A Persuasive Approach to Combating Misinformation

Safwan Hossain*
Harvard University

Andjela Mladenovic*
MILA, U de Montréal

Yiling Chen
Harvard University

Gauthier Gildel
MILA, U de Montréal

Information Design

How can an agent with informational advantage, strategically reveal this information to another agent to influence their behaviour?

0. Bayesian Persuasion

- Two player game between a sender, who gets to observe a world state $\theta \in \Theta$, and a receiver who gets to take an action [1].

- The utility of both players depend on this action along with the world state.
  - Complete Information - sender knows receiver utility

- Both players share a common prior belief μ about the possible world states θ.

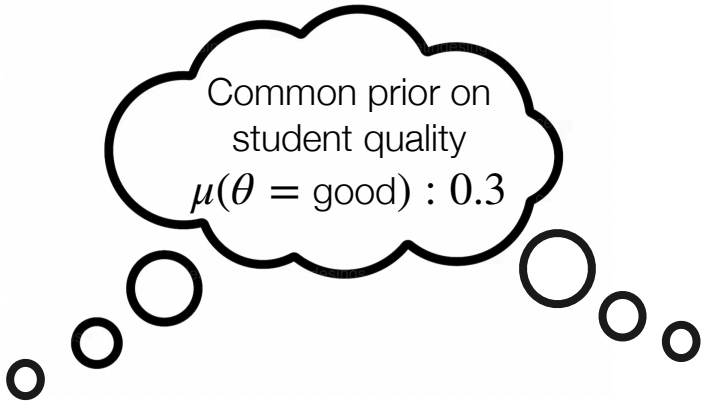
- The sender can commit to strategically revealing her knowledge of the world state through signaling.

- θ - student quality $\{\text{good}, \text{bad}\}$
- $a \in \{\text{hire}, \text{not hire}\}$

- Utility $u(a, \theta)$: +1 if student is hired
- Utility $w(a, \theta)$: +1 if hiring good student or not hiring bad ones

Sender - Professor

Receiver - Hiring Manager

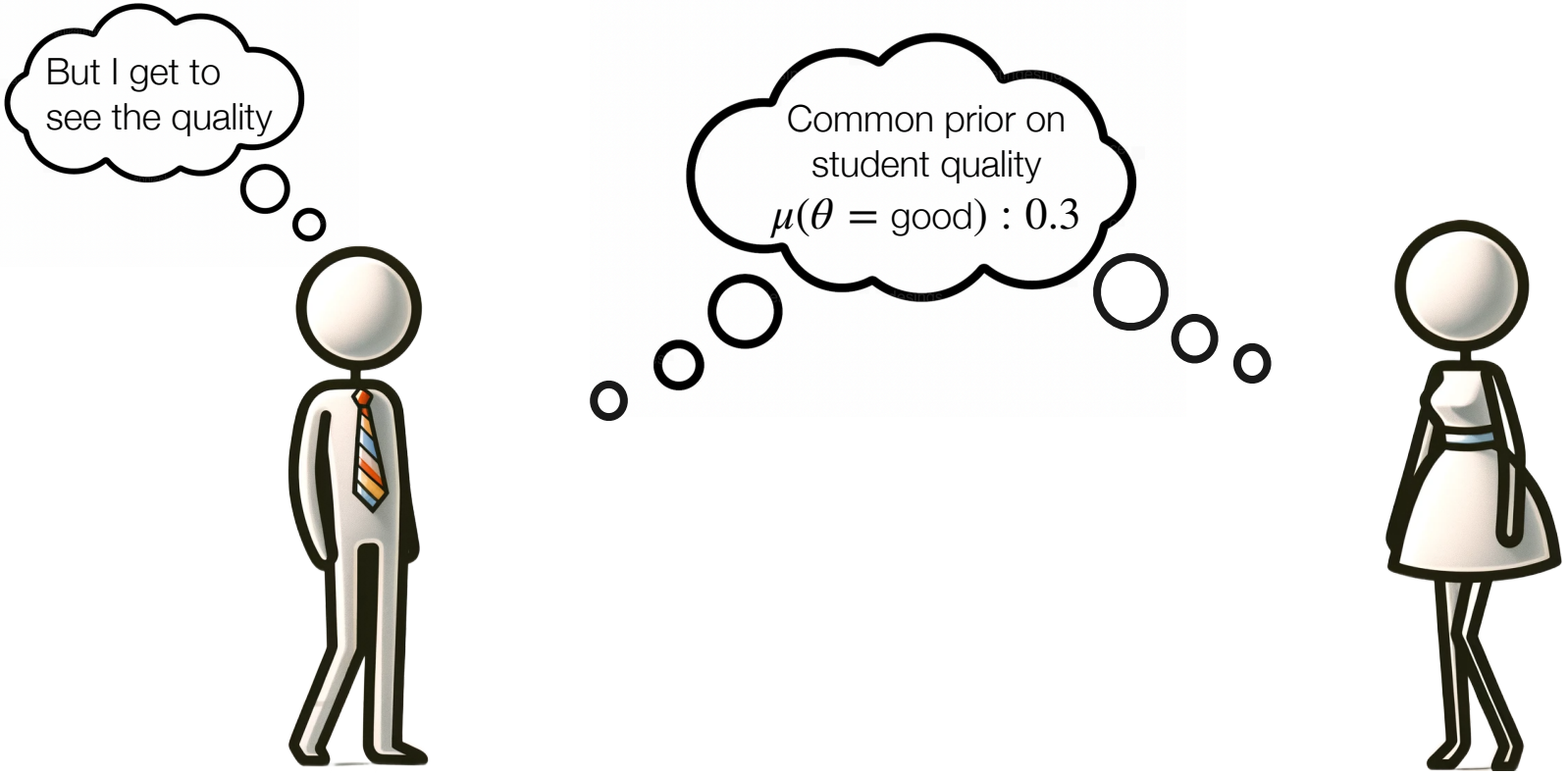


Sender - Professor

$u(a, \theta)$: +1 if student hired

Receiver - Hiring Manager

$w(a, \theta)$: +1 if right decision is made

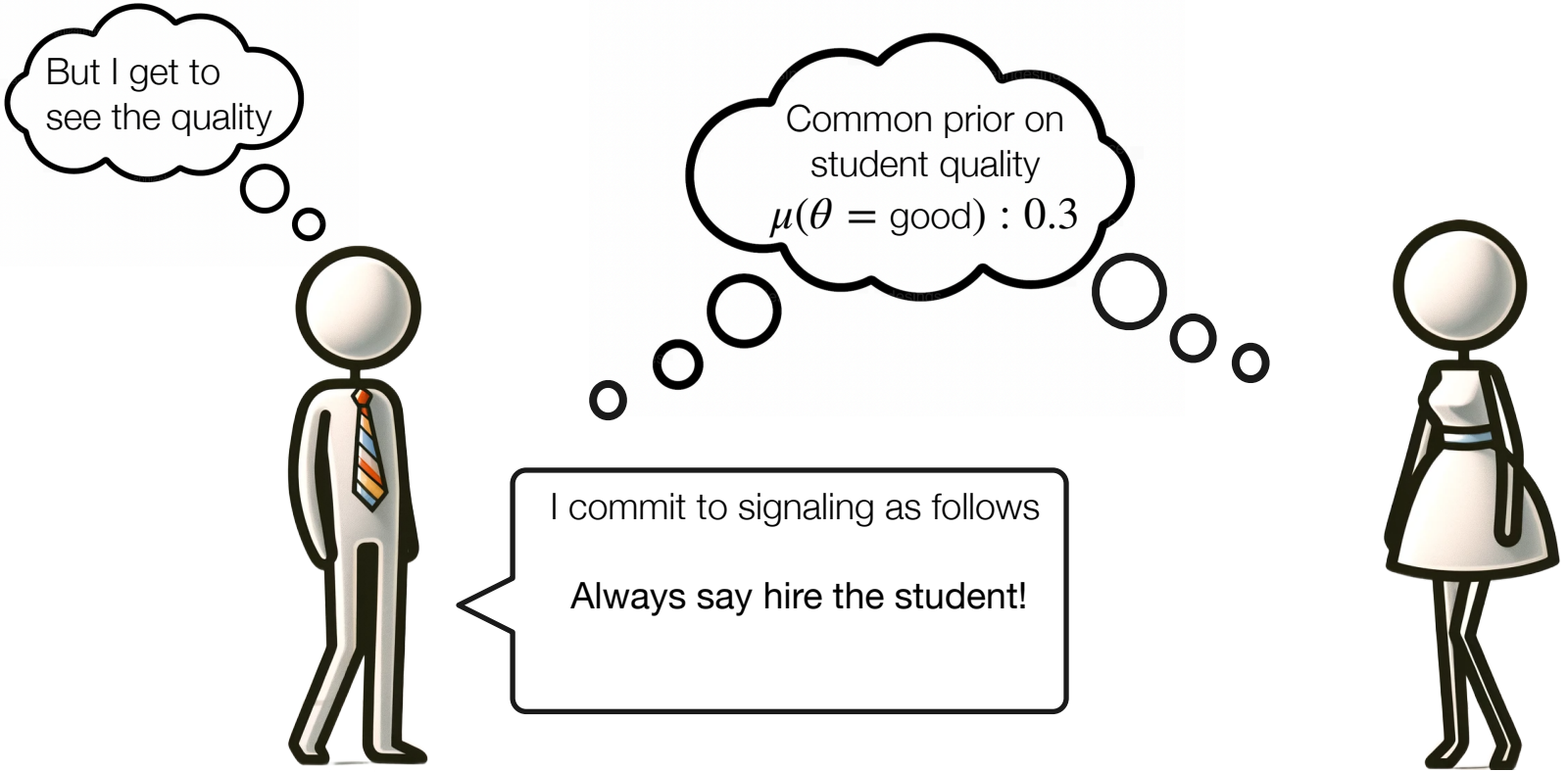


Sender - Professor

$u(a, \theta)$: +1 if student hired

Receiver - Hiring Manager

$w(a, \theta)$: +1 if right decision is made



Sender - Professor

$u(a, \theta)$: +1 if student hired

Receiver - Hiring Manager

$w(a, \theta)$: +1 if right decision is made

But I get to see the quality

Common prior on student quality

$\mu(\theta = \text{good}) : 0.3$

That's totally uninformative!

I commit to signaling as follows

Always say hire the student!

$\mathbb{E}[u(a^*, \theta)] = 0$

$a* = \text{not hire}$

$\mathbb{E}[w(a^*, \theta)] = 0.7$

Sender - Professor

$u(a, \theta)$: +1 if student hired

Receiver - Hiring Manager

$w(a, \theta)$: +1 if right decision is made

But I get to see the quality

Common prior on student quality

$\mu(\theta = \text{good}) : 0.3$

That's fully informative!

I commit to signaling as follows

$\theta = \text{good} \implies$ say hire

$\theta = \text{bad} \implies$ don't hire

$\mathbb{E}[u(a^*, \theta)] = 0.3$

$a^* =$ hire if s=hire

$a^* =$! hire if s= ! hire

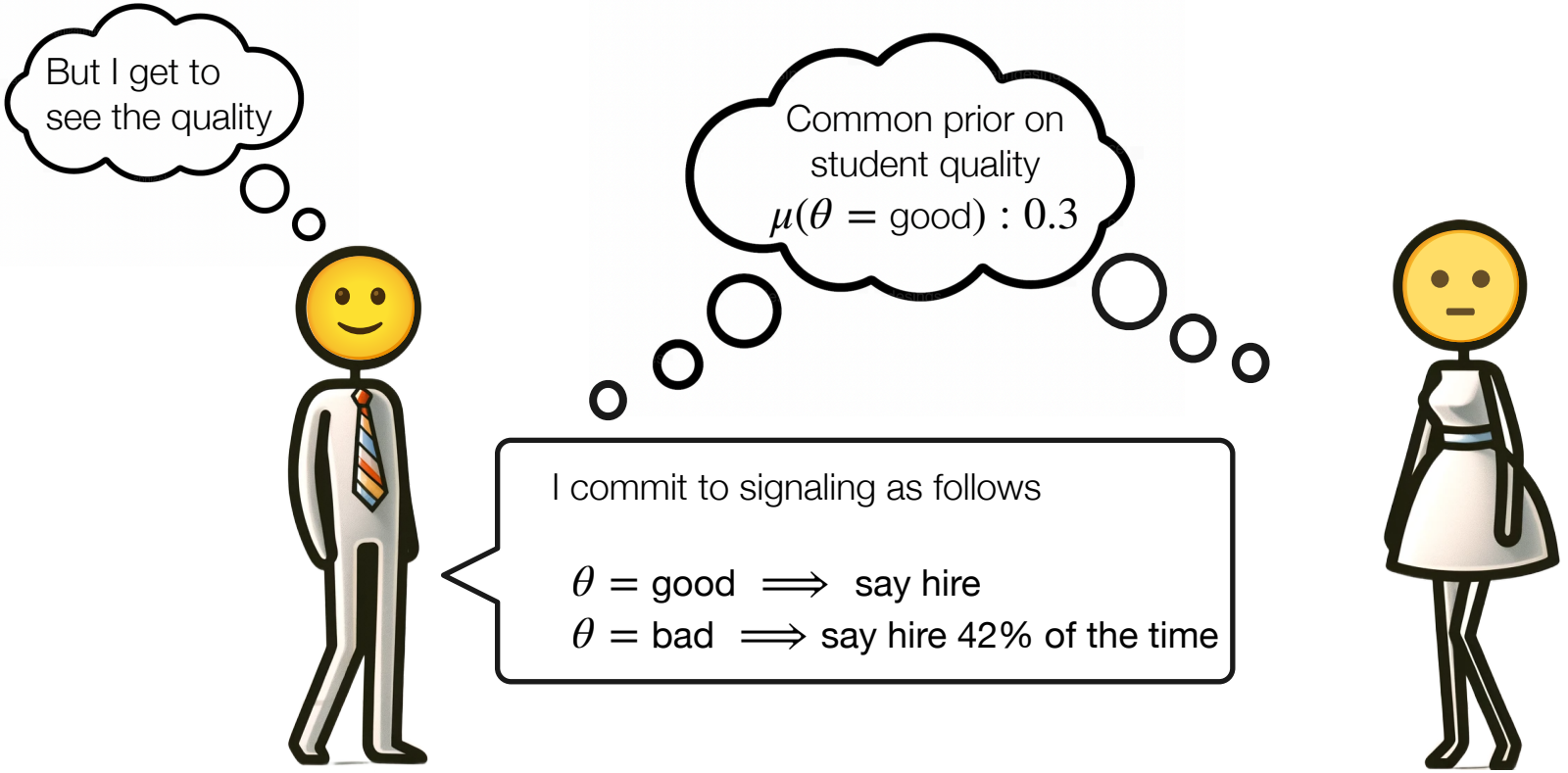
$\mathbb{E}[w(a^*, \theta)] = 1.0$

Sender - Professor

$u(a, \theta)$: +1 if student hired

Receiver - Hiring Manager

$w(a, \theta)$: +1 if right decision is made

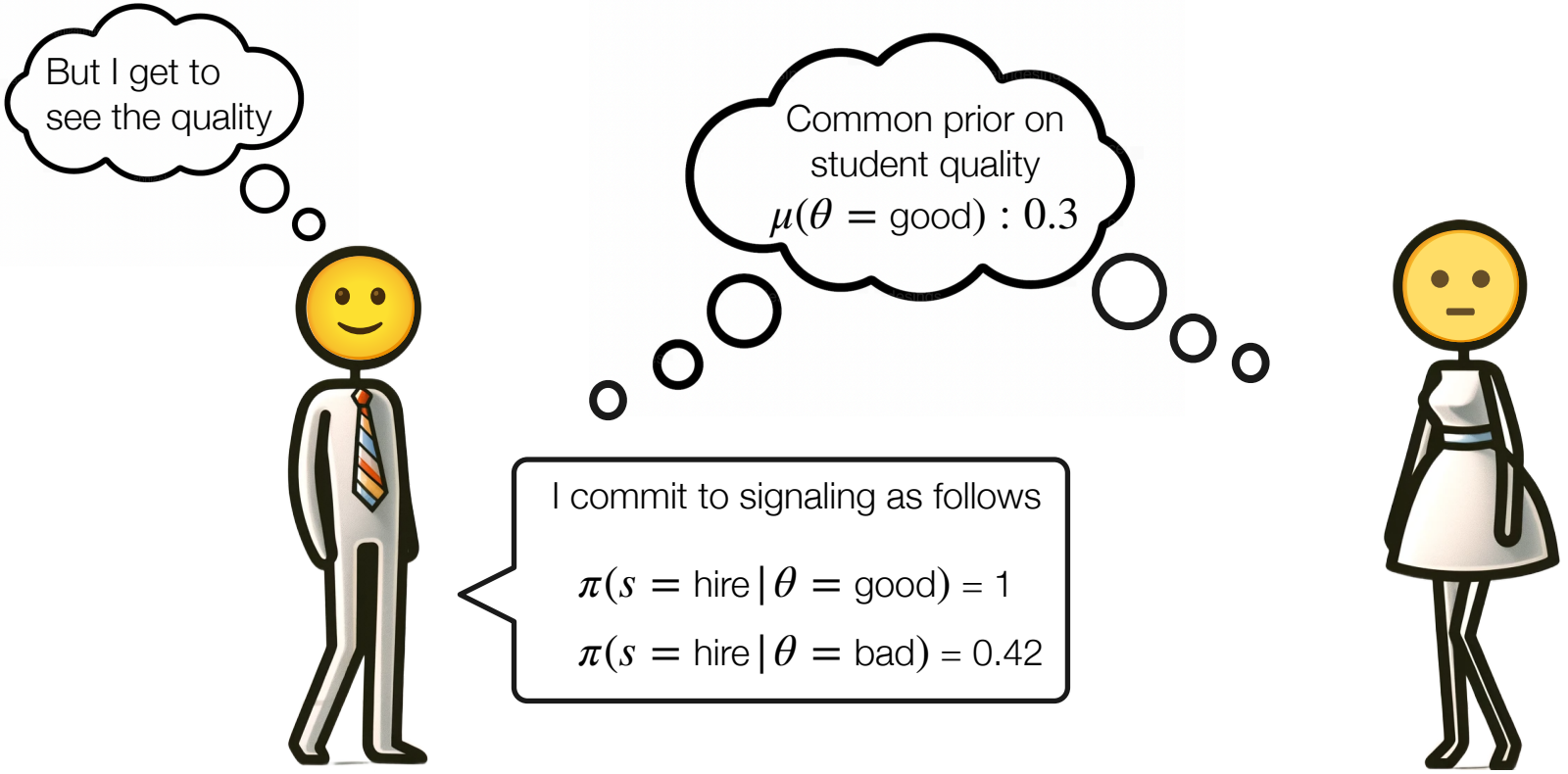


Sender - Professor

$u(a, \theta)$: +1 if student hired

Receiver - Hiring Manager

$w(a, \theta)$: +1 if right decision is made

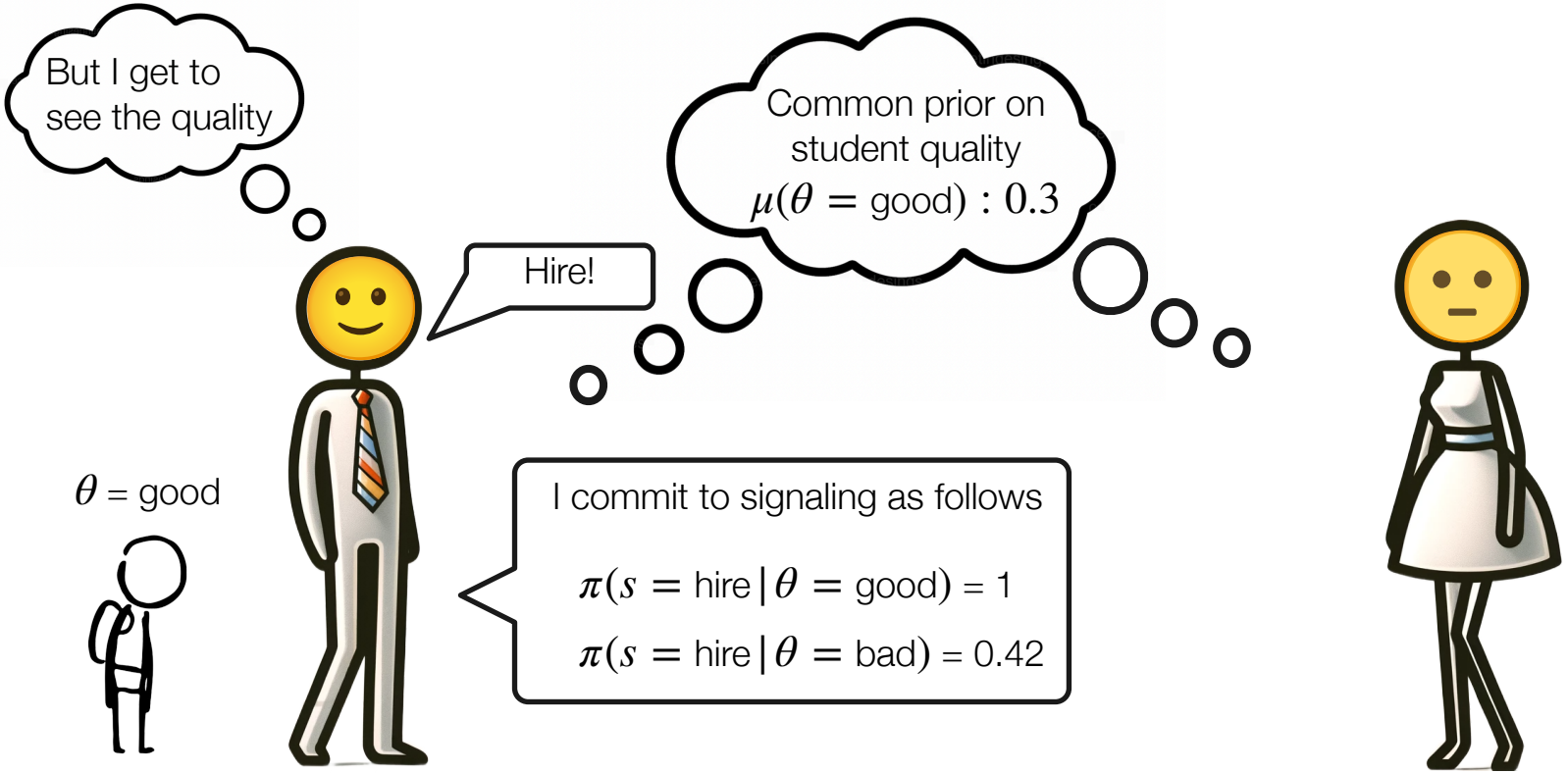


Sender - Professor

$u(a, \theta)$: +1 if student hired

Receiver - Hiring Manager

$w(a, \theta)$: +1 if right decision is made

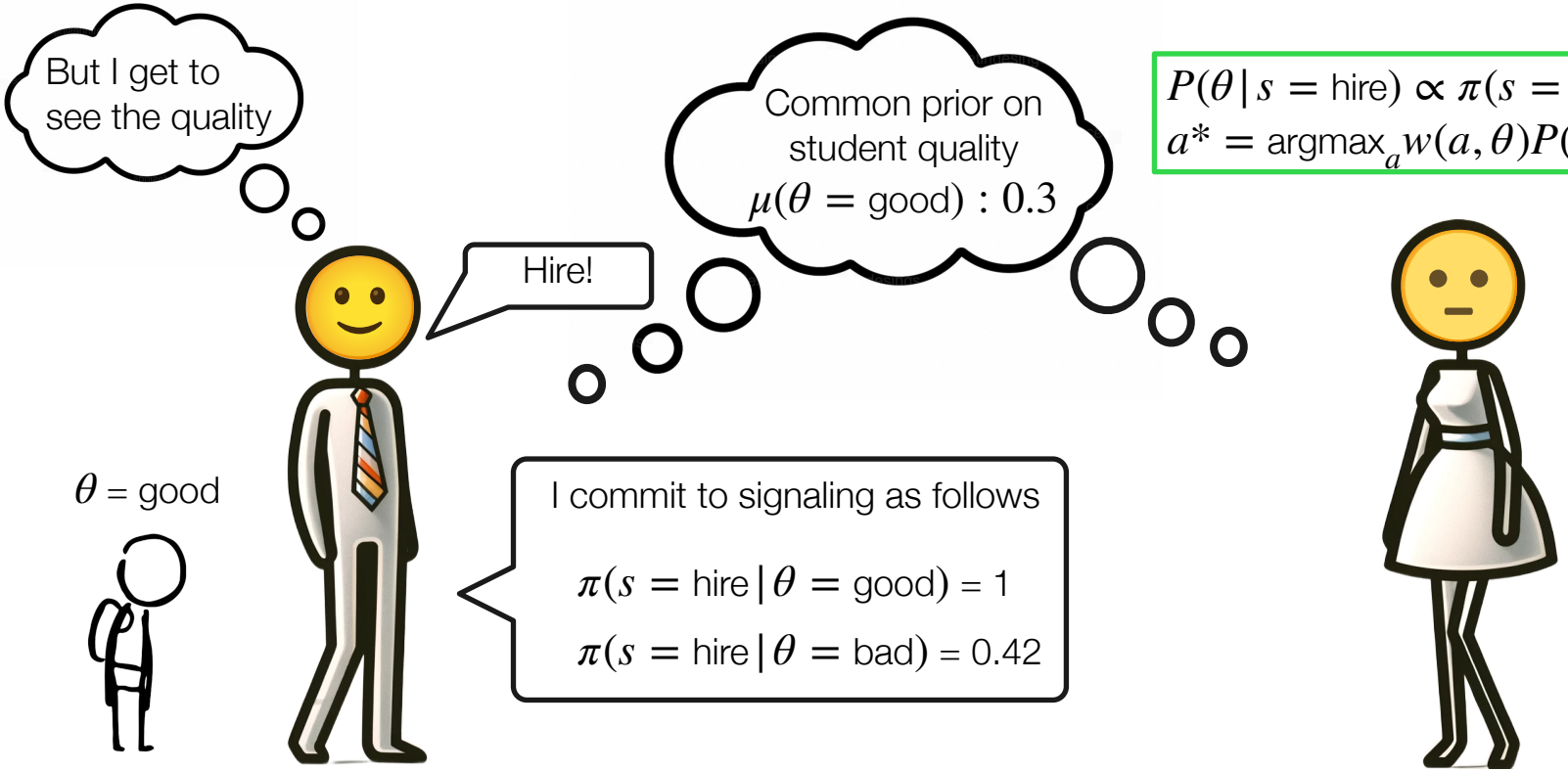


Sender - Professor

$u(a, \theta)$: +1 if student hired

Receiver - Hiring Manager

$w(a, \theta)$: +1 if right decision is made



Sender - Professor

$u(a, \theta)$: +1 if student hired

Receiver - Hiring Manager

$w(a, \theta)$: +1 if right decision is made

But I get to see the quality

Common prior on student quality

$\mu(\theta = \text{good}) : 0.3$

$P(\theta \mid s = \text{hire}) \propto \pi(s = \text{hire} \mid \theta)\mu(\theta)$

$a^* = \text{argmax}_a w(a, \theta) P(\theta \mid s = \text{hire})$

Hire!

$u(\theta = \text{good}, a^*)$

$\mathbb{E}[u(a^*, \theta)] = 0.6$

θ = good

I commit to signaling as follows

$\pi(s = \text{hire} \mid \theta = \text{good}) = 1$

$\pi(s = \text{hire} \mid \theta = \text{bad}) = 0.42$

$w(\theta = \text{good}, a^*)$

$\mathbb{E}[w(a^*, \theta)] = 0.7$

Sender - Professor

$u(a, \theta)$: +1 if student hired

Receiver - Hiring Manager

$w(a, \theta)$: +1 if right decision is made

0. Bayesian Persuasion - Details

- Sender must commit to signaling scheme before realization

- When sender is designing/choosing signaling scheme they have no more information than receiver.
    - Chooses a scheme to maximize expected ex-ante utility

- In the standard setting, under mild assumptions optimal scheme can be solved using a linear program [2].

1. Motivations

- Misinformation on social media is of enormous societal concern.
  - Platform design encourages users to seek validation - irrespective of veracity.

- Current approaches like tagging or censoring fall short.
  - Those spreading misinformation may not agree with platform's opinion on it.
  - Censorship can be abused by platforms and threaten freedom of speech.

[Pew Research]: While most Americans support tackling misinformation, more than half agreed that "freedom of information should be prioritized over … restricting false information online"

Can we convince users not to share misinformation in the first place, leveraging information they care about

Popularity of their post/ validation it will receive (which platform can estimate)

Signaling to change user's belief about their post, thus naturally altering their action

Persuasion

1. Model - Setup

- We model the interaction between a social media platform (sender) and a user (receiver) on the platform who has drafted content and is considering sharing.

- The post has a popularity feature v and misinformation feature m: $\theta = (m, v)$

- Both platform and user has a prior μ over these features based on user's past interactions and outcomes on the platform.

- User can take action: {share, not share}
  - User utility given by $w(\theta, a)$, and platform utility is $u(\theta, a)$

1. Model - Noisy Persuasion

- Platform can predict these features $\hat{\theta} = (\hat{m}, \hat{v})$ with some accuracy.
  - $\mathcal{Q}^{\theta}$ denotes the joint confusion matrix of these classifiers

- Platform can signal users conditioned on these imperfect observations.
  - $\pi(s \mid \hat{\theta}) = \pi(s \mid \hat{m}, \hat{v})$.
  - Users are Bayesian and update their optimal action accordingly.

- Platform signaling changes user behaviour and subsequently their future belief.
  - Performative model - will be expanded later.

1. Related Works

Impact of classifier accuracy on optimal utility and signaling structure?

How does signaling affect content distribution - long term effects of persuasion?

- No prior works on persuasion with noisy observations. Spiritually related:
  - [3] Robust persuasion with external signals; [4] persuasion over noisy channels.

- No prior works on persuasion from a performative lens
  - Standard performative prediction [5] results make strong technical assumptions that do not hold here.

- [6] study persuasion in social networks, but under a very different model.

1. Example

$m \in \{0/\text{True}, 1/\text{False}\}$; $v \in \{0/\text{Unpopular}, 1/\text{Popular}\}$; $a \in \{0/\text{discard}, 1/\text{share}\}$

m \ v	0	1
0	0.35	0.35
1	0.15	0.15

Prior μ

m \ v	0	1
0	-1.0	1.0
1	-1.0	1.0

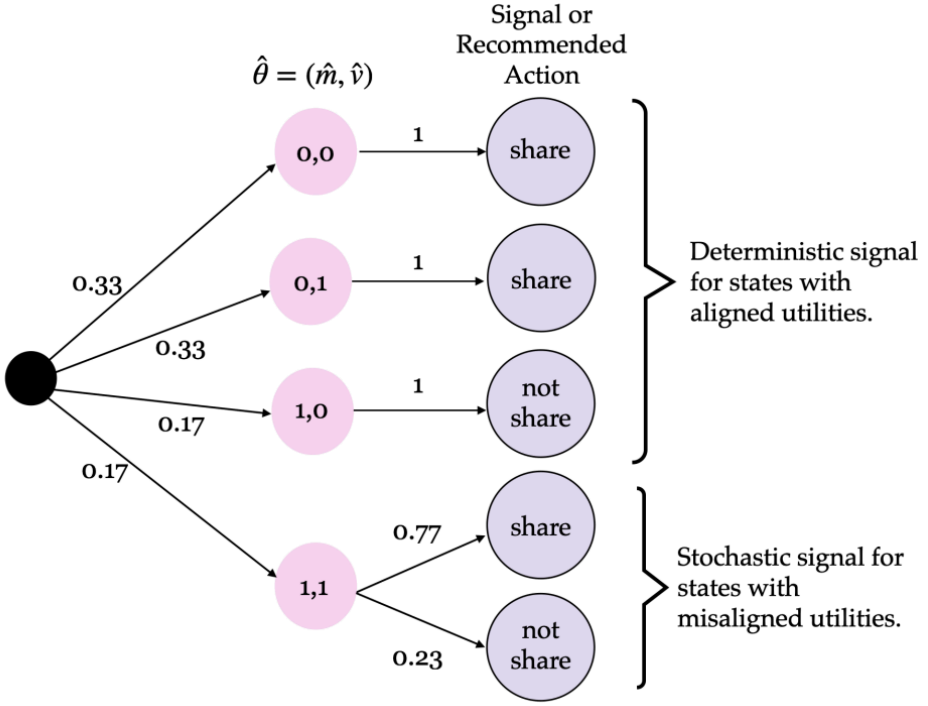
User Utility (Sharing)
$w(\theta, a = 1)$

m \ v	0	1
0	1.0	2.0
1	-1.0	**-3.0**

Platform Utility (Sharing)
$u(\theta, a = 1)$

1. Example

- Q^{Θ} / Classifier Accuracy: 90%
- Edges represent probability



Outcomes

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

1. Preliminaries (1)

- Each signal realization induces a posterior belief $\rho_s(\theta)$ onto the receiver.
- In standard BP, any set of posteriors can be induced insofar as $\Sigma P(s)\rho_s = \mu$

- Imperfect classifier limits the beliefs that can be induced onto the user.
  - Ex: If platform does not perfectly know $v = 1$, $\rho_s(v = 1) \neq 1$

- Can express effective signaling over true observations as:

$$\tilde{\pi}(s \mid \theta) = \sum_{\hat{\theta}} Q^{\Theta}_{\hat{\theta},\theta} \pi(s \mid \hat{\theta})$$

1. Preliminaries (2)

- Lemma: Similar to classic result, $|S| = |A|$ suffices to achieve optimal utility.
  - Signaling can be interpreted as action recommendation - $\pi(a \mid \hat{\theta})$

- Proposition: If receiver utility is independent of m:
  - Suffices for user to know marginal prior $\mu(v)$ and marginal scheme $\pi(a \mid \hat{v})$

- Noisy persuasion can never decrease user utility - mutually beneficial

Focusing on a single round of persuasion:

What is the optimal signaling scheme?

Is there an ordering for classifier accuracy w.r.t to optimal platform utility?

How does changing classifier accuracy affect optimal platform utility?

1. Optimal Signaling Scheme

$$\begin{aligned} \max \quad & \sum_{a_i}^{|\mathcal{A}|} \sum_{\theta} u(a_i, \theta) \mu(\theta) \tilde{\pi}(s = a_i | \theta) \\ \text{s.t.} \quad & \sum_{\theta} \Delta w_{ij}(\theta) \mu(\theta) \tilde{\pi}(s = a_i | \theta) \geq 0 \qquad \forall a_i, a_j \end{aligned}$$

$$\tilde{\pi}(s = a_i | \theta) = \sum_{\widehat{\theta}} \pi(s = a_i | \widehat{\theta}) Q^{\Theta}_{\widehat{\theta}, \theta} \qquad \forall\, a_i, \theta$$

$$\begin{aligned} \sum_{a_i} \pi(s = a_i | \widehat{\theta}) &= 1 && \forall \widehat{\theta} \\ \pi(s = a_i | \widehat{\theta}) &\geq 0 && \forall\, a_i, \widehat{\theta} \end{aligned}$$

- Standard persuasion LP with IC constraints with effective signaling
- Constrain effective signaling space based on accuracy of observations.
- Simplex constraints on signaling

Noisy Persuasion can be interpreted as optimizing the same objective but under a more restricted feasible region.

1. Classifier Ordering

- Clearly, a “better” classifier would lead to higher optimal utility under persuasion.
  - But what notion of “better”: Entropy, Precision, Recall, F1 score?
  - Since classification accuracy can be improved, this is also operational important

Theorem: For symmetric confusion matrices Q_1^Θ, Q_2^Θ*, optimal utility from signaling* $u_I^*(Q_2^\Theta) \geq u_I^*(Q_1^\Theta)$ *if and only if convexHull(rows of* Q_1^Θ*)* $\subseteq$ *convexHull(rows of* Q_2^Θ*)*

1. Classifier Ordering

Theorem: *For symmetric confusion matrices* Q_1^Θ, Q_2^Θ*, optimal utility from signaling* $u_I^*(Q_2^\Theta) \geq u_I^*(Q_1^\Theta)$ *if and only if convexHull(rows of* Q_1^Θ*)* $\subseteq$ *convexHull(rows of* Q_2^Θ*).*

$\Longrightarrow$ Any feasible effective signaling scheme under Q_1^Θ is also feasible under Q_2^Θ

$\Longleftarrow$ If $\exists$ row i of $Q_1^\Theta \notin$ convexHull(rows of Q_2^Θ), an instance exists wherein optimal utility under Q_2^Θ is strictly better than optimal utility under Q_1^Θ.

1. Geometric Intuition

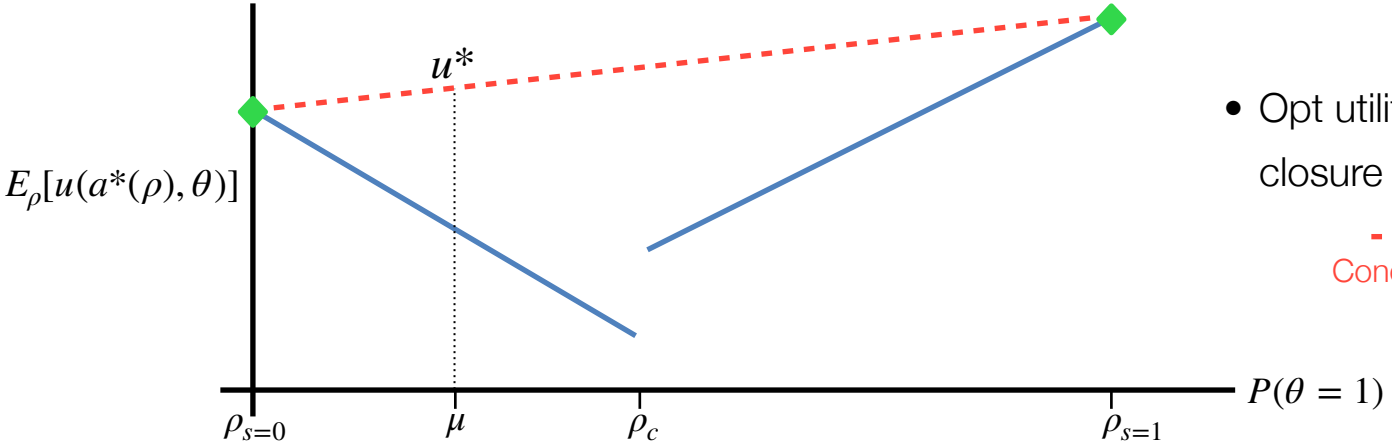
- Understand platform and user utility at any belief $\rho(\theta)$.
  - Belief space is the $\Delta^{|\Theta|}$ simplex.

- $\mathbb{E}_{\theta\sim\rho}[w(a,\theta)]$: user's expected utility (w.r.t ρ) for taking action a. It is a linear function of ρ.
- At any belief ρ, there is an optimal action $a^* = \text{argmax}_a \mathbb{E}_{\theta\sim\rho}[w(a,\theta)]$
  - ρ_c: belief where user has multiple optimal action; threshold where optimal action changes

- Similarly, $\mathbb{E}_{\theta\sim\rho}[u(a^*,\theta)]$ is the platform expected utility for the user taking optimal action a^*.
  - This is a piece-wise linear function.

1. Geometric Intuition

- Let $|\Theta| = 2$, so the belief ρ can be represented on the line capturing $P(\theta = 1)$
- Plot the platform expected utility at ρ under user's optimal action: $E_{\rho(\theta)}[u(a^*(\rho), \theta)]$
- Signaling induces a set of beliefs $\{\rho_s\}$ such that $\Sigma P(s)\rho_s = \mu$

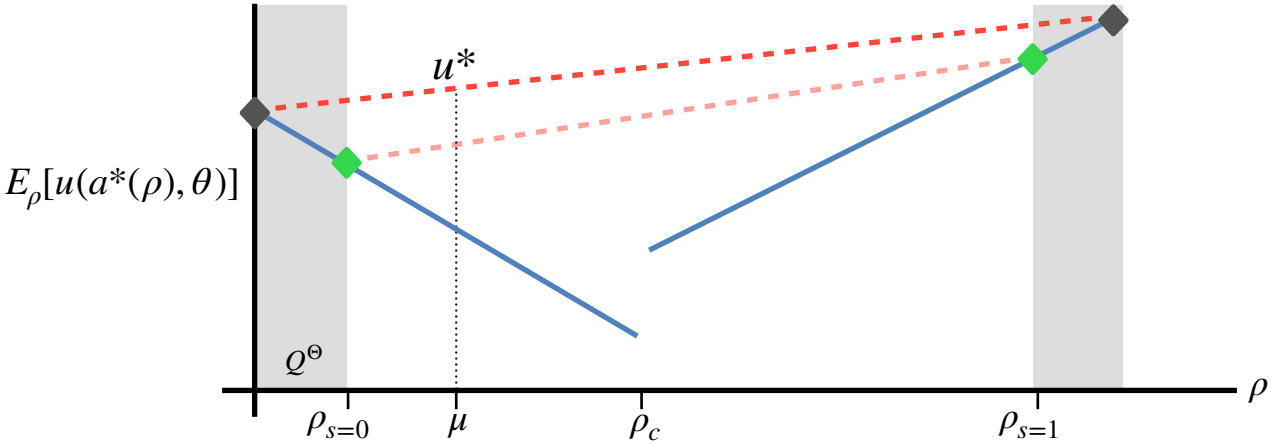


- Opt utility is the concave closure value at prior μ

Concave closure

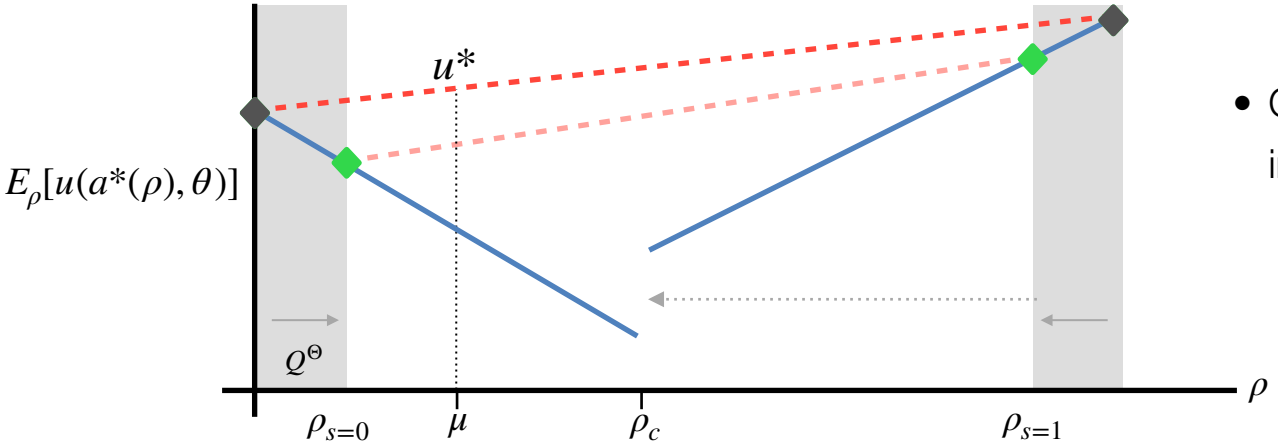
1. Geometric Noisy Persuasion

- Noisy persuasion restricts the space of inducible posteriors.
  - Can change the concave closure and thus the optimal utility.



1. Continuity

- Theorem: The optimal platform utility is continuous in Q^Θ except possibly when there exists a $\hat{\rho}$ on the $\Delta^{|\Theta|}$ simplex boundary such that $V\hat{\rho} = \rho_c$.
  - V captures $P(\theta \mid \hat{\theta})$ and is easily computable from Q^Θ and μ. It is invertible under light assumptions.*



- Confusion matrix restricts inducible posteriors.

Concave closure

What is the long term effect of applying persuasion?

1. Performative Model

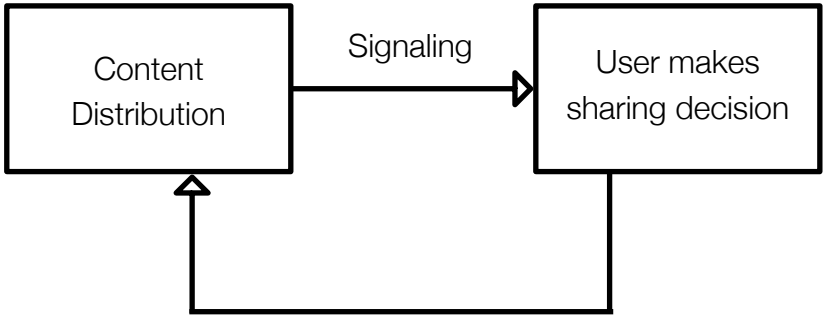
- Over time, the content distribution of a user skews toward the content they shared.

- Without persuasion, user's take their optimal action for the drafted content based on their prior. We naturally assume this action to be "share".
    - Distribution of content shared on platform remains the same as the prior.

- Applying persuasion means user actions changes based on signal.
    - Shifts the content distribution

- Platform must update their signaling scheme due to this changing belief.

2.1 Performative Model (2)



- For an instance $I = (u, w, \mu_0)$ and joint classifier confusion matrix Q^{Θ}.
  - At round t with prior μ_t, the platform chooses optimal signaling scheme $\pi_t(a \mid \hat{\theta})$.
  - User takes their optimal action based on realized signal.
  - Next round distribution is $\mu_{t+1} = \lambda\mu_t + (1-\lambda)\rho(\theta \mid a = 1)$

2.1 Performative Questions

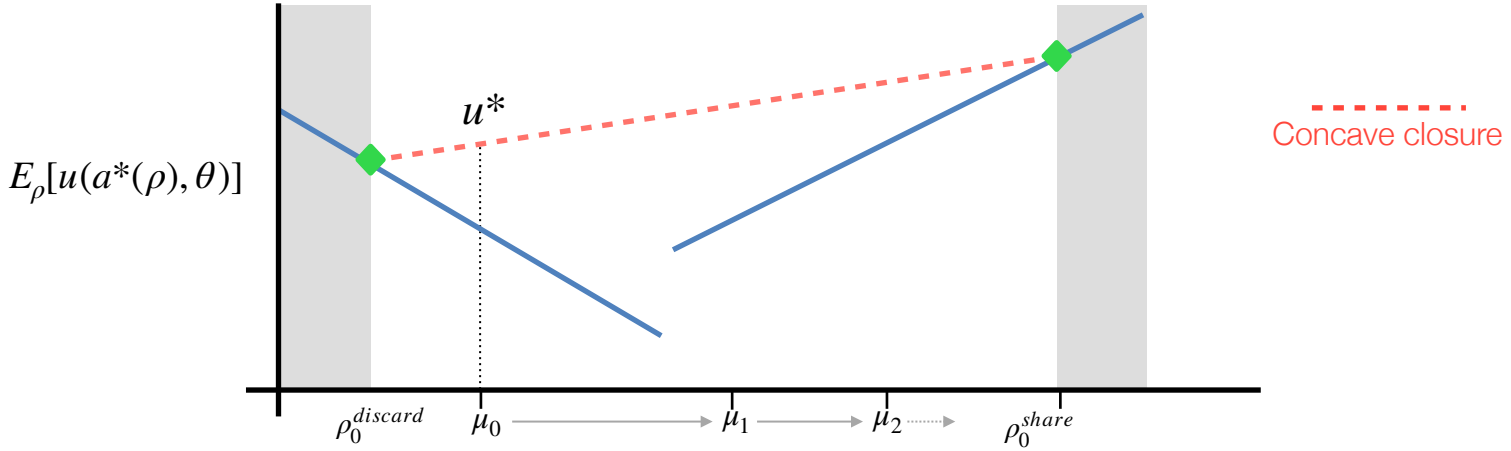
- The performative process converges to distribution μ^* if for any $\epsilon > 0$, there exists a T_c such that for $t > T_c$, $||\mu^* - \mu_t|| \leq \epsilon$.

- A distribution μ^s is stable if for some t, $\mu_t = \mu^s \implies \mu_{t+1} = \mu^s$.

What are the convergence and stability properties of this performative process?

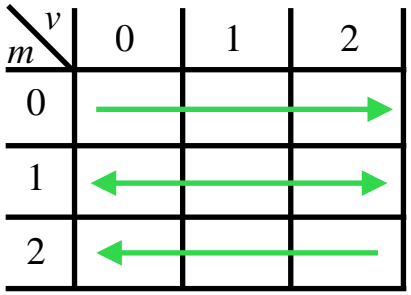
2.1 Performative Convergence

- Theorem: For $\lambda \neq 0$, the performative process converges to the first round's optimal "share" posterior.
  - Key: Induced posteriors from optimal signaling remain the same despite changing priors.
  - Implies: Platform utility due to signaling is monotonically increasing each round.
  - Convergent distribution point is also stable

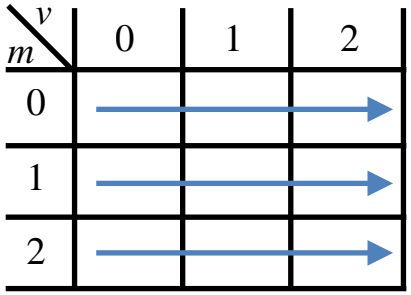


2.1 Experiments - Setup

- Lack of publicly available user data; led to synthetic dataset.
- 3 misinformation and validation states: 0 = unpopular/true; 2 = popular/false; 1= in between



Platform Utility (Sharing)



User Utility (Sharing)

- Utilities are samples randomly
- User is indifferent to misinfo

What is the reduction in misinformation due to signaling with noisy classifier?
How does misinformation at performative stable point compare to original prior?

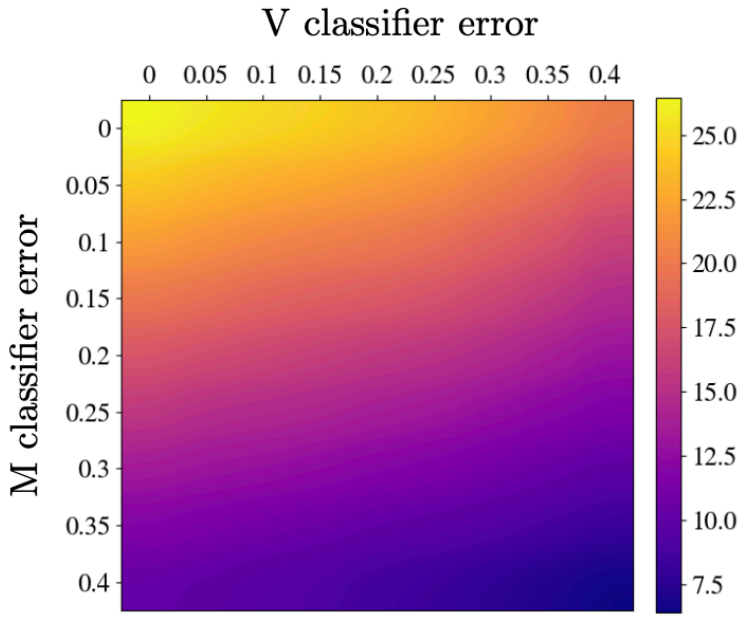


Figure 2: Avg % decrease in misinformation shared due to single application of noisy persuasion.

- Average of 100 random instances plotted.
- 90% confidence interval of 4%

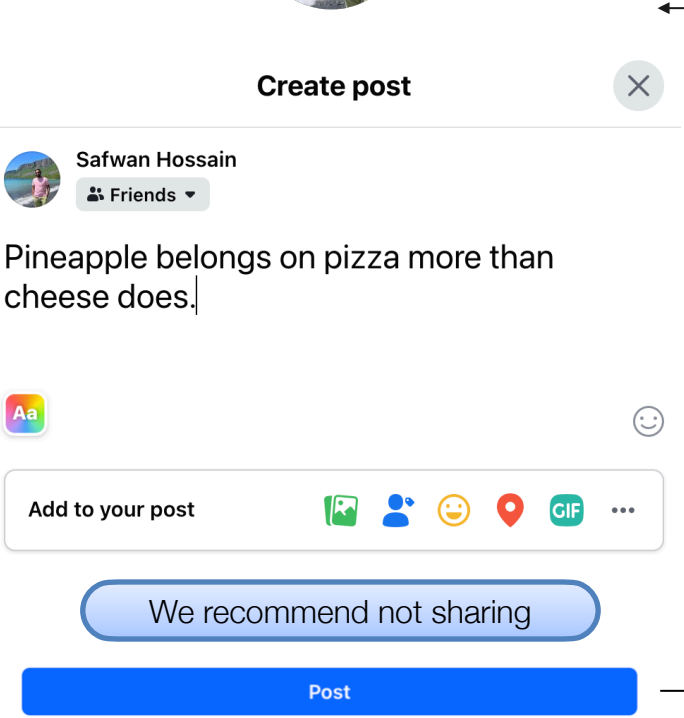
Facebook can predict popularity and misinformation with 80% accuracy

For content predicted to be popular, we will recommend sharing 70% of the time. For those predicted unpopular, 20%.

Q^{θ}

$\pi_t(s \mid \hat{\theta})$

$\sim s$



Thank you!

References:

[1] Matthew Gentzkow, Emir Kamenica. **Bayesian Persuasion**. American Economic Review. 2011

[2] Shaddin Dughmi, Haifeng Xu. **Algorithmic Bayesian persuasion**. ACM symposium on Theory of Computing (STOC). 2016

[3] Piotr Dworczak, Alessandro Pavan. **Preparing for the worst but hoping for the best: Robust (Bayesian) persuasion.** Econometrica. 2022

[4] Tsakas, Elias, Nikolas Tsakas. **Noisy persuasion.** Games and Economic Behavior. 2021

[5]: Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, Moritz Hardt. **Performative Prediction.** ICML. 2020

[6] Candogan, Ozan, Kimon Drakopoulos. **Optimal signaling of content accuracy: Engagement vs. misinformation.** Operations Research. 2020