Confessions

Francisco Poggi (Mannheim)

European Workshop on Market Design
June, 2025

Introduction

- **Setting**: (multiple) senders hold private information, and a receiver must take an action.
- Governing how information flows is a powerful tool to shape outcomes.
  - Communication Protocol (mechanism): who speaks when, what messages are allowed, how these messages are processed, aggregated, etc.
- However, off-protocol communication may undermine the effectiveness of the protocol.
- **Question**: how can we design communication protocols that are *robust* to off-protocol communication?

Introduction

- **Setting**: (multiple) senders hold private information, and a receiver must take an action.
- Governing how information flows is a powerful tool to shape outcomes.
  - Communication Protocol (mechanism): who speaks when, what messages are allowed, how these messages are processed, aggregated, etc.
- However, off-protocol communication may undermine the effectiveness of the protocol.
- **Question**: how can we design communication protocols that are *robust* to off-protocol communication?

Introduction

- **Setting**: (multiple) senders hold private information, and a receiver must take an action.
- Governing how information flows is a powerful tool to shape outcomes.
    - Communication Protocol (mechanism): who speaks when, what messages are allowed, how these messages are processed, aggregated, etc.
- However, off-protocol communication may undermine the effectiveness of the protocol.
- **Question**: how can we design communication protocols that are *robust* to off-protocol communication?

Introduction

- **Setting**: (multiple) senders hold private information, and a receiver must take an action.
- Governing how information flows is a powerful tool to shape outcomes.
    - Communication Protocol (mechanism): who speaks when, what messages are allowed, how these messages are processed, aggregated, etc.
- However, off-protocol communication may undermine the effectiveness of the protocol.
- **Question**: how can we design communication protocols that are *robust* to off-protocol communication?

In this Talk

- We define a property of communication protocols that captures robustness to off-protocol communication.
    - This concept is closely related to neologism-proofness in cheap talk games.
- Discuss properties of this concept (existence, revelation principle, etc.)
- Apply it to the problem of eliciting expert information when experts have career concerns.
    - To induce efficient information revelation, protocol must be anonymous.
    - But this anonymity, may generate incentives for sabotage with off-protocol communication.
    - A robust protocol partially reveals information about expertise.

In this Talk

- We define a property of communication protocols that captures robustness to off-protocol communication.
  - This concept is closely related to neologism-proofness in cheap talk games.
- Discuss properties of this concept (existence, revelation principle, etc.)
- Apply it to the problem of eliciting expert information when experts have career concerns.
  - To induce efficient information revelation, protocol must be anonymous.
  - But this anonymity, may generate incentives for sabotage with off-protocol communication.
  - A robust protocol partially reveals information about expertise.

In this Talk

- We define a property of communication protocols that captures robustness to off-protocol communication.
    - This concept is closely related to neologism-proofness in cheap talk games.
- Discuss properties of this concept (existence, revelation principle, etc.)
- Apply it to the problem of eliciting expert information when experts have career concerns.
    - To induce efficient information revelation, protocol must be anonymous.
    - But this anonymity, may generate incentives for sabotage with off-protocol communication.
    - A robust protocol partially reveals information about expertise.

In this Talk

- We define a property of communication protocols that captures robustness to off-protocol communication.
  - This concept is closely related to neologism-proofness in cheap talk games.
- Discuss properties of this concept (existence, revelation principle, etc.)
- Apply it to the problem of eliciting expert information when experts have career concerns.
  - To induce efficient information revelation, protocol must be anonymous.
  - But this anonymity, may generate incentives for sabotage with off-protocol communication.
  - A robust protocol partially reveals information about expertise.

Example 1

- Single Sender observes state $\theta \in \{L, R\}$ uniform.
- Receiver must choose an action $a \in \{l, r, \text{safe}\}$.

Payoff matrix

Action	$\theta = L$	$\theta = R$
l	$(3,3)$	$(0,0)$
r	$(0,0)$	$(3,3)$
safe	$(2,2)$	$(2,2)$

- **Cheap talk**: Two equilibria, informative and babbling.
- **Protocol**: partial revelation can be induced.
  - E.g.: Sender reports θ and Receiver observes a signal of the report with $2/3$ precision.
- Neologism: "The state is left."

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 $\tau(m_1, m_1, ..., m_N)$.
- 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 $\tau(m_1, m_1, ..., m_N)$.
- 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.

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 $\tau(m_1, m_1, ..., m_N)$.
- 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, τ) where
  - $T \subseteq P_i$
  - $\tau : M_0 \to \Delta(A)$ is a suggested transformation.
- Fix an equilibrium. A confession is *credible* iff, given the strategy of $-i$,
  - the transformation τ is only profitable for type-message combinations in T (strictly for some).
  - transformation τ 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, τ) where
  - $T \subseteq P_i$
  - $\tau : M_0 \to \Delta(A)$ is a suggested transformation.
- Fix an equilibrium. A confession is *credible* iff, given the strategy of $-i$,
  - the transformation τ is only profitable for type-message combinations in T (strictly for some).
  - transformation τ 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, τ) where
  - $T \subseteq P_i$
  - $\tau : M_0 \to \Delta(A)$ is a suggested transformation.
- Fix an equilibrium. A confession is *credible* iff, given the strategy of $-i$,
  - the transformation τ is only profitable for type-message combinations in T (strictly for some).
  - transformation τ 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, τ) where
  - $T \subseteq P_i$
  - $\tau : M_0 \to \Delta(A)$ is a suggested transformation.
- Fix an equilibrium. A confession is *credible* iff, given the strategy of $-i$,
  - the transformation τ is only profitable for type-message combinations in T (strictly for some).
  - transformation τ 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, τ) where
  - $T \subseteq P_i$
  - $\tau : M_0 \to \Delta(A)$ is a suggested transformation.
- Fix an equilibrium. A confession is *credible* iff, given the strategy of $-i$,
  - the transformation τ is only profitable for type-message combinations in T (strictly for some).
  - transformation τ is a best response, conditional on every $t \in T$.

Definition

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

Back to Example 1

Payoff matrix

Action	$\theta = L$	$\theta = R$
l	$(3, 3)$	$(0, 0)$
r	$(0, 0)$	$(3, 3)$
$safe$	$(2, 2)$	$(2, 2)$

- Consider the protocol with 2/3 precision.
- Credible confession: $(\{(A, A), (A, B)\}, \tau)$ where $\tau(\hat{a}) = A$.
- The only protocols that are neologism-proof are perfectly revealing.

Example 2: No Neologism Proof Cheap-Talk Equilibrium

- Single Sender observes state $\theta \in \{L, R\}$ uniform.
- Receiver must choose an action $a \in \{l, r, \text{safe}\}$. Same payoffs as Example 1.

Sender's payoff

Action	$\theta = L$	$\theta = R$
l	2	1
r	-1	0
safe	0	2

- There is no cheap talk equilibrium that is neologism proof.
- However, we can construct a neologism-proof protocol.

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.

Existence

Whenever there is a N-P Cheap Talk Equilibrium, there is a N-P Protocol.

- Any cheap talk communication can be replicated within the protocol.

Application: Experts with Career Concerns

- Payoff-relevant state: $\omega \in \{\text{Left}, \text{Right}\}$.
- Two senders (experts) observe conditionally independent signals $s_i \in \{\text{Left}, \text{Right}\}$.
- *Expertise* of the experts:
  - **Good expert**: Signal matches the state with probability $q_H > 1/2$.
  - **Bad expert**: Signal matches the state with probability $q_L \in (1/2, q_H)$.
  - There is exactly one good expert.
  - The identity $e \in \{1, 2\}$ of the good expert is unknown to the receiver.
  - Ex-ante identical probabilities.
- The Receiver must:
  - take an action: $a_0 \in \{\text{Left}, \text{Right}\}$
  - nominate one of the experts for a promotion: $m \in \{1, 2\}$.

Application: Experts with Career Concerns

- Payoff-relevant state: $\omega \in \{\text{Left}, \text{Right}\}$.
- Two senders (experts) observe conditionally independent signals $s_i \in \{\text{Left}, \text{Right}\}$.
- *Expertise* of the experts:
    - **Good expert**: Signal matches the state with probability $q_H > 1/2$.
    - **Bad expert**: Signal matches the state with probability $q_L \in (1/2, q_H)$.
    - There is exactly one good expert.
    - The identity $e \in \{1, 2\}$ of the good expert is unknown to the receiver.
    - Ex-ante identical probabilities.
- The Receiver must:
    - take an action: $a_0 \in \{\text{Left}, \text{Right}\}$
    - nominate one of the experts for a promotion: $m \in \{1, 2\}$.

Application: Experts with Career Concerns

- Payoff-relevant state: $\omega \in \{\text{Left}, \text{Right}\}$.
- Two senders (experts) observe conditionally independent signals $s_i \in \{\text{Left}, \text{Right}\}$.
- *Expertise* of the experts:
  - **Good expert**: Signal matches the state with probability $q_H > 1/2$.
  - **Bad expert**: Signal matches the state with probability $q_L \in (1/2, q_H)$.
  - There is exactly one good expert.
  - The identity $e \in \{1, 2\}$ of the good expert is unknown to the receiver.
  - Ex-ante identical probabilities.
- The Receiver must:
  - take an action: $a_0 \in \{\text{Left}, \text{Right}\}$
  - nominate one of the experts for a promotion: $m \in \{1, 2\}$.

Application: Experts with Career Concerns

- Payoff-relevant state: $\omega \in \{\text{Left}, \text{Right}\}$.
- Two senders (experts) observe conditionally independent signals $s_i \in \{\text{Left}, \text{Right}\}$.
- *Expertise* of the experts:
  - **Good expert**: Signal matches the state with probability $q_H > 1/2$.
  - **Bad expert**: Signal matches the state with probability $q_L \in (1/2, q_H)$.
  - There is exactly one good expert.
  - The identity $e \in \{1, 2\}$ of the good expert is unknown to the receiver.
  - Ex-ante identical probabilities.

- The Receiver must:
  - take an action: $a_0 \in \{\text{Left}, \text{Right}\}$
  - nominate one of the experts for a promotion: $m \in \{1, 2\}$.

Application: Experts with Career Concerns

- Payoff-relevant state: $\omega \in \{\text{Left}, \text{Right}\}$.
- Two senders (experts) observe conditionally independent signals $s_i \in \{\text{Left}, \text{Right}\}$.
- *Expertise* of the experts:
  - **Good expert**: Signal matches the state with probability $q_H > 1/2$.
  - **Bad expert**: Signal matches the state with probability $q_L \in (1/2, q_H)$.
  - There is exactly one good expert.
  - The identity $e \in \{1, 2\}$ of the good expert is unknown to the receiver.
  - Ex-ante identical probabilities.
- The Receiver must:
  - take an action: $a_0 \in \{\text{Left}, \text{Right}\}$
  - nominate one of the experts for a promotion: $m \in \{1, 2\}$.

Application: Experts with Career Concerns

- Payoff-relevant state: $\omega \in \{\text{Left}, \text{Right}\}$.
- Two senders (experts) observe conditionally independent signals $s_i \in \{\text{Left}, \text{Right}\}$.
- *Expertise* of the experts:
    - **Good expert**: Signal matches the state with probability $q_H > 1/2$.
    - **Bad expert**: Signal matches the state with probability $q_L \in (1/2, q_H)$.
    - There is exactly one good expert.
    - The identity $e \in \{1, 2\}$ of the good expert is unknown to the receiver.
    - Ex-ante identical probabilities.

- The Receiver must:
    - take an action: $a_0 \in \{\text{Left}, \text{Right}\}$
    - nominate one of the experts for a promotion: $m \in \{1, 2\}$.

Application: Payoffs

- Experts and DM want the action to match the state.
- The expert that is promoted obtains a bonus B.
- The DM prefers to promote the good expert.

$$\pi_0 = 1_{\{a=\omega\}} \;+\; C \cdot 1_{\{m=e\}}$$
$$\pi_i = 1_{\{a=\omega\}} \;+\; B \cdot 1_{\{m=i\}}$$

Application: Payoffs

- Experts and DM want the action to match the state.
- The expert that is promoted obtains a bonus B.
- The DM prefers to promote the good expert.

$$\pi_0 = 1_{\{a=\omega\}} \; + \; C \cdot 1_{\{m=e\}}$$
$$\pi_i = 1_{\{a=\omega\}} \; + \; B \cdot 1_{\{m=i\}}$$

Communication Design to Match the State

- Optimal action given experts' information: s_θ.
  - Probability of matching the state with optimal action: q_H.
- Is it possible to have a communication protocol that implements this action?
- **DRP**:
  - experts report their information $\hat{\theta}_i = (\hat{e}_i, \hat{s}_i)$
  - the protocol recommends an action $\hat{a} = (\hat{a}_0, \hat{m})$.

Communication Design to Match the State

- Optimal action given experts' information: s_θ.
  - Probability of matching the state with optimal action: q_H.
- Is it possible to have a communication protocol that implements this action?

- **DRP**:
  - experts report their information $\hat{\theta}_i = (\hat{e}_i, \hat{s}_i)$
  - the protocol recommends an action $\hat{a} = (\hat{a}_0, \hat{m})$.

Communication Design to Match the State

- Optimal action given experts' information: s_θ.
  - Probability of matching the state with optimal action: q_H.
- Is it possible to have a communication protocol that implements this action?

- **DRP**:
  - experts report their information $\hat{\theta}_i = (\hat{e}_i, \hat{s}_i)$
  - the protocol recommends an action $\hat{a} = (\hat{a}_0, \hat{m})$.

Family of DRP and IC

- Consider the following family of DRP:
  - If the reports coincide on identity of the good expert, i.e., $\hat{\theta}_1 = \hat{\theta}_2 = \hat{\theta}$,
    - protocol recommends action $\hat{s}_{\hat{\theta}}$.
    - protocol recommends to promote $\hat{\theta}$ with probability $g \geq 1/2$.
  - If reports do not coincide, protocol recommends
    - An action equal to the reported signals when these coincide.
    - a random action when reported signals don't coincide.
    - a (uniformly) random promotion.

- Let p_L be the probability of matching the state when signals are aggregated with the same weight. The bad expert reports truthfully if

$$q_H + (1-g) \cdot B \geq p_L + \frac{1}{2} \cdot B \qquad \Rightarrow \qquad g \leq \frac{1}{2} + \frac{q_H - p_L}{B}$$

Family of DRP and IC

- Consider the following family of DRP:
  - If the reports coincide on identity of the good expert, i.e., $\hat{\theta}_1 = \hat{\theta}_2 = \hat{\theta}$,
    - protocol recommends action $\hat{s}_{\hat{\theta}}$.
    - protocol recommends to promote $\hat{\theta}$ with probability $g \geq 1/2$.
  - If reports do not coincide, protocol recommends
    - An action equal to the reported signals when these coincide.
    - a random action when reported signals don't coincide.
    - a (uniformly) random promotion.
- Let p_L be the probability of matching the state when signals are aggregated with the same weight. The bad expert reports truthfully if

$$q_H + (1-g) \cdot B \geq p_L + \frac{1}{2} \cdot B \qquad \Rightarrow \qquad g \leq \frac{1}{2} + \frac{q_H - p_L}{B}$$

Family of DRP and IC

- Consider the following family of DRP:
  - If the reports coincide on identity of the good expert, i.e., $\hat{\theta}_1 = \hat{\theta}_2 = \hat{\theta}$,
    - protocol recommends action $\hat{s}_{\hat{\theta}}$.
    - protocol recommends to promote $\hat{\theta}$ with probability $g \geq 1/2$.
  - If reports do not coincide, protocol recommends
    - An action equal to the reported signals when these coincide.
    - a random action when reported signals don't coincide.
    - a (uniformly) random promotion.

- Let p_L be the probability of matching the state when signals are aggregated with the same weight. The bad expert reports truthfully if

$$q_H + (1-g) \cdot B \geq p_L + \frac{1}{2} \cdot B \qquad \Rightarrow \qquad g \leq \frac{1}{2} + \frac{q_H - p_L}{B}$$

Family of DRP and IC

- Consider the following family of DRP:
  - If the reports coincide on identity of the good expert, i.e., $\hat{\theta}_1 = \hat{\theta}_2 = \hat{\theta}$,
    - protocol recommends action $\hat{s}_{\hat{\theta}}$.
    - protocol recommends to promote $\hat{\theta}$ with probability $g \geq 1/2$.
  - If reports do not coincide, protocol recommends
    - An action equal to the reported signals when these coincide.
    - a random action when reported signals don't coincide.
    - a (uniformly) random promotion.

- Let p_L be the probability of matching the state when signals are aggregated with the same weight. The bad expert reports truthfully if

$$q_H + (1-g) \cdot B \geq p_L + \frac{1}{2} \cdot B \qquad \Rightarrow \qquad g \leq \frac{1}{2} + \frac{q_H - p_L}{B}$$

Good Expert Might Sabotage the Mechanism

Consider the following deviation of the good expert i which obtains $s_i = \text{Left}$.

- Instead of sending report (i, Left), he sends (i, Right).
- He approaches the DM and says:
    - "I am the good expert and gave an incorrect report, thus you should not follow the recommended action from the mechanism.
    - You should instead do the opposite of what the mechanism recommends.
    - By the way, there is no way that the bad expert benefits from you switching the action, so you should trust that I'm the good expert.
    - Therefore, you should give me that promotion"

Good Expert Might Sabotage the Mechanism

Consider the following deviation of the good expert i which obtains $s_i = \text{Left}$.

- Instead of sending report (i, Left), he sends (i, Right).
- He approaches the DM and says:
  - "I am the good expert and gave an incorrect report, thus you should not follow the recommended action from the mechanism.
  - You should instead do the opposite of what the mechanism recommends.
  - By the way, there is no way that the bad expert benefits from you switching the action, so you should trust that I'm the good expert.
  - Therefore, you should give me that promotion"

Good Expert Might Sabotage the Mechanism

- Is it beneficial for the good expert? **YES.**

$$q_H + B \qquad \text{vs} \qquad q_H + g \cdot B$$

- Is it be beneficial for the bad expert? **Depends.**

$$(1 - \underline{p}) + B \qquad \text{vs} \qquad q_H + (1 - g) \cdot B$$

- $\underline{p}$: minimum probability of having the mechanism recommend the correct action that can be induced by the bad expert.
- Beneficial iff $g > \frac{q_H - (1 - \underline{p})}{B}$

Good Expert Might Sabotage the Mechanism

- Is it beneficial for the good expert? **YES.**

$$q_H + B \qquad \text{vs} \qquad q_H + g \cdot B$$

- Is it be beneficial for the bad expert? **Depends.**

$$(1 - \underline{p}) + B \qquad \text{vs} \qquad q_H + (1 - g) \cdot B$$

- $\underline{p}$: minimum probability of having the mechanism recommend the correct action that can be induced by the bad expert.
- Beneficial iff $g > \frac{q_H - (1 - \underline{p})}{B}$

When can the DM take optimal actions

Proposition

The optimal recommendation can be implemented with a NP mechanism iff the career concerns are not too high: $B \leq 2 \cdot (p_L - (1 - \underline{p}))$.

- We just need that it exists a g small enough so that bad expert want to report his expertise truthfully, but high enough so that bad expert would like to sabotage the mechanism.

$$\frac{q_H - (1 - \underline{p})}{B} > \frac{1}{2} + \frac{q_H - p_L}{B}$$

When can the DM take optimal actions

Proposition

The optimal recommendation can be implemented with a NP mechanism iff the career concerns are not too high: $B \leq 2 \cdot (p_L - (1 - \underline{p}))$.

- We just need that it exists a g small enough so that bad expert want to report his expertise truthfully, but high enough so that bad expert would like to sabotage the mechanism.

$$\frac{q_H - (1 - \underline{p})}{B} > \frac{1}{2} + \frac{q_H - p_L}{B}$$

Conclusion

- When mechanism participants have a common language, mechanism 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 share a common language with the DM and can communicate outside of the mechanism.
- We find that
  - To induce an optimal action, a mechanism must aggregate the experts' recommendations, and not promote the good expert too often. However, good experts have incentives to sabotage the mechanism in an attempt to signal their type.
  - When career concerns are sufficiently high, it is not possible to implement the optimal action in a way that is robust to off-protocol communication.