=\mathsf{Down}$, where the equality on the right side of (2) follows because under an unblocked signal the receiver would (correctly) believe the true state of the world is Down (with some key), and so take $a_\omega=\mathsf{Down}$ with payoff 0. The states $\theta\in\Theta'$ with $\theta_\omega=\widehat{\theta_\omega}=\mathsf{Down}$ have positive measure by supposition, so we have (2) with strictly positive probability; hence, the messaging policy μ is not benevolent.

Now if instead we have $\widehat{\theta_\omega}=\mathsf{Up}$, i.e., signals are blocked with strictly positive probability in a strictly positive measure of Up-states, then we can conduct an analogous argument under the reverse belief construction. We let $\hat{p}_{\mathfrak{B}\to\mathsf{Down}}$ be the belief for the receiver such that whenever the receiver sees an unblocked signal, they believe the true state of the world is Up, and whenever the receiver sees a partially or fully blocked signal,

they believe the true state of the world is Down (with some key). Under beliefs $\hat{p}_{\mathfrak{B}\to\mathsf{Down}}$, the receiver's payoff under μ is 0 whenever they see a blocked signal, because whenever the receiver sees a blocked signal, they believe the true state of the world is actually Down (with some key), and so take $a_\omega = \mathsf{Down}$. But this means that

$$U_{\hat{p}_{\mathfrak{B}\to\mathsf{Down}}}(\theta, m(\theta, s(\theta))) = 0 < 1 = U_{\hat{p}_{\mathfrak{B}\to\mathsf{Down}}}(\theta, s(\theta)) \tag{3}$$

with strictly positive probability over $s \sim \sigma(\theta)$ and $m \sim \mu(s)$ for all states $\theta \in \Theta'$ with $\theta_\omega = \widehat{\theta_\omega} = \mathsf{Up}$, where the equality on the right side of (3) follows because under an unblocked signal the receiver would (correctly) believe the true state of the world is Up (with some key), and so take $a_\omega = \mathsf{Up}$ with payoff 1. Just as in the argument for the preceding case, the states $\theta \in \Theta'$ with $\theta_\omega = \widehat{\theta_\omega}$ have positive measure by supposition, so we see that the messaging policy μ is not benevolent. □

To return to an example from the introduction, blocking the sender from making a false claim (the signal) about the safety of genetically modified foods (the state of the world) may backfire by convincing a skeptical receiver that the false claim must be true—if not, why block it?

We have assumed that the receiver knows when a signal has been blocked (instead of never sent). As the following remark explains, the conclusion of Proposition 3 continues to hold in a model in which the receiver cannot directly tell whether a signal has been blocked; our other results extend to this case as well.

Remark 1. Generalize the sender's signal space to

$$\mathcal{S}^{\oslash} = \left[\{\mathsf{Down}, \mathsf{Up}, \mathfrak{B}\} \times \left[[0,1] \cup \{\mathfrak{B}\}\right]\right]$$

and observe that this implies that the receiver cannot tell whether a signal was blocked by the moderator (or was originally $\mathfrak{B}$). Then the conclusion of Proposition 3 follows immediately, observing that any belief p about the sender's and moderator's strategies that is possible in the original model—including those beliefs used in the proof of Proposition 3—remains possible in the model with the more general signal space. The conclusions of our other results extend to this alternative model as well.

The extension discussed in Remark 1 shows a sense in which moderation is, if anything, *more* difficult when the receiver does not know whether content has been blocked:

in that case, the receiver may treat the absence of a signal as evidence that a signal was blocked, even if no signal was ever sent.

4.3 Blocking General Signals

The contrast between Proposition 2 and Proposition 3 illustrates an important distinction between content that may enable harmful acts (Proposition 2) and content that may encourage false beliefs (Proposition 3). Here we sharpen that distinction by considering the general case of arbitrary γ and ϵ. For general $\gamma < 1$, moderating key information has what we might loosely think of as a "belief externality": the receiver can interpret the blocking of the key as evidence for a particular state of the world, in a way that causes harm to the extent that the payoff function puts weight on matching the state of the world relative to matching the key.

Proposition 4. *If the strategies are opaque, then there exists an ϵ-benevolent messaging policy that blocks part of the sender's signal with positive probability if and only if $\epsilon \geq 1 - \gamma$. More precisely, if the strategies are fully opaque to the receiver, then:*

(a) For $\epsilon < 1 - \gamma$, the only ϵ-benevolent messaging policies are those that never block any part of the sender's signal with positive probability.

(b) For $\epsilon \geq 1 - \gamma$:

(i) It is ϵ-benevolent for the moderator to block the sender's signal about the key if the state of the world is Down*, and not to block otherwise.*

(ii) If the sender's signal policy ***transmits the true key*** *in the sense that $s_\kappa(\theta) = \theta_\kappa$ everywhere, then any ϵ-benevolent messaging policy must have zero probability of blocking the sender's signal about the state of the world without also blocking the sender's signal about the key.*

Part a of Proposition 4 says that it is more difficult to achieve nontrivial, almost-benevolent content moderation when γ is small, i.e., when beliefs about the state of the world are important for the payoff. This part of the proposition follows directly from Lemma A.1 in the appendix, with a proof similar to that of Proposition 3.

Part b.i of Proposition 4 says that it is easier to achieve nontrivial, almost-benevolent content moderation when γ is large, in which case knowledge of the key is important

for the payoff. This part of the proposition follows directly from Lemma A.2 in the appendix, with a proof similar to that of Proposition 2.

Part b.ii of Proposition 4 elaborates on the scope for almost-benevolent content moderation when γ is large. If the sender's signal always includes the true key, then the moderator must block the signal about the key anytime the moderator blocks the signal about the state of the world. Intuitively, when strategies are opaque, blocking the signal about the state of the world may convince the receiver that the state of the world is Up even when it is actually Down. To avoid enabling a harmful act in that event, the moderator must hide the key. This part of the result follows from Lemma A.3 in the appendix, the proof of which again uses ideas from the proof of Proposition 3.

To return to an example from the introduction, suppose that the receiver must decide whether to denounce the public official (a_ω) and, if so, where to do it (a_κ). If denunciation is consequential only when it occurs near the official's home (γ is large), then there is scope for benevolent moderation that blocks the official's home address (key). If denunciation is consequential regardless of where it occurs (γ is small), then there is limited scope for nontrivial, benevolent moderation.

5 Discussion

Our analysis reveals important distinctions between moderation that is transparent and moderation that is opaque, and between content that is harmful because it *enables harmful acts* and content that is harmful because it *encourages false beliefs*. We find that when moderation is opaque, the former type of content is easier to moderate than the latter.

Many forms of potentially harmful content—such as doxxing (revealing personal information) or mobilization (e.g., posting the time and place of a riot or insurrection)—are potentially harmful precisely because they may enable harmful acts. Others are potentially harmful in and of themselves. Our analysis shows that blocking such content can be effective at preventing harm, even when the moderation process is opaque.

Other forms of potentially harmful content—such as misinformation (incorrect information) and disinformation (deliberately misleading information)—are potentially harmful only (or mainly) because they may encourage false beliefs. Our analysis shows that when the moderation process is opaque, blocking such content can be counterproductive.

We therefore provide a motivation for platforms to adopt especially stringent curbs on content that directly enables harmful acts. At the same time, our analysis provides a reason for skepticism about platforms' ability to limit the spread of false beliefs.

Of course, our analysis is cast in a model that is deliberately stark so as to highlight these conclusions and intuitions. In practice, platforms may be able to establish a reputation for honesty, or to implement technological solutions for exposing their content moderation policies to the public. Our analysis shows a reason why, if such developments can enable more transparent moderation, they can also enable more successful moderation of misleading signals.

Importantly, however, our analysis shows that transparency entails more than simply exposing the mechanics of content moderation—for example by posting a public list of policies or even implementing moderation via immutable code on an open blockchain. To be effective, transparency requires that the receiver can interpret the platform's policies and knows that they are being applied faithfully. Moreover, transparency requires that the receiver can trust how the moderator's policies depend on the truth, something that no technology (known to us) can guarantee.

References

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius. Forthcoming. "A Model of Online Misinformation." *Review of Economic Studies.*

Agarwal, Saharsh, Uttara M. Ananthakrishnan, and Catherine E. Tucker. 2023. "Content Moderation at the Infrastructure Layer: Evidence from Parler." *Available at SSRN 4232871.*

Andres, Raphaela, and Olga Slivko. 2023. "Combating Online Hate Speech: The Impact of Legislation on Twitter." *ZEW-Centre for European Economic Research Discussion Paper*, 21(103).

Beknazar-Yuzbashev, George, Rafael Jiménez-Durán, Jesse McCrosky, and Mateusz Stalinski. 2023. "Toxic Content and User Engagement on Social Media: Evidence from a Field Experiment." *Available at SSRN 4307346.*

Candogan, Ozan, and Kimon Drakopoulos. 2020. "Optimal Signaling of Content Accuracy: Engagement vs. Misinformation." *Operations Research*, 68(2): 497–515.

Dwork, Cynthia, Chris Hays, Jon Kleinberg, and Manish Raghavan. 2023. "Content Moderation and the Formation of Online Communities: A Theoretical Framework." *arXiv preprint arXiv:2310.10573*.

Fung, Brian. 2023. "Supreme Court Delays Considering Florida and Texas Laws that Force Social Media Platforms to Host Content." *CNN*. https://web.archive.org/web/20230430005239/https://www.cnn.com/2023/01/23/politics/supreme-court-delay-texas-florida-social-media-laws/index.html. Accessed October 2023.

Gentzkow, Matthew, Jesse M. Shapiro, and Daniel F. Stone. 2015. "Media Bias in the Marketplace: Theory." In *Handbook of Media Economics*. Vol. 1, ed. Simon P. Anderson, Joel Waldfogel and David Strömberg, Chapter 14, 623–645. Elsevier.

Gentzkow, Matthew, Michael B. Wong, and Allen T. Zhang. 2023. "Ideological Bias and Trust in Information Sources." *Working Paper, Stanford University*.

Guynn, Jessica. 2023. "Biden Administration Coerced Social Media Giants into Possible Free Speech Violations: Court." *USA Today*. https://web.archive.org/web/20231001204508/https://www.usatoday.com/story/money/2023/09/08/biden-administration-coerced-facebook-court-rules/70800723007/. Accessed October 2023.

Hossain, Safwan, Andjela Mladenovic, Yiling Chen, and Gauthier Gidel. 2023. "A Persuasive Approach to Combating Misinformation." *arXiv preprint arXiv:2310.12065*.

Jackson, Matthew O., Suraj Malladi, and David McAdams. 2022. "Learning through the Grapevine and the Impact of the Breadth and Depth of Social Networks." *Proceedings of the National Academy of Sciences*, 119(34): e2205549119.

Jiménez-Durán, Rafael. 2023. "The Economics of Content Moderation: Theory and Experimental Evidence from Hate Speech on Twitter." *George J. Stigler Center for the Study of the Economy & the State Working Paper*.

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*.

Klinenberg, Danny. Forthcoming. "Does Deplatforming Work?" *Journal of Conflict Resolution*.

Liu, Yi, Pinar Yildirim, and Z. John Zhang. 2022. "Implications of Revenue Models and Technology for Content Moderation Strategies." *Marketing Science*, 41(4): 831–847.

Lomborg, Bjorn. 2022. "Facebook, other Tech Giants Censor Inconvenient Facts about Climate Change." *New York Post*. https://web.archive.org/web/20231026004136/https://nypost.com/2022/02/07/facebook-other-tech-giants-censor-facts-about-climate-change/, Accessed October 2023.

Madio, Leonardo, and Martin Quinn. 2023. "Content Moderation and Advertising in Social Media Platforms." *Available at SSRN 3551103*.

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

Meta. 2023*b*. "How Meta Enforces its Policies: Detecting Violations." *https://transparency.fb.com/enforcement/detecting-violations/*, Accessed October 2023.

Mostagir, Mohamed, and James Siderius. 2022. "Naive and Bayesian Learning with Misinformation Policies." *Working paper, University of Michigan and Massachusetts Institute of Technology*.

Mostagir, Mohamed, and James Siderius. 2023. "When Should Platforms Break Echo Chambers?" *Working paper, University of Michigan and Massachusetts Institute of Technology*.

Müller, Karsten, and Carlo Schwarz. 2023. "The Effects of Online Content Moderation: Evidence from President Trump's Account Deletion." *Available at SSRN 4296306*.

Papanastasiou, Yiangos. 2020. "Fake News Propagation and Detection: A Sequential Model." *Management Science*, 66(5): 1826–1846.

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.

Post Editorial Board. 2021. "Facebook Admits the Truth: 'Fact Checks' are Really Just (Lefty) Opinion." *New York Post*. https://web.archive.org/web/20230919155331/https://nypost.com/2021/12/14/facebook-admits-the-truth-fact-checks-are-really-just-lefty-opinion/, Accessed October 2023.

Rauchfleisch, Adrian, and Jonas Kaiser. 2021. "Deplatforming the Far-right: An Analysis of YouTube and BitChute." *Available at SSRN 3867818*.

Ribeiro, Manoel Horta, Justin Cheng, and Robert West. 2023. "Automated Content Moderation Increases Adherence to Community Guidelines." *arXiv preprint arXiv:2210.10454*.

Shapiro, Jesse M. 2016. "Special Interests and the Media: Theory and an Application to Climate Change." *Journal of Public Economics*, 144: 91–108.

Soave, Robby. 2022. "Bari Weiss Twitter Files Reveal Systematic 'Blacklisting' of Disfavored Content." *Reason*. https://web.archive.org/web/20230731223519/https://reason.com/2022/12/09/bari-weiss-twitter-files-elon-musk-blacklist-shadow-banning/. Accessed October 2023.

Stewart, Elizabeth. 2021. "Detecting Fake News: Two Problems for Content Moderation." *Philosophy & Technology*, 34(4): 923–940.

Thompson, Alex. 2020. "Why the Right Wing has a Massive Advantage on Facebook." *Politico*. https://web.archive.org/web/20230426125729/https://www.politico.com/news/2020/09/26/facebook-conservatives-2020-421146. Accessed October 2023.

Twitter. 2023. "Twitter Recommendation Algorithm: Initial GitHub Commit." *https://github.com/twitter/the-algorithm/commit/ef4c5eb65e6e04fac4f0e1fa8bbeff56b75c1f98*, Accessed October 2023.

Uscinski, Joseph E., and Joseph M. Parent. 2014. *American Conspiracy Theories.* Oxford University Press.

Vanian, Jonathan. 2023. "Europe gives Elon Musk 24 Hours to Respond about Israel-Hamas War Misinformation and Violence on X, Formerly Twitter." *CNBC.* https://web.archive.org/web/20231011042119/https://www.cnbc.com/2023/10/10/elon-musk-warned-about-misinformation-violent-content-on-x-by-eu.html. Accessed October 2023.

Vu, Anh V., Alice Hutchings, and Ross Anderson. 2023. "No Easy Way Out: The Effectiveness of Deplatforming an Extremist Forum to Suppress Hate and Harassment." *arXiv preprint arXiv:2304.07037.*

Wiggers, Kyle. 2023. "Twitter Reveals Some of its Source Code, Including its Recommendation Algorithm." *TechCrunch.* https://web.archive.org/web/20230609054801/https://techcrunch.com/web/20230609054801/https://techcrunch.com/2023/03/31/twitter-reveals-some-of-its-source-code-including-its-recommendation-algorithm/, Accessed October 2023.

X. 2023. "Notices on X and What they Mean." *https://web.archive.org/web/20231112193516/https://help.twitter.com/en/rules-and-policies/notices-on-x*, Accessed November 2023.

Yang, Ya-Ting, Tao Li, and Quanyan Zhu. 2023. "Designing Policies for Truth: Combating Misinformation with Transparency and Information Design." *arXiv preprint arXiv:2304.08588.*

Zhang, Jiding, Ken Moon, and Senthil K. Veeraraghavan. 2022. "Does Fake News Create Echo Chambers?" *Available at SSRN 4144897.*

A Proofs Omitted from the Main Text

Proof of Proposition 4

The result follows from combining three lemmata:

Lemma A.1. *If strategies are fully opaque to the receiver, then for any $\epsilon < 1-\gamma$, in any ϵ-benevolent messaging policy $\mu(\cdot)$, we must have* $\text{Prob}_{\theta,\, s\sim\sigma(\theta),\, m\sim\mu(s)}[m(\theta, s(\theta)) = s(\theta)] = 1$. *That is, when strategies are opaque and $\epsilon < 1-\gamma$, no ϵ-benevolent policy can block any part of the sender's signal with positive probability.*

Lemma A.2. *The messaging policy $m^{\ominus}(\cdot)$ defined in Proposition 2 is ϵ-benevolent for any $\epsilon \geq 1-\gamma$. That is, when $\epsilon \geq 1-\gamma$, it is ϵ-benevolent for the moderator to block the sender's signal about the key if the state of the world is* Down*, and not to block otherwise.*

Lemma A.3. *If strategies are fully opaque to the receiver and the sender's signal policy transmits the true key, then for any $\epsilon < 1$, in any ϵ-benevolent messaging policy $\mu(\cdot)$, we must have* $\text{Prob}_{\theta,\, s\sim\sigma(\theta),\, m\sim\mu(s)}[m(\theta, s(\theta)) = (\mathfrak{B}, s_\kappa(\theta))] = 0$. *That is, when strategies are opaque and the sender's signal policy transmits the true key, under any ϵ-benevolent policy, there must be zero probability that the moderator blocks the sender's signal about the true state of the world without also blocking the sender's signal about the key.*

Proof of Lemma A.1

The proof closely follows the argument used to prove Proposition 3. We consider some messaging policy μ such that $\text{Prob}_{\theta,\sigma,\mu}[m(\theta, s) = s] < 1$; for such a policy, there is a positive-measure set of states $\Theta' \subseteq \Theta$ such that for all $\theta \in \Theta'$, we have

$$m(\theta, s(\theta)) \in \{(s_\omega(\theta), \mathfrak{B}), (\mathfrak{B}, s_\kappa(\theta)), (\mathfrak{B}, \mathfrak{B})\}$$

with strictly positive probability over $s \sim \sigma(\theta)$ and $m \sim \mu(s)$. Because θ_ω takes only two possible values, there must be at least one $\widehat{\theta_\omega} \in \{\mathsf{Down}, \mathsf{Up}\}$ such that a strictly positive measure of the states $\theta \in \Theta'$ have $\theta_\omega = \widehat{\theta_\omega}$.

First, we consider the case in which $\widehat{\theta_\omega} = \mathsf{Down}$, i.e., signals are (at least partially) blocked with strictly positive probability in a strictly positive measure of Down-states. We let $\hat{p}_{\mathfrak{B}\to\mathsf{Up}}$ be the belief for the receiver such that whenever the receiver sees an

unblocked signal, they believe the true state of the world is Down, and whenever the receiver sees a partially or fully blocked signal, they believe the true state of the world is Up (with some key). Under belief $\hat{p}_{\mathfrak{B}\to\mathsf{Up}}$, the receiver's payoff when they see a blocked signal in states $\theta \in \Theta'$ with $\theta_\omega = \widehat{\theta_\omega} = \mathsf{Down}$ is given by

$$(-1)^{\mathbf{1}_{\theta_\omega=\mathsf{Down}}} \cdot ((1-\gamma) + \gamma \mathbf{1}_{a_\kappa=\theta_\kappa}) \leq -(1-\gamma) \tag{4}$$

because when the receiver sees a blocked signal, they believe the true state of the world is actually Up (with some key), and so take $a_\omega = \mathsf{Up}$. Meanwhile,

$$U_{\hat{p}_{\mathfrak{B}\to\mathsf{Up}}}(\theta, s(\theta)) = 0 \tag{5}$$

for all states $\theta \in \Theta'$ with $\theta_\omega = \widehat{\theta_\omega} = \mathsf{Down}$, because under an unblocked signal the receiver would (correctly) believe the true state of the world is Down (with some key), and so take $a_\omega = \mathsf{Down}$ with payoff 0. Combining (4) and (5), we see that for all states $\theta \in \Theta'$ with $\theta_\omega = \widehat{\theta_\omega} = \mathsf{Down}$, we have

$$U_{\hat{p}_{\mathfrak{B}\to\mathsf{Up}}}(\theta, m(\theta, s(\theta))) + (1-\gamma) \leq 0 = U_{\hat{p}_{\mathfrak{B}\to\mathsf{Up}}}(\theta, s(\theta))$$

with strictly positive probability over $s \sim \sigma(\theta)$ and $m \sim \mu(s)$ for all states $\theta \in \Theta'$ with $\theta_\omega = \widehat{\theta_\omega} = \mathsf{Down}$. It follows that in such states,

$$U_{\hat{p}_{\mathfrak{B}\to\mathsf{Up}}}(\theta, m(\theta, s(\theta))) + \epsilon < 0 = U_{\hat{p}_{\mathfrak{B}\to\mathsf{Up}}}(\theta, s(\theta)). \tag{6}$$

with strictly positive probability (over $s \sim \sigma(\theta)$ and $m \sim \mu(s)$) for any $\epsilon < 1-\gamma$. The states $\theta \in \Theta'$ with $\theta_\omega = \widehat{\theta_\omega} = \mathsf{Down}$ have positive measure by supposition, so we have (6) with strictly positive probability; hence, the messaging policy μ is not ϵ-benevolent for any $\epsilon < 1-\gamma$.

Now if instead we have $\widehat{\theta_\omega} = \mathsf{Up}$, i.e., signals are blocked with strictly positive probability in a strictly positive measure of Up-states, then we can conduct an analogous argument under the reverse belief construction. We let $\hat{p}_{\mathfrak{B}\to\mathsf{Down}}$ be the belief for the receiver such that whenever the receiver sees an unblocked signal, they believe the true state of the world is Up, and whenever the receiver sees a partially or fully blocked signal, they believe the true state of the world is Down (with some key). Under beliefs

$\hat{p}_{\mathfrak{B}\to\mathsf{Down}}$, the receiver's payoff whenever they see a blocked signal in states $\theta \in \Theta'$ with $\theta_\omega = \widehat{\theta}_\omega = \mathsf{Up}$ is 0 because whenever the receiver sees a blocked signal, they believe the true state of the world is actually Down (with some key), and so take $a_\omega = \mathsf{Down}$. But meanwhile,

$$U_{\hat{p}_{\mathfrak{B}\to\mathsf{Up}}}(\theta, s(\theta)) = (-1)^{\mathbf{1}_{\theta_\omega=\mathsf{Down}}} \cdot ((1-\gamma) + \gamma \mathbf{1}_{a_\kappa=\theta_\kappa}) \geq 1-\gamma$$

for all states $\theta \in \Theta'$ with $\theta_\omega = \widehat{\theta}_\omega = \mathsf{Up}$, where the equality follows because under an unblocked signal the receiver would (correctly) believe the true state of the world is Up (with some key), and so take $a_\omega = \mathsf{Up}$ with payoff at least $1-\gamma$. Just as in the argument for the preceding case, blocking occurs with strictly positive probability over $s \sim \sigma(\theta)$ and $m \sim \mu(s)$ in the states $\theta \in \Theta'$ with $\theta_\omega = \widehat{\theta}_\omega$, and those states have positive measure by supposition, so we see that the messaging policy μ is not ϵ-benevolent for any $\epsilon < 1-\gamma$.

Proof of Lemma A.2

The proof closely follows the argument used to prove Proposition 2. If $\theta = \mathsf{Up}$, then $m^{\ominus}(s) = s$ (for any realization of s) and so we have $U_p(\theta, m^{\ominus}(\theta, s(\theta))) = U_p(\theta, s(\theta))$ *a priori* for all $p \in \mathcal{P}$, so in particular we have

$$\text{Prob}_{\theta,\sigma}\left[U_p\left(\theta, m^{\ominus}(\theta, s)\right) = U_p\left(\theta, s\right)\right] = 1$$

for any σ.

When $\theta = \mathsf{Down}$, meanwhile, we have $U_p(\theta, s(\theta)) \leq 0$, so we can only have

$$U_p(\theta, m^{\ominus}(\theta, s(\theta))) + \epsilon < U_p(\theta, s(\theta))$$

(again, for any realization of s) if

$$U_p(\theta, m^{\ominus}(\theta, s(\theta))) < -\epsilon \leq -(1-\gamma), \tag{7}$$

where the second inequality follows from the hypothesis that $\epsilon \geq 1-\gamma$. But for (7) to happen, we must have both $a_\omega = \mathsf{Up}$ and $a_\kappa = \theta_\kappa$ with strictly positive probability over θ in response to the message $m^{\ominus}(\theta, s(\theta)) = (s_\omega, \mathfrak{B})$. The remainder of the argument

then follows exactly as in the proof of Proposition 2.

Proof of Lemma A.3

Again, the proof closely follows the argument used to prove Proposition 3. We consider some messaging policy μ such that $\text{Prob}_{\theta,\sigma,\mu}[m(\theta, s(\theta)) = (\mathfrak{B}, s_\kappa(\theta))] > 0$; for such a policy, there is a positive-measure set of states $\Theta' \subseteq \Theta$ such that for all $\theta \in \Theta'$, we have

$$m(\theta, s(\theta)) = (\mathfrak{B}, s_\kappa(\theta))$$

with strictly positive probability over $s \sim \sigma(\theta)$ and $m \sim \mu(s)$. Because θ_ω takes only two possible values, there must be at least one $\widehat{\theta_\omega} \in \{\mathsf{Down}, \mathsf{Up}\}$ such that a strictly positive measure of the states $\theta \in \Theta'$ have $\theta_\omega = \widehat{\theta_\omega}$.

First, we consider the case in which $\widehat{\theta_\omega} = \mathsf{Down}$. We let $\hat{p}_{\mathfrak{B}\to\mathsf{Up}}$ be the belief for the receiver such that whenever the receiver sees an unblocked signal, they believe the true state of the world is Down, and whenever the receiver sees a message in which the signal about the state of the world is blocked but the signal about the key is not, they believe the true state of the world is Up and that the signal about the key conveys the true key. Under belief $\hat{p}_{\mathfrak{B}\to\mathsf{Up}}$, when the receiver sees a blocked signal about the state of the world but an unblocked signal of the key, they believe the true state of the world is actually Up and that the key is $s_\kappa(\theta)$—and so take the action $a = (\mathsf{Up}, s_\kappa(\theta)) = (\mathsf{Up}, \theta_\kappa)$, where the second equality follows because the sender's signal policy transmits the true key. Hence, when the receiver sees a blocked signal about the state of the world but an unblocked signal of the key, in states $\theta \in \Theta'$ with $\theta_\omega = \widehat{\theta_\omega} = \mathsf{Down}$, they receive payoff

$$(-1)^{\mathbf{1}_{\theta_\omega = \mathsf{Down}}} \cdot ((1-\gamma) + \gamma \mathbf{1}_{a_\kappa = \theta_\kappa}) = (-1) \cdot ((1-\gamma) + \gamma) = -1.$$

But this means that

$$U_{\hat{p}_{\mathfrak{B}\to\mathsf{Up}}}(\theta, m(\theta, s(\theta))) = -1 = U_{\hat{p}_{\mathfrak{B}\to\mathsf{Up}}}(\theta, s(\theta)) - 1 \tag{8}$$

with strictly positive probability over $s \sim \sigma(\theta)$ and $m \sim \mu(s)$ for all states $\theta \in \Theta'$ with $\theta_\omega = \widehat{\theta_\omega} = \mathsf{Down}$, where the equality on the right side of (8) follows because under an unblocked signal the receiver would (correctly) believe the true state of the world is Down, and so take $a_\omega = \mathsf{Down}$ with payoff 0. The states $\theta \in \Theta'$ with $\theta_\omega = \widehat{\theta_\omega} = \mathsf{Down}$

have positive measure by supposition, so we have (8) with strictly positive probability; hence, the messaging policy μ is not ϵ-benevolent for any $\epsilon < 1$.

Now if instead we have $\widehat{\theta_\omega} = \mathsf{Up}$, then we can conduct an analogous argument under the reverse belief construction. We let $\hat{p}_{\mathfrak{B}\to\mathsf{Down}}$ be the belief for the receiver such that whenever the receiver sees an unblocked signal, they believe the true state of the world is Up and the signal about the key reflects the true key, and whenever the receiver sees a message in which the signal about the state of the world is blocked but the signal about the key is not, they believe the true state of the world is Down (with some key). Under beliefs $\hat{p}_{\mathfrak{B}\to\mathsf{Down}}$, the receiver's payoff when the receiver sees a blocked signal about the state of the world but an unblocked signal of the key is 0, because in those states the receiver believes the true state of the world is actually Down (with some key), and so takes $a_\omega = \mathsf{Down}$. But meanwhile under an unblocked signal the receiver would (correctly) believe the true state of the world is Up with key $s(\theta) = \theta_\kappa$ (where the equality follows from the hypothesis that the sender's signal policy transmits the true key); this would yield payoff

$$U_{\hat{p}_{\mathfrak{B}\to\mathsf{Down}}}(\theta, s(\theta)) = (-1)^{\mathbf{1}_{\theta_\omega=\mathsf{Down}}} \cdot ((1-\gamma) + \gamma \mathbf{1}_{a_\kappa=\theta_\kappa}) = (1-\gamma) + \gamma = 1.$$

Thus, for all states $\theta \in \Theta'$ with $\theta_\omega = \widehat{\theta_\omega} = \mathsf{Up}$, we have

$$U_{\hat{p}_{\mathfrak{B}\to\mathsf{Down}}}(\theta, m(\theta, s(\theta))) = 0 = U_{\hat{p}_{\mathfrak{B}\to\mathsf{Down}}}(\theta, s(\theta)) - 1$$

with strictly positive probability over $s \sim \sigma(\theta)$ and $m \sim \mu(s)$; because these states have positive measure by supposition, we see that the messaging policy μ is not ϵ-benevolent for any $\epsilon < 1$.