A Persuasive Approach to Combating Misinformation

Safwan Hossain [*1] **Andjela Mladenovic** [*2] **Yiling Chen** [1] **Gauthier Gidel** [2]

Abstract

Bayesian Persuasion is proposed as a tool for social media platforms to combat the spread of misinformation. Since platforms can use machine learning to predict the popularity and misinformation features of to-be-shared posts, and users are largely motivated to share popular content, platforms can strategically signal this informational advantage to change user beliefs and persuade them not to share misinformation. We characterize the optimal signaling scheme with imperfect predictions as a linear program and give sufficient and necessary conditions on the classifier to ensure optimal platform utility is non-decreasing and continuous. Next, this interaction is considered under a performative model, wherein platform intervention affects the user's future behaviour. The convergence and stability of optimal signaling under this performative process are fully characterized. Lastly, we experimentally validate that our approach significantly reduces misinformation in both the single round and performative setting.

1. Introduction

Spreading misinformation has been one of the major critiques of social media platforms. The structure of these platforms rewards users for sharing content to gain popularity (e.g. number of likes and future re-shares), often irrespective of its veracity. This has prompted debates on how platforms should police their content. The most common approach is using fact-checking to tag and censor untruthful content. However, such approaches can hardly keep up with the speed and scale of today's content generation and the validity of content may not be a binary true or false. Censorship moreover leads to thorny debates around the platform regulating freedom of speech. While most Americans support taking steps to restrict misinformation, half of Americans in 2021 agreed that "freedom of information should be prioritized over ... restricting false information online" (Amy & Walker, 2021).

We present information design as a viable approach, either as an alternative or complement, for addressing misinformation on platforms. Our approach leverages the information asymmetry between a platform and its users. The platform strategically reveals some information that users care about, who act in self-interest given the information. When properly designed, this strategic revelation can lead to harmful/poor content being shared less frequently. This approach acknowledges the different incentive structures for the user and platform and does not require the platform to police already shared posts; instead, the platform disincentivizes the sharing of misinformed content by the user.

Specifically, we model the platform-user interaction under the Bayesian Persuasion (BP) framework (Kamenica & Gentzkow, 2011). A post has a two-dimensional hidden state: its popularity if shared and its degree of misinformation. While a user has some prior belief over these states, she naturally does not know the true popularity of her to-be-shared content. She may also be unaware of her post's level of misinformation as it is often introduced unintentionally and driven by social-psychological factors such as a sense of belonging (Wardle, 2020), confirmation bias (Ecker et al., 2022), and habitual sharing (Ceylan et al., 2023). Moreover, the user may also be indifferent to the misinformation state and care only about the popularity their post achieves if shared. While the user may be indifferent or unaware of misinformation, the platform's utility for user action (sharing or not sharing) depends on the true realization of both states — the post's popularity and degree of misinformation. This misalignment of the user's and the platform's utility poses a challenge for the platform: the platform does not want the user to share misinformed posts, but the user derives high utility by sharing popular content, which may not be truthful. The platform, however, possesses an informational advantage due to its vast troves of historical user and content data. So while the platform, like the user, does not know with certainty the true state of the post, it can leverage this to build classification models to predict (with certain error rates) the post's popularity and degree

[*]Equal contribution [1]Harvard University [2]Mila, Université de Montréal. Correspondence to: Safwan Hossain <shossain@g.harvard.edu>, Andjela Mladenovic <andjela.mladenovic@mila.quebec>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

of misinformation. Based on the predictions, the platform can strategically reveal some stochastic information about the post's state, hoping to alter the user's belief about the post. The user, updating their belief based on the revealed information, can decide to share or not share the post to maximize her expected utility. The interaction of the platform and the user forms a Bayesian Stackelberg game, with the platform (leader) choosing an information revelation (signaling) scheme first, with the user (follower) deciding on an optimal action for herself based on the received information. The goal of the platform is to choose a signaling scheme to optimize the platform's expected utility at equilibrium and, by doing so, reduce misinformed content on the platform while maintaining engagement. The platform must be cognizant of the long-term dynamic of such interventions and its implicit effect on user behaviour over time.

Our contributions: We propose an information design approach for social media platforms to address the spread of misinformed content. Platforms predict the properties of a user's to-be-shared post and strategically reveal this to users to improve the quality of shared content. This *noisy persuasion* setting is formally defined in section 3. The lack of perfect information leads to an effective reduction of signaling power. Section 4 discusses this alongside other preliminaries that generalize known results for the standard BP setting. The exact effect of the platform's classification accuracy on its optimal signaling scheme and the resulting utility is then discussed in section 5. Specifically, we formulate this as a linear program and provide sufficient and necessary conditions on the classifier to ensure the optimal platform utility is monotone and continuous. In section 6, the platform-user interaction is viewed through a performative lens, wherein signaling affects the user's sharing behaviour and correspondingly, their future beliefs about their content. We give a complete characterization, proving the stability and convergence of this performative process. Our findings are experimentally validated in section 7, with technical and conceptual extensions for tackling misinformation using information design discussed in section 8.

2. Related Works

This work proposes a soft approach, not involving censorship or tagging, for combating online misinformation, an approach advocated in the literature (Howe et al., 2023; Jackson et al., 2022; Pennycook & Rand, 2022). Howe et al. (2023) proposed to cap the depth (how many times messages can be relayed) or width (how many people a message can be shared with) of a social network to improve information fidelity. Jackson et al. (2022) observed that allowing people to only share posts that they have indicated are true, what they termed self-censoring, reduces the spread of misinformation. Pennycook & Rand (2022) experimentally concluded that interventions that shifted users' attention toward the concept of accuracy could help reduce online misinformation sharing. These works, together with ours, all focus on reducing the sharing of misinformation. Yang et al. (2023) and Candogan & Drakopoulos (2020) respectively studied reducing the creation and the consumption of inaccurate information through signaling. Instead of considering an intervention, Acemoglu et al. (2021) modeled the propagation of misinformation as a game and analyzed its equilibrium, whereas Candogan & Drakopoulos (2020) focuses on optimization with respect to externality effects. While spiritually similar, both assume platforms have perfect knowledge and do not consider performative effects.

The seminal work on Bayesian persuasion (Kamenica & Gentzkow, 2011) has led to many follow-up studies: the computational complexity of the sender's optimization problem (Dughmi & Xu, 2016), the impact of restricting signaling schemes and exogenous information distortions on the informativeness and welfare of equilibrium Kosenko (2021); Tsakas & Tsakas (2021), and robust equilibrium concepts when sender has ambiguity over external information that receiver may have (Dworczak & Pavan, 2022). Information design has also been used for studying price discrimination (Bergemann et al., 2015), multiplayer games (Zhou et al., 2022), and revenue management (Drakopoulos et al., 2021). Bergemann & Morris (2019), Kamenica (2019) and Candogan (2020) offer comprehensive reviews on work in this area. Our work proposes information design as a tool for addressing misinformation on social media under a learned observation model. The insights within this model can also be considered spiritually similar to works on information ordering in economics literature (Bergin, 2005, Chapter 12).

Our dynamic setting in section 6 is inspired by the literature on performative prediction (Perdomo et al., 2020; Mendler-Dünner et al., 2020; Mofakhami et al., 2023). This models the phenomenon that when decision-making is influenced by predictions, the decision will affect future predictions. In our setting, it is not the predictions but rather the signaling process that leads to this dynamic. Our performative model here parallels the stateful one proposed by Brown et al. (2022) and naturally captures the notion that both the past distribution and the user's present sharing decisions (influenced by signaling) will affect content distribution on the platform. The platform should be cognizant of such long-term distribution changes when trying to influence user decisions. From a technical perspective, our results here are novel and not captured by existing results in the literature.

3. Model

Consider the interaction between a social media *platform*, and a *user*. Without any platform intervention, the user lacks additional information when they draft posts for submission;

thus, they take their default action to share. Our model considers a platform predicting features of this draft and committing to revealing or *signaling* this information to the user according to a randomized scheme. Signaling affects user's belief about their post's content, and thus their action to share or not. We now precisely define this, observing that this single-user model is without loss of generality since the platform could interact in such a way with multiple users.

States and Predictions: Let $\mathcal{M} = \{1, \ldots, m_{max}\}$ and $\mathcal{V} = \{1, \ldots, v_{max}\}$ denote the possible misinformation and validation/popularity states respectively. A post drafted by a user has some true joint feature state $\theta = (m, v)$ drawn from a prior distribution $\mu \sim \Delta^{|\Theta|}$, where $\Theta = \mathcal{M} \times \mathcal{V}$ and Δ^k denotes the k simplex. The prior encapsulates the distribution of user's shared posts. Both platform and users know this distribution since it is simply composed of past statistics of this user's content. However, the user does not observe the true state θ for their drafted content - i.e., they do not know the true popularity/misinformation of any drafted post a priori. The platform, however, can leverage their data and scale to predict these states of the draft content, with predictions denoted by $\widehat{\theta} = (\widehat{m}, \widehat{v})$. Unlike the canonical BP settings, we do not assume these predictions to be perfect. We capture the inaccuracy of the prediction models used in our *noisy persuasion* problem as follows:

Definition 1. *The platform's validation and misinformation classifier uncertainty is captured by the multi-class confusion matrices* $Q^{\mathcal{V}} \in [0, 1]^{|\mathcal{V}| \times |\mathcal{V}|}$ *and* $Q^{\mathcal{M}} \in [0, 1]^{|\mathcal{M}| \times |\mathcal{M}|}$. *Let* $Q^{\Theta} = Q^{\mathcal{M}} \otimes Q^{\mathcal{V}}$ *denote the combined confusion matrix, with* $\otimes$ *representing the Kronecker product. An element at index* $(\widehat{\theta}, \theta)$ *of* Q^{Θ}*, denoted by* $Q^{\Theta}_{\widehat{\theta}, \theta}$*, records* $P(\widehat{\theta}|\theta)$.

We assume the predictors $\widehat{m}$ and $\widehat{v}$ to be conditionally independent given m, v and at least as good as chance - i.e. $P(\widehat{m} = x | m = x) \geq \frac{1}{m_{max}}$ and same for $Q^{\mathcal{V}}$.

Actions and Utilities: Upon drafting content, the user can choose between one of two actions: $\mathcal{A} = \{0, 1\}$. The action 0 corresponds to the user *not sharing* the content, and 1 corresponds to them *sharing*. Both platform and user utility depend on this action and the underlying state of the content. We formally define them below:

Definition 2. *We denote the user utility function as* $w : \mathcal{A} \times \mathcal{M} \times \mathcal{V} \rightarrow \mathbb{R}$ *and the platform utility function as* $u : \mathcal{A} \times \mathcal{M} \times \mathcal{V} \rightarrow \mathbb{R}$. *We assume both utilities are bounded.*

Similar to the standard persuasion settings, we assume the platform to know the user's utility. This allows them to anticipate the user's best action and design a scheme accordingly. For users, a special utility structure is when they care only about the validation their posts receive and may be indifferent to or disagree with the platform's characterization of misinformation. Lastly, note that minimal restrictions are placed on platform utility and thus they are free to choose this to balance revenue/engagement and veracity as desired.

Signaling Scheme: The platform maintains a set of *signals* $\mathcal{S}$ and reveals a *signaling scheme* before the user drafts any content. This is simply a set of conditional distributions, denoted by $\pi(s|\widehat{m}, \widehat{v})$, which specifies the probability of sending signal s when the platform predicts the draft content to have state $\widehat{\theta} = (\widehat{m}, \widehat{v})$. Since scheme π must be committed to a priori, the platform goal is to design π to maximize their expected ex-ante utility, formally defined as:

Definition 3. *The platform's* ex-ante utility for a signaling scheme $\pi(s|\widehat{\theta})$ *is* $\sum_s P(s) \, \mathbb{E}_{\rho^s} [u(a^*, \theta)]$*, where* $\rho^s(\theta) = P(\theta|s)$ *is the posterior distribution induced by signal* s *and scheme* π*, and* $a^* = \arg\max_a \mathbb{E}_{\rho^s} [w(\theta, a)]$ *is the optimal receiver action for that posterior belief.*

We now summarize the platform-user interaction for an *instance* $\mathcal{I} = (u, w, \mu)$ as follows:

- Platform reveals a signaling scheme $\pi(s|\widehat{\theta})$
- User drafts content with unknown state $\theta = (m, v) \sim \mu$
- Platform uses learning models with joint confusion matrix Q^{Θ} to obtain prediction $\widehat{\theta} = (\widehat{m}, \widehat{v})$ for this post, and then samples signal $s \sim \pi(s|\widehat{\theta})$.
- User observes the signal, computes their posterior belief, and takes their optimal action a^*.
- User attains utility $w(a^*, \theta)$ and platform attains utility $u(a^*, \theta)$.

Game theoretically, this interaction outlines a Stakelberg game with the platform and user taking on the leader and follower roles respectively. The signaling scheme maximizing the platform's ex-ante utility given the user best-responds (definition 3) is the Stakelberg equilibrium strategy. A key thrust of our work is understanding how properties of the prediction accuracy (i.e. confusion matrix Q^{Θ}) affect this optimal signaling scheme and the resulting platform utility.

Performative Model: Without persuasion, the user takes their utility maximizing action based on their prior: $a^* = \arg\max_a \mathbb{E}_{\mu_0} \, w(a, \theta)$, which we naturally assume to be "share". As such, the distribution of content shared over time matches the original prior. Applying persuasion affects this, however, and the user's decision to share now depends on the signaling scheme; thus, the distribution of a user's shared posts changes over time due to this intervention. We model this through a performative angle. At round t, μ_t represents the distribution of the user's currently published content. The platform deploys a signaling scheme $\pi_t(s|\widehat{\theta})$, and the user observes recommendations from this whenever she drafts content. Note that draft posts that were persuaded to not be shared are absent from the platform, and older content loses relevance. As such, we model the content distribution for round $t + 1$ as a convex combination of the present prior μ_t and the content that was shared this round, $P_t(\theta|a^* = 1)$. This is formally presented in Section 6.

4. Preliminaries

We begin by observing that the noisy setting restricts the platform's signaling power compared to the standard setting. We then establish two characterizations, one on the signal space at optimal signaling and the other on optimal sender utility, which parallel the characterizations in standard Bayesian persuasion. We then give a didactic example that synthesizes these observations and shows that persuasion improves platform utility and reduces misinformation.

4.1. Noisy Persuasion is Less Powerful

Compared with standard BP where the sender observes the realization of θ perfectly and signals based on θ, our setting is noisy as the platform only learns and signals based on $\widehat{\theta}$. In this noisy setting, the set of *effective signaling* schemes - i.e. equivalent signaling schemes that are based on θ - the platform can use is a subset of those in standard noiseless BP settings. Indeed, for a given prior μ and confusion matrix Q^Θ, a noisy signaling scheme $\pi(s|\widehat{\theta})$ is equivalent to a signaling scheme $\widetilde{\pi}(s|\theta) = \sum_{\widehat{\theta}} \pi(s|\widehat{\theta}) Q^\Theta_{\widehat{\theta},\theta}$ in standard BP since both induce the same posterior beliefs over θ.

While any signaling scheme over noisy observations can be transformed into one over exact observations, the converse is not true. That is, there exists schemes $\widetilde{\pi}(s|\theta)$ that cannot be expressed in terms of $\pi(s|\widehat{\theta})$. To see this intuitively, when the platform observes the state θ perfectly, it can signal $\widetilde{\pi}(s = \theta|\theta) = 1$ to fully reveal the state to the user. However, if the platform itself does not observe θ perfectly, it can never induce such a certain belief in the user. In other words, noisy persuasion (when $Q^\Theta \neq \mathcal{I}$, since $\mathcal{I}$ corresponds to perfect predictions/standard BP) has a smaller set of effective signaling schemes at its disposal. While the comparison to standard persuasion is intuitive, a more prescient question is how does the signaling space and crucially the platform's optimal expected utility change between two arbitrary confusion matrices Q^Θ_1 and Q^Θ_2? This is a more involved question which we answer in Theorem 1 by giving a strict ordering for confusion matrices.

4.2. Simplifying the Signaling Scheme

The platform in general can use any set of signals. However, we next show in Proposition 1 that only $|\mathcal{A}|$ signals are needed to attain its best possible ex-ante expected utility in noisy persuasion (hereinafter referred to as *optimal platform utility*), just as in standard BP with perfect observations (Kamenica & Gentzkow, 2011; Dughmi & Xu, 2016). Since $|\mathcal{S}| = |\mathcal{A}|$ suffices, signals can, without loss of generality, be interpreted as recommending a specific action, providing significant operational simplicity (proofs of this section are in Appendix A).

Proposition 1. *For instance $\mathcal{I} = (u, w, \mu)$ and joint confusion matrix Q^Θ, let $u^*_{\mathcal{I}}(Q^\Theta)$ represent the optimal platform utility achievable with an arbitrary number of signals. Then it is also possible for the platform to achieve $u^*_{\mathcal{I}}(Q^\Theta)$ utility using exactly $|\mathcal{A}|$ signals (i.e. $|\mathcal{S}| = |\mathcal{A}|$).*

4.3. Geometry of Noisy Persuasion

For any scheme, the posterior distributions induced by the signal realization s, denoted by ρ^s, must always satisfy $\sum_s P(s)\rho^s = \mu$, a condition termed *Bayes Plausibility*. Kamenica & Gentzkow (2011) show that optimal signaling can be interpreted as inducing a platform-advantageous set of Bayes plausible beliefs. Formally, they construct a mapping from belief to expected sender utility and show that the optimal sender utility is equivalent to evaluating the concave closure of this function at the prior. We now show that this observation can be generalized to our setting with predicted states, a valuable insight for forthcoming results.

Definition 4. *Let $\widehat{\rho}$ denote a belief over predicted states $\widehat{\theta}$, and ρ a belief over true states θ. Then for instance $\mathcal{I}$, define $\overline{u}(\widehat{\rho}) : \Delta^{|\Theta|} \to \mathbb{R}$ as mapping from $\widehat{\rho}$ to the platform expected utility (w.r.t. to the corresponding true belief ρ) for the optimal user action at that belief: $\mathbb{E}_\rho[u(a^*(\rho), \theta)]$. Let $co(\widehat{\rho})$ denote the convex hull of the graph of $\overline{u}(\widehat{\rho})$, and $cl(\widehat{\rho}) = \sup\{\overline{u}(\widehat{\rho}) | (\widehat{\rho}, z) \in co(\widehat{\rho})\}$ its concave closure.*

To interpret this, for a belief over predicted states $\widehat{\rho}$, the corresponding belief over true states ρ can be computed through a linear map V^Θ (see Lemma 2). Then $a^*(\widehat{\rho})$ denotes the optimal user action for the corresponding true belief ρ. With $a^*(\widehat{\rho})$, one can compute the platform's expected utility $\overline{u}(\widehat{\rho})$ over the corresponding $\rho(\theta)$. $co(\widehat{\rho})$ denotes the convex hull of all such $(\widehat{\rho}, \overline{u}(\widehat{\rho}))$ points, with the concave closure $cl(\widehat{\rho})$ representing the boundary of this convex hull. Lastly, Bayes Plausibility can be re-stated as $\sum_s P(s)\widehat{\rho}(s) = \widehat{\mu}_0$, where $\widehat{\mu}$ is the corresponding belief over predicted states for prior μ. Proposition 2 now relates the optimal platform utility to the concave closure, generalizing the result of Kamenica & Gentzkow (2011) to the noisy setting.

Proposition 2. *The platform utility achieved by optimal signaling on instance $\mathcal{I}$ is equal to $cl(\widehat{\mu})$, where $\widehat{\mu} = \sum_\theta \mu(\theta) Q^\Theta_{\widehat{\theta},\theta}$ and $\mu(\theta)$ is the prior.*

Corollary 1. *The expected user utility can never decrease due to any signaling scheme.*

Now consider a function $w_a(\widehat{\rho})$ denoting the user's expected utility for taking action a at belief $\widehat{\rho}$. This is a linear function since the mapping from $\widehat{\rho}(\widehat{\theta})$ to $\rho(\theta)$ is linear (Lemma 1), and expectation is linear. By analogously defining $\overline{w}(\widehat{\rho}) = \max(w_0(\widehat{\rho}), w_1(\widehat{\rho}))$ for the user (their expected utility at a belief due to optimal action), we note this is a convex function. Due to Bayes plausibility, the expected user utility after signaling is $\sum_s P(s)\overline{w}(\widehat{\rho}_s)$, which by the convexity of $\overline{w}$ can never be lower than the utility at the prior (no signaling), leading to Corollary 1. It is also evident from the geometry that the platform's utility also does not decrease due to signaling. To illustrate this, consider an adversarial user who wishes to actively spread misinformation, which is at odds with the platform's goal. Since the platform and user utility are misaligned, the optimal signaling scheme is completely uninformative (ex: send the share and not share signal with equal probability, irrespective of the state). Under such an uninformative/uncorrelated scheme, the adversarial user (who knows the scheme) gains no more information about the state than what they had in their prior and can thus cause no additional harm. The linear program in section 5 finds the optimal signaling scheme for any user utility (including adversarial ones). Geometrically, strict improvement of utility due to signaling is equivalent to $\overline{u}(\widehat{\mu}) < cl(\widehat{\mu})$. Intuitively, if the platform utility $\overline{u}(\widehat{\mu})$ is aligned with the user along some direction in the belief space, then it can always signal to reveal more information along that direction and strictly improve. For any non-adversarial user, such a direction always naturally exists since both parties would prefer to share popular and true posts and not share false unpopular ones. Indeed, alignment is even easier if the user utility is also cognizant of misinformation, and not purely popularity-driven.

4.4. Example

Consider an instance with binary validation and misinformation states (0 is not true/not popular and 1 is false/popular), with 1 and 0 denoting user action "share" and "not share". We consider a popularity-driven user who is indifferent to misinformation, which as discussed is the harder setting. Let the prior μ and platform and user utilities u and w be:

$\theta = (m, v)$	$\mu(\theta)$	$u(\theta, 0)$	$u(\theta, 1)$	$w(\theta, 0)$	$w(\theta, 1)$
$(0, 0)$	0.35	0	1	0	-1
$(0, 1)$	0.35	-1	2	-2	1
$(1, 0)$	0.15	0	-1	0	-1
$(1, 1)$	0.15	0	-3	-2	1

Without signaling, the optimal user action at the prior is to share, whose outcomes are given in the first row of Table 1. Suppose the platform has a 90% accurate classifier for both m and v. The prior over predicted states $\widehat{\theta}$ are presented in Fig. 1 (first set of edges). Consider the signaling scheme specified here, with the platform deterministically signaling "share" or "not share" in the first three states and stochastically signaling in the last state. It can be shown when receiving either signal, the user's optimal action coincides with this recommendation. Further, when the user behaves optimally under this signaling scheme, their expected utility is unchanged; the platform's expected utility, however, and the fraction of shared content that is misinformation, drastically improve (Table 1).

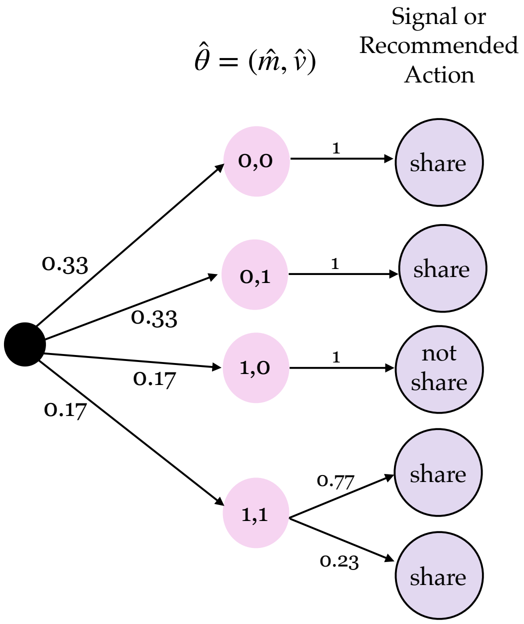

Figure 1. Signaling example. Edges represent probabilities.

	Platform utility	% of shared post that is misinfo.	User utility
Before persuasion	0.45	30	0
After persuasion	0.64	17	0

Table 1. Before and after persuasion.

5. Optimal Noisy Persuasion

While the geometric perspective gives us insight into the properties of signaling, it does not offer a way to compute an optimal signaling scheme. In this section, we begin by providing a linear program (LP) that computes the optimal signaling scheme under inaccurate predictions. We then characterize how the resulting optimal platform utility is affected by the confusion matrix Q^{Θ}. Specifically, we give an ordering of matrices Q^{Θ} as it relates to effective signaling space, provide necessary and sufficient conditions on Q^{Θ} for optimal platform utility to be non-decreasing, and show that this optimal utility is Lipschitz continuous in Q^{Θ}. These results are not only structural but also of operational significance since classifier accuracy is something platforms can modify and improve, making it important to understand the dynamic. Proofs for this section are in Appendix B.

As shown in Section 4.2, it suffices to focus on signaling schemes where $\mathcal{S} = \mathcal{A}$, with each signal interpreted as an action recommendation. When the platform commits to a signaling scheme $\pi(s|\widehat{\theta})$, its effective signaling scheme is $\widetilde{\pi}(s|\theta) = \sum_{\widehat{\theta}} \pi(s|\widehat{\theta}) Q^{\Theta}_{\widehat{\theta},\theta}$. Similar to Dughmi & Xu (2016) who formulated an LP for optimal signaling scheme in standard BP, we formulate the following LP for solving the optimal signaling scheme in our noisy persuasion setting:

$$\begin{aligned}
\max \quad & \textstyle\sum_{a_i}^{|\mathcal{A}|} \sum_{\theta} u(a_i,\theta)\mu(\theta)\widetilde{\pi}(s=a_i|\theta) && (1)\\
\text{s.t.} \quad & \textstyle\sum_{\theta} \Delta w_{ij}(\theta)\mu(\theta)\widetilde{\pi}(s=a_i|\theta) \geq 0 && \forall a_i, a_j \ (2)\\
& \widetilde{\pi}(s=a_i|\theta) = \textstyle\sum_{\widehat{\theta}} \pi(s=a_i|\widehat{\theta}) Q^{\Theta}_{\widehat{\theta},\theta} && \forall\, a_i, \theta \ (3)\\
& \textstyle\sum_{a_i} \pi(s=a_i|\widehat{\theta}) = 1 && \forall \widehat{\theta} \quad (4)\\
& \pi(s=a_i|\widehat{\theta}) \geq 0 && \forall\, a_i, \widehat{\theta} \ (5)
\end{aligned}$$

where $\Delta w_{ij}(\theta) = w(a_i,\theta) - w(a_j,\theta)$. (2) is the incentive compatibility constraint, which enforces that the recommended action has a higher expected utility under the induced posterior than any other action, making it optimal for the user. This phenomenon is often referred to as *optimal signaling is persuasive*. Objective (1) then captures the platform's ex-ante expected utility when $\widetilde{\pi}$ is the induced effective signaling scheme. (1) and (2) is the same as the LP for standard Bayesian persuasion. However, our LP for noisy persuasion requires additional constraints (3), (4), (5), which restrict $\widetilde{\pi}$ to the set of effective signaling schemes that can be induced under confusion matrix Q^{Θ}.

We now wish to understand how for a given instance $\mathcal{I}$, the quality of the platform classifier affects optimal achievable utility, $u^*_{\mathcal{I}}(Q^{\Theta})$. While it is natural that a better classifier would lead to higher utility, there is no unique notion of "better" for a multi-class classifier. For example, entropy, recall, and F1 score are all different notions of classifier quality. We precisely settle this question for symmetric Q^{Θ} in Theorem 1 (Proofs for this section are in Appendix B) by leveraging our LP to prove that the optimal platform utility is non-decreasing exactly with respect to the set inclusion of the convex hull of the columns $[Q^{\Theta}]_{:1}, \ldots, [Q^{\Theta}]_{:|\Theta|}$ of Q^{Θ}. Importantly, our proof shows that if the columns of Q^{Θ}_1 are contained within the convex hull of columns of Q^{Θ}_2, *any* effective signaling for Q^{Θ}_1 can also be achieved under Q^{Θ}_2. If this is not satisfied, then we show that an instance can always be constructed such that the *optimal signaling* under Q^{Θ}_1 corresponds to an effective signaling not possible under Q^{Θ}_2, and vice versa. By taking the contrapositive, it also gives an ordering of confusion matrices as it relates to the effective signaling space, formally presented in Corollary 2.

Theorem 1. *Given two symmetric confusion matrices Q^{Θ}_1 and Q^{Θ}_2 and any instance $\mathcal{I}$, the optimal sender utility, $u^*_{\mathcal{I}}(Q^{\Theta}_2) \geq u^*_{\mathcal{I}}(Q^{\Theta}_1)$, is non-decreasing if and only if $co([Q^{\Theta}_1]_{:1}, \ldots, [Q^{\Theta}_1]_{:|\Theta|}) \subseteq co([Q^{\Theta}_2]_{:1}, \ldots, [Q^{\Theta}_2]_{:|\Theta|})$, where $co([Q^{\Theta}_i]_{:1}, \ldots, [Q^{\Theta}_i]_{:|\Theta|})$ is the convex hull of the columns of Q^{Θ}_i.*

Corollary 2. *Let $\Phi_{\mathcal{I}}(Q^{\Theta})$ denote the set of all effective signaling schemes $\widetilde{\pi}$ that noisy persuasion with confusion matrix Q^{Θ} can achieve. Then for two symmetric confusion matrices $Q^{\Theta}_1, Q^{\Theta}_2$, we have $\Phi_{\mathcal{I}}(Q^{\Theta}_1) \subseteq \Phi_{\mathcal{I}}(Q^{\Theta}_2)$ if and only if $co([Q^{\Theta}_1]_{1:}, \ldots, [Q^{\Theta}_1]_{|\Theta|:}) \subseteq co([Q^{\Theta}_2]_{1:}, \ldots, [Q^{\Theta}_2]_{|\Theta|:})$.*

Theorem 1 is also operationally insightful since the conditions for optimal platform utility to monotonically increase can be easily parsed from the Q^{Θ} matrix, noting that convex-hull inclusion is easily verifiable. It also gives platforms a clear metric for comparing classifier quality for the noisy persuasion task. We next present the second key result of this section which shows exactly where the change in optimal utility due to confusion matrix Q^{Θ} is Lipschitz continuous. Indeed, it is operationally important for platforms to know if and when making slight modifications to the underlying classifier may abruptly affect the optimal persuasion utility and signaling scheme. The proof relies on the geometric insights developed in section 4. To sketch, we show that the concave closure function $cl(\widehat{\rho})$ is Lipschitz in $\widehat{\rho}$. Next, we show that the change from Q^{Θ}_1 to Q^{Θ}_2 leads to a bounded vertical and lateral shift in the closure function due to this property, except possibly if this change leads to the boundary of the effective inducible posterior to be where the user is indifferent.

Theorem 2. *For an instance $\mathcal{I}$, the mapping from confusion matrix to maximum platform utility, $Q^{\Theta} \mapsto u^*_{\mathcal{I}}(Q^{\Theta})$, is Lipschitz continuous everywhere except possibly when there exists an inducible distribution over predicted states, $\widehat{\rho}_d$, where (1) the user is indifferent between optimal actions and (2) $\widehat{\rho}_d$ lies on the $\Delta^{|\Theta|}$ simplex boundary.*

6. Performative Perspectives

We now consider the dynamics of noisy persuasion over time. Recall that without persuasion, users simply take the best action for their prior, which we naturally assume to be "share". As such, the user's content distribution on the platform is unchanged from the initial prior. With persuasive signaling, we now stochastically induce different beliefs in the user, and their decision to share is affected accordingly. Over time, the content distribution this user forms their belief upon will skew toward the type of content they shared. Formally, the prior distribution between two-time steps, μ_t and μ_{t+1}, is affected by the signaling employed by the platform.[1] Correspondingly, the platform's signaling at $t+1$ may change from that at time t. We study the effect of this phenomenon in this section; to focus our analysis, we

[1] The actual time between t and $t+1$ is unimportant for our technical analysis and we leave it for practitioners at deployment.

consider the utility function and predictions models fixed and look to model the interaction between signaling and the subsequent prior belief it gives rise to.

The nascent literature around performative prediction captures a similar tension in the classification setting (Perdomo et al., 2020; Mendler-Dünner et al., 2020; Mofakhami et al., 2023). Therein, an optimal classifier is deployed for a given distribution, which is in turn affected by the chosen classifier. While sharing several parallels, these works generally require the underlying optimization problem to be unconstrained and strongly convex, which while reasonable for classification, do not hold for our optimal signaling linear program. Further, they model the distribution update to depend solely on the optimization variable, where the past has no effect. Transition in our social media setting is not necessarily stateless since older posts still exist on the platform, albeit with diminished relevance. Inspired by Brown et al. (2022), we consider a stateful performative model where we interpolate between the earlier prior and the new distribution affected by signaling. This dynamic is also known in the literature as "geometrically decaying environments" (Ray et al., 2022), and models the environment (including strategic data sources within it) as having a memory shaped by previous interactions with the influence of past events diminishing over time at a geometric rate[2]. We precisely describe this interaction for our setting below:

Definition 5. *The performative persuasion process for an instance $\mathcal{I} = (u, w, \mu_0)$ with joint confusion matrix Q^Θ is defined as follows:*

- *At each round t, with prior $\mu_t(\theta)$, the platform chooses a signaling scheme $\pi_t(s|\widehat{\theta})$, $s \in \{0, 1\}$.*
- *Users take their optimal action (share or not share) for each realized signal.*
- *The next round's prior distribution is given by: $\mu_{t+1} = \lambda\mu_t + (1 - \lambda)\rho_t(\theta|a = 1)$, where $\rho_t(\theta|a = 1)$ is the distribution of content that was shared this round, and $\lambda \in [0, 1]$ is a hyper-parameter.*

While the standard performative analysis restricts itself to studying myopic/greedy agents, i.e. the platform selects the optimal signaling scheme for each round's prior μ_t, we study the process above in more generality and answer several questions that arise naturally. What is the effect of being myopic or far-sighted and how does this relate to the signaling scheme chosen at each round? Will this process converge and how should we qualify this convergent point? What are the *stable points* of this process, an important notion in performative literature that broadly corresponds to a fixed point of the process under myopic actions, and can they be reached from any initial prior? We begin by precisely defining these notions within our setting.

Definition 6. *The performative process converges to distribution μ^* under a sequence of signaling schemes $\{\pi_t\}$ if for any $\varepsilon > 0$, there exists a T_c such that for $t > T_c$, $\mu_t - \mu^* < \varepsilon$, where distribution μ_t arises from using signaling scheme π_{t-1}.*

Definition 7. *Let π_t^* denote the optimal signaling scheme for the prior distribution at round t, with μ_{t+1}^* denoting the resulting next round prior according to the performative update. A distribution μ^s is denoted stable if $\mu_t = \mu^s$ implies $\mu_{t+1}^* = \mu^s$.*

Stable points need not be unique, and our first result gives a sufficient and necessary characterization of all stable points in our persuasion process using the geometric insights of optimal signaling.

Proposition 3. *A prior belief μ_t at time t is stable if and only if $w_1(Q^\Theta\mu_t) \geq w_0(Q^\Theta\mu_t)$ and $cl(Q^\Theta\mu_t) = \overline{u}(Q^\Theta\mu_t)$.*

Since the above implies that the platform utility cannot increase through signaling at a stable point, there is a natural preference order for stable points. For two stable points μ_{st}^1 and μ_{st}^2, the platform prefers $\mu_{st}^1 \succ_p \mu_{st}^2$ if and only if $\overline{u}(\mu_{st}^1) > \overline{u}(\mu_{st}^2)$. We next formalize this optimal stable point and give the exact conditions wherein converging to this optimal stable point through signaling is possible from an initial prior.

Proposition 4. *The optimal stable point distribution is given by*

$$\rho_{max} = \underset{\rho | w_1(Q^\Theta\rho) \geq w_0(Q^\Theta\rho) \wedge cl(Q^\Theta\rho) = \overline{u}(Q^\Theta\rho)}{\arg\max} \overline{u}(Q^\Theta\rho) \quad (6)$$

For an initial prior μ_0, there exists a sequence of signaling schemes such that the performative persuasion process converges to this optimal achievable stable point if there exists a distribution ρ' wherein the optimal user action is to not share such that μ_0 lies in the line segment defined by (ρ', ρ_{max}).

The result above indicates that if platforms were to take a long-term view and deploy possibly non-optimal single-round signaling, then convergence to the optimal distribution may be possible. Nonetheless in many scenarios, taking such a long-term view may not be feasible for platforms. Modeling long-term performative effects is noisy, and imprecise modeling or unexpected behavioral changes may lead to a divergent outcome, with evidence of such phenomena known in the recommendation systems literature (Tennenholtz & Kurland, 2019). To that end, it becomes prescient to

[2] This is also justified by empirical observation in behavioral economics: people tend to measure their choices against a reference distribution, which can change over time to reflect a cumulative effect of past utilities. [Kahneman et al, 2013, Nar, et al. 2017]

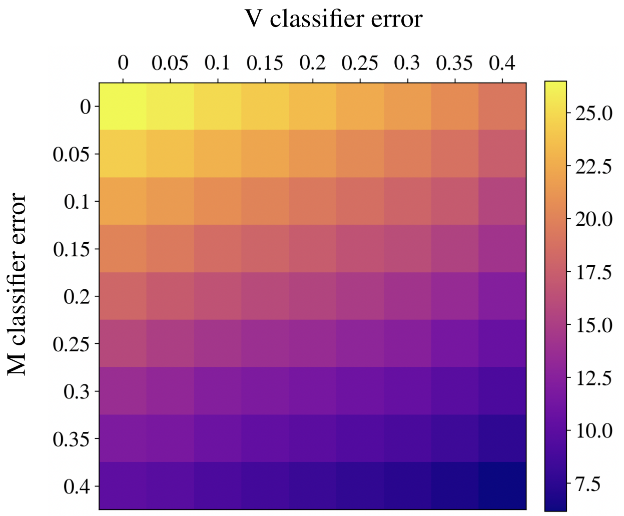

Figure 2. Avg % decrease in misinformation shared due to single application of noisy persuasion.

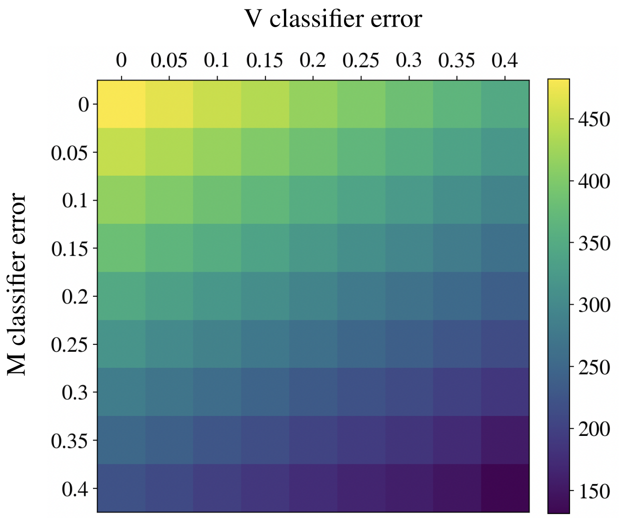

Figure 3. Avg % increase in platform utility between the prior and the myopic stable point.

also consider a more conservative platform that myopically chooses the optimal signaling scheme for each round's prior. Leveraging the suite of results about optimal signaling in noisy persuasion instances, we show that such myopic behaviour will always converge to the optimal posterior belief induced by the share signal (referred to as the share posterior) at the initial round 0. We also show this convergent point to be stable. This proof relies on the geometric insight developed in section 4.3; crucially, we leverage the fact that the performative process does not affect the concave closure function, and show that Bayes plausibility implies μ_t always lies between the first two optimal posteriors induced at the first round.

Theorem 3. *The performative process always converges to the best optimal posterior induced by the share signal in the first round: $\rho_0^1(\theta) = \sum_{\widehat{\theta}} \widehat{\rho}_0^1 V^{\Theta}_{\theta,\widehat{\theta}}$ [3]. Further, this convergent point is always stable.*

The convergence of this process has some nice properties. First, given an instance $\mathcal{I}$, and the respective classifier accuracies, a myopic platform can easily determine the utility at convergence (and thus the benefit of long-term persuasion) since it is simply the utility of the share posterior induced by optimal signaling in the first round. The fact that this convergent distribution is stable is additionally beneficial since it signifies that even in the stateless case ($\lambda = 0$), the process will terminate at the ρ_0^1 belief. We lastly ask whether this myopic stable point is more beneficial than the starting prior for the platform. To that end, we provide the following sufficient condition, which also implies myopic signaling leads to monotonically increasing utility for the platform. We experimentally observe the stable point to be meaningfully beneficial for the platform in practice (Fig. 3).

Proposition 5. *For $c_n = \max_\theta u(0, \theta)$, define a normalized platform utility $u'(a, \theta) = u(a, \theta) - c_n$.*[4] *Then a myopic signaling will converge to a stable point strictly better than the prior in a monotonic fashion if the sender's normalized utility at μ_0 is positive.*

7. Experiments

We now experimentally validate our approach: specifically, while we provide detailed theoretical results on the platform utility under optimal signaling, it is instructive to see how this translates into reducing misinformation sharing. Due to a lack of public data, we create a synthetic dataset for the three components of a noisy persuasion instance: the prior distribution, platform utility, and user utility. Three possible states are considered for validation and misinformation respectively, with 0 representing unpopular/true, 2 representing popular/false, and 1 representing a middle ground. We sample utilities uniformly under the following natural constraints: the platform has 0 utility for the "not share" action, a positive utility that is increasing in v for $m = 0$, a negative utility that is decreasing in v for $m = 2$, and $\forall v$, $u(a = 1, m = 0, v) > u(a = 1, m = 1, v) > u(a = 1, m = 2, v)$. Note that for $m = 1$, the platform utility could either increase or decrease with v, mirroring the possible trade-offs platforms could make between revenue and content quality when uncertain. The user is assumed to be purely popularity-driven with 0 utility for not sharing and increasing utility in v (from negative for $v = 0$ to positive for $v = 2$) for sharing; as discussed in section 4.3, this is

[3] $\widehat{\rho}_0^1$ is the posterior induced by signal 1 (share) at round 0. which has utility $\overline{u}(\widehat{\rho}_0^1)$

[4] Recall the 0 action is to *not share*

usually the harder setting for persuasion. For each instance, we vary the classification error between 0 to 0.4 (the error is equally divided amongst all the incorrect classes) and plot in Fig. 2 the decrease in the percentage of shared posts that are misinformation ($m = 2$) before and after persuasion.

Even when both classifiers perform poorly, a 10% misinformation reduction is achieved, increasing to around 20% when the classifier errors are below 0.15. We also note that the results are more sensitive to the accuracy of the misinformation classifier than the validation one. This is expected since the platform utility ordering essentially flips based on the m state, making it crucial for platforms to capture the true misinformation level more accurately.

Regarding the performative dynamic, note that Theorem 3 implies that myopic signaling leads the process to converge to the optimal induced posterior corresponding to the user sharing $(a = 1)$ at round $t = 0$. Thus, the misinformation shared at the myopically performative convergent belief is also captured by Fig. 2. In Fig. 3, we plot the average increase in platform utility between the starting prior, and this myopically achieved stable distribution. We note a substantial increase in platform utility, which can essentially be seen as a proxy for platform health/revenue. Further augmenting the result of proposition 5, this also suggests that for real-world instances, the myopic stable point is largely beneficial for the platform and that repeated application of persuasion has a long-term positive impact. Lastly, both plots have a 90% confidence interval of less than 4%.

8. Discussion

This work takes a softer approach toward addressing misinformation on social media platforms by leveraging an information design framework based on Bayesian persuasion. Our setting, wherein underlying states are not perfectly observed but predicted, generalizes this classical framework to a noisy and realistic setting. We rigorously characterize how the prediction accuracy affects optimal signaling and platform utility, providing operationally useful results to platforms while noting that user utility can never decrease due to this. Further, the techniques used provide significant insights on noisy persuasion from a geometric and optimization perspective which may be of broader interest. We also consider the long-term implications of such an intervention and model the platform-user interactions from a performative angle. We rigorously characterize the convergence and stability properties of this process; these results illustrate that persuasion can have a long-term positive impact on content quality and veracity on social media platforms.

Our work leaves open a number of technical and conceptual questions. Providing necessary conditions that ensure increased platform utility at a stable point would complement Theorem 5 and be an insightful result. As would augmenting our experiments with real user data and interactions to evaluate the real-world effectiveness of such an approach. Along a similar line, our model assumes a perfectly rational user with the sender not having any exogenous restrictions; generalizing this to consider a bounded rationality model (Jones, 1999; de Clippel & Zhang, 2022) or designing robust signaling under exogenous restrictions (Dworczak & Pavan, 2022; Kosenko, 2021) would be an intriguing research direction. The robust persuasion model of Dworczak & Pavan (2022) may be especially relevant for the misinformation setting since it considers an exogenous third party also influencing the user and proposes optimal signaling for the worst case. Another interesting direction is how persuasion impacts influence propagation and network effects within platform (Barbieri et al., 2013; Arieli et al., 2022). Lastly, developing broader socio-technical guidelines around information design for online interactions is a prescient and necessary direction which we leave for future work.

Acknowledgements

We thank the reviewers for the helpful comments. This work is supported by NSF under grant No. IIS-2147187, Amazon, as well as CIFAR, FRQNT, and Mila EDI funding.

Impact Statement

Misinformation on social media widely is considered one of the most pressing issues in contemporary society, affecting a wide-range of issues from democratic governance to public health (Persily & Tucker, 2020; Suarez-Lledo & Alvarez-Galvez, 2021). As a socio-technical issue, it is imperative that we as computer scientist look to mitigate this important problem.

While there is no debate about the negative outcomes associated with misinformation, exactly classifying misinformation is more challenging, especially at scale. This renders typical content moderation and censoring approaches difficult. Further, it raises concerns about social media platforms abusing their power to regulate discourse, often leading to the creation of echo chambers. As we presented in our work, public opinion is also wary of such approaches (Amy & Walker, 2021). Correspondingly, our work proposes a constructive information design approach that does not require platforms policing content - rather, it attempts to use information revelation to persuade rational users away from sharing misinformation. We make no unrealistic assumptions about the setting, and strongly believe this can be a useful first step toward tackling this issue. That said, there are several considerations to be aware of:

- The persuasion approach relies on user being, at worst, indifferent to misinformation in their content. As such, it will not be as successful in dealing with malicious actor who consciously and willingly spread misinformation. Standard approaches like identifying and banning such actors are still needed.
- In persuasion, the platform commits to a signaling scheme prior to observations, and honours that scheme upon observation. It requires the platform/sender to not lie. This assumption is considered largely innocuous in economics literature as platforms risk severe reputational damage. In practice however, we believe strong regulatory guidelines also need to be established to further discourage such behaviour.

References

Acemoglu, D., Ozdaglar, A., and Siderius, J. A model of online misinformation. Technical report, National Bureau of Economic Research, 2021.

Amy, M. and Walker, M. More americans now say government should take steps to restrict false information online than in 2018. *Pew Research Center, Washington, D.C.*, 2021.

Arieli, I., Gradwohl, R., and Smorodinsky, R. Herd design. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 871–872, 2022.

Barbieri, N., Bonchi, F., and Manco, G. Topic-aware social influence propagation models. *Knowledge and information systems*, 37:555–584, 2013.

Bergemann, D. and Morris, S. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, March 2019. doi: 10.1257/jel.20181489. URL https://www.aeaweb.org/articles?id=10.1257/jel.20181489.

Bergemann, D., Brooks, B., and Morris, S. The limits of price discrimination. *American Economic Review*, 105(3):921–57, March 2015. doi: 10.1257/aer.20130848. URL https://www.aeaweb.org/articles?id=10.1257/aer.20130848.

Bergin, J. *Microeconomic theory: a concise course*. Oxford University Press, 2005.

Boyd, S. P. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Brown, G., Hod, S., and Kalemaj, I. Performative prediction in a stateful world. In *International Conference on Artificial Intelligence and Statistics*, pp. 6045–6061. PMLR, 2022.

Candogan, O. *Information Design in Operations*, chapter 8, pp. 176–201. INFORMS, 2020. doi: 10.1287/educ.2020.0217. URL https://pubsonline.informs.org/doi/abs/10.1287/educ.2020.0217.

Candogan, O. and Drakopoulos, K. Optimal signaling of content accuracy: Engagement vs. misinformation. *Operations Research*, 68:497–515, 2020.

Ceylan, G., Anderson, I. A., and Wood, W. Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences*, 120(4):e2216614120, 2023. doi: 10.1073/pnas.2216614120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2216614120.

de Clippel, G. and Zhang, X. Non-bayesian persuasion. *Journal of Political Economy*, 130(10):2594–2642, 2022.

Drakopoulos, K., Jain, S., and Randhawa, R. Persuading customers to buy early: The value of personalized information provisioning. *Management Science*, 67(2):828–853, 2021. doi: 10.1287/mnsc.2020.3580. URL https://doi.org/10.1287/mnsc.2020.3580.

Dughmi, S. and Xu, H. Algorithmic bayesian persuasion. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 412–425, 2016.

Dworczak, P. and Pavan, A. Preparing for the worst but hoping for the best: Robust (bayesian) persuasion. *Econometrica*, 90(5):2017–2051, 2022.

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., and Amazeen, M. A. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29, 2022. doi: 10.1038/s44159-021-00006-y. URL https://doi.org/10.1038/s44159-021-00006-y.

Howe, P., Perfors, A., Ransom, K. J., Walker, B., Fay, N., Kashima, Y., and Saletta, M. Self-censorship appears to be an effective way of reducing the spread of misinformation on social media. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, 2023.

Jackson, M. O., Malladi, S., and McAdams, D. 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, 2022. doi: 10.1073/pnas.2205549119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2205549119.

Jones, B. D. Bounded rationality. *Annual review of political science*, 2(1):297–321, 1999.

Kamenica, E. Bayesian persuasion and information design. *Annual Review of Economics*, 11(1):249–272, 2019. doi: 10.1146/annurev-economics-080218-025739. URL https://doi.org/10.1146/annurev-economics-080218-025739.

Kamenica, E. and Gentzkow, M. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Kosenko, A. Noisy bayesian persuasion with private information. *mimeo*, 2021.

Mendler-Dünner, C., Perdomo, J., Zrnic, T., and Hardt, M. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33: 4929–4939, 2020.

Mofakhami, M., Mitliagkas, I., and Gidel, G. Performative prediction with neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 11079–11093. PMLR, 2023.

Pennycook, G. and Rand, D. G. Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, 13 (1):2333, 2022.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

Persily, N. and Tucker, J. A. Social media and democracy: The state of the field, prospects for reform. 2020.

Ray, M., Ratliff, L. J., Drusvyatskiy, D., and Fazel, M. Decision-dependent risk minimization in geometrically decaying dynamic environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8081–8088, 2022.

Suarez-Lledo, V. and Alvarez-Galvez, J. Prevalence of health misinformation on social media: systematic review. *Journal of medical Internet research*, 23(1):e17187, 2021.

Tennenholtz, M. and Kurland, O. Rethinking search engines and recommendation systems: a game theoretic perspective. *Communications of the ACM*, 62(12):66–75, 2019.

Tsakas, E. and Tsakas, N. Noisy persuasion. *Games and Economic Behavior*, 130:44–61, 2021.

Wardle, C. Understanding information disorder, 2020. URL https://firstdraftnews.org/long-form-article/understanding-information-disorder/. Accessed on: October 11, 2023.

Yang, Y.-T., Li, T., and Zhu, Q. Designing policies for truth: Combating misinformation with transparency and information design. *arXiv preprint arXiv:2304.08588*, 2023.

Zhou, C., Nguyen, T. H., and Xu, H. Algorithmic information design in multi-player games: Possibilities and limits in singleton congestion. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, pp. 869, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391504. doi: 10.1145/3490486.3538238. URL https://doi.org/10.1145/3490486.3538238.

A. Appendix A

Lemma 1. *For a signaling scheme $\pi(s|\widehat{\theta})$, the probability of observing a signal s given true realization θ is given by $P(s|\theta) = \sum_{\widehat{\theta}} \pi(s|\widehat{\theta}) Q^{\Theta}_{\widehat{\theta},\theta}$. Further, $P(s|m,v) = \sum_{\widehat{m},\widehat{v}} \pi(s|\widehat{m},\widehat{v}) Q^{\mathcal{M}}_{\widehat{m},m} Q^{\mathcal{V}}_{\widehat{v},v}$*

Proof. The following is due to total probability law: $P(s|\theta) = \sum_{\widehat{\theta}} P(s|\theta,\widehat{\theta}) P(\widehat{\theta}|\theta)$. Note that given $\widehat{\theta}$, signal s is conditionally independent of θ since signaling is directly specified by the former. We also note that the classification predictors $\widehat{m}, \widehat{v}$ are assumed to be conditionally independent given m, v, with Q^{Θ} being a Kronecker product. Thus, $P(s|m,v) = \sum_{\widehat{m},\widehat{v}} \pi(s|\widehat{m},\widehat{v}) P(\widehat{m}|m) P(\widehat{v}|v)$. □

Lemma 2. *Given belief over true states $\rho(\theta)$, the corresponding belief over predicted states is $\widehat{\rho}(\widehat{\theta}) = \sum_{\theta} \rho(\theta) Q^{\Theta}_{\widehat{\theta},\theta}$. Similarly, for belief $\widehat{\rho}(\widehat{\theta})$, the corresponding belief over true states is: $\rho(\theta) = \sum_{\widehat{\theta}} \widehat{\rho}(\widehat{\theta}) V^{\Theta}_{\theta,\widehat{\theta}}$, where the $V^{\Theta}_{\theta,\widehat{\theta}} = \frac{Q^{\Theta}_{\widehat{\theta},\theta}\mu(\theta)}{\sum_{\theta'} Q^{\Theta}_{\widehat{\theta},\theta'}\mu(\theta')}$.*

Proof. First, it is easy to see that $\widehat{\rho}(\widehat{\theta}) = \sum_{\theta} P(\widehat{\theta}|\theta)\rho(\theta) = \sum_{\theta} \rho(\theta) Q^{\Theta}_{\widehat{\theta},\theta}$. For the other direction, note that $\rho(\theta) = \sum_{\widehat{\theta}} \widehat{\rho}(\widehat{\theta}) P(\theta|\widehat{\theta})$. Next, by Bayes rule, it is evident that $P(\theta|\widehat{\theta}) = \frac{Q^{\Theta}_{\widehat{\theta},\theta}\mu(\theta)}{\sum_{\theta'} Q^{\Theta}_{\widehat{\theta},\theta'}\mu(\theta')} \triangleq V^{\Theta}_{\theta,\widehat{\theta}}$. □

Proof of Proposition 1

Proof. Let π^* denote the optimal unrestricted signaling scheme with a total of ℓ signals. We can state the posterior over true states $\theta = (m, v)$ for signal realization s as: $P(\theta|s) = \frac{1}{P(s)}\mu(\theta) \sum_{\widehat{\theta}} \pi^*(s|\widehat{\theta}) Q^{\Theta}_{\widehat{\theta},\theta}$. For each a, let S_a denote the set of signals whose induced posterior under π^* leads to optimal action a. Next, consider a signaling scheme that directly recommends an action, satisfying $|\mathcal{S}| = |\mathcal{A}|$. Define this as follows: $\pi'(a|\widehat{\theta}) = \sum_{s\in S_a} \pi^*(s|\widehat{\theta})$. Next, observe that utility at this optimal scheme, denoted by $u^*_{\mathcal{I}}(Q^{\Theta}) = \sum_a \sum_{s\in S_a} P(s) \sum_{\theta} u(a,\theta) P(\theta|s) = \sum_{a,\theta} \sum_{s\in S_a} u(a,\theta)\mu(\theta) P(s|\theta)$. We next use lemma 1 and write this as equal to:

$$\sum_{a,\theta,\widehat{\theta}} \mu(\theta) u(a,\theta) Q^{\Theta}_{\widehat{\theta},\theta} \sum_{s\in S_a} \pi^*(s|\widehat{\theta}) = \sum_{a,\theta} \mu(\theta) u(a,\theta) \sum_{\widehat{\theta}} Q^{\Theta}_{\widehat{\theta},\theta} \pi'(a|\widehat{\theta})$$

We note that the inner summand (over $\widehat{\theta}$) is equivalent to $P(s = a|\theta)$ under the new action-recommending signaling scheme. Thus we can write the expected utility under the new signaling is: $\sum_a P(s = a) \sum_{\theta} u(a, m, v) P(\theta|a) = u^*_{\mathcal{I}}(Q^{\Theta})$ completing the proof. □

Proof of Proposition 2

Proof. Given a prior $\mu(\theta)$ over true states θ, $\widehat{\mu}$, which we call the noisy prior, can be interpreted as the corresponding belief over predicted states. The Bayes plausibility condition, which immediately follows from Bayes rule, implies that for any signaling scheme π, $\widehat{\mu}(\widehat{\theta}) = \sum_s P(s)\widehat{\rho}(\widehat{\theta}|s)$, where $\widehat{\rho}(\widehat{\theta}|s)$ (also denoted by $\widehat{\rho}^s$) is the induced belief over predicted states upon receiving signal s. Thus, the expected utility of any signaling scheme must be in the set $\{\overline{u}(\widehat{\mu}) | (\widehat{\mu}, \overline{u}(\widehat{\mu})) \in co(\widehat{P})\}$, since this includes all convex combinations of induced beliefs that equal the noisy prior, and their corresponding expected platform utility (which is simply the same convex combination of expected utilities at those beliefs). Thus $z^* = \sup\{\overline{u}(\widehat{\mu}) | (\widehat{\mu}, \overline{u}(\widehat{\mu})) \in co(\widehat{P})\} = cl(\widehat{\mu})$ is the maximum utility achievable by any signaling scheme with an arbitrary number of signals. By proposition 1 we know there exists a signaling scheme with $|S| = |A|$ that can also achieve this utility. □

Using the notation from section 4, recall $w_a(\widehat{\rho})$ denotes the user's expected utility for taking action a at belief $\widehat{\rho}$. Then for a belief $\widehat{\rho}$, if $w_0(\widehat{\rho}) = w_1(\widehat{\rho})$, then $\widehat{\rho}$ represents the threshold where the receiver's optimal action changes. This threshold is a hyperplane since it is the intersection of two hyperplanes, and we call this the *indifference plane* $\mathcal{D}$, since the user is indifferent to both actions at this point. [5]. Since the $\overline{u}(\widehat{\rho})$ function is based on the optimal user action at belief $\widehat{\rho}$, $\overline{u}(\widehat{\rho})$ is possibly discontinuous over this indifference plane since this is where the optimal action for the user changes. We now show

[5] As standard in persuasion literature, we assume when the user is tied, it is broken in favour of the platform

in Lemma 3 that posteriors induced by a strictly optimal signaling scheme (strictly improves upon the platform utility at the prior belief) are inextricably linked to the indifference plane.

Lemma 3. *The concave closure $cl(\widehat{\rho})$ is continuous, piecewise linear, and possibly non-differentiable over the indifference plane. Further, the posteriors $\widehat{\rho}^s$ induced by strictly optimal signaling is either on the simplex boundary or the indifference plane, and $cl(\widehat{\rho}^s) = \overline{u}(\widehat{\rho}^s)$.*

Proof. We observe that $\overline{u}(\widehat{\rho})$ is piece-wise linear due to the mapping from $\widehat{\rho}(\widehat{\theta})$ to $\rho(\theta)$ being linear and expectation being a linear operator, with possible discontinuities at beliefs wherein the receiver is indifferent. Since $co(\widehat{\rho})$ is the convex hull of this piecewise linear function defined for all $\widehat{\rho}$, and $cl(\widehat{\mu})$ corresponds to the boundary of this convex hull, this must be continuous, piecewise linear and possibly non-differentiable over the indifference planes.

Next, invoking lemma 4 in Kamenica & Gentzkow (2011) implies that at any induced posterior $\widehat{\rho}^s$ for an optimal scheme, either (1) the belief is on the boundary, (2) the receiver must be indifferent to multiple actions at this belief, or (3) for any other belief wherein the receiver optimal action is not $a^*(\widehat{\rho}^s)$, it is strictly better for the platform that $a^*(\widehat{\rho}^s)$ is taken. Consider the posterior induced by signal 1, $\widehat{\rho}^1$, and suppose the user optimal action is a. Condition (3) implies that for all beliefs wherein the optimal receiver action is *not share*, the platform would strictly prefer the share action. However, recall that in section 3, we assumed that for each action a, there is a state (and thus a corresponding belief) where the user and platform both prefer this action. Thus, only the first two can hold.

Lastly, to show $cl(\widehat{\rho}^s) = \overline{u}(\widehat{\rho}^s)$, consider the line segment ℓ_1 connecting the induced posteriors of the optimal signaling scheme: $(\widehat{\rho}^0, \overline{u}(\widehat{\rho}^0))$ and $(\widehat{\rho}^1, \overline{u}(\widehat{\rho}^1))$. Observe that the optimal sender utility is obtained by evaluating this line segment at the noisy prior $\widehat{\mu}$. If the line connecting these two points (including the end-points) is part of $cl(\widehat{\rho})$, then our claim holds. If not, then this line must be in the interior of the convex hull since $cl(\widehat{\rho})$ represents the boundary of this convex hull. Then, there exist points $(\widehat{\rho}'_0, \overline{u}(\widehat{\rho}'_0))$ and $(\widehat{\rho}'_1, \overline{u}(\widehat{\rho}'_1))$ such that the line between these two is strictly above ℓ_1. Evaluating this line segment at the prior $\widehat{\mu}$ (and thus satisfying Bayes plausibility) yields a strictly higher value than at ℓ_1, contradicting our original claim that we start with an optimal signaling scheme.

□

As discussed in section 4, the harder scenario is when the user is purely popularity-driven and indifferent to misinformation since there is less alignment with the platform. Such users may indeed not even agree with the platform's characterization of misinformation. In these cases, we show below that while the platform can still signal conditioned on both $\widehat{\theta} = (\widehat{m}, \widehat{v})$, it suffices to reveal the signaling over $\widehat{v}$ to the user without loss of generality. This allows a nice operational simplicity.

Proposition 6. *For a user whose utility only depends on the validation state, it suffices for the platform to reveal their marginal signaling scheme $\pi(s|\widehat{v})$ and $Q^{\mathcal{V}}$ to the user, for them to compute their true posterior $P(v|s) = \frac{1}{P(s)}\mu(v)\sum_{\widehat{v}}\pi(s|\widehat{v})Q^{\mathcal{V}}_{\widehat{v},v}$.*

Proof. Consider the platform revealing the marginal scheme $\pi(s|\widehat{v}) = \sum_{\widehat{m}} \pi(s|\widehat{m}, \widehat{v})$. Since the signal only depends on $\widehat{v}$, users can compute $P(s|v) = \sum_{\widehat{v}} P(s|v, \widehat{v})Q^{\mathcal{V}}_{\widehat{v},v} = \sum_{\widehat{v}} \pi(s|\widehat{v})Q^{\mathcal{V}}_{\widehat{v},v}$, and then compute the desired posterior over v using the prior as follows: $P(v|s) = \frac{1}{P(s)}P(s|v)\mu(v)$. □

B. Appendix B

Proof of Theorem 1

Proof. We first note that in our setting with binary actions, this optimal persuasion LP can be tersely expressed as follows, with c and B only depending on $\mathcal{I}$ (and not on Q^Θ), and $\boldsymbol{\pi}$ and $\widetilde{\boldsymbol{\pi}}$ denoting row vectors capturing the probability of sending signal 1 conditioned on noisy and true observations.

$$\begin{aligned}&\text{maximize:}\langle c, \widetilde{\boldsymbol{\pi}}\rangle\\ &\text{subject to: } \widetilde{\boldsymbol{\pi}}B \geq \mathbf{0} \text{ and } \boldsymbol{\pi}Q^\Theta = \widetilde{\boldsymbol{\pi}} \text{ and } \mathbf{0} \leq \boldsymbol{\pi} \leq \mathbf{1}\end{aligned}$$

The set $\{\boldsymbol{\pi}Q^\Theta \mid \mathbf{0} \leq \boldsymbol{\pi} \leq \mathbf{1}\}$ can be interpreted geometrically. This set corresponds to the *parallelepiped* induced by the rows of Q^Θ. In other words, for a set of vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$, we define $paral(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ as the set of vectors that can be expressed as: $\beta_1\boldsymbol{v}_1 + \cdots + \beta_k\boldsymbol{v}_k$, with $\beta_i \in [0, 1]$. One interpretation of this reformulated LP is that Bayesian persuasion with noisy observations can be seen as standard Bayesian persuasion with an additional constraint that the signaling scheme over the true observations (also referred to as effective signaling) $\widetilde{\boldsymbol{\pi}}$ has to belong to the parallelepiped of the rows of the confusion matrix Q^Θ. We now prove the stated theorem:
$\Rightarrow$ We start by considering the sufficient condition for the optimal platform utility to increase wherein the columns of Q_1^Θ can be written as a convex combination of the columns of Q_2^Θ. Notice since both Q_1^Θ and Q_2^Θ this condition is equivalent to the rows of Q_1^Θ can be written as a convex combination of the rows of Q_2^Θ. Now, denote the convex weights by row vectors $\boldsymbol{\lambda_1}, \ldots, \boldsymbol{\lambda_{|\Theta|}}$. Then the following holds: $LQ_2^\Theta = Q_1^\Theta$, where L is defined as $L = \begin{bmatrix}\boldsymbol{\lambda_1}\\ \vdots\\ \boldsymbol{\lambda_{|\Theta|}}\end{bmatrix}$. Notice that L is a doubly-stochastic matrix. Specifically, observe that $\mathbf{1}LQ_2^\Theta = \mathbf{1}Q_1^\Theta = \mathbf{1}$, which implies $\mathbf{1}L = \mathbf{1}({Q_2^\Theta}^{-1}) = \mathbf{1}$. Let $(\widetilde{\boldsymbol{\pi_1}}, \boldsymbol{\pi_1})$ denote any feasible solution for LP(Q_1^Θ). Specifically, $\widetilde{\boldsymbol{\pi_1}}$ is a row vector whose coordinates are $\pi(1|\theta)$ and $\boldsymbol{\pi_1}$ is a row vector whose coordinates are $\widetilde{\pi}(1|\widehat{\theta})$. We will show that $(\widetilde{\boldsymbol{\pi_2}}, \boldsymbol{\pi_2}) = (\widetilde{\boldsymbol{\pi_1}}, \boldsymbol{\pi_1}L)$ is a feasible solution for LP(Q_2^Θ). Notice that all constraints and objective of LP(Q_2^Θ) depend only on $\widetilde{\boldsymbol{\pi}}$, except following constraints: $\widetilde{\boldsymbol{\pi_2}} = \boldsymbol{\pi_2}Q_2^\Theta$ and $\mathbf{0} \leq \boldsymbol{\pi_2} \leq \mathbf{1}$. Note the first constraint is satisfied by $\boldsymbol{\pi_2}Q_2^\Theta = \boldsymbol{\pi_1}LQ_2^\Theta = \boldsymbol{\pi_1}Q_1^\Theta = \widetilde{\boldsymbol{\pi_1}} = \widetilde{\boldsymbol{\pi_2}}$, while second constraint is

$$\boldsymbol{\pi_2} = \boldsymbol{\pi_1}L = \begin{pmatrix}\langle\boldsymbol{\pi_1}, [L]_{1:}\rangle\\ \vdots\\ \langle\boldsymbol{\pi_1}, [L]_{|\Theta|:}\rangle\end{pmatrix} \tag{7}$$

satisfied by the fact that L is double stochastic matrix. Now, we can conclude that for optimal solution $(\widetilde{\boldsymbol{\pi_1}}^*, \boldsymbol{\pi_1^*})$ for LP(Q_1^Θ), we have that solution $(\widetilde{\boldsymbol{\pi_2}}^*, \boldsymbol{\pi_2^*}) = (\widetilde{\boldsymbol{\pi_1}}^*, \boldsymbol{\pi_1^*}L)$ is feasible solution for LP(Q_2^Θ) and that optimal platform utility for LP(Q_1^Θ) is always achievable when solving LP(Q_2^Θ).

$\Leftarrow$ We now show the stated condition is necessary. Again, we note that the condition wherein the columns of Q_1^Θ can be written as convex combination of columns of Q_2^Θ, is equivalent to condition wherein the rows of Q_1^Θ can be written as convex combination of rows of Q_2^Θ. Now we show that if the condition is not satisfied, there always exist instances wherein the utility is not monotone. For symmetric confusion matrices, we first show that the rows of $[Q_1^\Theta]_{i:}$ belong to

$$paral([Q_2^\Theta]_{1:}, \ldots, [Q_2^\Theta]_{|\Theta|:}) := \{\beta_1[Q_2^\Theta]_{1:} + \cdots + \beta_{|\Theta|}[Q_2^\Theta]_{|\Theta|:} \;:\; \beta_k \in [0, 1]\} \tag{8}$$

also belong to $[Q_1^\Theta]_{i:} \in co([Q_2^\Theta]_{1:}), \ldots, [Q_2^\Theta]_{|\Theta|:})$. Observe that $\sum_j [Q_1^\Theta]_{i:j} = 1$ and further, $\sum_j \sum_\ell \beta_\ell [Q_2^\Theta]_{\ell:j} = \sum_\ell \beta_\ell \sum_j [Q_2^\Theta]_{\ell:j} = \sum_\ell \beta_\ell = 1$, which is exactly the convex hull condition. Thus, the convex hull structure is equivalent to the parallelepiped structure when Q_1^Θ and Q_2^Θ are symmetric stochastic matrices. Now, assume there exists a column of matrix Q_1^Θ, such that $[Q_1^\Theta]_{i:} \notin paral([Q_2^\Theta]_{1:}, \ldots, [Q_2^\Theta]_{|\Theta|:})$ then we aim to show that there exists an instance $\mathcal{I} = (u, w, \mu)$ such that $u_{\mathcal{I}}^*(Q_1^\Theta) > u_{\mathcal{I}}^*(Q_2^\Theta)$, violating the monotone condition. Consider an instance wherein the receiver is indifferent to both actions at all states. Thus, the first two constraints of LP(Q^Θ) are always satisfied. Then for Q_1^Θ, let ϕ_1 denote the set of any $\widetilde{\boldsymbol{\pi}}$ such that $\boldsymbol{\pi}Q_1^\Theta = \widetilde{\boldsymbol{\pi}}$, with $\mathbf{0} < \boldsymbol{\pi} < \mathbf{1}$, define ϕ_2 similarly for Q_2^Θ. Since there exists a column of matrix Q_1^Θ, such that $[Q_1^\Theta]_{i:} \notin paral([Q_2^\Theta]_{1:}, \ldots, [Q_2^\Theta]_{|\Theta|:})$ we know that there exists a point, denoted by $\widetilde{\widetilde{\boldsymbol{\pi}}}_{\mathbf{1}}$ which belongs to $paral([Q_1^\Theta]_{1:}, \ldots, [Q_1^\Theta]_{|\Theta|:})$, but not $paral([Q_2^\Theta]_{1:}, \ldots, [Q_2^\Theta]_{|\Theta|:})$. Since $\widetilde{\widetilde{\boldsymbol{\pi}}}_{\mathbf{1}} \in paral([Q_1^\Theta]_{1:}, \ldots, [Q_1^\Theta]_{|\Theta|:})$ we know that $\exists\boldsymbol{\pi_1}$ whose values are between $\mathbf{0}$ and $\mathbf{1}$ such that $\boldsymbol{\pi_1}Q_1^\Theta = \widetilde{\widetilde{\boldsymbol{\pi}}}_{\mathbf{1}}$. On contrary, since $\widetilde{\widetilde{\boldsymbol{\pi}}}_{\mathbf{1}} \notin paral([Q_2^\Theta]_{1:}, \ldots, [Q_2^\Theta]_{|\Theta|:})$ we know that $\nexists \boldsymbol{\pi_2}$ whose values are between 0 and 1 such that $\boldsymbol{\pi_2} Q_2^{\Theta} = \widetilde{\widetilde{\boldsymbol{\pi}}}_{\mathbf{1}}$. Therefore, we know that $\exists \widetilde{\widetilde{\boldsymbol{\pi}}}_{\mathbf{1}}$ such that $\widetilde{\widetilde{\boldsymbol{\pi}}}_{\mathbf{1}} \in \phi_1$ and $\widetilde{\widetilde{\boldsymbol{\pi}}}_{\mathbf{1}} \notin \phi_2$. Now, by Hyperplane Separation Theorem (Boyd & Vandenberghe, 2004, Exercise 2.22), we know that $\exists \boldsymbol{b}$ such that $\langle \boldsymbol{b}, \widetilde{\widetilde{\boldsymbol{\pi_1}}} \rangle > c_1$ and $\langle \boldsymbol{b}, \widetilde{\boldsymbol{\pi_2}} \rangle < c_2, \forall \widetilde{\boldsymbol{\pi_2}} \in \Phi_2$ such that $c_1 > c_2$. Now notice that we can simply achieve our objective expression by $(u(a_1, \theta) - u(a_2, \theta))\mu(\theta) = b_\theta$ where $\theta = 1 \ldots |\Theta|$, and b_θ is the θ-th coordinate of vector $\boldsymbol{b}$. Notice that in the edge case, where $\mu(\theta) = 0$ for some θ values the initial LP problem reduces to the same LP structure with a smaller dimension, and the proof is exactly the same. Therefore, we see that we can find examples where $u^*_{\mathcal{I}}(Q_1^{\Theta}) > u^*_{\mathcal{I}}(Q_2^{\Theta})$. □

Proof of Theorem 2

Proof. We wish to show that for any instance $\mathcal{I}$ and any pair $(Q_1^{\Theta}, Q_2^{\Theta})$, there exists a constant L such that: $|u^*_{\mathcal{I}}(Q_1^{\Theta}) - u^*_{\mathcal{I}}(Q_2^{\Theta})| \leq L||Q_1^{\Theta} - Q_2^{\Theta}||_\infty$, where we use the ℓ_∞ matrix norm. We first show that the concave closure function $cl(\widehat{\rho})$ is Lipschitz in $\widehat{\rho}$. Next, we show that the change from Q_1^{Θ} to Q_2^{Θ} leads to a bounded vertical shift in the closure function due to this Lipschitz property. Lastly, the point at which we evaluate the concave closure function to determine the optimal sender utility also changes in a bounded manner. The combined effects are all bounded and give rise to our Lipschitz constant.

From lemma 3, we know that on either side of the plane of indifference, $cl(\widehat{\rho})$ is a continuous linear function. Since we have 2 actions, there is a single plane of indifference. We will aim to tightly upper bound the directional derivative for each joint state $\widehat{\theta} = (\widehat{m}, \widehat{v})$ of the linear regions on either side of this indifference plane. Since the concave closure function is linear on either side of the indifference plane, it suffices to consider the maximum value of $\frac{\Delta u}{\Delta \widehat{\rho}_{\widehat{\theta}}}$ from the indifference plane, where $\widehat{\rho}_{\widehat{\theta}}$ denotes the coordinate corresponding to $\widehat{\theta}$. Pick an arbitrary belief $\widehat{\rho}$ on the indifference plane. The slope along a direction $\widehat{\theta}$ is naturally upper bounded by $\frac{u_s^{max} - u_s^{min}}{\min(\widehat{\rho}_{\widehat{\theta}}, 1 - \widehat{\rho}_{\widehat{\theta}})}$. We now look to tighten this. Indeed, if there exists another belief $\widehat{\rho}'$ on the indifference plane such that $\widehat{\rho}'_{\widehat{\theta}}$ is closer to 0.5, this would imply the change in utility is larger than $u^{max} - u^{min}$, which is not possible. To tighten this, let $\widehat{\rho}_{\widehat{\theta}}^{min}$ and $\widehat{\rho}_{\widehat{\theta}}^{max}$ denote the smallest and largest values of coordinate $\widehat{\theta}$ for beliefs $\widehat{\rho}$ that lie on the indifference plane. Then the largest value of the directional derivative of the closure function along the $\widehat{\theta}$ direction is given by $c_{\widehat{\theta}} = \max\left(\frac{u^{max} - u^{min}}{1 - \widehat{\rho}_{\widehat{\theta}}^{min}}, \frac{u^{max} - u^{min}}{\widehat{\rho}_{\widehat{\theta}}^{max}}\right)$. Thus, when belief changes from $\widehat{\rho}_1$ to $\widehat{\rho}_2$, the maximum change in the concave closure function is upper-bounded by $|\widehat{\rho}_1 - \widehat{\rho}_2| \sum_{\widehat{\theta}} c_{\widehat{\theta}}$, with $c = \sum_{\widehat{\theta}} c_{\widehat{\theta}}$ being the Lipschitz constant.

For a belief over predicted states $P(\widehat{\theta}) = \widehat{\rho}$, the corresponding belief over true states $\rho = P(\theta)$ changes under the different confusion matrices. This affects the closure graph in two ways. First, it vertically shifts the underlying $\overline{u}$ function since $\overline{u}(\widehat{\rho}) = \mathbb{E}_{\rho(\theta)}[u(a^*, m, v)]$. More formally, for a given belief $\widehat{\rho}$ the change due to the shift from Q_1^{Θ} to Q_2^{Θ} is given by: $\overline{u}(\widehat{\rho}; Q_1^{\Theta}) - \overline{u}(\widehat{\rho}; Q_2^{\Theta})$

$$= \sum_{\theta} u(a^*, \theta) \sum_{\widehat{\theta}} \widehat{\rho}(\widehat{\theta})(V_1^{\Theta}(\theta, \widehat{\theta}) - V_2^{\Theta}(\widehat{\theta}, \theta)) \leq |\Theta| u^{max} ||V_1^{\Theta} - V_2^{\Theta}||$$

which follows due to lemma 2. Next, exploiting the relationship between matrices V^{Θ} and Q^{Θ} as outlined in lemma 2, the following holds for any element $(\theta, \widehat{\theta})$ of matrix V^{Θ}:

$$V_{\theta,\widehat{\theta}}^{\Theta} = \frac{\mu(\theta) Q_{\widehat{\theta},\theta}^{\Theta}}{\sum_{\theta'} \mu(\theta') Q_{\widehat{\theta},\theta'}^{\Theta}} \implies \left| \frac{\partial V_{\theta,\widehat{\theta}}^{\Theta}}{\partial Q_{\widehat{\theta},\theta_i}^{\Theta}} \right| \leq \frac{\mu(\theta_i)\mu(\theta) Q_{\widehat{\theta},\theta}^{\Theta}}{\left(\sum_{\theta'} \mu(\theta') Q_{\widehat{\theta},\theta'}^{\Theta}\right)^2} \leq \frac{|\Theta|^2}{\mu_{min}^2} \tag{9}$$

where the last inequality follows since we consider classifiers to at at least as good as chance, the diagonals of Q^{Θ} are always at least $\frac{1}{|\Theta|}$. We also note that $\mu_{min} > 0$ since if any state occurs with 0 probability, we can without loss of generality, reformulate the problem to exclude that state. With the dependence of an element of V^{Θ} established to be Lipschitz in an element of Q^{Θ}, it follows that $||V_1^{\Theta} - V_2^{\Theta}|| \leq M||Q_1^{\Theta} - Q_2^{\Theta}||$ where $M = \frac{|\Theta|^3}{\mu_{min}^2}$.

Second, let ρ_c denote a belief wherein the user has equal expected utility for both actions - in other words, the user is indifferent. The predicted induced belief $\widehat{\rho}_c$ that corresponds to this indifferent belief may also shift due to the change from Q_1^{Θ} to Q_2^{Θ}, affecting the closure. Formally, let the row vector $\widetilde{\pi}$ denote a signaling scheme over true states θ (suffices to consider only the "share" signal) that induces ρ_c. To satisfy the feasibility constraints (see equations 1), we need a signaling over predicted states $\widehat{\theta}$ such that: $\pi_i = \widetilde{\pi} Q_i^{\Theta}$ and $\pi_i \in \Delta^{|\Theta|}$ for $i \in \{1, 2\}$. When this is not feasible for both, this indifference point is not reachable by signaling and the change from Q_1^{Θ} to Q_2^{Θ} has no effect on the closure with respect to the indifference plane. When this is feasible for one but not the other, it maps exactly to the possible discontinuity mentioned in the statement. Consequently, we consider the last case - both π_1 and π_2 are feasible. We have that:

$$\widehat{\rho}_1 = \frac{\pi_1 \odot \widehat{\mu}}{||\widetilde{\pi}||_1} = \frac{\widetilde{\pi} {Q_1^{\Theta}}^{-1} \operatorname{Diag}(Q_1^{\Theta} \mu)}{||\widetilde{\pi}||_1} \tag{10}$$

$$\implies ||\widehat{\rho}_1 - \widehat{\rho}_2|| \leq \frac{\widetilde{\pi}}{||\widetilde{\pi}||} ||{Q_1^{\Theta}}^{-1} \operatorname{Diag}(Q_1^{\Theta} \mu) - {Q_2^{\Theta}}^{-1} \operatorname{Diag}(Q_2^{\Theta} \mu)|| \tag{11}$$

By adding and subtracting ${Q_1^{\Theta}}^{-1} \operatorname{Diag}(Q_2^{\Theta} \mu)$ and using the triangle inequality, we have that:

$$\leq ||{Q_1^{\Theta}}^{-1} \operatorname{Diag}(Q_1^{\Theta} \mu) - {Q_1^{\Theta}}^{-1} \operatorname{Diag}(Q_2^{\Theta} \mu)|| + ||{Q_1^{\Theta}}^{-1} \operatorname{Diag}(Q_2^{\Theta} \mu) - {Q_2^{\Theta}}^{-1} \operatorname{Diag}(Q_2^{\Theta} \mu)|| \tag{12}$$

Using the sub-multiplicative property of the matrix norm, the fact that $|| \operatorname{Diag}(v)|| = ||v||$, and that μ belongs to the simplex, the first term can be bounded by: $||{Q_1^{\Theta}}^{-1}|| \cdot ||Q_1^{\Theta} - Q_2^{\Theta}||$. Similarly, the second term of equation 12 can be upper bounded by: $||{Q_1^{\Theta}}^{-1} - {Q_2^{\Theta}}^{-1}|| \cdot ||Q_2^{\Theta}||$. Therefore, we can bound the distance between the two indifference points with respect to the predicted states by:

$$||\widehat{\rho}_1 - \widehat{\rho}_2|| \leq ||{Q_1^{\Theta}}^{-1}|| \cdot ||Q_1^{\Theta} - Q_2^{\Theta}|| + ||{Q_1^{\Theta}}^{-1} - {Q_2^{\Theta}}^{-1}|| \cdot ||Q_2^{\Theta}|| \tag{13}$$

Since we consider non-degenerate confusion matrices Q^{Θ} whose inverses satisfy $\|{Q^{\Theta}}^{-1}\| \leq L$ (for instance, if $\|Q^{\Theta} - I_d\| \leq r < 1$, we would have $L \leq \frac{1}{1-r}$) we can appeal to the sub-multiplicavity of the matrix norm and state: $||{Q_1^{\Theta}}^{-1} - {Q_2^{\Theta}}^{-1}|| \leq L^2 ||Q_1^{\Theta} - Q_2^{\Theta}||$. Thus, we have that:

$$||\widehat{\rho}_1 - \widehat{\rho}_2|| \leq L ||Q_1^{\Theta} - Q_2^{\Theta}|| + L^2 ||Q_1^{\Theta} - Q_2^{\Theta}|| \tag{14}$$

There is thus a region of length $\lambda ||Q_1^{\Theta} - Q_2^{\Theta}||$ (λ surmises the constants above) where the optimal action is different under $cl(\widehat{\rho}; Q_1^{\Theta})$ and $cl(\widehat{\rho}; Q_2^{\Theta})$. The largest difference between the two closure functions for a belief $\widehat{\rho}$ occurs in this region since one function could be increasing and the other decreasing (for a belief outside of this region, the optimal action is the same thus, both functions are both increasing or decreasing). However, since the closure function is lipschitz with constant c, the maximum difference at any belief $\widehat{\rho}$ is given by: $|cl(\widehat{\rho}; Q_1^{\Theta}) - cl(\widehat{\rho}; Q_2^{\Theta})| \leq c\lambda ||Q_1^{\Theta} - Q_2^{\Theta}|| + |\Theta| M u^{max} ||Q_1^{\Theta} - Q_2^{\Theta}||$.

Lastly, recall that by theorem 2, the optimal platform utility is equal to evaluating the closure function at the noisy prior $\widehat{\mu}(\widehat{\theta})$. For a prior $\mu(\theta)$, we have established that the predicted belief shifts at most $|\Theta| ||Q_1^{\Theta} - Q_2^{\Theta}||_1$. Since the closure function is Lipschitz, the change in the evaluation point from $\widehat{\mu}(\widehat{\theta}\; Q_1^{\Theta})$ to $\widehat{\mu}(\widehat{\theta}\; Q_2^{\Theta})$ leads to a difference of at most $c|\Theta| ||Q_1^{\Theta} - Q_2^{\Theta}||_1$. Combining this with the fact that the closure functions are vertically by at most $(\lambda c + M u^{max}) |\Theta| ||Q_1^{\Theta} - Q_2^{\Theta}||_1$ at any belief, implies that the total change in optimal utility due to change from Q_1^{Θ} to Q_2^{Θ} is at most $|\Theta|(2\lambda c + M u^{max}) ||Q_1^{\Theta} - Q_2^{\Theta}||_1$. Thus, the optimal sender utility for an instance $\mathcal{I}$ as a function of confusion matrix Q^{Θ} is $|\Theta|(2\lambda c + M u^{max})$-Lipschitz. □

C. Appendix C

Proof of Propostion 3

Proof. We note that for any prior belief μ_t over true states θ, $Q^{\Theta} \mu_t$ is the corresponding noisy prior belief over predicted states $\widehat{\theta}$, which we notationally denote as $\widehat{\mu}$. The first condition states that for belief μ_t, the optimal action for the receiver is to share. Observe that if this is not the case then either (1) optimal signaling with induce a share posterior which cannot be equal to μ_t and thus violate stability or (2) optimal signaling will only induce "not share" posteriors, in which case the user never shares and the performative process becomes ill-defined and the user essentially detaches from the platform.

Next, from proposition 2 we know that the optimal utility from signaling is equal to $cl(\widehat{\mu})$. If $cl(\widehat{\mu}_t) = cl(Q^{\Theta} \mu_t) = \overline{u}(Q^{\Theta} \mu_t) = \overline{u}(\widehat{\mu}_t)$, with the latter representing the platform utility without signaling, we confirm that signaling cannot increase utility and μ_t is thus stable. For the reverse direction, it suffices to note that: $cl(Q^{\Theta} \mu_t) \neq \overline{u}(Q^{\Theta} \mu_t) \implies cl(Q^{\Theta} \mu_t) > \overline{u}(Q^{\Theta} \mu_t)$ (due to cl representing the concave closure of $\overline{u}$), and there thus exist distinct posteriors inducible through signaling such that the expected utility is equal to $cl(Q^{\Theta} \mu_t)$. Due to Bayes plausibility, these posteriors cannot be equal to the prior; thus, the posterior corresponding to user's sharing is distinct from the noisy prior $\widehat{\mu}_t$; correspondingly, μ_t is not stable. □

C.1. Proof of Proposition 4

Proof. We consider the equivalent optimal stable distribution over predicted states as $\widehat{\rho}_{max} = \arg\max_{\rho | w_1(\widehat{\rho}) \geq w_0(\widehat{\rho}) \wedge cl(Q^\Theta \rho) = \overline{u}(Q^\Theta \rho)} \overline{u}(\widehat{\rho})$. Note that the constraints here correspond exactly to the stability conditions outlined in proposition 3. Thus, it is evidently the optimal stable point of the process. Next, we note that if such a ρ' does exist, then let $\widehat{\rho}' = Q^\Theta \rho'$ and there must exist parameter α such that $\alpha_0 \widehat{\rho}_{max} + (1 - \alpha_0)\widehat{\rho}' = Q^\Theta \mu_0 = \widehat{\mu}_0$. In other words, Bayes plausibility is satisfied, implying that this pair of posteriors is inducible through signaling.

The performative dynamic implies the next prior, $\widehat{\mu}_1$ is a convex combination of $\widehat{\mu}_0$ and $\widehat{\rho}_{max}$. Since $\widehat{\mu}_0$ itself is a convex combination of $\widehat{\rho}_{max}$ and $\widehat{\rho}'$, we have that:

$$\widehat{\mu}_1 = \lambda \widehat{\mu}_0 + (1 - \lambda)\widehat{\rho}_{max} = \lambda \left[\alpha_0 \widehat{\rho}_{max} + (1 - \alpha_0)\widehat{\rho}'\right] + (1 - \lambda)\widehat{\rho}_{max} \tag{15}$$

$$= \widehat{\rho}_{max}[\lambda \alpha_0 + 1 - \lambda] + \lambda(1 - \alpha_0)\widehat{\rho}' \tag{16}$$

Note that inducing posteriors $\widehat{\rho}_{max}$ and $\widehat{\rho}'$ is still Bayes plausible in round 1 - albeit with different weights. As such, choose signaling scheme π_1 to induce these posteriors with the corresponding weights. We will choose signaling for all subsequent rounds to induce these posteriors and inductively show it is possible. Observe that if at round t, there exists an α_t such that: $\widehat{\mu}_t = \alpha_t \widehat{\rho}_{max} + (1 - \alpha_t)\widehat{\rho}'$, then $\widehat{\mu}_{t+1} = \widehat{\rho}_{max}[\lambda \alpha_t + 1 - \lambda] + \lambda(1 - \alpha_t)\widehat{\rho}'$. Note that (1) $\widehat{\mu}_{t+1}$ is closer to $\widehat{\rho}_{max}$ than $\widehat{\mu}_t$ since the total weight on $\widehat{\rho}'$ decreases between the two rounds and (2) inducing posteriors $\widehat{\rho}_{max}$ and $\widehat{\rho}'$ is still feasible at round $t + 1$ due to Bayes plausibility being satisfied. We choose π_{t+1} to induce this accordingly. Thus, we note that as $t \to \infty$, this sequence of signaling schemes will lead the process to converge to $\widehat{\rho}_{max}$, the optimal stable point. □

C.2. Proof of Theorem 3

Proof. We first express the performative dynamics with respect to the observed states $\widehat{\theta} = (\widehat{m}, \widehat{v})$. Observe the following: $\lambda \widehat{\mu}_t(\widehat{\theta}) + (1 - \lambda)\widehat{\rho}_t(\widehat{\theta} | s = 1) = \sum_\theta Q^\Theta_{\widehat{\theta},\theta}[\lambda \mu_t(\theta) + (1 - \lambda)\rho_t(\theta | s = 1)]$, which follows due to lemma 2. Next, by invoking the performative dynamics and lemma 2 again, we have that this is equivalent to $\sum_\theta \mu_{t+1}(\theta) Q^\Theta_{\widehat{\theta},\theta} = \widehat{\mu}_{t+1}(\widehat{\theta})$. In other words: $\widehat{\mu}_{t+1}(\widehat{\theta}) = \lambda \widehat{\mu}_t(\widehat{\theta}) + (1 - \lambda)\widehat{\rho}_t(\widehat{\theta} | s = 1)$.

Next, observe that the performative process only changes the prior (and thus the optimal signaling scheme), but does not affect the platform belief to expected utility function $\overline{u}(\widehat{\rho})$ nor its concave closure $cl(\widehat{\rho})$. Both of these depend only on the platform and user utility. Next, consider the scenario at $t = 0$ with noisy prior $\widehat{\mu}_0(\widehat{\theta})$, whereupon the platform commits to an optimal signaling scheme $\pi^*_0(s|\widehat{\theta})$. If there are multiple optimal schemes (thus multiple optimal posteriors that can be induced) and since this is the first round, let the platform break the tie by choosing the pair for whom the corresponding $\overline{u}(\widehat{\rho}_0(\widehat{\theta} | s = 1))$ is the largest[6]. Denote this pair of optimal signaling schemes by $\{\widehat{\rho}^0_0, \widehat{\rho}^1_0\}$. Next, for $\alpha \in [0, 1]$, consider the line segment $\ell(\alpha) = (1 - \alpha)z_0 + \alpha z_1$ which connects the points $z_0 = [\widehat{\rho}^0_0, cl(\widehat{\rho}^0_0)]$ and $z_1 = [\widehat{\rho}^1_0, cl(\widehat{\rho}^1_0)]$. By lemma 3, we know that the endpoints of this line segment, corresponding to the optimal induced posteriors, also lie on the $\overline{u}(\widehat{\rho})$. In other words, $[\widehat{\rho}^x_0, cl(\widehat{\rho}^x_0)] = [\widehat{\rho}^x_0, \overline{u}(\widehat{\rho}^x_0)]$ for $x \in \{0, 1\}$. By Bayes plausibility, there exists an α_0 such that the first element of $\ell(\alpha_0)$, denoted by $\ell(\alpha_0)_0 = \widehat{\mu}_0$. Theorem 2 further states that the expected platform utility under the optimal scheme is given by $cl(\widehat{\mu}_0) = \sum_{x \in 0,1} P_0(x)\overline{u}(\widehat{\rho}^x_0)$, where $P_0(x)$ is the probability of signal x at round 0. Observe that this point is on the line segment ℓ – i.e. $\ell(\alpha_0) = [\widehat{\mu}_0, cl(\widehat{\mu}_0)]$. Thus, we have that the line segment ℓ matches the $cl(\widehat{\rho})$ at both the endpoints and one interior point. Since the $cl(\widehat{\rho})$ function is concave piece-wise linear continuous, it must imply $cl(\widehat{\rho})$ coincides with the line segment ℓ between beliefs $\widehat{\rho}^0_0$ and $\widehat{\rho}^1_0$.

The performative dynamic implies the next prior, $\widehat{\mu}_1$ is a convex combination of $\widehat{\mu}_0$ and $\widehat{\rho}^1_0$. Thus, there exists an α_1 such that $\ell(\alpha_1)_0 = \widehat{\mu}_1$, since the performative process moves to a point between the earlier prior and the $s = 1$ posterior, both of which are on the ℓ line segment. More specifically, $\alpha_1 = (1 - \lambda) + \lambda \alpha_0$. The value of optimal signaling at $\widehat{\mu}_1$ is again equivalent to $cl(\widehat{\mu}_1)$ by theorem 2. Since the curve $\ell(\cdot)$ coincides with $cl(\cdot)$ between $\widehat{\rho}^0_0$ and $\widehat{\rho}^1_0$, and $\widehat{\mu}_1$ lies in this region, the value of optimal signaling is equal to $\ell(\alpha_1)_1 = cl(\widehat{\mu}_1)$. Thus, $\widehat{\rho}^0_0$ and $\widehat{\rho}^1_0$ are still optimal induced posteriors for prior at $\widehat{\mu}_1$, and recall we break ties by choosing posteriors closest to the last round. In other words, $\widehat{\rho}^x_1 = \widehat{\rho}^x_0$, with $P_1(s = 0) = 1 - \alpha_1$ and $P_1(s = 1) = \alpha_1$ is an optimal scheme for the platform at $\widehat{\mu}_1$. It is evident that this invariant will be maintained throughout this process. Precisely, for each μ_t in this process, there exists a $\alpha_t \in [0, 1]$ such that $\ell(\alpha_t)_0 = \widehat{\mu}_t$, which implies $cl(\mu_t) = \ell(\alpha_t)_1$, which implies signaling such that $\widehat{\rho}^x_t = \widehat{\rho}^x_0$, and $P_t(s = 0) = 1 - \alpha_t$ and $P_1(s = 1) = \alpha_t$

[6] If multiple optimal signaling schemes exist at subsequent rounds, we assume ties are broken by choosing the scheme whose posteriors are closest to the earlier round's posterior.

is optimal for the platform at μ_t. This in turn means μ_{t+1} can be expressed as $\ell(\alpha_{t+1})_0$, with $\alpha_{t+1} = (1-\lambda) + \lambda\alpha_t$. Since $\alpha_t, \lambda \in [0,1]$, the following always holds: $\alpha_t < \alpha_{t+1} < 1$; thus, as $t \to \infty$, $\alpha_t \to 1$ and $\widehat{\mu}_t \to \widehat{\rho}_0^1$.

Lastly, we note from lemma 3 that any posterior distribution induced by an optimal signaling scheme satisfies $cl(\widehat{\rho}_0^1) = \overline{u}(\widehat{\rho}_0^1)$. Proposition 3 immediately implies this convergent distribution is also stable. □

C.3. Proof of Proposition 5

Proof. First, note that scaling the platform utility by an additive constant does not affect optimal signaling, and thus we can solve the normalized instance without loss of generality. Note that in the first round, we assume the user's optimal action to be "share". If the initial is not a stable point, then platform utility can always improve due to signaling. Thus:

$$\mathbb{E}_{\mu_t}[u'(a_t,\theta)] < P_t(s=0)\,\mathbb{E}_{\rho_t^0}[u'(0,\theta)] + P_t(s=1)\,\mathbb{E}_{\rho_t^1}[u'(1,\theta)]$$

Since platform utility for action 0 (not share) is always less than or equal to 0 for any state under the theorem conditions: $\mathbb{E}_{\mu_t}[u'(a_t,\theta)] < P_t(s=1)\,\mathbb{E}_{\rho_t^1}[u'(1,\theta)]$, which implies

$$\frac{1}{P_t(s=1)}\,\mathbb{E}_{\mu_t}[u'(a_t,\theta)] < \mathbb{E}_{\rho_t^1}[u'(1,\theta)] = \mathbb{E}_{\mu_\infty}[u'(1,\theta)]$$

where the last term is the utility at the myopic stable point. Lastly, the proof of theorem 3 shows that the myopic process monotonically approaches ρ_1^0. Since the above implies $cl(Q^\Theta\rho_1^0) > cl(Q^\Theta\mu_0)$ and the utility of optimal signaling always lies on this line segment of the closure function cl, the utility is monotonically increasing over time as well. □