Strategically Controlling Worldviews

Cuimin Ba (Ts'uei-min Ba)

joint with Danil Dmitriev, Freddie Papazyan, Victoria Hang

Nov 18, 2025

The act of persuasion sits at the core of human interaction

political campaign

advertising

court argument ...

The act of persuasion sits at the core of human interaction

political campaign

advertising

court argument ...

Persuaders selectively share information and shape how people make sense of them

Economic theory studies persuasive behavior through **two distinct approaches**

Economic theory studies persuasive behavior through **two distinct approaches**

Information-based persuasion

Bayesian persuasion

- Kamenica & Gentzkow (2011)

Cheap talk

- Crawford & Sobel (1982)

Verifiable disclosure

- Grossman & Hart (1980)

Economic theory studies persuasive behavior through **two distinct approaches**

Information-based persuasion

Bayesian persuasion

- Kamenica & Gentzkow (2011)

Cheap talk

- Crawford & Sobel (1982)

Verifiable disclosure

- Grossman & Hart (1980)

Model-based persuasion

Sender persuades by proposing alternative models/narratives/interpretations/worldviews

- Schwartzstein & Sunderam (2021)
- Barron & Fries (2024)
- Aina (2025) ...

Economic theory studies persuasive behavior through **two distinct approaches**

Information-based persuasion

Bayesian persuasion

- Kamenica & Gentzkow (2011)

Cheap talk

- Crawford & Sobel (1982)

Verifiable disclosure

- Grossman & Hart (1980)

Model-based persuasion

Sender persuades by proposing alternative models/narratives/interpretations/worldviews

- Schwartzstein & Sunderam (2021)
- Barron & Fries (2024)
- Aina (2025) ...

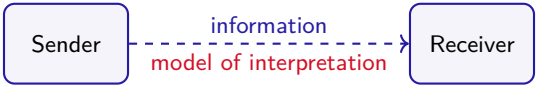
No unifying framework to study how persuaders **combine information and narratives**

Questions

- How does a persuader optimally combine both persuasion techniques?
- Do they interact to deliver outcomes neither could achieve alone?
- How does the answer depend on the information technology available?

Framework

We study a strategic communication game that allows for model switching



The receiver can update not only beliefs but also model of interpreting data

We show that...

Despite that:

- Receiver starts correctly specified and Bayesian updates within models
- Receiver perfectly understands Sender's strategic motives
- Sender faces information constraints, e.g., hard information only or no commitment power

In many cases, Sender **always implements their preferred action**

- **Strategic complementarity** between controlling info. and how it is interpreted

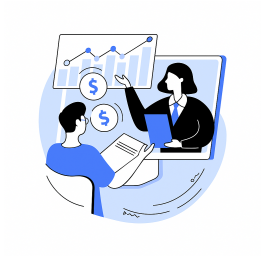
Implication

Results help understand **the Disinformation Puzzle** (Stiglitz and Kosenko, 2024)

> *"But, apart from a very limited literature within economics on fraud, little attention has been paid to concerns about mis and disinformation. This is perhaps **not surprising**, given economists' predilection for rationality and rational expectations. Indeed, the success of mis and **disinformation represents a puzzle** for standard economics, which begins by assuming individual rationality, including an individual's ability to rationally evaluate the accuracy of information and update priors, using Bayes' theorem.*
>
> *...presumably, information from a source that repeatedly provided mis and disinformation would lose credibility and therefore would play no role in decision making - and so would not be a problem...*
>
> *Not surprisingly, much of this dis/misinformation **goes beyond the Bayesian framework** that has been central to the analysis so far."*

Illustrative Example

Invest or not invest?



- A financial advisor wants a client to invest in a risky asset
- The asset pays $+1$ in state G and -2 in state B
- Hence investment is optimal if $\Pr(G) \geq 2/3$
- Advisor observes the realized state but Client doesn't
- Assuming a uniform distribution, Client will **not** invest

Can Providing Information Help?

- It depends on the info technology available

Can Providing Information Help?

- It depends on the info technology available
- Consider the **best case for Advisor**: she can send cheap-talk messages and has commitment power

Can Providing Information Help?

- It depends on the info technology available
- Consider the **best case for Advisor**: she can send cheap-talk messages and has commitment power
- Canonical Bayesian persuasion (BP) solution:

	state G	state B
send	g	g with prob. $\frac{1}{2}$ b with prob. $\frac{1}{2}$

- Investment happens with prob. $3/4$

Can Providing a Model Help?

- Suppose Advisor sticks to this messaging strategy:

	state G	state B
send	g	g with prob. $\frac{1}{2}$ b with prob. $\frac{1}{2}$

- Client also initially holds this correct model

Can Providing a Model Help?

- Suppose Advisor sticks to this messaging strategy:

	state G	state B
send	g	g with prob. $\frac{1}{2}$ b with prob. $\frac{1}{2}$

- Client also initially holds this correct model
- But before sending messages, Advisor can propose an alternative model to interpret the messages
- Client will switch if it fits the data **twice as well**

Can Providing a Model Help?

- Suppose Advisor sticks to this messaging strategy:

	state G	state B
send	g	g with prob. $\frac{1}{2}$ b with prob. $\frac{1}{2}$

- Client also initially holds this correct model
- But before sending messages, Advisor can propose an alternative model to interpret the messages
- Client will switch if it fits the data **twice as well**
- Should Advisor provide a different interpretation?

Can Providing a Model Help?

- Suppose Advisor sticks to this messaging strategy:

	state G	state B
send	g	g with prob. $\frac{1}{2}$ b with prob. $\frac{1}{2}$

- Client also initially holds this correct model
- But before sending messages, Advisor can propose an alternative model to interpret the messages
- Client will switch if it fits the data **twice as well**
- Should Advisor provide a different interpretation? **Yes**

Can Providing a Model Help?

- Client initially holds the correct model (θ_0):

	state G	state B
send	g	g with prob. $\frac{1}{2}$ b with prob. $\frac{1}{2}$

Can Providing a Model Help?

- Client initially holds the correct model (θ_0):

	state G	state B
send	g	g with prob. $\frac{1}{2}$ b with prob. $\frac{1}{2}$

- Suppose Advisor proposes this alternative model (θ_1):

	state G	state B
send	b	g

Can Providing a Model Help?

- Client initially holds the correct model (θ_0):

	state G	state B
send	g	g with prob. $\frac{1}{2}$ b with prob. $\frac{1}{2}$

- Suppose Advisor proposes this alternative model (θ_1):

	state G	state B
send	b	g

Can Providing a Model Help?

- Client initially holds the correct model (θ_0):

	state G	state B
send	g	g with prob. $\frac{1}{2}$ b with prob. $\frac{1}{2}$

- Suppose Advisor proposes this alternative model (θ_1):

	state G	state B
send	b	g

- When Advisor sends b, Client switches to θ_1 since $\Pr(b|\theta_1) = 1/2 \geq 2 \times 1/4 = 2\times \Pr(b|\theta_0)$, and she **invests**

Can Providing a Model Help?

- Client initially holds the correct model (θ_0):

	state G	state B
send	g	g with prob. $\frac{1}{2}$ b with prob. $\frac{1}{2}$

- Suppose Advisor proposes this alternative model (θ_1):

	state G	state B
send	b	g

- When Advisor sends b, Client switches to θ_1 since $\Pr(b|\theta_1) = 1/2 \geq 2 \times 1/4 = 2\times \Pr(b|\theta_0)$, and she **invests**
- When Advisor sends g, Client sticks to θ_0 since $\Pr(g|\theta_1) = 1/2 < 2 \times 3/4 = 2\times \Pr(g|\theta_0)$, *but* she still **invests**

Lessons

Advisor ensures the Client invests with prob. 1

Full manipulation achieved by jointly designing information structure and model of interpretation

Strategy works even if Advisor is a cheap talker (w/o commitment power)

Lessons

Advisor ensures the Client invests with prob. 1

Full manipulation achieved by jointly designing information structure and model of interpretation

Strategy works even if Advisor is a cheap talker (w/o commitment power)

Yet, the strategy construction shown in the example is not generally optimal

Lessons

Advisor ensures the Client invests with prob. 1

Full manipulation achieved by jointly designing information structure and model of interpretation

Strategy works even if Advisor is a cheap talker (w/o commitment power)

Yet, the strategy construction shown in the example is not generally optimal

Main result

Simple characterization of the sender's optimal info + model strategy in general environments

Framework

Setup

- Two players: a sender (S, she) and a receiver (R, he)
- R chooses between a risky action ($a = 1$) and a safe action ($a = 0$)
- The safe action pays $\omega_0 \in \mathbb{R}$
- The risky action's payoff $=$ state of the world $\omega \in \Omega \subset \mathbb{R}$

Setup

- Two players: a sender (S, she) and a receiver (R, he)
- R chooses between a risky action ($a = 1$) and a safe action ($a = 0$)
- The safe action pays $\omega_0 \in \mathbb{R}$
- The risky action's payoff $=$ state of the world $\omega \in \Omega \subset \mathbb{R}$
- Nature draws ω from a full-support common prior F
- Assume $\mathbb{E}(\omega) < \omega_0$ s.t. the optimal action is $a = 0$ without any information
- State space Ω contains values $\underline{\omega}$ and $\overline{\omega}$ such that with $\underline{\omega} < \omega_0 < \overline{\omega}$

Providing information

- S has state-independent preference: she gets 1 whenever $a = 1$ and 0 otherwise
- S privately observes ω and then sends R a message $m \in M(\omega) \subseteq M$
  - Assume M is sufficiently rich
  - This general setup allows us to accommodate a variety of environments
  - Information is unverifiable if $M(\omega) \equiv M, \forall \omega \in \Omega$
  - Information is perfectly verifiable if $M(\omega) = \{\omega, \varnothing\}$, where $\varnothing$ is a null message
- Denote S's chosen **messaging strategy** as $\sigma : \Omega \to \Delta M$ and all feasible policies as Σ

Providing a model

- R initially holds a correctly specified model of S, denoted as θ_0
  - He understands S's incentives and decision environment
  - If R keeps this model, he correctly anticipates S's true messaging strategy in equilibrium

Providing a model

- R initially holds a correctly specified model of S, denoted as θ_0
  - He understands S's incentives and decision environment
  - If R keeps this model, he correctly anticipates S's true messaging strategy in equilibrium
- S may propose an alternative model about how she communicates, denoted as θ_1
  - S proposes a behavioral model that she will use messaging strategy $\sigma' \in \Sigma$
  - We assume S proposes the model **before** ω realizes to shut down any signaling effect

Model-switching rule

Given m, R chooses which model to believe in using a **Bayes factor** rule (Ba, 2025):

- Let $\alpha \geq 1$ be a switching threshold. R calculates

$$\text{Bayes factor } \lambda(m) \coloneqq \frac{\Pr(m|\theta_1)}{\Pr(m|\theta_0)} = \frac{\int_\Omega \sigma'(m|\omega)dF}{\int_\Omega \sigma(m|\omega)dF}$$

- If $\lambda(m) \geq \alpha$, R switches to θ_1; otherwise, R sticks to θ_0
- Threshold α serves as a measure of R's resistance

Timeline

1. S **proposes model** $\theta_1 = \sigma'$
2. Nature draws ω and S observes its realization
3. S **sends message** m according to messaging strategy σ
4. Upon receiving m, R **decides whether to switch** to θ_1
5. R **updates belief** about ω based on chosen model and chooses an action

Timeline

1. S **proposes model** $\theta_1 = \sigma'$
2. Nature draws ω and S observes its realization
3. S **sends message** m according to messaging strategy σ
4. Upon receiving m, R **decides whether to switch** to θ_1
5. R **updates belief** about ω based on chosen model and chooses an action

- We leave unspecified yet if S has commitment power in messaging

Results

Preview of results

- Theorem 1: conditions for **full manipulation** with **unverifiable information**
- Propositions 1 & 2: **Cheap talk** & **Bayesian persuasion**

- Work in progress: extending Theorem 1 to allow constraints in info technologies
- Proposition 3: **Perfectly verifiable information**

Understand the demand and budget for a model switch

Absent model switching, strategy σ gives rise to winning messages and losing messages:

$$\text{winning} \quad M_+(\sigma) = \{m \in M : \mathbb{E}(\omega|\sigma, m) \geq \omega_0\}$$
$$\text{losing} \quad M_-(\sigma) = \{m \in M : \mathbb{E}(\omega|\sigma, m) < \omega_0\}$$

Understand the demand and budget for a model switch

Absent model switching, strategy σ gives rise to winning messages and losing messages:

$$\begin{aligned} \text{winning} \quad & M_+(\sigma) = \{m \in M : \mathbb{E}(\omega|\sigma, m) \geq \omega_0\} \\ \text{losing} \quad & M_-(\sigma) = \{m \in M : \mathbb{E}(\omega|\sigma, m) < \omega_0\} \end{aligned}$$

Total prob. of losing under strategy σ:

$$D(\sigma) = \int_\Omega \sigma(M_-(\sigma)|\omega) dF$$

Understand the demand and budget for a model switch

Absent model switching, strategy σ gives rise to winning messages and losing messages:

$$\begin{aligned} \text{winning} \quad & M_+(\sigma) = \{m \in M : \mathbb{E}(\omega|\sigma, m) \geq \omega_0\} \\ \text{losing} \quad & M_-(\sigma) = \{m \in M : \mathbb{E}(\omega|\sigma, m) < \omega_0\} \end{aligned}$$

Total prob. of losing under strategy σ (call this Demand for model persuasion):

$$D(\sigma) = \int_\Omega \sigma(M_-(\sigma)|\omega) dF$$

Understand the demand and budget for a model switch

Absent model switching, strategy σ gives rise to winning messages and losing messages:

$$\begin{aligned} \text{winning} \quad & M_+(\sigma) = \{m \in M : \mathbb{E}(\omega|\sigma, m) \geq \omega_0\} \\ \text{losing} \quad & M_-(\sigma) = \{m \in M : \mathbb{E}(\omega|\sigma, m) < \omega_0\} \end{aligned}$$

Total prob. of losing under strategy σ (call this Demand for model persuasion):

$$D(\sigma) = \int_\Omega \sigma(M_-(\sigma)|\omega) dF$$

Total prob. of winning under proposed strategy σ':

$$B(\sigma') = \int_\Omega \sigma'(M_+(\sigma')|\omega) dF$$

Understand the demand and budget for a model switch

Absent model switching, strategy σ gives rise to winning messages and losing messages:

$$\begin{aligned} \text{winning} \quad & M_+(\sigma) = \{m \in M : \mathbb{E}(\omega|\sigma, m) \geq \omega_0\} \\ \text{losing} \quad & M_-(\sigma) = \{m \in M : \mathbb{E}(\omega|\sigma, m) < \omega_0\} \end{aligned}$$

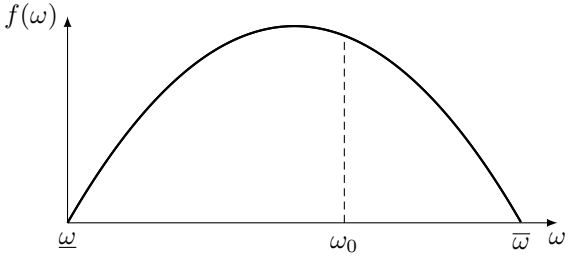
Total prob. of losing under strategy σ (call this Demand for model persuasion):

$$D(\sigma) = \int_\Omega \sigma(M_-(\sigma)|\omega) dF$$

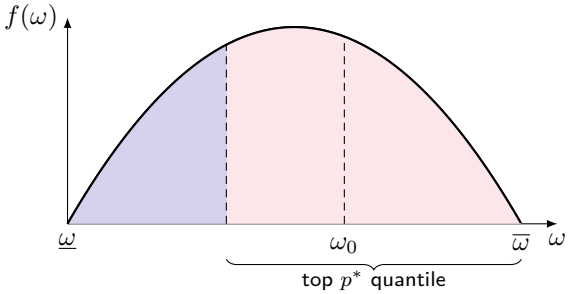
Total prob. of winning under proposed strategy σ' (call this Budget for model persuasion):

$$B(\sigma') = \int_\Omega \sigma'(M_+(\sigma')|\omega) dF$$

Example



Example



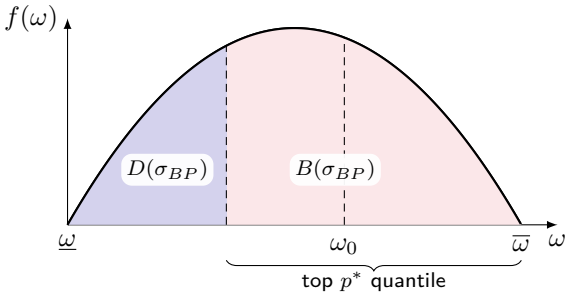
Suppose $\sigma = \sigma' = \sigma_{BP}$:

Let $\hat{\omega}(p) :=$ mean of top p quantile

Find $p^* \in (0,1)$ s.t. $\hat{\omega}(p^*) = \omega_0$

Group top p^* quantile and bottom $1 - p^*$ quantile into two messages

Example



Suppose $\sigma = \sigma' = \sigma_{BP}$:

Let $\hat{\omega}(p) :=$ mean of top p quantile

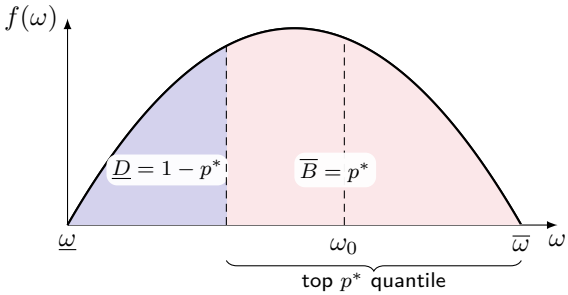
Find $p^* \in (0,1)$ s.t. $\hat{\omega}(p^*) = \omega_0$

Group top p^* quantile and bottom $1 - p^*$ quantile into two messages

S loses in $(1 - p^*)$ share of states and wins in p^* share of states

$$D(\sigma_{BP}) = 1 - p^*$$
$$B(\sigma_{BP}) = p^*$$

Example



Suppose $\sigma = \sigma' = \sigma_{BP}$:

Let $\hat{\omega}(p) :=$ mean of top p quantile

Find $p^* \in (0,1)$ s.t. $\hat{\omega}(p^*) = \omega_0$

Group top p^* quantile and bottom $1 - p^*$ quantile into two messages

S loses in $(1 - p^*)$ share of states and wins in p^* share of states

$$\underline{D} := \min_{\sigma \in \Sigma} D(\sigma) = D(\sigma_{BP}) = 1 - p^*$$

$$\overline{B} := \max_{\sigma' \in \Sigma} B(\sigma') = B(\sigma_{BP}) = p^*$$

Full manipulation

Theorem 1

Suppose S can send unverifiable information, then full manipulation is achievable *iff*

$$\overline{B} \geq \alpha \times \underline{D}$$

which simplifies to $p^* \geq \alpha/(1+\alpha)$.

Full manipulation

Theorem 1

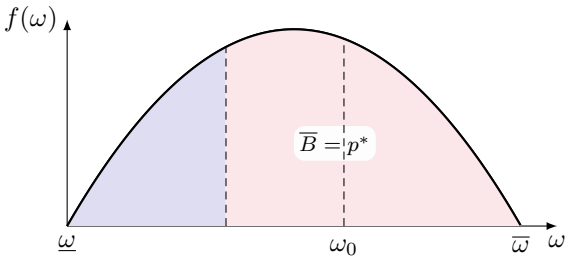
Suppose S can send unverifiable information, then full manipulation is achievable *iff*

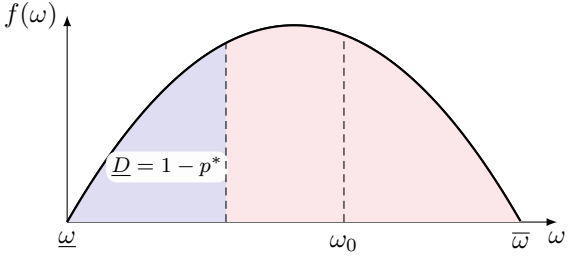
$$\overline{B} \geq \alpha \times \underline{D},$$

max budget $\geq$ price $\times$ min demand

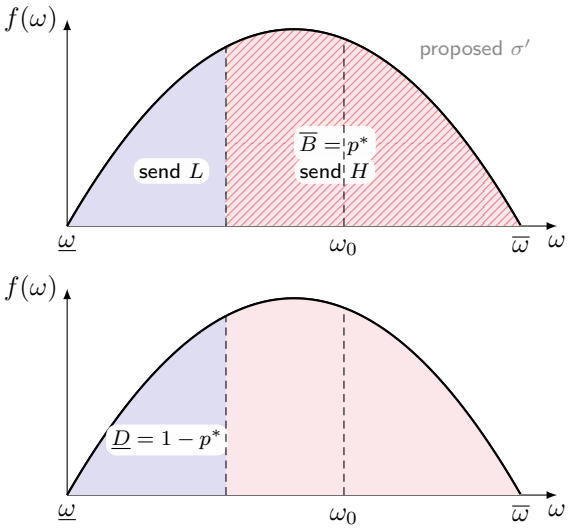
which simplifies to $p^* \geq \alpha/(1+\alpha)$.

Intuition


$f(\omega)$
$\overline{B} = p^*$
$\underline{\omega}$
ω_0
$\overline{\omega}$
ω


$f(\omega)$
$\underline{D} = 1 - p^*$
$\underline{\omega}$
ω_0
$\overline{\omega}$
ω

Intuition

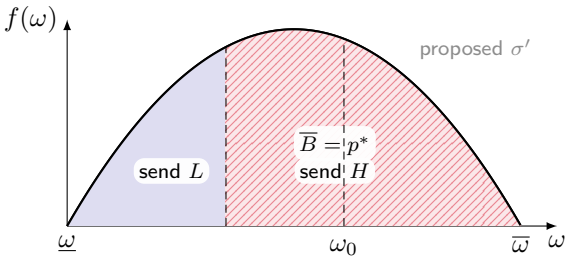


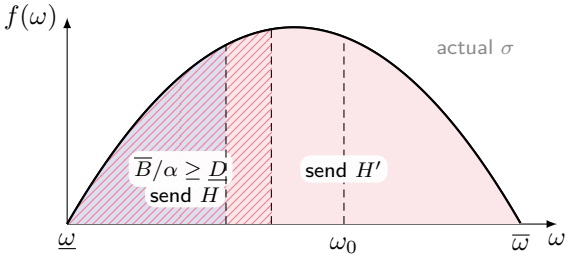
S proposes θ_1:

"I am a Kamenica-Gentzkow sender

L for low states and H for high states"

Intuition





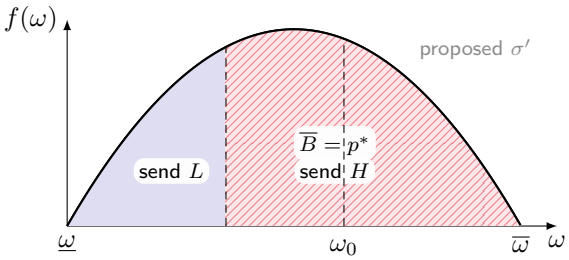
S proposes θ_1:

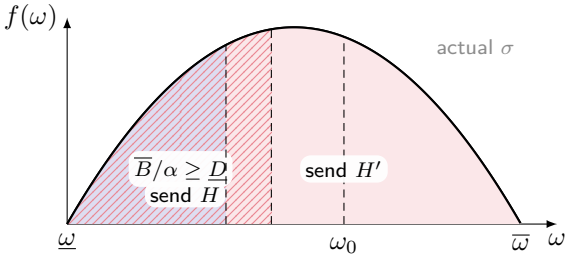
"I am a Kamenica-Gentzkow sender L for low states and H for high states"

but then behaves according to θ_0:

"I say it's H when the states are low and H' when states are truly high"

Intuition





S proposes θ_1:

"I am a Kamenica-Gentzkow sender L for low states and H for high states"

but then behaves according to θ_0:

"I say it's H when the states are low and H' when states are truly high"

Strategy only works if $\overline{B} \geq \alpha \times \underline{D}$

Most efficient use of budget for demand

Observations

Theorem 1

Suppose S can send unverifiable information, then full manipulation is achievable *iff* $\overline{B} \geq \alpha \times \underline{D}$, which simplifies to $p^* > \alpha/(1+\alpha)$.

Observations

Theorem 1

Suppose S can send unverifiable information, then full manipulation is achievable *iff* $\overline{B} \geq \alpha \times \underline{D}$, which simplifies to $p^* \geq \alpha/(1+\alpha)$.

- Condition for full manipulation concerns only the fraction of "winning states" in the BP policy, but the "losing states" can be arbitrarily bad

Observations

Theorem 1

Suppose S can send unverifiable information, then full manipulation is achievable *iff* $\overline{B} \geq \alpha \times \underline{D}$, which simplifies to $p^* \geq \alpha/(1+\alpha)$.

- Condition for full manipulation concerns only the fraction of "winning states" in the BP policy, but the "losing states" can be arbitrarily bad $\rightarrow$ extreme welfare loss

Observations

Theorem 1

Suppose S can send unverifiable information, then full manipulation is achievable *iff* $\overline{B} \geq \alpha \times \underline{D}$, which simplifies to $p^* \geq \alpha/(1+\alpha)$.

- Condition for full manipulation concerns only the fraction of "winning states" in the BP policy, but the "losing states" can be arbitrarily bad $\rightarrow$ extreme welfare loss
- Full manipulation is harder as model switching requires more evidence

Observations

Theorem 1

Suppose S can send unverifiable information, then full manipulation is achievable *iff* $\overline{B} \geq \alpha \times \underline{D}$, which simplifies to $p^* \geq \alpha/(1+\alpha)$.

- Condition for full manipulation concerns only the fraction of "winning states" in the BP policy, but the "losing states" can be arbitrarily bad $\rightarrow$ extreme welfare loss
- Full manipulation is harder as model switching requires more evidence

  When $\alpha = 1$, full manipulation $\Leftrightarrow p^* \geq 1/2$, i.e., average of the best 50% states exceeds ω_0

  W/o model persuasion, S can only guarantee $a = 1$ if the unconditional mean exceeds ω_0

Next: What happens if full manipulation is not achievable?

- It depends on what S can do about the remaining states
- **Case 1**: Cheap talk w/o commitment

  In a Perfect Bayesian equilibrium*, $\sigma(\cdot|\omega)$ maximizes S's payoff for every ω

  (* adapted to allow model-switching)
- **Case 2**: Cheap talk with commitment (aka. Bayesian persuasion)

  S chooses σ beforehand and commits to it

Cheap talk: either full or no manipulation

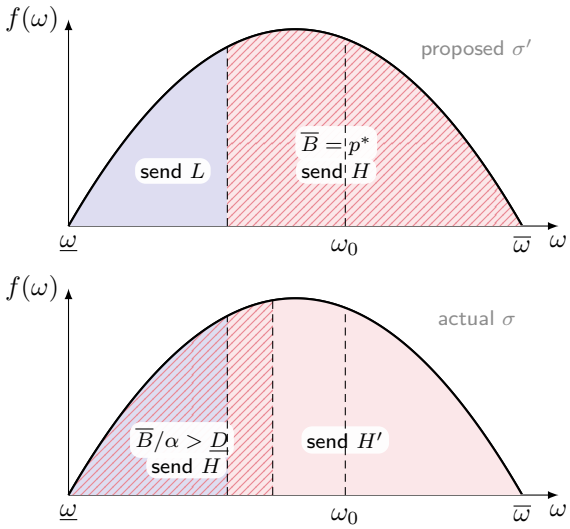
Proposition 1

Suppose S can send unverifiable information and does not have commitment power. In any PBE:

(i) If $\overline{B} \geq \alpha \times \underline{D}$, S **fully manipulates** R.

(ii) If $\overline{B} < \alpha \times \underline{D}$, S **cannot manipulate** R, and R must take $a = 1$ with probability 0.

Cheap talk: either full or no manipulation



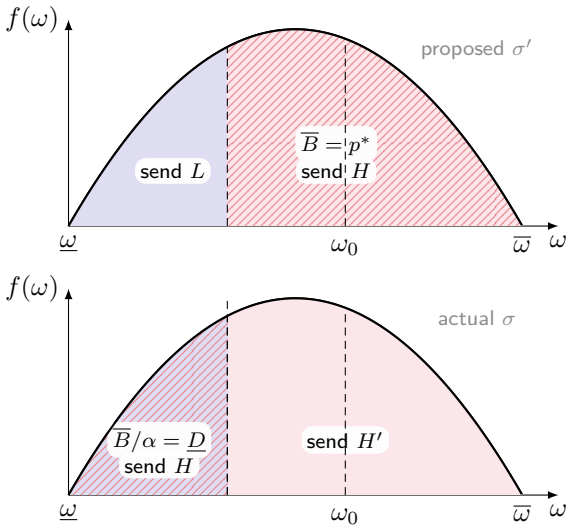
Proposition 1

Suppose S can send unverifiable information and does not have commitment power. In any PBE:

(i) If $\overline{B} \geq \alpha \times \underline{D}$, S **fully manipulates** R.

(ii) If $\overline{B} < \alpha \times \underline{D}$, S **cannot manipulate** R, and R must take $a = 1$ with probability 0.

Cheap talk: either full or no manipulation



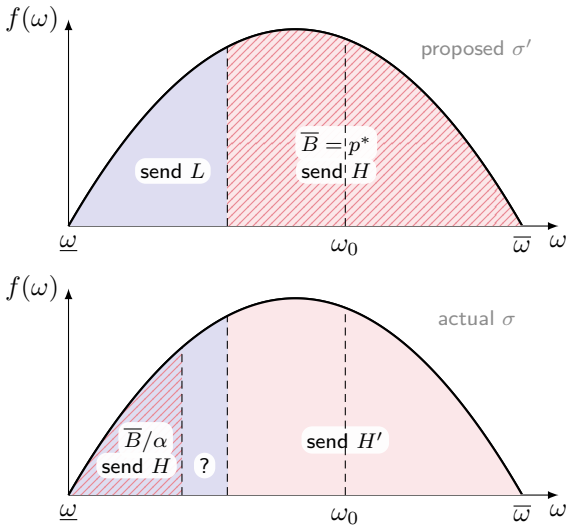
Proposition 1

Suppose S can send unverifiable information and does not have commitment power. In any PBE:

(i) If $\overline{B} \geq \alpha \times \underline{D}$, S **fully manipulates** R.

(ii) If $\overline{B} < \alpha \times \underline{D}$, S **cannot manipulate** R, and R must take $a = 1$ with probability 0.

Cheap talk: either full or no manipulation



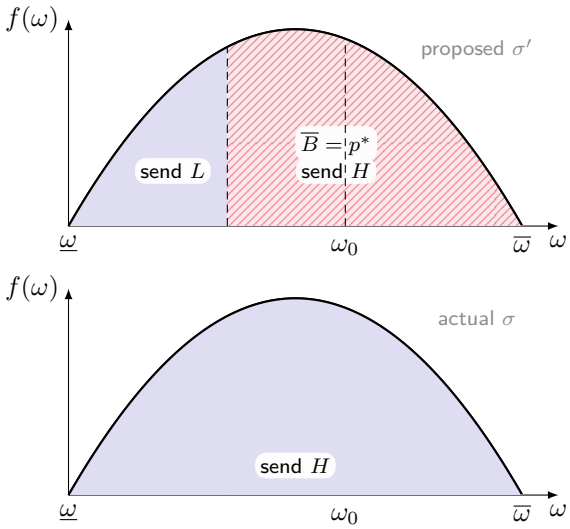
Proposition 1

Suppose S can send unverifiable information and does not have commitment power. In any PBE:

(i) If $\overline{B} \geq \alpha \times \underline{D}$, S **fully manipulates** R.

(ii) If $\overline{B} < \alpha \times \underline{D}$, S **cannot manipulate** R, and R must take $a = 1$ with probability 0.

Cheap talk: either full or no manipulation



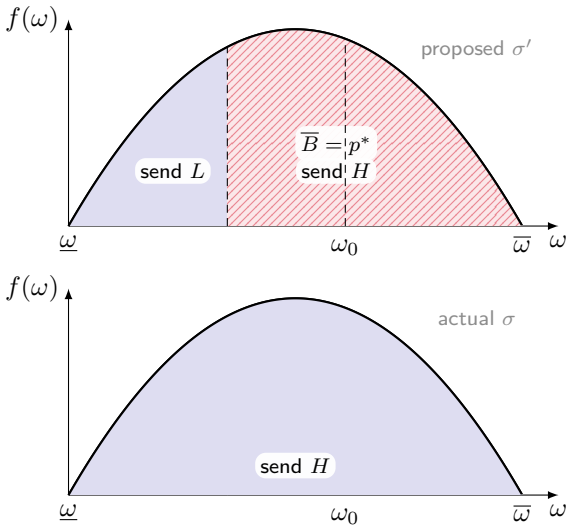
Proposition 1

Suppose S can send unverifiable information and does not have commitment power. In any PBE:

(i) If $\overline{B} \geq \alpha \times \underline{D}$, S **fully manipulates** R.

(ii) If $\overline{B} < \alpha \times \underline{D}$, S **cannot manipulate** R, and R must take $a = 1$ with probability 0.

Cheap talk: either full or no manipulation



Proposition 1

Suppose S can send unverifiable information and does not have commitment power. In any PBE:

(i) If $\overline{B} \geq \alpha \times \underline{D}$, S **fully manipulates** R.

(ii) If $\overline{B} < \alpha \times \underline{D}$, S **cannot manipulate** R, and R must take $a = 1$ with probability 0.

Remark. Model persuasion has no bite unless the cheap talker can pass as a KG sender with commitment within R's reasonable doubt

Cheap talk with commitment: partial manipulation

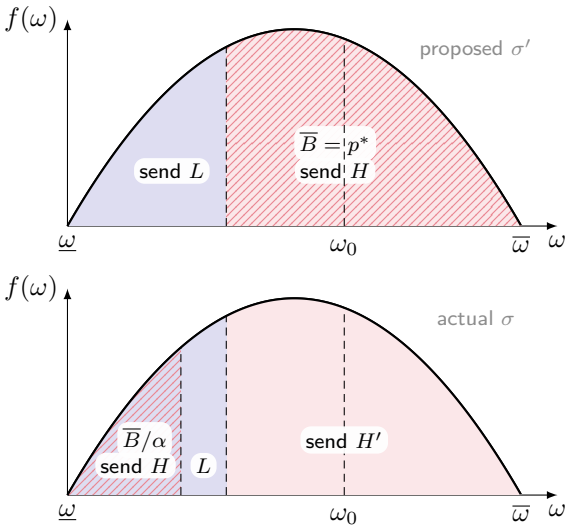
Proposition 2

Suppose S can send unverifiable info and has commitment power. In any PBE:

(i) If $\overline{B} \geq \alpha \times \underline{D}$, S **fully manipulates** R.

(ii) If $\overline{B} < \alpha \times \underline{D}$, S **partially manipulates** R using the optimal strategy described in the left.

Cheap talk with commitment: partial manipulation



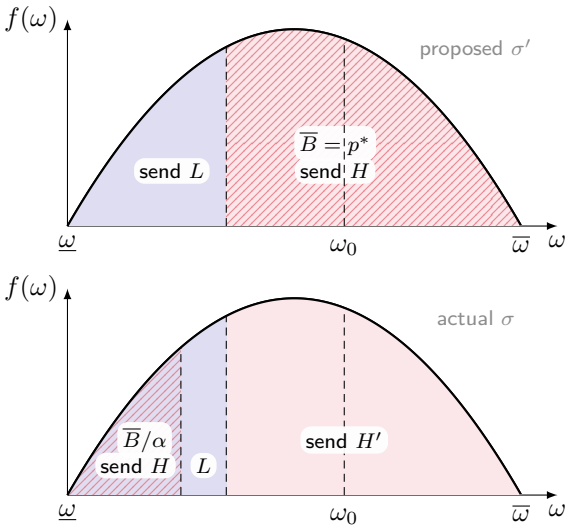
Proposition 2

Suppose S can send unverifiable info and has commitment power. In any PBE:

(i) If $\overline{B} \geq \alpha \times \underline{D}$, S **fully manipulates** R.

(ii) If $\overline{B} < \alpha \times \underline{D}$, S **partially manipulates** R using the optimal strategy described in the left.

Cheap talk with commitment: partial manipulation



Proposition 2

Suppose S can send unverifiable info and has commitment power. In any PBE:

(i) If $\overline{B} \geq \alpha \times \underline{D}$, S **fully manipulates** R.

(ii) If $\overline{B} < \alpha \times \underline{D}$, S **partially manipulates** R using the optimal strategy described in the left. R takes $a = 1$ with ex ante prob. $p^* + p^*/\alpha$.

Remark. When $\alpha = 1$, model persuasion **doubles** the prob. of the risky action.

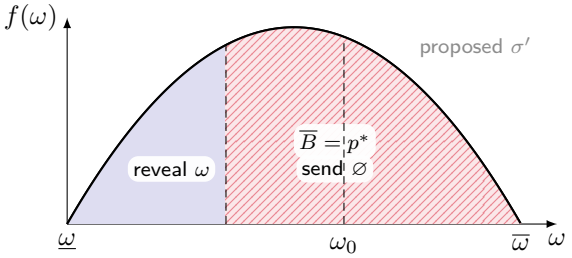
Constraints in information technology

- When S is **constrained** in her information technology, $M(\omega) \subsetneq M$ for some $\omega \in \Omega$ and we may have $M(\omega) \neq M(\omega')$ for $\omega \neq \omega'$
- The budget for model persuasion may not be **fungible** for the demand
- However, a version of Theorem 1 still applies for **perfect verifiable information**

Perfect verifiable information

Suppose in each state $\omega \in \Omega$, S can only send $M(\omega) = \{\omega, \varnothing\}$

Perfect verifiable information

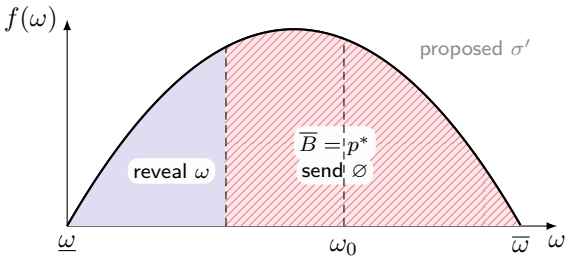


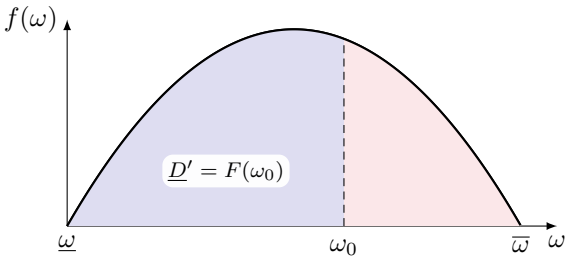
Suppose in each state $\omega \in \Omega$, S can only send $M(\omega) = \{\omega, \varnothing\}$

S proposes θ_1:

"I sound the alarm only when things go wrong; no news is good news"

Perfect verifiable information



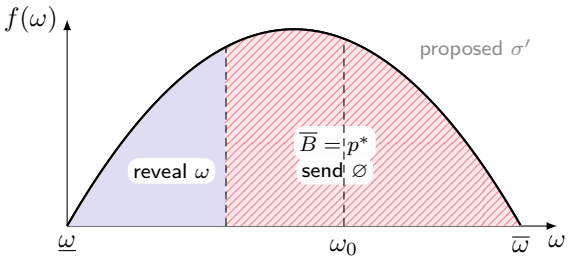


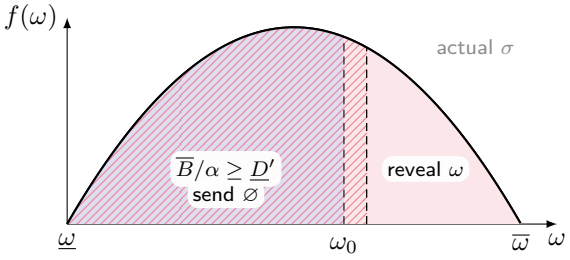
Suppose in each state $\omega \in \Omega$, S can only send $M(\omega) = \{\omega, \varnothing\}$

S proposes θ_1:

"I sound the alarm only when things go wrong; no news is good news"

Perfect verifiable information





Suppose in each state $\omega \in \Omega$, S can only send $M(\omega) = \{\omega, \varnothing\}$

S proposes θ_1:

"I sound the alarm only when things go wrong; no news is good news"

but then behaves according to θ_0:

"I stay quiet about bad states, but I make sure to share good ones"

Strategy only works if $\overline{B} \geq \alpha \times \underline{D}'$

Perfect verifiable info: either full manipulation or unraveling

Proposition 3

Suppose S can send either reveal the state or stay silent. In any PBE:

(i) If $\overline{B} \geq \alpha \times \underline{D}'$, S **fully manipulates** R.

(ii) If $\overline{B} < \alpha \times \underline{D}'$, S **cannot manipulate** R. Moreover, R takes $a = 1$ when $\omega \geq \omega_0$.

- Since $\underline{D}' \geq \underline{D}$, full manipulation is harder than under unverifiable info
- When full manipulation is impossible, model persuasion has no bite, and **unraveling** is the unique equilibrium

Comparing information technologies

Without the model proposal stage, S prefers

$$\text{Bayesian persuasion} \succ \text{verifiable disclosure} \succ \text{cheap talk}$$

Comparing information technologies

Without the model proposal stage, S prefers

$$\text{Bayesian persuasion} \succ \text{verifiable disclosure} \succ \text{cheap talk}$$

Corollary 1

With the model proposal stage:

If $\overline{B} \geq \alpha \underline{D}'$, Bayesian persuasion $\sim$ verifiable disclosure $\sim$ cheap talk

info technology is irrelevant

Comparing information technologies

Without the model proposal stage, S prefers

$$\text{Bayesian persuasion} \succ \text{verifiable disclosure} \succ \text{cheap talk}$$

Corollary 1

With the model proposal stage:

If $\overline{B} \geq \alpha \underline{D}'$, Bayesian persuasion $\sim$ verifiable disclosure $\sim$ cheap talk

info technology is irrelevant

If $\overline{B} \in [\alpha \underline{D}, \alpha \underline{D}')$, Bayesian persuasion $\sim$ cheap talk $\succ$ verifiable disclosure

soft information leaves room for manipulation

Comparing information technologies

Without the model proposal stage, S prefers

$$\text{Bayesian persuasion} \succ \text{verifiable disclosure} \succ \text{cheap talk}$$

Corollary 1

With the model proposal stage:

If $\overline{B} \geq \alpha \underline{D}'$, Bayesian persuasion $\sim$ verifiable disclosure $\sim$ cheap talk

info technology is irrelevant

If $\overline{B} \in [\alpha \underline{D}, \alpha \underline{D}')$, Bayesian persuasion $\sim$ cheap talk $\succ$ verifiable disclosure

soft information leaves room for manipulation

If $\overline{B} < \alpha \underline{D}$, Bayesian persuasion $\succ$ verifiable disclosure $\succ$ cheap talk

hard information wins credibility

Comparing information technologies

Without the model proposal stage, S prefers

$$\text{Bayesian persuasion} \succ \text{verifiable disclosure} \succ \text{cheap talk}$$

Corollary 1

With the model proposal stage:

If $\overline{B} \geq \alpha \underline{D}'$, Bayesian persuasion $\sim$ verifiable disclosure $\sim$ cheap talk

info technology is irrelevant

If $\overline{B} \in [\alpha \underline{D}, \alpha \underline{D}')$, Bayesian persuasion $\sim$ cheap talk $\succ$ verifiable disclosure

soft information leaves room for manipulation

If $\overline{B} < \alpha \underline{D}$, Bayesian persuasion $\succ$ verifiable disclosure $\succ$ cheap talk

hard information wins credibility

- Connect back to the Disinformation Puzzle

Conclusion

Literature

Strategic communication

- Cheap talk: Crawford and Sobel (1982), Lipnowski and Ravid (2020)
- Verifiable disclosure: Grossman (1981), Milgrom (1981), Dye (1985)
- Information design and persuasion: Kamenica and Gentzkow (2011)
- Limited commitment: Fréchette, Lizzeri, and Perego (2022), Lin and Liu (2024)
- Boundedly rational receiver: De Clippel & Zhang (2019), Bloedel & Segal (2021)

Model persuasion and misspecification

- Galperti (2019), Schwartzstein and Sunderam (2021, 2024), Ichihashi and Meng (2021), Jain (2023), Ambuehl and Thysen (2024), Barron and Fries (2024), Eliaz and Spiegler (2025), Aina (2025), Ba (2025)

Reputation games

- Jehiel and Samuelson (2012), Best and Quigley (2024)

Summary

- **Unifying framework** that combines information-based and model-based persuasion
- **Characterization** of when full manipulation is possible and how to achieve it
- **Simple strategy**: claim to be a KG sender, then deviate (a formalization of gaslighting?)
- **Comparing soft v.s. hard information**: wiggle room for manipulation v.s. credibility
- Possible explanation for **the Disinformation Puzzle**

Thank you for listening!

Feedback is greatly appreciated :D

Email: bacuimin@gmail.com

Perfect Bayesian Equilibrium (PBE)

go back

Definition (PBE)

1. S proposes **alt. model σ'^*** to maximize their ex ante expected payoff, anticipating their messaging strat. σ^* and R's behavior (model choice θ^*, belief π^*, and action a^*).
2. Given any proposed alt. model, S chooses **message $\sigma^*(\omega)$** to maximize their expected payoff in every state $\omega \in \Omega$, anticipating R's behavior (θ^*, π^*, a^*).
3. R selects **model θ^*** $\in \{\sigma'^*, \sigma^*\}$ using the Bayes factor rule, whenever possible.
4. R updates **beliefs π^***, under each model using Bayes' rule, whenever possible.
5. R chooses **action a^*** $\in \{0, 1\}$ that maximizes their expected payoff, given their beliefs and chosen model, $\pi^*(\cdot|\cdot, \theta^*)$.

Definition (PBE) – Formal Details

1. S proposes **alt. model σ'^{*}** to maximize their ex ante expected payoff, anticipating their messaging strat. $\sigma^{*}_{\cdot\cdot}$ and R's behavior (model choice $\theta^{*}_{\cdot\cdot}$, belief π^{*}, and action a^{*}).

$$\sigma'^{*} \in \arg\max_{\sigma' \in \Sigma} \mathbb{E}_{\omega \sim F} \mathbb{E}_{m \sim \sigma^{*}_{\cdot\cdot}(\cdot|\omega)} u_s\Big(a^{*}\Big(\pi^{*}\big(\cdot\,|m, \theta^{*}_{\cdot\cdot}(m)\big)\Big)\Big).$$

2. Given any proposed alt. model, S chooses **messaging strategy $\sigma^{*}_{\cdot\cdot}$** to maximize their expected payoff in every state $\omega \in \Omega$, anticipating R's behavior ($\theta^{*}_{\cdot\cdot}$, π^{*}, a^{*}).

$$\sigma^{*}_{\cdot\cdot}(\cdot\,|\omega) \in \arg\max_{\sigma_{\cdot}(\cdot|\omega) \in \Delta M} \mathbb{E}_{m \sim \sigma(\cdot|\omega)} u_s\Big(a^{*}\Big(\pi^{*}\big(\cdot\,|m, \theta^{*}_{\cdot\cdot}(m)\big)\Big)\Big) \qquad \forall \omega \in \Omega$$

3. R selects **model $\theta^{*}_{\cdot\cdot} \in \{\sigma'^{*}, \sigma^{*}\}$** using the Bayes factor rule, whenever possible.

$$\theta^{*}_{\cdot\cdot}(m) = \sigma'^{*} \Leftrightarrow \Pr\big(m|\sigma'^{*}\big)/\Pr(m|\sigma^{*}) \geq \alpha \qquad \forall m \in M$$

4. R updates **beliefs π^{*}**, under each model using Bayes' rule, whenever possible.

$$\pi^{*}(\omega|m, \hat{\sigma}) = \frac{\hat{\sigma}(m|\omega) f(\omega)}{\int_{\Omega} \hat{\sigma}(m|\tilde{\omega}) f(\tilde{\omega}) d\tilde{\omega}} \qquad \forall \omega \in \Omega,\ m \in M,\ \hat{\sigma} \in \{\sigma'^{*}, \sigma^{*}\}$$

5. R chooses optimal **action $a^{*} \in \{0, 1\}$** given their beliefs and chosen model, $\pi^{*}(\cdot|\cdot, \theta^{*}_{\cdot\cdot})$.

$$a^{*}\Big(\pi^{*}\big(\cdot\,|m, \theta^{*}_{\cdot\cdot}(m)\big)\Big) = \arg\max_{a \in \{0,1\}} \mathbb{E}_{\omega \sim \pi^{*}\left(\cdot|m, \theta^{*}_{\cdot\cdot}(m)\right)}[u_R(a, \omega)] \qquad \forall m \in M$$

Definition (PBE) – Formal Details with Commitment Messaging

1. S proposes **alt. model σ'^*** to maximize their ex ante expected payoff, anticipating their messaging strat. $\sigma^*_{\cdot\cdot}$ and R's behavior (model choice $\theta^*_{\cdot\cdot}$, belief π^*, and action a^*).

$$\sigma'^* \in \arg\max_{\sigma' \in \Sigma} \mathbb{E}_{\omega\sim F}\mathbb{E}_{m\sim\sigma'^*(\cdot|\omega)} u_s\Big(a^*\Big(\cdot\,\Big|\,\pi^*\big(\cdot\,\big|m,\theta^*_{\cdot\cdot}(m)\big)\Big)\Big).$$

2. Given any proposed alt. model, S chooses **messaging strategy $\sigma^*_{\cdot\cdot}$** to maximize their expected payoff in every state $\omega \in \Omega$, anticipating R's behavior ($\theta^*_{\cdot\cdot}$, π^*, a^*).

$$\sigma^*_{\cdot\cdot}(\,\cdot\,|\,\cdot\,) \in \arg\max_{\sigma\in\Sigma} \mathbb{E}_{\omega\sim F}\mathbb{E}_{m\sim\sigma(\cdot|\cdot)} u_s\Big(a^*\Big(\pi^*\big(\cdot\,\big|m,\theta^*_{\cdot\cdot}(m)\big)\Big)\Big)$$

3. R selects **model $\theta^*_{\cdot\cdot}$** $\in \{\sigma'^*, \sigma^*\}$ using the Bayes factor rule, whenever possible.

$$\theta^*_{\cdot\cdot}(m) = \sigma'^* \Leftrightarrow \Pr\big(m|\sigma'^*\big)/\Pr(m|\sigma^*) \geq \alpha \qquad \forall m \in M$$

4. R updates **beliefs π^***, under each model using Bayes' rule, whenever possible.

$$\pi^*(\omega|m,\hat{\sigma}) = \frac{\hat{\sigma}(m|\omega)f(\omega)}{\int_\Omega \hat{\sigma}(m|\tilde{\omega})f(\tilde{\omega})d\tilde{\omega}} \qquad \forall\omega\in\Omega,\ m\in M,\ \hat{\sigma}\in\{\sigma'^*,\sigma^*\}$$

5. R chooses optimal **action a^*** $\in \{0,1\}$ given their beliefs and chosen model, $\pi^*(\cdot|\cdot,\theta^*)$.

$$a^*\Big(\pi^*\big(\cdot\,\big|m,\theta^*(m)\big)\Big) = \arg\max_{a\in\{0,1\}} \mathbb{E}_{\omega\sim\pi^*\left(\cdot\,\middle|\,m,\theta^*(m)\right)}[u_R(a,\omega)] \qquad \forall m\in M$$