Confessions

Francisco Poggi (Mannheim)

April 24, 2026

"I cannot prove to them that I confess the truth;
yet those whose ears love opens to me will believe me."

— Augustine, *Confessions*, Book X, Ch. 3

Motivation

- Organizations design *how information flows*.
  - **Communication protocol** (mechanism): who speaks when, what messages are allowed, how they are processed and aggregated.
  - E.g. internal reports, peer review, court procedure, expert advice.
- However, participants can sometimes talk *off protocol*, potentially undermining the mechanism.

Question

- When does off-protocol communication *undermine* the protocol?
- How can we design protocols that are robust to off-protocol communication?

Motivation

- Organizations design *how information flows*.
  - **Communication protocol** (mechanism): who speaks when, what messages are allowed, how they are processed and aggregated.
  - E.g. internal reports, peer review, court procedure, expert advice.
- However, participants can sometimes talk *off protocol*, potentially undermining the mechanism.

Question

- When does off-protocol communication *undermine* the protocol?
- How can we design protocols that are robust to off-protocol communication?

Motivation

- Organizations design *how information flows*.
  - **Communication protocol** (mechanism): who speaks when, what messages are allowed, how they are processed and aggregated.
  - E.g. internal reports, peer review, court procedure, expert advice.
- However, participants can sometimes talk *off protocol*, potentially undermining the mechanism.

Question

- When does off-protocol communication *undermine* the protocol?
- How can we design protocols that are robust to off-protocol communication?

Motivation

- Organizations design *how information flows*.
  - **Communication protocol** (mechanism): who speaks when, what messages are allowed, how they are processed and aggregated.
  - E.g. internal reports, peer review, court procedure, expert advice.
- However, participants can sometimes talk *off protocol*, potentially undermining the mechanism.

Question

- When does off-protocol communication *undermine* the protocol?
- How can we design protocols that are robust to off-protocol communication?

Motivation

- Organizations design *how information flows*.
  - **Communication protocol** (mechanism): who speaks when, what messages are allowed, how they are processed and aggregated.
  - E.g. internal reports, peer review, court procedure, expert advice.
- However, participants can sometimes talk *off protocol*, potentially undermining the mechanism.

Question

- When does off-protocol communication *undermine* the protocol?
- How can we design protocols that are robust to off-protocol communication?

Motivation

- Organizations design *how information flows*.
  - **Communication protocol** (mechanism): who speaks when, what messages are allowed, how they are processed and aggregated.
  - E.g. internal reports, peer review, court procedure, expert advice.
- However, participants can sometimes talk *off protocol*, potentially undermining the mechanism.

Question

- When does off-protocol communication *undermine* the protocol?
- How can we design protocols that are robust to off-protocol communication?

Experts with career concerns

Players. Two senders ($i = 1, 2$) and a decision-maker (DM).

- One sender is the *expert*, the other is a *quack*. The DM does not know which is which (uniform prior over assignments).
- State $\omega \in \{L, R\}$, uniform prior.
- Each sender observes a conditionally independent binary signal with precision p_E and p_Q respectively.
  - For exposition: $p_Q = \frac{1}{2}$ (the quack is uninformed).
- The DM chooses an action $a \in \{\ell, r\}$ and a promotion $\pi \in \{1, 2\}$.

Payoffs.

- Everyone receives 1 if $a = \omega$.
- Sender i receives an additional bonus $B > 0$ if promoted.
- The DM wants to promote the expert.

Experts with career concerns

Players. Two senders ($i = 1, 2$) and a decision-maker (DM).

- One sender is the *expert*, the other is a *quack*. The DM does not know which is which (uniform prior over assignments).
- State $\omega \in \{L, R\}$, uniform prior.
- Each sender observes a conditionally independent binary signal with precision p_E and p_Q respectively.
  - For exposition: $p_Q = \frac{1}{2}$ (the quack is uninformed).
- The DM chooses an action $a \in \{\ell, r\}$ and a promotion $\pi \in \{1, 2\}$.

Payoffs.

- Everyone receives 1 if $a = \omega$.
- Sender i receives an additional bonus $B > 0$ if promoted.
- The DM wants to promote the expert.

Experts with career concerns

Players. Two senders ($i = 1, 2$) and a decision-maker (DM).

- One sender is the *expert*, the other is a *quack*. The DM does not know which is which (uniform prior over assignments).
- State $\omega \in \{L, R\}$, uniform prior.
- Each sender observes a conditionally independent binary signal with precision p_E and p_Q respectively.
  - For exposition: $p_Q = \frac{1}{2}$ (the quack is uninformed).
- The DM chooses an action $a \in \{\ell, r\}$ and a promotion $\pi \in \{1, 2\}$.

Payoffs.

- Everyone receives 1 if $a = \omega$.
- Sender i receives an additional bonus $B > 0$ if promoted.
- The DM wants to promote the expert.

Experts with career concerns

Players. Two senders ($i = 1, 2$) and a decision-maker (DM).

- One sender is the *expert*, the other is a *quack*. The DM does not know which is which (uniform prior over assignments).
- State $\omega \in \{L, R\}$, uniform prior.
- Each sender observes a conditionally independent binary signal with precision p_E and p_Q respectively.
  - For exposition: $p_Q = \frac{1}{2}$ (the quack is uninformed).
- The DM chooses an action $a \in \{\ell, r\}$ and a promotion $\pi \in \{1, 2\}$.

Payoffs.

- Everyone receives 1 if $a = \omega$.
- Sender i receives an additional bonus $B > 0$ if promoted.
- The DM wants to promote the expert.

Experts with career concerns

Players. Two senders ($i = 1, 2$) and a decision-maker (DM).

- One sender is the *expert*, the other is a *quack*. The DM does not know which is which (uniform prior over assignments).
- State $\omega \in \{L, R\}$, uniform prior.
- Each sender observes a conditionally independent binary signal with precision p_E and p_Q respectively.
  - For exposition: $p_Q = \frac{1}{2}$ (the quack is uninformed).
- The DM chooses an action $a \in \{\ell, r\}$ and a promotion $\pi \in \{1, 2\}$.

Payoffs.

- Everyone receives 1 if $a = \omega$.
- Sender i receives an additional bonus $B > 0$ if promoted.
- The DM wants to promote the expert.

Experts with career concerns

Players. Two senders ($i = 1, 2$) and a decision-maker (DM).

- One sender is the *expert*, the other is a *quack*. The DM does not know which is which (uniform prior over assignments).
- State $\omega \in \{L, R\}$, uniform prior.
- Each sender observes a conditionally independent binary signal with precision p_E and p_Q respectively.
  - For exposition: $p_Q = \frac{1}{2}$ (the quack is uninformed).
- The DM chooses an action $a \in \{\ell, r\}$ and a promotion $\pi \in \{1, 2\}$.

Payoffs.

- Everyone receives 1 if $a = \omega$.
- Sender i receives an additional bonus $B > 0$ if promoted.
- The DM wants to promote the expert.

Experts with career concerns

Players. Two senders ($i = 1, 2$) and a decision-maker (DM).

- One sender is the *expert*, the other is a *quack*. The DM does not know which is which (uniform prior over assignments).
- State $\omega \in \{L, R\}$, uniform prior.
- Each sender observes a conditionally independent binary signal with precision p_E and p_Q respectively.
  - For exposition: $p_Q = \frac{1}{2}$ (the quack is uninformed).
- The DM chooses an action $a \in \{\ell, r\}$ and a promotion $\pi \in \{1, 2\}$.

Payoffs.

- Everyone receives 1 if $a = \omega$.
- Sender i receives an additional bonus $B > 0$ if promoted.
- The DM wants to promote the expert.

Experts with career concerns

Players. Two senders ($i = 1, 2$) and a decision-maker (DM).

- One sender is the *expert*, the other is a *quack*. The DM does not know which is which (uniform prior over assignments).
- State $\omega \in \{L, R\}$, uniform prior.
- Each sender observes a conditionally independent binary signal with precision p_E and p_Q respectively.
  - For exposition: $p_Q = \frac{1}{2}$ (the quack is uninformed).
- The DM chooses an action $a \in \{\ell, r\}$ and a promotion $\pi \in \{1, 2\}$.

Payoffs.

- Everyone receives 1 if $a = \omega$.
- Sender i receives an additional bonus $B > 0$ if promoted.
- The DM wants to promote the expert.

Direct mechanisms

A **direct mechanism** asks each sender to report:

- whether they are the expert,
- their binary signal.

It gives an action recommendation (a, π) to the DM.

We focus on a *symmetric* family parameterized by

$$x = \Pr(a = \omega \mid \text{truthful reports}),$$

$$y = \Pr(\pi \text{ is the expert} \mid \text{truthful reports}).$$

Inconsistent reports. When both (or neither) senders claim to be the expert, the mechanism draws (a, π) uniformly at random.

Feasibility: $x \leq p_E$: the recommendation cannot be more informative than the expert's signal (using $p_Q = \frac{1}{2}$).

Direct mechanisms

A **direct mechanism** asks each sender to report:

- whether they are the expert,
- their binary signal.

It gives an action recommendation (a, π) to the DM.

We focus on a *symmetric* family parameterized by

$$x = \Pr(a = \omega \mid \text{truthful reports}),$$

$$y = \Pr(\pi \text{ is the expert} \mid \text{truthful reports}).$$

Inconsistent reports. When both (or neither) senders claim to be the expert, the mechanism draws (a, π) uniformly at random.

Feasibility: $x \leq p_E$: the recommendation cannot be more informative than the expert's signal (using $p_Q = \frac{1}{2}$).

Direct mechanisms

A **direct mechanism** asks each sender to report:

- whether they are the expert,
- their binary signal.

It gives an action recommendation (a, π) to the DM.

We focus on a *symmetric* family parameterized by

$$x \;=\; \Pr(a = \omega \mid \text{truthful reports}),$$

$$y \;=\; \Pr(\pi \text{ is the expert} \mid \text{truthful reports}).$$

Inconsistent reports. When both (or neither) senders claim to be the expert, the mechanism draws (a, π) uniformly at random.

Feasibility: $x \leq p_E$: the recommendation cannot be more informative than the expert's signal (using $p_Q = \frac{1}{2}$).

Direct mechanisms

A **direct mechanism** asks each sender to report:

- whether they are the expert,
- their binary signal.

It gives an action recommendation (a, π) to the DM.

We focus on a *symmetric* family parameterized by

$$x \;=\; \Pr(a = \omega \mid \text{truthful reports}),$$

$$y \;=\; \Pr(\pi \text{ is the expert} \mid \text{truthful reports}).$$

Inconsistent reports. When both (or neither) senders claim to be the expert, the mechanism draws (a, π) uniformly at random.

Feasibility: $x \leq p_E$: the recommendation cannot be more informative than the expert's signal (using $p_Q = \frac{1}{2}$).

Expert truthfulness and obedience

Obedience (DM). The DM follows the recommendation iff

$$x \geq \tfrac{1}{2}, \qquad y \geq \tfrac{1}{2}.$$

Truthfulness (expert). Given the quack reports truthfully:

$$\underbrace{x + yB}_{\text{truthful}} \geq \underbrace{\tfrac{1}{2} + \tfrac{B}{2}}_{\substack{\text{claim non-expert} \\ \text{(no one claims)}}} \quad \text{and} \quad \underbrace{x + yB}_{\text{truthful}} \geq \underbrace{(1 - x) + yB}_{\text{flip signal}}$$

Both slack under obedience.

Expert truthfulness and obedience

Obedience (DM). The DM follows the recommendation iff

$$x \geq \tfrac{1}{2}, \qquad y \geq \tfrac{1}{2}.$$

Truthfulness (expert). Given the quack reports truthfully:

$$\underbrace{x + yB}_{\text{truthful}} \geq \underbrace{\tfrac{1}{2} + \tfrac{B}{2}}_{\substack{\text{claim non-expert} \\ \text{(no one claims)}}} \quad \text{and} \quad \underbrace{x + yB}_{\text{truthful}} \geq \underbrace{(1 - x) + yB}_{\text{flip signal}}$$

Both slack under obedience.

Quack truthfulness

Truthfulness (quack).

$$\underbrace{x + (1-y)B}_{\text{truthful}} \;\geq\; \underbrace{\tfrac{1}{2} + \tfrac{B}{2}}_{\substack{\text{claim expert} \\ \text{(both claim)}}} \iff (2y-1)\,B \;\leq\; 2x - 1.$$

Cases.

- $B = 0$: no rent from promotion, quack is indifferent.
- $y = \frac{1}{2}$: promotion carries no information about expertise, quack is promoted with probability $\frac{1}{2}$ either way.
- The constraint binds when B is large *and* y is close to 1.

Quack truthfulness

Truthfulness (quack).

$$\underbrace{x + (1-y)B}_{\text{truthful}} \;\geq\; \underbrace{\tfrac{1}{2} + \tfrac{B}{2}}_{\substack{\text{claim expert} \\ \text{(both claim)}}} \iff (2y-1)\,B \;\leq\; 2x - 1.$$

Cases.

- $B = 0$: no rent from promotion, quack is indifferent.
- $y = \frac{1}{2}$: promotion carries no information about expertise, quack is promoted with probability $\frac{1}{2}$ either way.
- The constraint binds when B is large *and* y is close to 1.

Quack truthfulness

Truthfulness (quack).

$$\underbrace{x + (1-y)B}_{\text{truthful}} \;\geq\; \underbrace{\tfrac{1}{2} + \tfrac{B}{2}}_{\substack{\text{claim expert} \\ \text{(both claim)}}} \iff (2y-1)\,B \;\leq\; 2x - 1.$$

Cases.

- $B = 0$: no rent from promotion, quack is indifferent.
- $y = \frac{1}{2}$: promotion carries no information about expertise, quack is promoted with probability $\frac{1}{2}$ either way.
- The constraint binds when B is large *and* y is close to 1.

Quack truthfulness

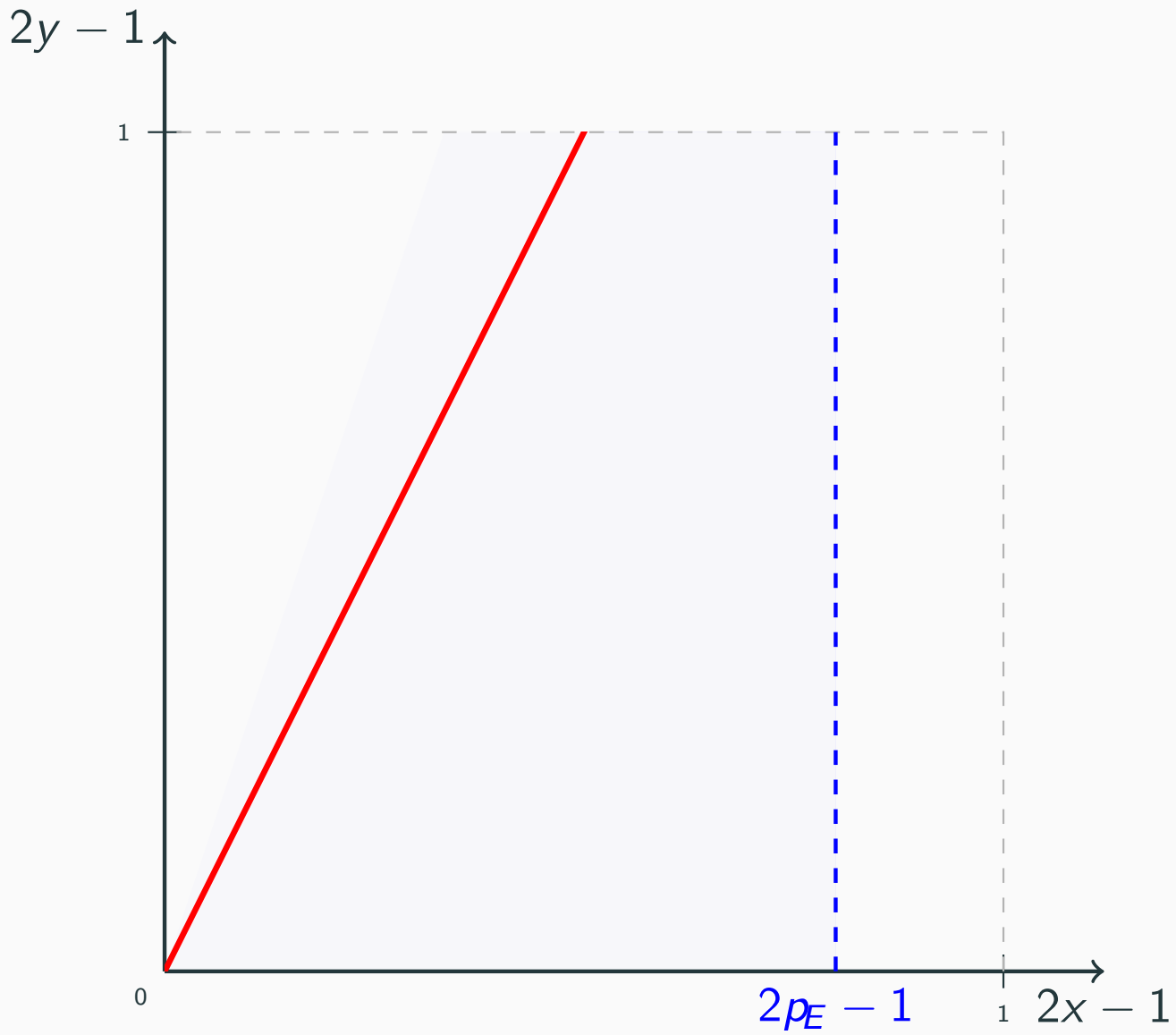
Truthfulness (quack).

$$\underbrace{x + (1-y)B}_{\text{truthful}} \;\geq\; \underbrace{\tfrac{1}{2} + \tfrac{B}{2}}_{\substack{\text{claim expert} \\ \text{(both claim)}}} \quad\Longleftrightarrow\quad (2y-1)\,B \;\leq\; 2x-1.$$

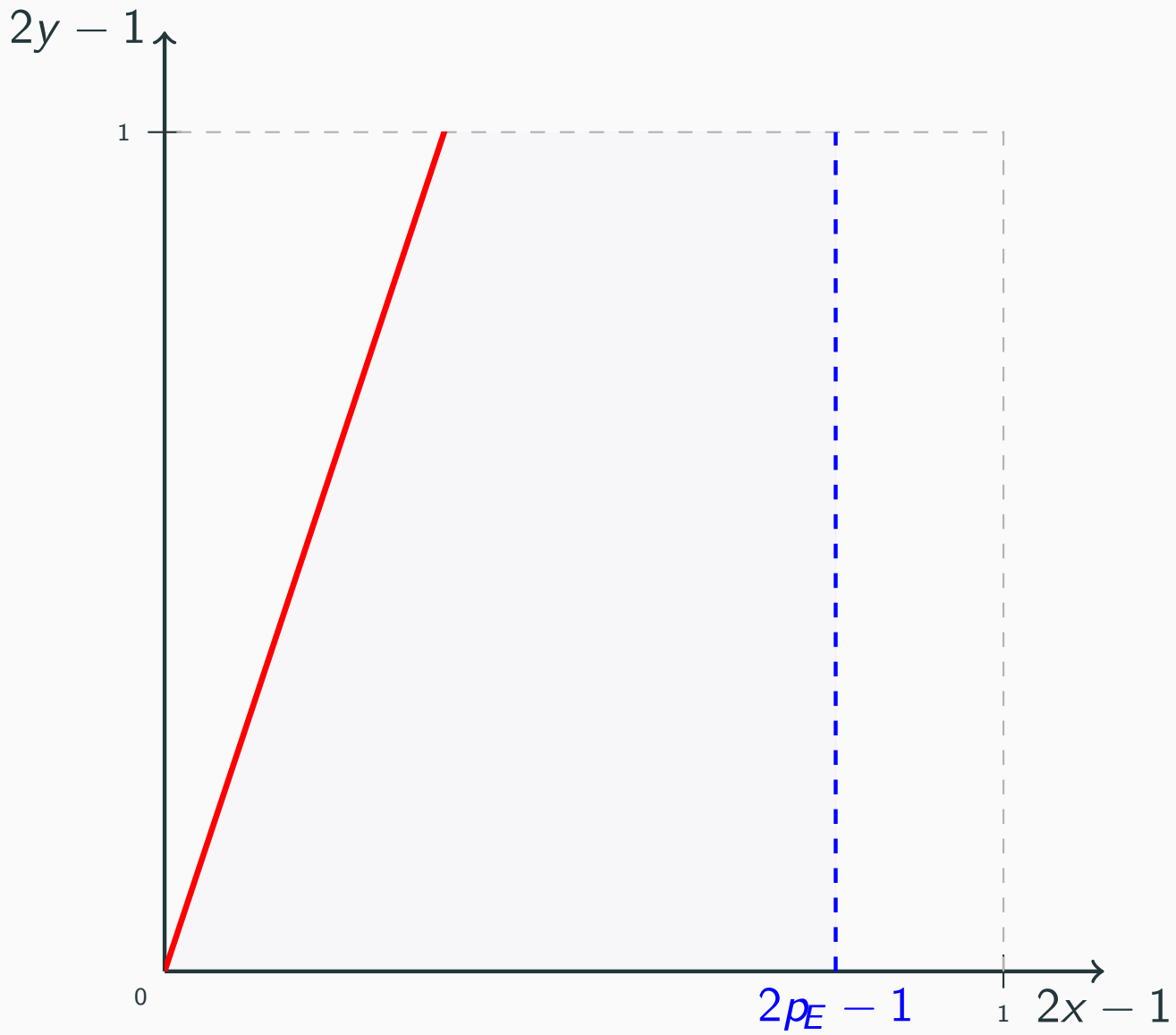
Cases.

- $B = 0$: no rent from promotion, quack is indifferent.
- $y = \frac{1}{2}$: promotion carries no information about expertise, quack is promoted with probability $\frac{1}{2}$ either way.
- The constraint binds when B is large *and* y is close to 1.

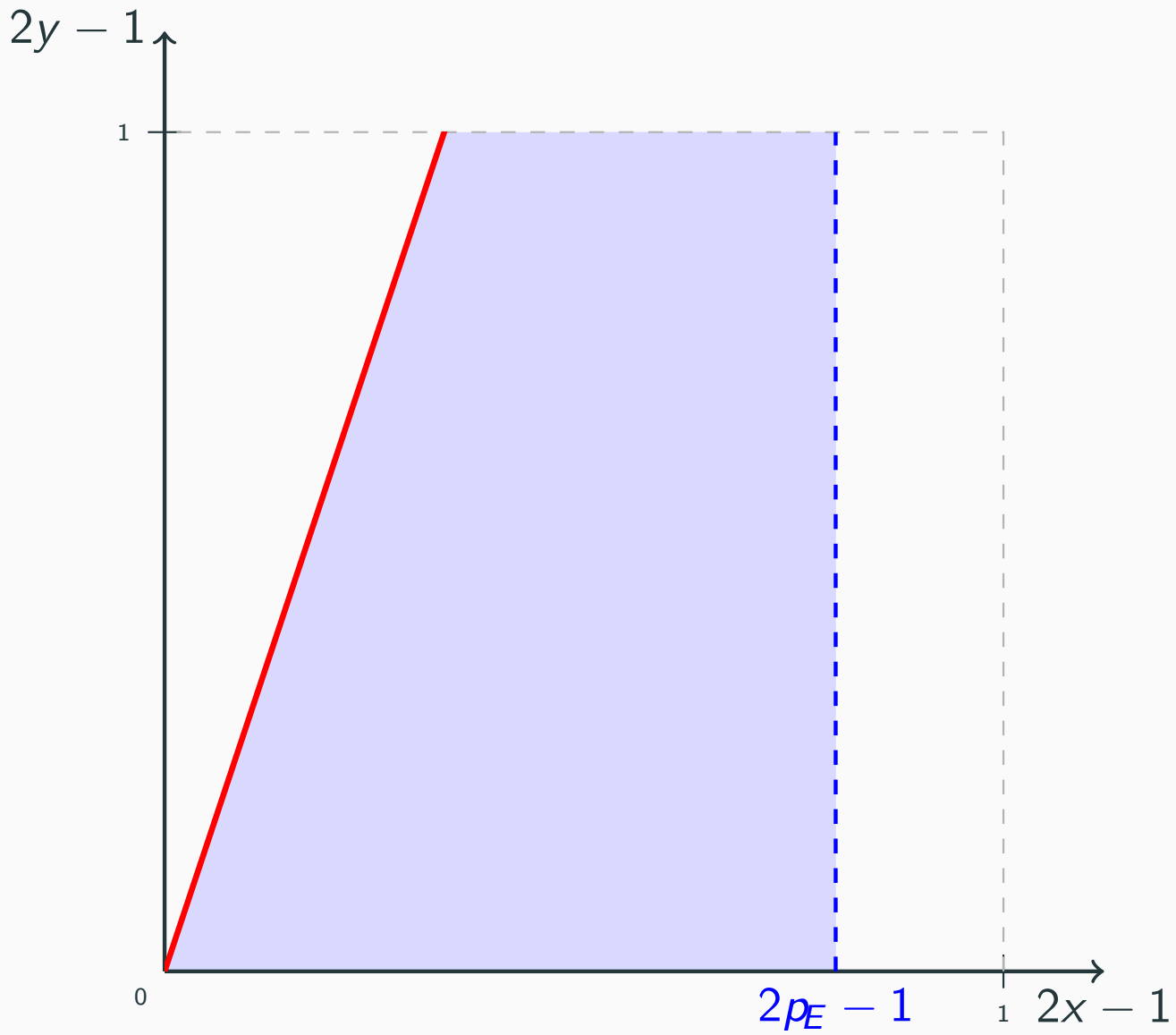
IC Direct Revelation Mechanisms



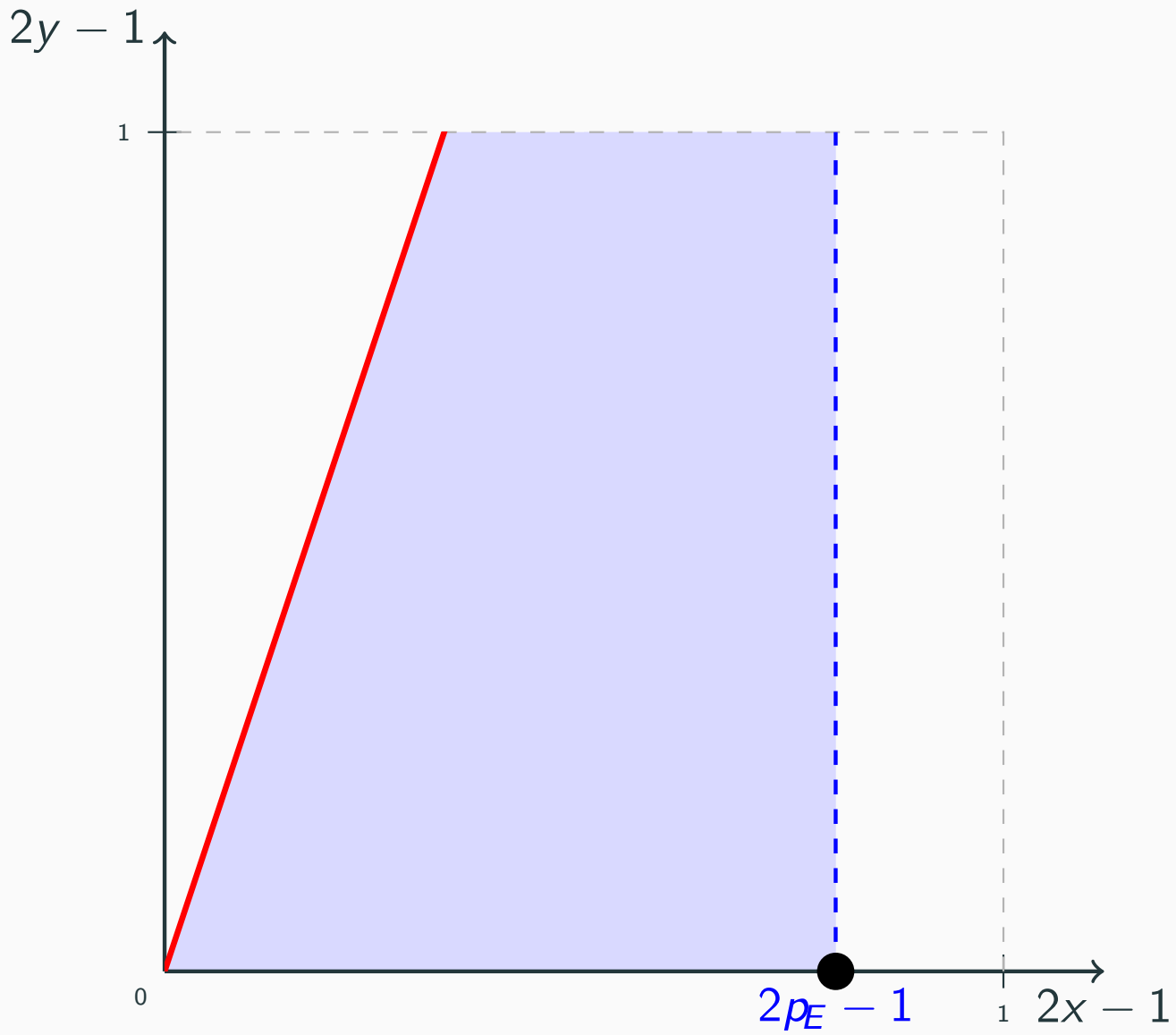
IC Direct Revelation Mechanisms



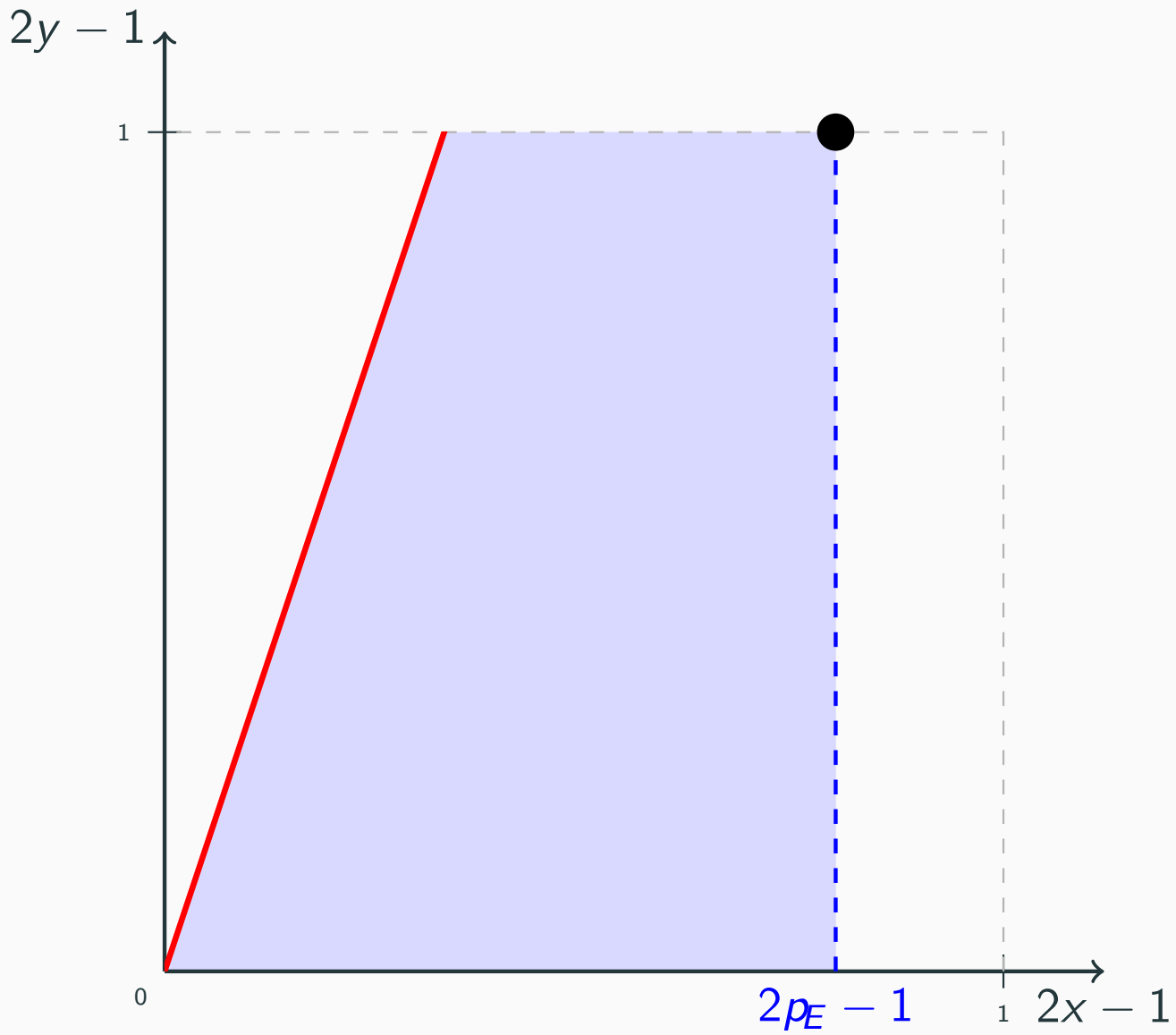
IC Direct Revelation Mechanisms



IC Direct Revelation Mechanisms



IC Direct Revelation Mechanisms



Expert might sabotage the mechanism

Timing.

Reports $\rightarrow$ Recommendation $\rightarrow$ Confession $\rightarrow$ Action, promotion

Fix a mechanism with $y < 1$.
Consider the following double deviation by the expert.

- Sabotage: Swap the reported signal ($L \leftrightarrow R$).
- Confess to the DM:

  "I am the expert and I lied in my report. Flip the recommended action. A quack would never benefit from asking you to flip — so you can trust me, and promote me."

- Intuitive criterion: Cho & Kreps (1987); Neologism-proofness: Farrell (1993).

Expert might sabotage the mechanism

Timing.

Reports $\rightarrow$ Recommendation $\rightarrow$ Confession $\rightarrow$ Action, promotion

Fix a mechanism with $y < 1$.
Consider the following double deviation by the expert.

- **Sabotage:** Swap the reported signal ($L \leftrightarrow R$).
- **Confess** to the DM:

  "I am the expert and I lied in my report. Flip the recommended action. A quack would never benefit from asking you to flip — so you can trust me, and promote me."

- Intuitive criterion: Cho & Kreps (1987); Neologism-proofness: Farrell (1993).

Expert might sabotage the mechanism

Timing.

Reports $\rightarrow$ Recommendation $\rightarrow$ Confession $\rightarrow$ Action, promotion

Fix a mechanism with $y < 1$.
Consider the following double deviation by the expert.

- **Sabotage:** Swap the reported signal ($L \leftrightarrow R$).
- **Confess** to the DM:

  "I am the expert and I lied in my report. Flip the recommended action. A quack would never benefit from asking you to flip — so you can trust me, and promote me."

- Intuitive criterion: Cho & Kreps (1987); Neologism-proofness: Farrell (1993).

Expert might sabotage the mechanism

Timing.

Reports $\rightarrow$ Recommendation $\rightarrow$ Confession $\rightarrow$ Action, promotion

Fix a mechanism with $y < 1$.
Consider the following double deviation by the expert.

- **Sabotage:** Swap the reported signal ($L \leftrightarrow R$).
- **Confess** to the DM:

  "I am the expert and I lied in my report. Flip the recommended action. A quack would never benefit from asking you to flip — so you can trust me, and promote me."

- Intuitive criterion: Cho & Kreps (1987); Neologism-proofness: Farrell (1993).

Expert might sabotage the mechanism

Does it work?

If the DM believes the confession, the expert's payoff is

$$x + B \text{ instead of } x + y\,B$$

The claim "a quack would not benefit" is what we need to verify.

If the DM believes the confession, the quack maximizes his payoff by claiming to be the expert in the mechanism and lying in the confession.

$$\frac{1}{2} + B \qquad \text{vs} \qquad x + (1 - y)\,B$$

Beneficial iff $(2y - 1) \geq \frac{2x-1}{B} - 1$.

Expert might sabotage the mechanism

Does it work?

If the DM believes the confession, the expert's payoff is

$$x + B \text{ instead of } x + y\,B$$

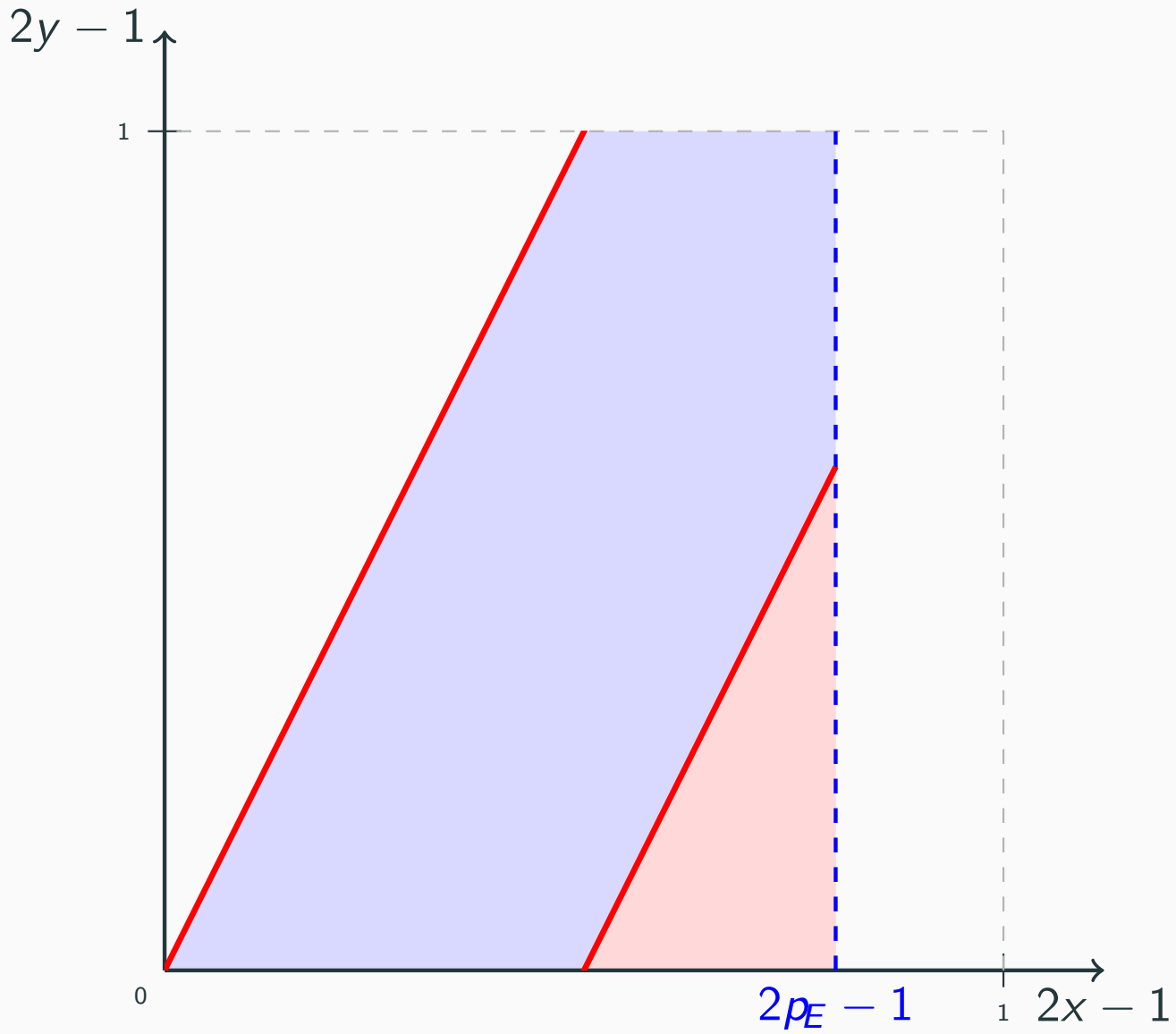
The claim "a quack would not benefit" is what we need to verify.

If the DM believes the confession, the quack maximizes his payoff by claiming to be the expert in the mechanism and lying in the confession.

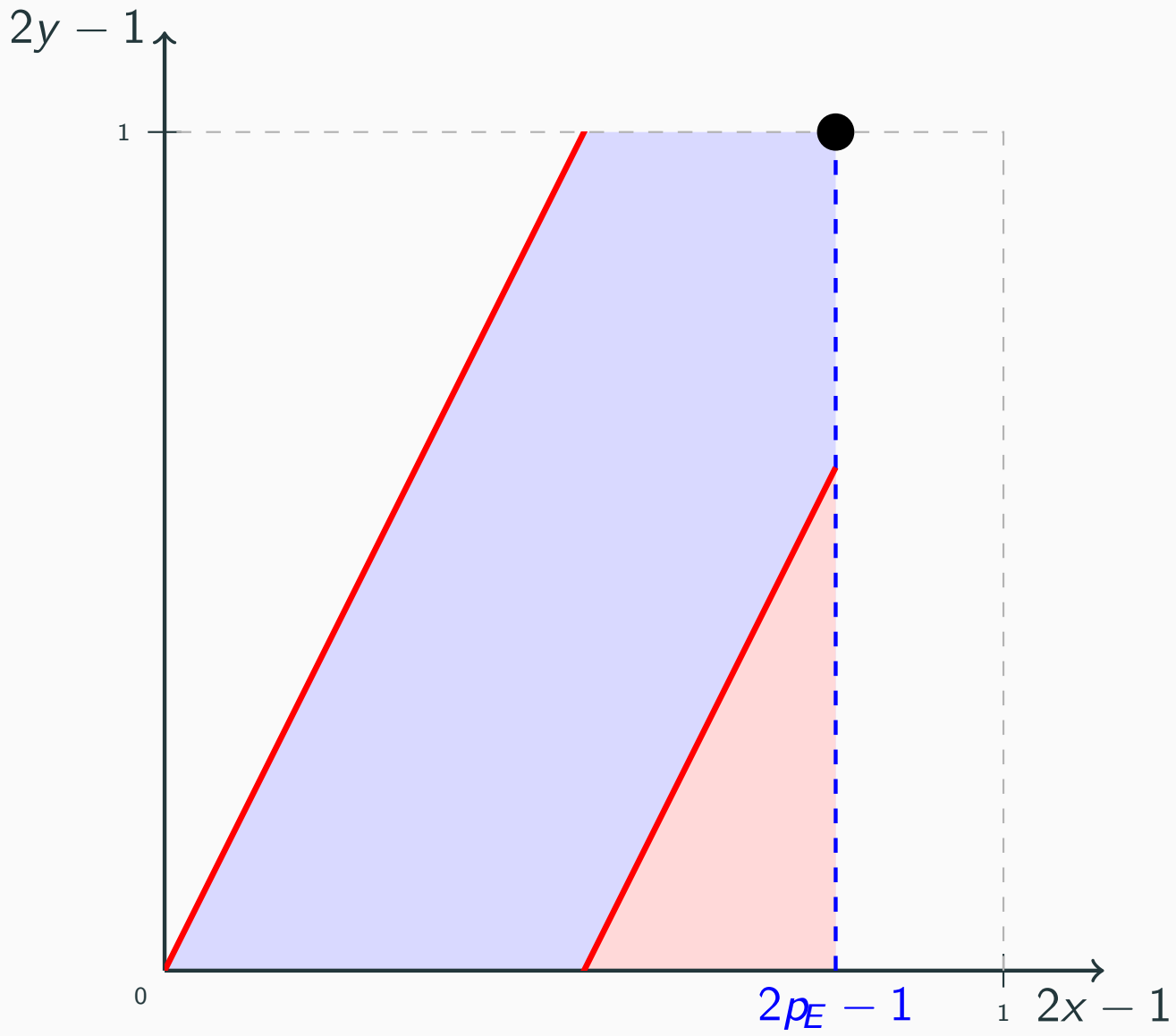
$$\frac{1}{2} + B \qquad \text{vs} \qquad x + (1 - y)\,B$$

Beneficial iff $(2y - 1) \geq \frac{2x-1}{B} - 1$.

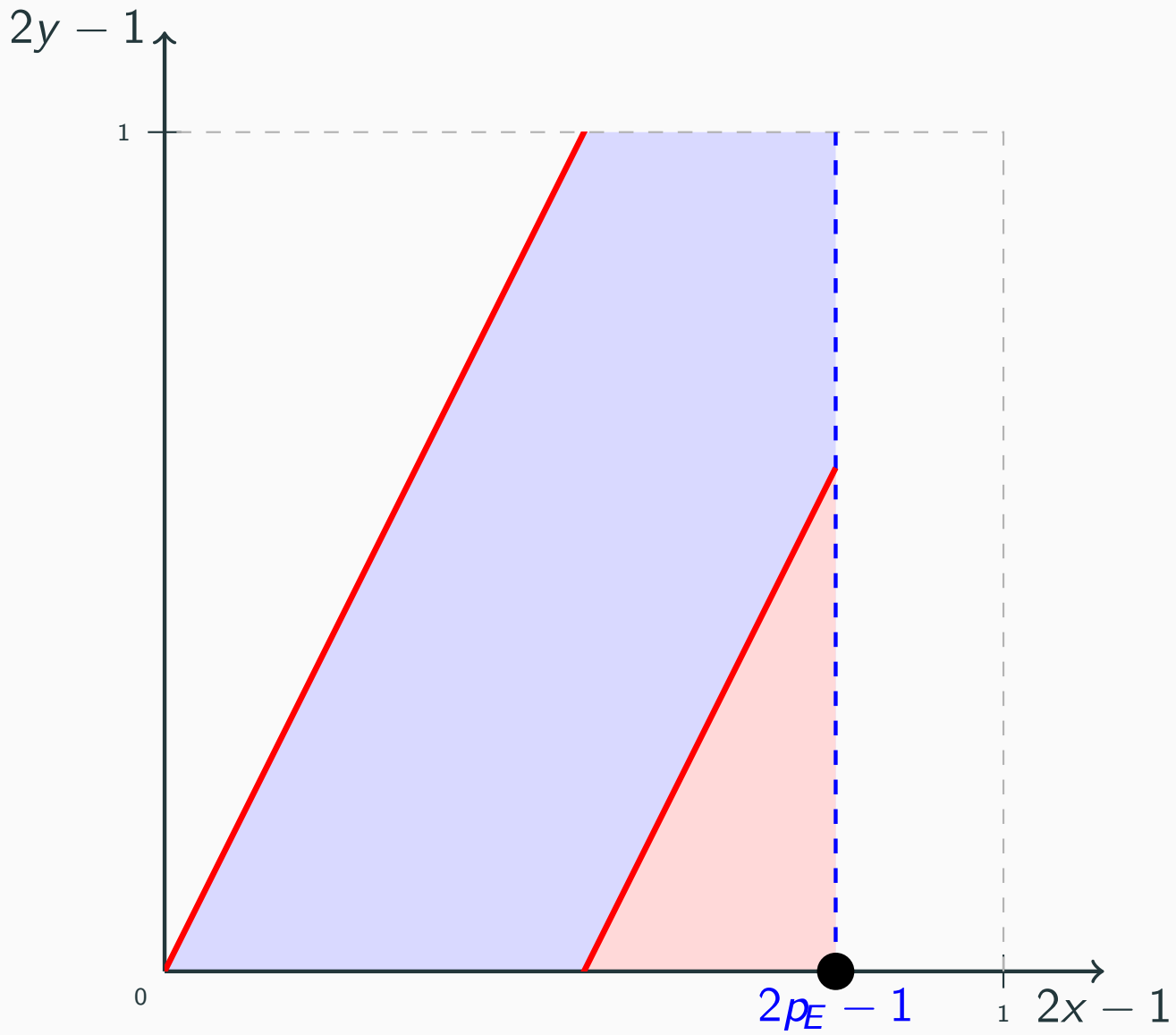
Mechanisms that are robust to off-protocol communication



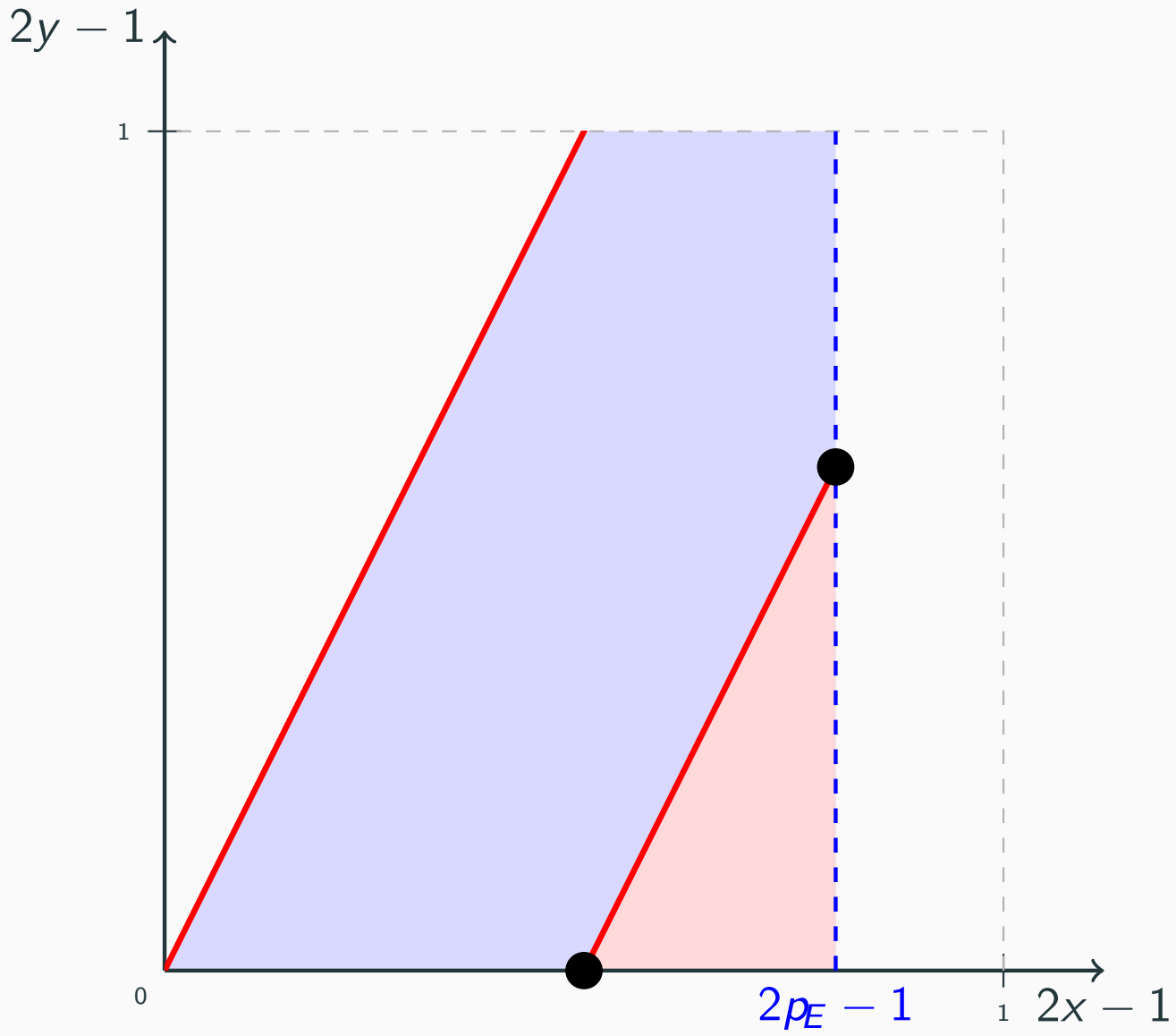
Mechanisms that are robust to off-protocol communication



Mechanisms that are robust to off-protocol communication



Mechanisms that are robust to off-protocol communication



Intuition

Designer's preferred mechanism (absent confessions): high accuracy, low promotion informativeness.

$$x = p_E, \qquad y = \tfrac{1}{2}.$$

IC holds with slack — the quack is promoted half the time regardless of reports, so has nothing to gain from deviating in the protocol.

But this mechanism is not robust. The expert's confession is maximally credible: the quack would strictly lose from sabotaging (accuracy loss exceeds promotion gain whenever B is not too large).

So the designer must distort. Two options:

- **Raise** y: make promotion informative enough that the quack, currently promoted half the time, is tempted to also sabotage — killing credibility.
- **Lower** x: reduce the accuracy cost the quack pays for sabotaging, again tempting him — again killing credibility.

Intuition

Designer's preferred mechanism (absent confessions): high accuracy, low promotion informativeness.

$$x = p_E, \qquad y = \tfrac{1}{2}.$$

IC holds with slack — the quack is promoted half the time regardless of reports, so has nothing to gain from deviating in the protocol.

But this mechanism is not robust. The expert's confession is maximally credible: the quack would strictly lose from sabotaging (accuracy loss exceeds promotion gain whenever B is not too large).

So the designer must distort. Two options:

- **Raise** y: make promotion informative enough that the quack, currently promoted half the time, is tempted to also sabotage — killing credibility.
- **Lower** x: reduce the accuracy cost the quack pays for sabotaging, again tempting him — again killing credibility.

Intuition

Designer's preferred mechanism (absent confessions): high accuracy, low promotion informativeness.

$$x = p_E, \qquad y = \tfrac{1}{2}.$$

IC holds with slack — the quack is promoted half the time regardless of reports, so has nothing to gain from deviating in the protocol.

But this mechanism is not robust. The expert's confession is maximally credible: the quack would strictly lose from sabotaging (accuracy loss exceeds promotion gain whenever B is not too large).

So the designer must distort. Two options:

- **Raise** y: make promotion informative enough that the quack, currently promoted half the time, is tempted to also sabotage — killing credibility.
- **Lower** x: reduce the accuracy cost the quack pays for sabotaging, again tempting him — again killing credibility.

General Model

Model

- $n+1$ **players**: Receiver ($i=0$) and n Senders ($i=1,...,n$).
- Senders have a private type $\theta_i \in \Theta_i$.
  - $\theta = (\theta_1, ..., \theta_n)$ is drawn from $\mu \in \Delta(\Theta)$.
- Receiver chooses action $a \in A$.
- For every player i, the final payoff is given by

$$\pi_i : A \times \Theta \to \mathbb{R}$$

- **Timing**:
  - Senders observe their private type.
  - **Communication**
  - Receiver takes action.
  - Payoffs are realized.

Model

- $n+1$ **players**: Receiver ($i=0$) and n Senders ($i=1,...,n$).
- Senders have a private type $\theta_i \in \Theta_i$.
  - $\theta = (\theta_1, ..., \theta_n)$ is drawn from $\mu \in \Delta(\Theta)$.
- Receiver chooses action $a \in A$.
- For every player i, the final payoff is given by

$$\pi_i : A \times \Theta \to \mathbb{R}$$

- **Timing**:
  - Senders observe their private type.
  - **Communication**
  - Receiver takes action.
  - Payoffs are realized.

Model

- $n+1$ **players**: Receiver ($i=0$) and n Senders ($i=1,...,n$).
- Senders have a private type $\theta_i \in \Theta_i$.
  - $\theta = (\theta_1, ..., \theta_n)$ is drawn from $\mu \in \Delta(\Theta)$.
- Receiver chooses action $a \in A$.
- For every player i, the final payoff is given by

$$\pi_i : A \times \Theta \to \mathbb{R}$$

- **Timing**:
  - Senders observe their private type.
  - **Communication**
  - Receiver takes action.
  - Payoffs are realized.

Model

- $n+1$ **players**: Receiver ($i=0$) and n Senders ($i=1,...,n$).
- Senders have a private type $\theta_i \in \Theta_i$.
  - $\theta = (\theta_1, ..., \theta_n)$ is drawn from $\mu \in \Delta(\Theta)$.
- Receiver chooses action $a \in A$.
- For every player i, the final payoff is given by

$$\pi_i : A \times \Theta \to \mathbb{R}$$

- **Timing**:
  - Senders observe their private type.
  - **Communication**
  - Receiver takes action.
  - Payoffs are realized.

Model

- $n+1$ **players**: Receiver ($i=0$) and n Senders ($i=1,...,n$).
- Senders have a private type $\theta_i \in \Theta_i$.
  - $\theta = (\theta_1, ..., \theta_n)$ is drawn from $\mu \in \Delta(\Theta)$.
- Receiver chooses action $a \in A$.
- For every player i, the final payoff is given by

$$\pi_i : A \times \Theta \to \mathbb{R}$$

- **Timing**:
  - Senders observe their private type.
  - **Communication**
  - Receiver takes action.
  - Payoffs are realized.

Protocols

Communication Protocol

A protocol $(\{M_i\}_{i=0}^N, \tau)$ is a set of messages for each agent and a function $\tau : M_1 \times M_2 ... \times M_N \to \Delta(M_0)$.

- The interpretation is that senders $1, ..., N$ simultaneously submit a message $m_i \in M_i$ to a black box that sends the receiver a message in M_0 according to $\tau(m_1, m_2, ..., m_N)$.
- A *direct revelation protocol* (DRP) is a communication protocol with $M_0 = A$ and $M_i = \Theta_i$ for $i = 1, ..., N$.
- We use $\Gamma : \Theta \to \Delta(A)$ to denote DRP.

Protocols

Communication Protocol

A protocol $(\{M_i\}_{i=0}^N, \tau)$ is a set of messages for each agent and a function $\tau : M_1 \times M_2 ... \times M_N \to \Delta(M_0)$.

- The interpretation is that senders $1, ..., N$ simultaneously submit a message $m_i \in M_i$ to a black box that sends the receiver a message in M_0 according to $\tau(m_1, m_2, ..., m_N)$.
- A *direct revelation protocol* (DRP) is a communication protocol with $M_0 = A$ and $M_i = \Theta_i$ for $i = 1, ..., N$.
- We use $\Gamma : \Theta \to \Delta(A)$ to denote DRP.

Protocols

Communication Protocol

A protocol $(\{M_i\}_{i=0}^N, \tau)$ is a set of messages for each agent and a function $\tau : M_1 \times M_2 ... \times M_N \to \Delta(M_0)$.

- The interpretation is that senders $1, ..., N$ simultaneously submit a message $m_i \in M_i$ to a black box that sends the receiver a message in M_0 according to $\tau(m_1, m_2, ..., m_N)$.
- A *direct revelation protocol* (DRP) is a communication protocol with $M_0 = A$ and $M_i = \Theta_i$ for $i = 1, ..., N$.
- We use $\Gamma : \Theta \to \Delta(A)$ to denote DRP.

Direct Revelation Mechanisms

Following Myerson (1982), we define obedience and truthfulness:

Obedience

DRP Γ is *obedient* if the Receiver finds it optimal to follow the recommendation, assuming truthful reporting.

Truthful

A DRP Γ is *truthful* if reporting truthfully is a BNE, assuming obedience.

- **Revelation Principle** [Myerson (1982)]: it is without loss to focus on truthful and obedient direct revelation protocols.

Confessions

- We now turn to analyzing off-protocol communication.
  - Fix a protocol $(\{M_i\}, \tau)$.
  - Let $P_i := \Theta_i \times M_i$.
- A *confession* for Sender i is a tuple $(T, \hat{\tau})$ where
  - $T \subseteq P_i$
  - $\hat{\tau} : M_0 \to \Delta(A)$ is a suggested action.
- Fix an equilibrium. A confession is *credible* iff, given the strategy of $-i$,
  - $\hat{\tau}$ is only profitable for type-message combinations in T (strictly for some).
  - $\hat{\tau}$ is a best response, conditional on every $t \in T$.

Definition

A protocol and equilibrium is N-P if there are no credible confessions.

Confessions

- We now turn to analyzing off-protocol communication.
  - Fix a protocol $(\{M_i\}, \tau)$.
  - Let $P_i := \Theta_i \times M_i$.
- A *confession* for Sender i is a tuple $(T, \hat{\tau})$ where
  - $T \subseteq P_i$
  - $\hat{\tau} : M_0 \to \Delta(A)$ is a suggested action.
- Fix an equilibrium. A confession is *credible* iff, given the strategy of $-i$,
  - $\hat{\tau}$ is only profitable for type-message combinations in T (strictly for some).
  - $\hat{\tau}$ is a best response, conditional on every $t \in T$.

Definition

A protocol and equilibrium is N-P if there are no credible confessions.

Confessions

- We now turn to analyzing off-protocol communication.
  - Fix a protocol $(\{M_i\}, \tau)$.
  - Let $P_i := \Theta_i \times M_i$.
- A *confession* for Sender i is a tuple $(T, \hat{\tau})$ where
  - $T \subseteq P_i$
  - $\hat{\tau} : M_0 \to \Delta(A)$ is a suggested action.
- Fix an equilibrium. A confession is *credible* iff, given the strategy of $-i$,
  - $\hat{\tau}$ is only profitable for type-message combinations in T (strictly for some).
  - $\hat{\tau}$ is a best response, conditional on every $t \in T$.

Definition

A protocol and equilibrium is N-P if there are no credible confessions.

Confessions

- We now turn to analyzing off-protocol communication.
  - Fix a protocol $(\{M_i\}, \tau)$.
  - Let $P_i := \Theta_i \times M_i$.
- A *confession* for Sender i is a tuple $(T, \hat{\tau})$ where
  - $T \subseteq P_i$
  - $\hat{\tau} : M_0 \to \Delta(A)$ is a suggested action.
- Fix an equilibrium. A confession is *credible* iff, given the strategy of $-i$,
  - $\hat{\tau}$ is only profitable for type-message combinations in T (strictly for some).
  - $\hat{\tau}$ is a best response, conditional on every $t \in T$.

Definition

A protocol and equilibrium is N-P if there are no credible confessions.

Confessions

- We now turn to analyzing off-protocol communication.
  - Fix a protocol $(\{M_i\}, \tau)$.
  - Let $P_i := \Theta_i \times M_i$.
- A *confession* for Sender i is a tuple $(T, \hat{\tau})$ where
  - $T \subseteq P_i$
  - $\hat{\tau} : M_0 \to \Delta(A)$ is a suggested action.
- Fix an equilibrium. A confession is *credible* iff, given the strategy of $-i$,
  - $\hat{\tau}$ is only profitable for type-message combinations in T (strictly for some).
  - $\hat{\tau}$ is a best response, conditional on every $t \in T$.

Definition

A protocol and equilibrium is N-P if there are no credible confessions.

General Results

Revelation Principle

Everything that can be implemented with a protocol that is N-P, can be implemented with a Direct Protocol that is Obedient, Truthful, and N-P.

Conclusion

- Protocol designers have to account for the incentives to confess deviations.

- We study the problem of designing communication protocols to elicit experts' information when
  - Experts have career concerns.
  - Experts can communicate outside of the protocol.

- We find that
  - When the career concerns are large, protocol that maximizes chances of action cannot promote the expert too often.
  - When career concerns are small, a protocol that maximizes the chances of action matching the state must promote the expert sufficiently often.

Literature

Expert advice with career concerns. Reputational incentives distort advice and aggregation: Scharfstein & Stein (1990), Prendergast & Stole (1996), Ottaviani & Sørensen (2001, 2006). *We design the protocol given these incentives.*

Mechanism design for expert information. Wolinsky (2002), Gerardi, McLean & Postlewaite (2009); mediated communication: Myerson (1986), Forges (1986). *We allow participants to speak outside the mechanism.*

Robust and credible mechanisms. Bergemann & Morris (2005); credible auctions: Akbarpour & Li (2020); renegotiation-proofness: Maskin & Moore (1999). *Robustness to what participants may say, not to their beliefs.*

Credibility of off-path messages. Intuitive criterion: Cho & Kreps (1987); credible beliefs: Grossman & Perry (1986); neologism-proofness: Farrell (1993). *Confession test is in this tradition.*