Motivated Misspecification

Mingzi Niu

Department of Economics
Rice University

February 21, 2024

Motivation

How does a project manager motivate a worker to exert effort?

Motivation

How does a project manager motivate a worker to exert effort?

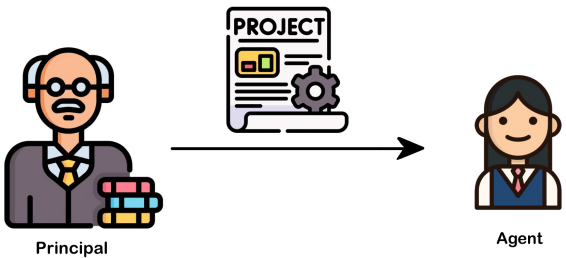
- Monetary incentives (e.g., Holmström 1979)
  - give a bonus for good performance
- Informational incentives (e.g., Ely & Szydlowski 2020)
  - reveal information over time to induce incremental effort

Motivation

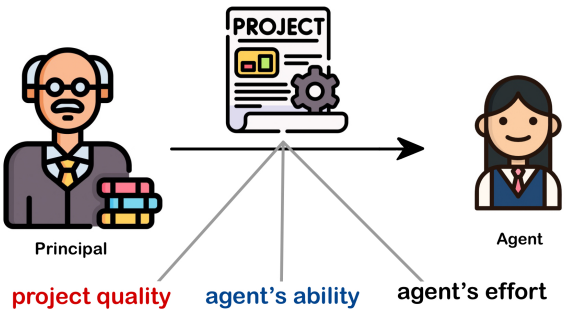
How does a project manager motivate a worker to exert effort?

- Monetary incentives (e.g., Holmström 1979)
  - give a bonus for good performance
- Informational incentives (e.g., Ely & Szydlowski 2020)
  - reveal information over time to induce incremental effort
- This paper: **Perception Manipulation**
  - influence how to interpret observations

Motivation

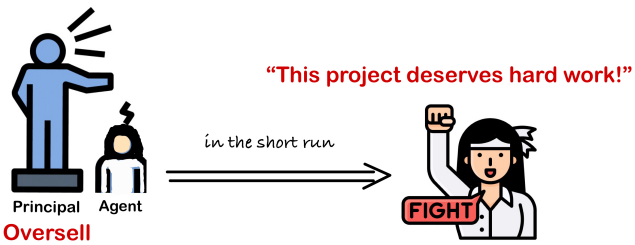


Motivation


PROJECT
Principal
Agent
project quality
agent's ability
agent's effort

Key Tension

- Initially, high expectation inflates perceived return to effort $\Rightarrow$ stimulates effort

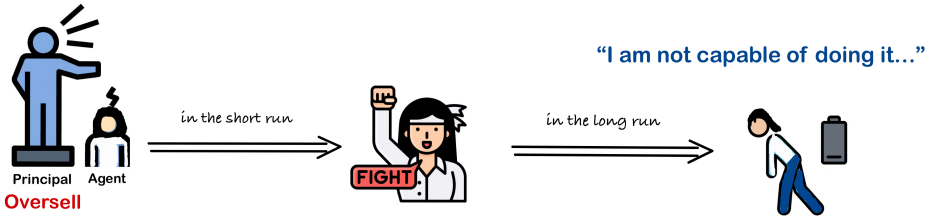


Key Tension

- Initially, high expectation **stimulates effort** by inflating perceived return to effort
- However, high expectation can **backfire** in the long run by triggering misinference
  - over time, agent notices outputs are not as high as expected

  $\Rightarrow$ agent thus doubts her own ability to do this job

  $\Rightarrow$ potentially demotivate agent's effort



Question

How should the principal set the agent's expectations in the first place?

Question

How should the principal set the agent's expectations in the first place?

- Perception: learning model to interpret observations
- Misperception: learning under a misspecified model
- Principal's manipulation strategy determines agent's perception
- **What type of misperception would arise in a certain environment?**

Main Results

- If principal is sufficiently impatient, he oversells project quality

Main Results

- If principal is sufficiently impatient, he oversells project quality
- Otherwise, output function matters
  1. production features sufficient statistic: manipulation cannot affect long-run effort
  2. ability-laden production: principal oversells $\Rightarrow$ makes agent overly optimistic
  3. quality-laden production: principal undersells $\Rightarrow$ makes agent overly pessimistic

Related Literature: Misspecified Learning

- *Original concepts*: Nyarko (1991), **Esponda & Pouzo (2016)**
- *Techniques to analyze asymptotic properties*: Fudenberg, Romanyuk & Strack (2017), Bohren & Hauser (2021), Esponda & Pouzo (2021), Heidhues, Kőszegi & Strack (2021), Esponda, Pouzo & Yamamoto (2021), Fudenberg, Lanzani & Strack (2021,2023), Frick, Iijima & Ishii (2023)
- *Implications of misspecified learning*: Andreoni & Mylovanov (2012), Bohren (2016), **Heidhues, Kőszegi & Strack**(**2018**, 2023), Frick, Iijima & Ishii (2020), Ba & Gindin (2023), Gagnon-Bartsch & Bushong (2022), He (2022), Kőszegi, Loewenstein & Murooka (2022), Levy, Razin & Young (2022)
- *Model selection*: He & Libgober (2020), Montiel Olea, Ortoleva, Pai & Prat (2022), Ba (2023), Fudenberg & Lanzani (2023)

Related Literature

- **Misspecified learning** (e.g., Esponda & Pouzo 2016)
  - individuals update their beliefs from observations using a misspecified model

Related Literature

- **Misspecified learning** (e.g., Esponda & Pouzo 2016)
  - individuals update their beliefs from observations using a misspecified model

→ Endogenize model misspecification

Related Literature

- **Misspecified learning** (e.g., Esponda & Pouzo 2016)
  - individuals update their beliefs from observations using a misspecified model

→ Endogenize model misspecification

- **Motivated belief** (e.g., Bénabou & Tirole 2016)
  - individual's belief results from a tradeoff between accuracy and preference

- **Bayesian persuasion** (e.g., Kamenica & Gentzkow 2011)
  - sender controls d.g.p and induced posteriors must follow Bayesian plausibility

Related Literature

- **Misspecified learning** (e.g., Esponda & Pouzo 2016)
  - individuals update their beliefs from observations using a misspecified model

→ Endogenize model misspecification

- **Motivated belief** (e.g., Bénabou & Tirole 2016)
  - individual's belief results from a tradeoff between accuracy and preference

→ Manipulate belief in an interpersonal setting

- **Bayesian persuasion** (e.g., Kamenica & Gentzkow 2011)
  - sender controls d.g.p and induced posteriors must follow Bayesian plausibility

Related Literature

- **Misspecified learning** (e.g., Esponda & Pouzo 2016)
  - individuals update their beliefs from observations using a misspecified model

→ Endogenize model misspecification

- **Motivated belief** (e.g., Bénabou & Tirole 2016)
  - individual's belief results from a tradeoff between accuracy and preference

→ Manipulate belief in an interpersonal setting

- **Bayesian persuasion** (e.g., Kamenica & Gentzkow 2011)
  - sender controls d.g.p and induced posteriors must follow Bayesian plausibility

→ Frame the learning model (i.e., how to interpret observations)

Agenda

1. Model
2. Immediate Effect and Long-Run Effect
3. Applications
   - Mentorship
   - Abusive Relationship
4. Extensions
   - Sophisticated Agent
   - Short-Run Incentives

Agenda

1. Model
2. Immediate Effect and Long-Run Effect
3. Applications
   - Mentorship
   - Abusive Relationship
4. Extensions
   - Sophisticated Agent
   - Short-Run Incentives

Model

- Time horizon: $t = 0, 1, 2, 3 \ldots$
- Players: Principal & Agent
- True project **quality** is $Q \in [\underline{q}, \bar{q}]$; true agent's **ability** is $A \in \mathbb{R}$
- Principal: influences agent's perceived project quality $\tilde{q} \in [\underline{q}, \bar{q}]$ at $t = 0$
  - principal is *truth-telling* if $\tilde{q} = Q$
  - principal is *overselling* if $\tilde{q} > Q$
  - principal is *underselling* if $\tilde{q} < Q$
- Agent: exerts **effort** $e_t \geq 0$ in each period

Information Structure

- Common prior belief on agent's ability π_0
- Only principal knows the true project quality Q
- Project output at time t: $y_t = Y(\boldsymbol{e_t}, \boldsymbol{Q}, \boldsymbol{A}) + \epsilon_t$
  - $\boldsymbol{e_t} \geq 0$: agent's **effort** exerted at time t
  - $\boldsymbol{Q} \in [\underline{q}, \bar{q}]$: true project **quality**
  - $\boldsymbol{A} \in \mathbb{R}$: true agent's **ability**
  - $\epsilon_t \sim N(0, \sigma_\epsilon^2)$: random noise i.i.d. across periods

Information Structure

- Common prior belief on agent's ability π_0
- Only principal knows the true project quality Q
- Project output at time t: $y_t = Y(\boldsymbol{e_t}, \boldsymbol{Q}, \boldsymbol{A}) + \epsilon_t$
  - $\boldsymbol{e_t} \geq 0$: agent's **effort** exerted at time t
  - $\boldsymbol{Q} \in [\underline{q}, \bar{q}]$: true project **quality**
  - $\boldsymbol{A} \in \mathbb{R}$: true agent's **ability**
  - $\epsilon_t \sim N(0, \sigma_\epsilon^2)$: random noise i.i.d. across periods
- Agent's learning dynamics: learns only about her own ability
  - fixes her perceived project **quality** at $\tilde{q}$
  - learns about her own **ability** upon output observations in a Bayesian manner

Stage-Game Payoffs

- Principal: $u_t^P = y_t$
- Agent: $u_t^A = y_t - c(e_t)$
  - $c(e_t) \geq 0$: the cost of effort

Assumptions: $Y(\cdot), c(\cdot)$ are twice continuously differentiable, and

1. three contributors to output: $Y_e, Y_q, Y_a > 0$
2. decreasing marginal product and increasing marginal cost of effort: $Y_{ee} \leq 0, c_{ee} > 0$
3. mean output varies flexibly: $\lim_{a \to -\infty} Y(e, q, a) = -\infty, \lim_{a \to \infty} Y(e, q, a) = \infty$
4. **effort and project quality are complements**: $Y_{eq} > 0$

Agent's Payoff and Strategy

Myopic agent

$$U_t^A(e_t; \tilde{q}, \pi_{t-1}) = E[Y(e_t, \tilde{q}, a)|\pi_{t-1}] - c(e_t)$$

$e(\tilde{q}, \pi)$: agent's effort strategy

- perceived project quality is $\tilde{q}$
- belief on her own ability is π

Principal's Payoff and Strategy

Forward-looking principal

Technical difficulty in tracking down effort trajectories over periods

Principal's Payoff and Strategy

Forward-looking principal

Technical difficulty in tracking down effort trajectories over periods

1. $e_1 = e(\tilde{q}, \pi_0) \Rightarrow y_1 | e_1 \sim N(Y(e_1, Q, A), \sigma_\epsilon^2) \Rightarrow \pi_1$

Principal's Payoff and Strategy

Forward-looking principal

Technical difficulty in tracking down effort trajectories over periods

1. $e_1 = e(\tilde{q}, \pi_0) \Rightarrow y_1 | e_1 \sim N(Y(e_1, Q, A), \sigma_\epsilon^2) \Rightarrow \pi_1$
2. $e_2 = e(\tilde{q}, \pi_1) \Rightarrow y_2 | e_2 \sim N(Y(e_2, Q, A), \sigma_\epsilon^2) \Rightarrow \pi_2$

Principal's Payoff and Strategy

Forward-looking principal

Technical difficulty in tracking down effort trajectories over periods

1. $e_1 = e(\tilde{q}, \pi_0) \Rightarrow y_1 | e_1 \sim N(Y(e_1, Q, A), \sigma_\epsilon^2) \Rightarrow \pi_1$
2. $e_2 = e(\tilde{q}, \pi_1) \Rightarrow y_2 | e_2 \sim N(Y(e_2, Q, A), \sigma_\epsilon^2) \Rightarrow \pi_2$
3. $e_3 = e(\tilde{q}, \pi_2) \Rightarrow \cdots$

Principal's Payoff and Strategy

For tractability

- Principal weighs the immediate and long-run payoff

$$U^P(\tilde{q}; Q, \pi_0, \gamma) = (1-\gamma)U_1 + \gamma U_\infty$$

  - immediate payoff: $U_1 = E[Y(e_1, Q, a)|\pi_0]$
  - long-run payoff: $U_\infty = E[Y(e_\infty, Q, a)|\pi_0]$
    - long-run (stable) state: **Berk-Nash Equilibrium** (*Esponda & Pouzo, 2016*)
  - $\gamma \in [0, 1]$: weight assigned to long-run payoff

- $\tilde{q} \in [\underline{q}, \bar{q}]$: principal's manipulation strategy

Long-Run Stable State: Berk-Nash Equilibrium

Definition (Berk-Nash Equilibrium)

A pair of effort strategy and belief on ability (e_∞, π_∞) is a Berk-Nash eqm if

1. $e_\infty = e^*(\tilde{q}, \pi_\infty)$
2. $\pi_\infty \in \Delta(\Theta(e_\infty))$ where $\Theta(e_\infty) = \text{argmin}_{a' \in \mathbb{R}} K(f, f_{a'})$

1. Effort e_∞ is optimal under belief on ability π_∞ and perceived project quality $\tilde{q}$
2. Belief π_∞ assigns probability 1 among ability levels that yield a perceived output distribution **"closest"** to the true output distribution
   - true output distribution: $f(y|e_\infty) \sim N(Y(e_\infty, Q, A), \sigma_\epsilon^2)$
   - perceived output distribution: $f_{a'}(y|e_\infty) \sim N(Y(e_\infty, \tilde{q}, a'), \sigma_\epsilon^2)$
   - **"distance"** between distributions is measured by Kullback-Leibler divergence

Equilibrium of the Manipulation Game

A pair of manipulation strategy and effort strategy $(\tilde{q}^*, e^*)$ such that

- Principal chooses agent's perceived project quality

$$\tilde{q}^* \in \underset{q \in [\underline{q}, \bar{q}]}{\operatorname{argmax}} (1-\gamma) E[Y(e_1, Q, a)|\pi_0] + \gamma E[Y(e_\infty, Q, a)|\pi_0]$$

- Agent exerts her myopically optimal effort

$$e^*(\tilde{q}, \pi_{t-1}) \in \underset{e \geq 0}{\operatorname{argmax}} E[Y(e_t, \tilde{q}, a)|\pi_{t-1}] - c(e_t),$$

given her perceived project quality $\tilde{q}$ and belief π_{t-1} about ability at time t

Agenda

1. Model
2. Immediate Effect and Long-Run Effect
3. Applications
   - Mentorship
   - Abusive Relationship
4. Extensions
   - Sophisticated Agent
   - Short-Run Incentives

Benchmark

Complete info: agent knows true project quality Q and her true ability A

$$e^{FB} = e^*(Q, A)$$

Benchmark

Complete info: agent knows true project quality Q and her true ability A

$$e^{FB} = e^*(Q, A)$$

Proposition (Correct Learning)

If principal is truth-telling, then agent eventually

- learns her own ability correctly
- exerts her first-best effort

Benchmark

Complete info: agent knows true project quality Q and her true ability A

$$e^{FB} = e^*(Q, A)$$

If principal is truth-telling, then stable state replicates the benchmark

Proposition (Correct Learning)

For any true state, prior belief, signal precision $(Q, A, \pi_0, \sigma_\epsilon^2) \in [\underline{q}, \bar{q}] \times \mathbb{R} \times \Delta(\mathbb{R}) \times \mathbb{R}_{++}$, if $\tilde{q} = Q$, then $\pi_\infty = \delta_A$ and $e_\infty = e^{FB}$

Benchmark

Complete info: agent knows true project quality Q and her true ability A

$$e^{FB} = e^*(Q, A)$$

If principal is truth-telling, then stable state replicates the benchmark

Proposition (Correct Learning)

For any true state, prior belief, signal precision $(Q, A, \pi_0, \sigma_\epsilon^2) \in [\underline{q}, \bar{q}] \times \mathbb{R} \times \Delta(\mathbb{R}) \times \mathbb{R}_{++}$, if $\tilde{q} = Q$, then $\pi_\infty = \delta_A$ and $e_\infty = e^{FB}$

Can principal induce higher effort by manipulating perceived project quality?

Preview of Main Results

A threshold on the weight that principal assigns to long-run welfare: $\hat{\gamma} \in [0, 1]$

- Principal is impatient ($\gamma < \hat{\gamma}$): overselling
- Principal is sufficiently patient ($\gamma > \hat{\gamma}$): depend on output function $Y(\cdot)$

This holds true for any true state (Q, A), prior belief π_0, signal precision σ_ϵ^2

Immediate Effect ($\gamma = 0$)

$$e_1 = e^*(\tilde{q}, \pi_0)$$

Effort and product quality are complements ($Y_{eq} > 0$)

$\Longrightarrow$ Overselling stimulates agent's immediate effort (e_1 strictly increases in $\tilde{q}$)

Immediate Effect ($\gamma = 0$)

$$e_1 = e^*(\tilde{q}, \pi_0)$$

Effort and product quality are complements ($Y_{eq} > 0$)

$\Longrightarrow$ Overselling stimulates agent's immediate effort (e_1 strictly increases in $\tilde{q}$)

Proposition (Immediate Effect of Manipulation)

Principal oversells the project if he only cares about immediate welfare ($\gamma = 0$)

Long-Run Effect ($\gamma = 1$): Stable Belief

$$K(f, f_{a'}) = \frac{[Y(e_\infty, \tilde{q}, a') - Y(e_\infty, Q, A)]^2}{2\sigma_\epsilon^2}$$

- Unique $\tilde{a} \in \mathbb{R}$: agent's **perceived mean output** matches **actual mean output**

$$\boldsymbol{Y(e_\infty, \tilde{q}, \tilde{a}) = Y(e_\infty, Q, A)}$$

  - this is the unique ability level that minimizes $K(f, f_{a'})$:

$$\Theta(e_\infty) = \underset{a' \in \mathbb{R}}{\operatorname{argmin}} K(f, f_{a'}) = \{\tilde{a}\}$$

Long-Run Effect ($\gamma = 1$): Stable Belief

$$K(f, f_{a'}) = \frac{[Y(e_\infty, \tilde{q}, a') - Y(e_\infty, Q, A)]^2}{2\sigma_\epsilon^2}$$

- Unique $\tilde{a} \in \mathbb{R}$: agent's **perceived mean output** matches **actual mean output**

$$\boldsymbol{Y(e_\infty, \tilde{q}, \tilde{a}) = Y(e_\infty, Q, A)}$$

  - this is the unique ability level that minimizes $K(f, f_{a'})$:

$$\Theta(e_\infty) = \underset{a' \in \mathbb{R}}{\operatorname{argmin}} K(f, f_{a'}) = \{\tilde{a}\}$$

- Stable belief on ability degenerate at ability $\tilde{a}$: $\pi_\infty = \delta_{\tilde{a}}$
  - in the long run, agent concludes with certainty that her own ability is $\tilde{a}$

Long-Run Effect ($\gamma = 1$): Stable Belief

$$\underbrace{Y(e_\infty, \tilde{q}, \tilde{a})}_{\text{perceived mean output}} = \underbrace{Y(e_\infty, Q, A)}_{\text{actual mean output}}$$

Lemma (Attribution Error)

Agent eventually underestimates her ability ($\tilde{a} < A$) if principal oversells ($\tilde{q} > Q$)

Long-Run Effect ($\gamma = 1$): Stable Belief

$$\underbrace{Y(e_\infty, \tilde{q}, \tilde{a})}_{\text{perceived mean output}} = \underbrace{Y(e_\infty, Q, A)}_{\text{actual mean output}}$$

Lemma (Attribution Error)

Agent eventually underestimates her ability ($\tilde{a} < A$) if principal oversells ($\tilde{q} > Q$)

The stable state is **self-confirming**

- Agent perceives project quality as $\tilde{q}$ and concludes that her ability is $\tilde{a}$
- Under such perception, she exerts the myopically optimal effort $e_\infty = e^*(\tilde{q}, \tilde{a})$, which yields an output distribution that matches her perceived output distribution

Long-Run Effect ($\gamma = 1$): Stable Effort

$$e_\infty = e^*(\tilde{q}, \tilde{a})$$

Long-Run Effect ($\gamma = 1$): Stable Effort

$$e_\infty = e^*(\tilde{q}, \tilde{a})$$

$$e'_\infty(\tilde{q}) = \underbrace{\frac{\partial e^*(\tilde{q}, \tilde{a})}{\partial \tilde{q}}}_{\text{direct effect}} + \underbrace{\frac{\partial e^*(\tilde{q}, \tilde{a})}{\partial \tilde{a}} \frac{\partial \tilde{a}}{\partial \tilde{q}}}_{\text{indirect effect}}$$

Dual Effects of Manipulation on Stable Effort

$$e_\infty = e^*(\tilde{q}, \tilde{a})$$

$$e'_\infty(\tilde{q}) = \underbrace{\frac{\partial e^*(\tilde{q}, \tilde{a})}{\partial \tilde{q}}}_{(+)} + \underbrace{\frac{\boldsymbol{\partial e^*(\tilde{q}, \tilde{a})}}{\boldsymbol{\partial \tilde{a}}}}_{\textbf{(?)}} \underbrace{\frac{\partial \tilde{a}}{\partial \tilde{q}}}_{(-)}$$

Ability and Effort Are Substitutes ($Y_{ea} < 0$)

$$e'_{\infty}(\tilde{q}) = \underbrace{\frac{\partial e^*(\tilde{q}, \tilde{a})}{\partial \tilde{q}}}_{(+)} + \underbrace{\frac{\boldsymbol{\partial e^*(\tilde{q}, \tilde{a})}}{\boldsymbol{\partial \tilde{a}}}}_{\boldsymbol{(-)}} \underbrace{\frac{\partial \tilde{a}}{\partial \tilde{q}}}_{(-)} > 0$$

Ability and Effort Are Substitutes ($Y_{ea} < 0$)

$$e'_\infty(\tilde{q}) = \underbrace{\frac{\partial e^*(\tilde{q},\tilde{a})}{\partial \tilde{q}}}_{(+)} + \underbrace{\frac{\partial e^*(\tilde{q},\tilde{a})}{\partial \tilde{a}}}_{(-)} \underbrace{\frac{\partial \tilde{a}}{\partial \tilde{q}}}_{(-)} > 0$$

Lemma (Self-Reinforcing Learning)

If $Y_{ea} \leq 0$, stable effort e_∞ strictly increases in perceived project quality $\tilde{q}$

- **Heidhues & Kőszegi & Strack (2018)**: (exogenously) overconfident agent

Ability and Effort Are Substitutes ($Y_{ea} < 0$)

$$e'_\infty(\tilde{q}) = \underbrace{\frac{\partial e^*(\tilde{q}, \tilde{a})}{\partial \tilde{q}}}_{(+)} + \underbrace{\frac{\partial e^*(\tilde{q}, \tilde{a})}{\partial \tilde{a}}}_{(-)} \underbrace{\frac{\partial \tilde{a}}{\partial \tilde{q}}}_{(-)} > 0$$

Lemma (Self-Reinforcing Learning)

If $Y_{ea} \leq 0$, stable effort e_∞ strictly increases in perceived project quality $\tilde{q}$

- **Heidhues & Kőszegi & Strack (2018)**: (exogenously) overconfident agent
- Endogenize overconfidence: when ability and effort are substitutes, overconfidence stimulates effort in both short and long run

Ability and Effort Are Complements ($Y_{ea} > 0$)

$$e'_{\infty}(\tilde{q}) = \underbrace{\frac{\partial e^*(\tilde{q}, \tilde{a})}{\partial \tilde{q}}}_{(+)} + \underbrace{\frac{\boldsymbol{\partial e^*(\tilde{q}, \tilde{a})}}{\boldsymbol{\partial \tilde{a}}}}_{\boldsymbol{(+)}} \underbrace{\frac{\partial \tilde{a}}{\partial \tilde{q}}}_{(-)}$$

How stable effort varies with agent's perceived quality $\tilde{q}$ is undetermined a priori

Three Classes of Production

1. *Sufficient Statistic*: $Y(e, q, a) = V(e, S(q, a))$ where $V_S > 0$ and $V_{eS} > 0$
   - project quality and ability are not jointly identifiable

Examples

- Cobb–Douglas production function: $Y(e, q, a) = ae^{\beta} q^{\alpha}$
  - $S(q, a) = aq^{\alpha}$
- CES production function: $Y(e, q, a) = \phi \left[\alpha_1 e^{\beta} + \alpha_2 q^{\beta} + \alpha_3 a^{\beta}\right]^{\frac{1}{\beta}}$
  - $S(q, a) = \alpha_2 q^{\beta} + \alpha_3 a^{\beta}$

Three Classes of Production

1. *Sufficient Statistic*: $Y(e, q, a) = V(e, S(q, a))$
   - $V_S > 0, V_{eS} > 0$

- e.g., $Y(e, q, a) = qae$
  - $S(q, a) = qa$

Three Classes of Production

1. *Sufficient Statistic*: $Y(e, q, a) = V(e, S(q, a))$
   - $V_S > 0, V_{eS} > 0$

   ▶ e.g., $Y(e, q, a) = qae$
   - $S(q, a) = qa$

2. *Ability-Laden*: $Y(e, q, a) = V(e, S(q, a), a)$
   - $V_S > 0, V_a > 0, V_{eS} > 0, V_{ea} \leq 0$

   ▶ e.g., $Y(e, q, a) = qae + a$
   - $S(q, a) = qa$

Three Classes of Production

1. *Sufficient Statistic*: $Y(e,q,a) = V(e, S(q,a))$
   - $V_S > 0, V_{eS} > 0$

   ▶ e.g., $Y(e,q,a) = qae$
   - $S(q,a) = qa$

2. *Ability-Laden*: $Y(e,q,a) = V(e, S(q,a), a)$
   - $V_S > 0, V_a > 0, V_{eS} > 0, V_{ea} \leq 0$

   ▶ e.g., $Y(e,q,a) = qae + a$
   - $S(q,a) = qa$

3. *Quality-Laden*: $Y(e,q,a) = V(e, S(q,a), q)$
   - $V_S > 0, V_q > 0, V_{eS} > 0, V_{eq} \leq 0$

   ▶ e.g., $Y(e,q,a) = qae + q$
   - $S(q,a) = qa$

Three Classes of Production

1. *Sufficient Statistic*: $Y(e,q,a) = V(e, S(q,a))$
   - $V_S > 0, V_{eS} > 0$

   ▶ e.g., $Y(e,q,a) = qae$
   - $S(q,a) = qa$

2. *Ability-Laden*: $Y(e,q,a) = V(e, S(q,a), a)$
   - $V_S > 0, V_a > 0, V_{eS} > 0, V_{ea} \leq 0$

   ▶ e.g., $Y(e,q,a) = qae + a$
   - $S(q,a) = qa$

3. *Quality-Laden*: $Y(e,q,a) = V(e, S(q,a), q)$
   - $V_S > 0, V_q > 0, V_{eS} > 0, V_{eq} \leq 0$

   ▶ e.g., $Y(e,q,a) = qae + q$
   - $S(q,a) = qa$

- jointly exclusive but not exhaustive

Three Classes of Production

1. *Sufficient Statistic*: $Y(e, q, a) = V(e, S(q, a))$
   - $V_S > 0, V_{eS} > 0$

2. *Ability-Laden*: $Y(e, q, a) = V(e, S(q, a), a)$
   - $V_S > 0, V_a > 0, V_{eS} > 0, V_{ea} \leq 0$

3. *Quality-Laden*: $Y(e, q, a) = V(e, S(q, a), q)$
   - $V_S > 0, V_q > 0, V_{eS} > 0, V_{eq} \leq 0$

- jointly exclusive but not exhaustive

- e.g., $Y(e, q, a) = qae$
  - $S(q, a) = qa$
- e.g., $Y(e, q, a) = qae + a$
  - $S(q, a) = qa$
- e.g., $Y(e, q, a) = qae + q$
  - $S(q, a) = qa$

- common marginal incentive

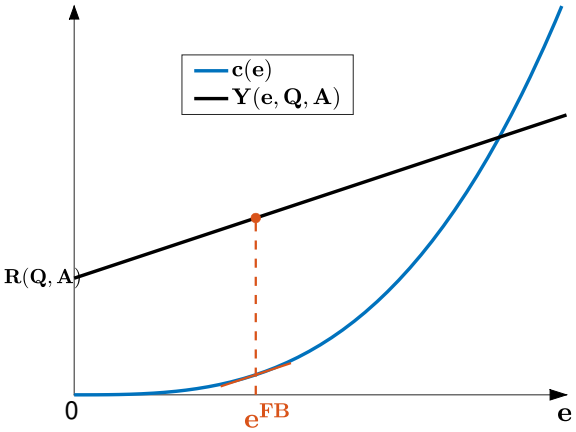
Long-Run Effect

Proposition (Long-Run Effect of Manipulation)

1. If the production features a sufficient statistic, then $e_\infty = e^{FB}$
2. If the production is ability-laden, then $e_\infty > e^{FB}$ if principal is overselling
3. If the production is quality-laden, then $e_\infty > e^{FB}$ if principal is underselling

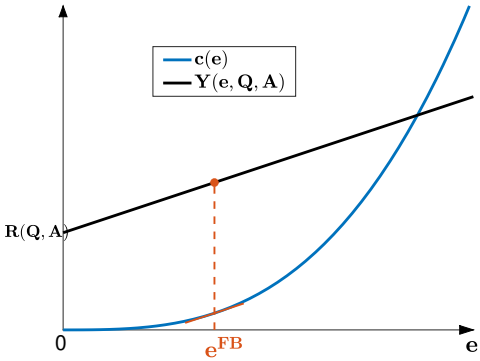
Illustrative Example

$$Y(e, q, a) = S(q, a)e + R(q, a)$$



Illustrative Example

$$Y(e, q, a) = S(q, a)e + R(q, a)$$

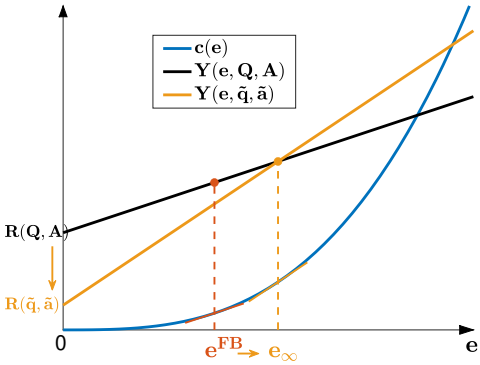


To stimulate stable effort

1. higher perceived marginal return to effort
2. perceived mean output = actual mean output

Illustrative Example

$$Y(e, q, a) = S(q, a)e + R(q, a)$$

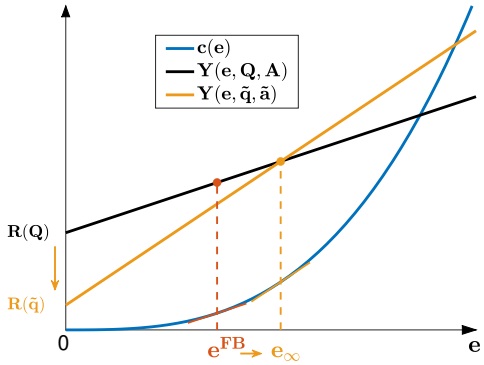


To stimulate stable effort

1. higher perceived marginal return to effort
2. perceived mean output = actual mean output

Illustrative Example

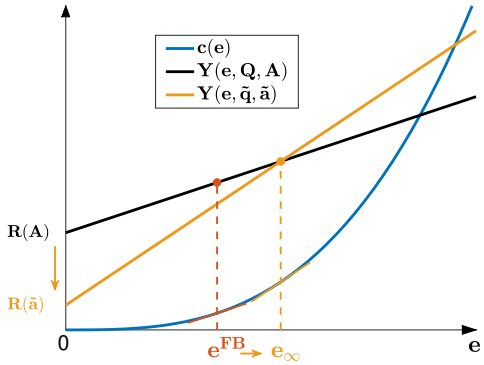
$$Y(e, q, a) = S(q, a)e + R(q, a)$$



- Quality-laden: $R(q, a) = R(q)$
  - undersell the project to decrease $R(q)$

Illustrative Example

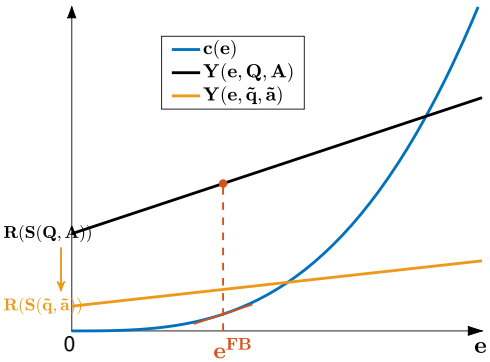
$$Y(e, q, a) = S(q, a)e + R(q, a)$$



- Quality-laden: $R(q, a) = R(q)$
  - undersell the project to decrease $R(q)$
- Ability-laden: $R(q, a) = R(a)$
  - oversell the project to decrease $R(a)$

Illustrative Example

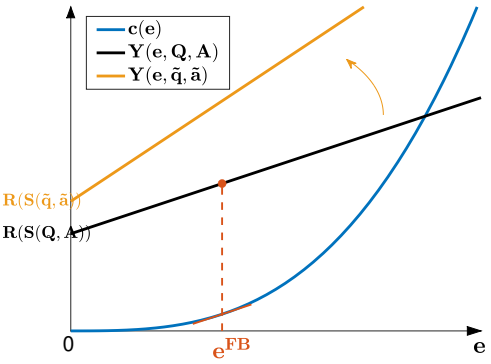
$$Y(e, q, a) = S(q, a)e + R(q, a)$$



- Quality-laden: $R(q, a) = R(q)$
  - undersell the project to decrease $R(q)$
- Ability-laden: $R(q, a) = R(a)$
  - oversell the project to decrease $R(a)$
- Sufficient statistic: $R(q, a) = R(S(q, a))$
  - no lasting manipulation impact on effort

Illustrative Example

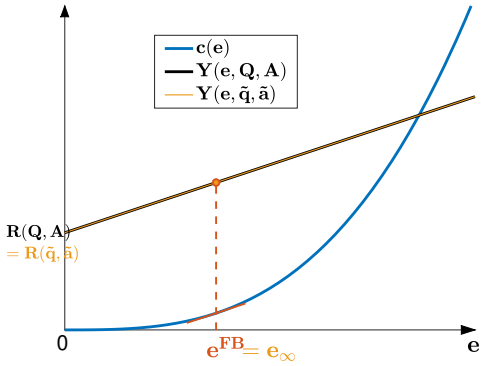
$$Y(e, q, a) = S(q, a)e + R(q, a)$$



- Quality-laden: $R(q, a) = R(q)$
  - undersell the project to decrease $R(q)$
- Ability-laden: $R(q, a) = R(a)$
  - oversell the project to decrease $R(a)$
- Sufficient statistic: $R(q, a) = R(S(q, a))$
  - no lasting manipulation impact on effort

Illustrative Example

$$Y(e, q, a) = S(q, a)e + R(q, a)$$



- Quality-laden: $R(q, a) = R(q)$
  - undersell the project to decrease $R(q)$
- Ability-laden: $R(q, a) = R(a)$
  - oversell the project to decrease $R(a)$
- Sufficient statistic: $R(q, a) = R(S(q, a))$
  - no lasting manipulation impact on effort

Generalization (Basic Intuition)

1. **Sufficient Statistic**: $Y(e, q, a) = V(e, S(q, a))$
   - fixing effort, agent learns $S(q, a)$ correctly $\Rightarrow e_\infty = e^{FB}$
   - direct effect of misspecified quality is offset by indirect effect of inferred ability

Generalization (Basic Intuition)

1. **Sufficient Statistic**: $Y(e, q, a) = V(e, S(q, a))$
   - fixing effort, agent learns $S(q, a)$ correctly $\Rightarrow e_\infty = e^{FB}$
   - direct effect of misspecified quality is offset by indirect effect of inferred ability

2. **Ability-Laden**: $Y(e, q, a) = V(e, S(q, a), a)$
   - overselling $\Rightarrow$ lower perceived ability $\Rightarrow$ higher inferred $S(q, a) \Rightarrow$ higher e_∞
   - direct effect $>$ indirect effect

Generalization (Basic Intuition)

1. **Sufficient Statistic**: $Y(e, q, a) = V(e, S(q, a))$
   - fixing effort, agent learns $S(q, a)$ correctly $\Rightarrow e_\infty = e^{FB}$
   - direct effect of misspecified quality is offset by indirect effect of inferred ability

2. **Ability-Laden**: $Y(e, q, a) = V(e, S(q, a), a)$
   - overselling $\Rightarrow$ lower perceived ability $\Rightarrow$ higher inferred $S(q, a) \Rightarrow$ higher e_∞
   - direct effect $>$ indirect effect

3. **Quality-Laden**: $Y(e, q, a) = V(e, S(q, a), q)$
   - overselling $\Rightarrow$ higher perceived quality $\Rightarrow$ lower inferred $S(q, a) \Rightarrow$ lower e_∞
   - direct effect $<$ indirect effect

Summary

Theorem (Determination of Manipulation)

For any true state, prior belief, signal precision $(Q, A, \pi_0, \sigma_\epsilon^2)$

1. If production features a sufficient statistic
   - principal is indifferent if he is fully patient $\gamma = 1$
   - principal oversells otherwise $\gamma \in [0, 1)$

Summary

Theorem (Determination of Manipulation)

For any true state, prior belief, signal precision $(Q, A, \pi_0, \sigma_\epsilon^2)$

1. If production features a sufficient statistic
   - principal is indifferent if he is fully patient $\gamma = 1$
   - principal oversells otherwise $\gamma \in [0, 1)$
2. If production is ability-laden, principal oversells the project

Summary

Theorem (Determination of Manipulation)

For any true state, prior belief, signal precision $(Q, A, \pi_0, \sigma_\epsilon^2)$

1. If production features a sufficient statistic
   - principal is indifferent if he is fully patient $\gamma = 1$
   - principal oversells otherwise $\gamma \in [0, 1)$
2. If production is ability-laden, principal oversells the project
3. If production is quality-laden, there exists a threshold $\hat{\gamma} \in (0, 1)$
   - principal oversells if he is impatient $\gamma < \hat{\gamma}$
   - principal undersells if he is sufficiently patient $\gamma > \hat{\gamma}$

Summary

Theorem (Determination of Manipulation)

For any true state, prior belief, signal precision $(Q, A, \pi_0, \sigma_\epsilon^2)$

1. If production features a sufficient statistic
   - principal is indifferent if he is fully patient $\gamma = 1$
   - principal oversells otherwise $\gamma \in [0, 1)$
2. If production is ability-laden, principal oversells the project
3. If production is **quality-laden**, there exists a threshold $\hat{\gamma} \in (0, 1)$
   - principal oversells if he is impatient $\gamma < \hat{\gamma}$
   - principal **undersells** if he is **sufficiently patient** $\gamma > \hat{\gamma}$

Agenda

1. Model
2. Immediate Effect and Long-Run Effect
3. Applications
   - Mentorship
   - Abusive Relationship
4. Extensions
   - Sophisticated Agent
   - Short-Run Incentives

Mentorship

$$y_t = Y(e_t, q, a) + \epsilon_t$$

- e.g., lab advisor (principal) – graduate student (agent)
  - y: findings from a long-term research project
  - q: value/potential of the research question
  - a: ability of the mentee and/or her coworkers
  - e: mentee's effort
- Assume return to mentee's effort increases with her ability and question value

Mentorship

$$y_t = Y(e_t, q, a) + \epsilon_t$$

- e.g., lab advisor (principal) – graduate student (agent)
  - y: findings from a long-term research project
  - q: value/potential of the research question
  - a: ability of the mentee and/or her coworkers
  - e: mentee's effort
- Assume return to mentee's effort increases with her ability and question value

Expectation management strategies vary with nature of the project

Mentorship: Question-Based Project

All work is directed toward a specific preregistered question

$$Y(e, q, a) = q(e + a + \psi ae)$$

- Value of research question is decisive in yielding research findings
- $\psi \geq 0$: synergy of agent's ability and effort

Mentorship: Question-Based Project

$$Y(e,q,a) = q(e + a + \psi ae) = S(q,a)\left(e + \frac{a}{1+\psi a}\right)$$

- $S(q,a) = q(1+\psi a)$: intrinsic motivation to study
- This type of research project is ability-laden

Mentor, regardless of his patience, should oversell value of the project

$\Rightarrow$ mentee underestimates her own ability

$\Rightarrow$ humility raises mentee's perceived return to effort

$$S(\tilde{q},\tilde{a}) > S(Q,A)$$

$\Rightarrow$ stimulates mentee's stable effort and promotes research advances

Mentorship: Ability-Based Project

Agent's ability is a dominant factor in determining research findings

$$Y(e,q,a) = a(e+q+\psi qe) = S(q,a)\left(e+\frac{q}{1+\psi q}\right)$$

Patient mentor should undersell the project quality

$\Rightarrow$ reduce mentee's self-blaming for low outputs, boost her self-confidence

$\Rightarrow$ overconfidence increases mentee's perceived return to effort

$$S(\tilde{q},\tilde{a}) > S(Q,A)$$

$\Rightarrow$ stimulates mentee's stable effort and promotes research advances

Abusive Relationship

$$Y = S(q, a)e + R(q)$$

- e.g., gaslighting in professional or intimate relationships
  - Y: individual wellbeing
  - q: individual's personal quality (e.g., sense of humor, sensibility, judgement)
  - a: potential of the relationship (e.g., how good the match is)
  - e: effort
  - $S(q, a)$: attachment/affection to the relationship
- Individual's personal quality has an independent value beyond her relational benefit

Abusive Relationship

$$Y = S(q, a)e + R(q)$$

- Manipulative partner aims to extract excessive effort from the individual
- Emotional abuse tactics (e.g., blame-shifting, overprotection, neglect, isolation, and criticism) to manipulate how the individual perceives herself
- Manipulator erodes the individual's self-image by emotional abuse
  $\Rightarrow$ individual perceives the relationship to be vitally important
  $\Rightarrow$ motivates unduly high sacrifice and sensitivity to the relationship

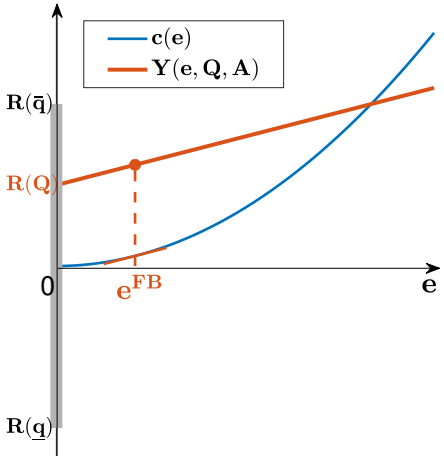
Abusive Relationship

$$Y = S(q, a)e + R(q)$$

- Manipulative partner aims to extract excessive effort from the individual
- Emotional abuse tactics (e.g., blame-shifting, overprotection, neglect, **isolation**, and criticism) to manipulate how the individual perceives herself
- Manipulator erodes the individual's self-image by emotional abuse
  $\Rightarrow$ individual perceives the relationship to be vitally important
  $\Rightarrow$ motivates unduly high sacrifice and sensitivity to the relationship

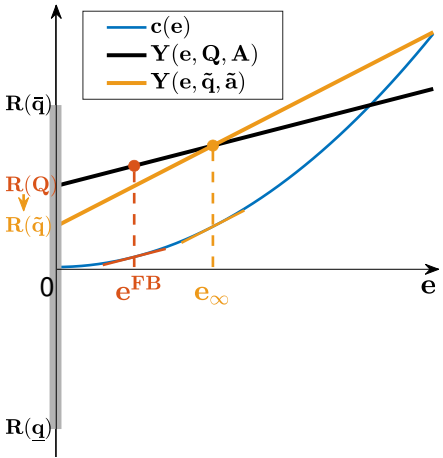
How Does Isolation Facilitate Perception Manipulation?

$$Y(e, q, a) = S(q, a)e + R(q)$$



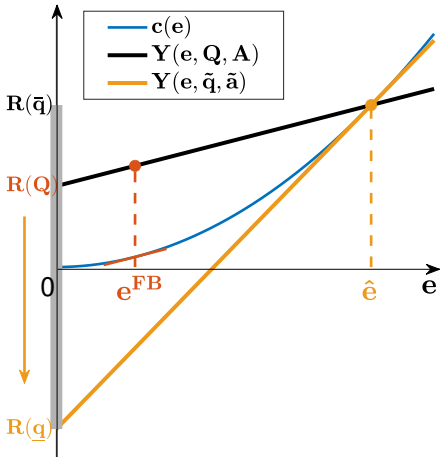
How Does Isolation Facilitate Perception Manipulation?

$$Y(e, q, a) = S(q, a)e + R(q)$$

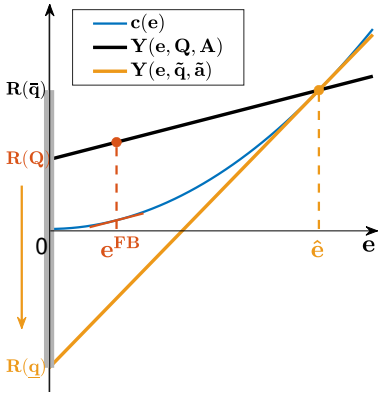


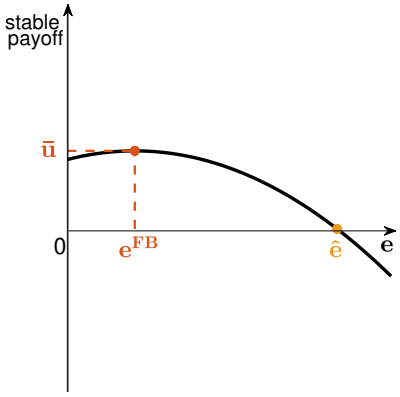
How Does Isolation Facilitate Perception Manipulation?

$$Y(e, q, a) = S(q, a)e + R(q)$$



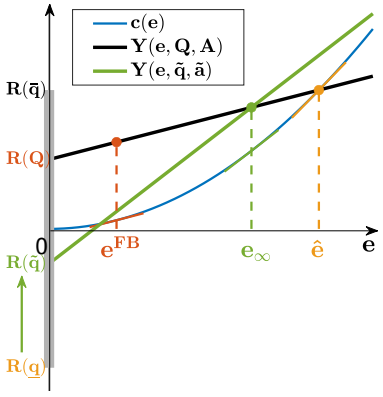
How Does Isolation Facilitate Perception Manipulation?

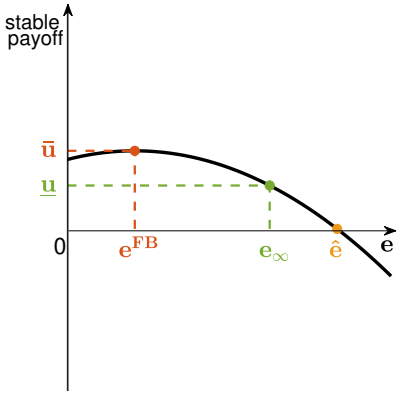




Truth-telling maximizes agent's stable payoff vs. **Optimal manipulation** minimizes it

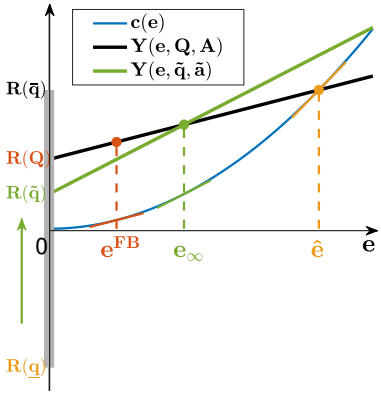
Value of Outside Option

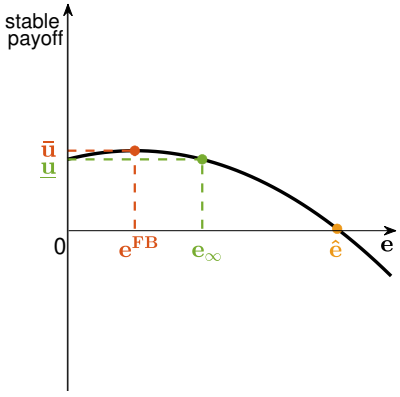




With an **outside option** $\underline{u}$, principal's manipulation strategy shifts towards truth-telling

Value of Outside Option





Manipulation is attenuated as agent improves her outside option (i.e., a higher $\underline{u}$)

Isolation Facilitates Perception Manipulation

- If agent can quit, principal would manipulate agent but not too much
- Outside options restrict the extent to which principal can manipulate agent
- One role of isolation: make the agent unaware of/learn less about outside options

Agenda

1. Model
2. Immediate Effect and Long-Run Effect
3. Applications
   - Mentorship
   - Abusive Relationship
4. Extensions
   - Sophisticated Agent
   - Short-Run Incentives

Perception Manipulation for Sophisticated Agent

Previously, agent fully trusts principal (e.g., naive audience vs. trusted authority)

Can principal manipulate a sophisticated agent's perception?

Perception Manipulation for Sophisticated Agent

Previously, agent fully trusts principal (e.g., naive audience vs. trusted authority)

Can principal manipulate a sophisticated agent's perception?

Uncertainty about the principal's type enables manipulation

Perception Manipulation for Sophisticated Agent

- Ability-laden production
$$Y(e, q, a) = qae + a$$
- Principal privately knows project quality $Q \in \{0, 1\}$ and reports $\tilde{q} \in \{0, 1\}$:
  - honest (H): always truth-telling ($\tilde{q} = Q$)
  - manipulative (M): always overselling ($\tilde{q} = 1$)
- Agent's prior on:
  - project quality being $q = 1$: $p_q = 1/2$
  - principal being honest: $p_h \in [0, 1]$
- Both players share the same prior on agent's ability: $\pi_0 = U[0, 2]$
- Cost function: $c(e) = e^2/2$

Perception Manipulation for Sophisticated Agent

Agent's myopically optimal effort is given by her expected project quality

Without any information from the principal

$$e^*(p_q, \pi_0) = p_q = 1/2$$

Perception Manipulation for Sophisticated Agent

Agent's myopically optimal effort is given by her expected project quality

Without any information from the principal

$$e^*(p_q, \pi_0) = p_q = 1/2$$

After observing principal's report $\tilde{q} = 1$

$$e^*(p'_q, \pi_0) = p'_q = \frac{p_q}{p_q + (1 - p_q)(1 - p_h)} = \frac{1}{2 - p_h}$$

- strictly increases in the perceived probability of meeting an honest principal p_h
  - naive: $p_h = 1 \quad \Rightarrow \; p'_q = 1$
  - skeptic: $p_h = 0 \quad \Rightarrow \; p'_q = p_q = 1/2$
  - possibly honest: $p_h > 0 \quad \Rightarrow \; p'_q > p_q = 1/2$

Perception Manipulation for Sophisticated Agent

- Agent is vulnerable to manipulation when she is uncertain about principal's type
- Sophistication level: affects the extent of manipulation, but not the direction
- Naivete makes forces involved stark

Two-Period Model with Noiseless Observations

- $t = 1$
  - principal sets agent's perceived project quality $\tilde{q}$
  - agent exerts effort $e_1 = e^*(\tilde{q}, \pi_0)$
- $t = 2$
  - agent observes the output $Y_1 = Y(e_1, Q, A)$, updates her ability belief to π_1
  - agent exerts effort $e_2 = e^*(\tilde{q}, \pi_1)$

Agent's posterior upon observing output in period 1 degenerates at $\tilde{a}_1$

$$Y(e_1, Q, A) = Y(e_1, \tilde{q}, \tilde{a}_1)$$

Two-Period Model with Noiseless Observations

- $t = 1$
  - principal sets agent's perceived project quality $\tilde{q}$
  - agent exerts effort $e_1 = e^*(\tilde{q}, \pi_0)$
- $t = 2$
  - agent observes the output $Y_1 = Y(e_1, Q, A)$, updates her ability belief to π_1
  - agent exerts effort $e_2 = e^*(\tilde{q}, \pi_1)$

Agent's posterior upon observing output in period 1 degenerates at $\tilde{a}_1$

$$Y(e_1, Q, A) = Y(e_1, \tilde{q}, \tilde{a}_1)$$

Main results hold

Two-Period Model with Noiseless Observations

- The simplified model captures the key tension in expectation management

Two-Period Model with Noiseless Observations

- The simplified model captures the key tension in expectation management
- Two periods of noiseless output observations suffice to detect manipulation
- Baseline model illustrates how manipulation persists over time
  - noisy observations enable manipulation in a long-term relationship
  - principal triggers misspecified learning which results in a self-confirming stable state

Future Work

- Break free from manipulation
  - forward-looking & sophisticated: experiment with models
- Include long-run consideration in communication
  - cheap talk/signaling/bayesian persuasion + misspecified learning
  - how much manipulation can be sustained by a specific communication protocol

Conclusion

- I propose a model of perception manipulation as a new approach to incentivize effort, and study how the form of misspecification is endogenously determined
- I identify three classes of output functions
  1. sufficient statistics: manipulation cannot affect long-run effort
  2. ability-laden: overselling stimulates long-run effort
  3. quality-laden: underselling stimulates long-run effort
- *Key mechanism*
  - principal downplays contribution to output that is unaffected by agent's effort
  - agent misattributes output more to her effort and inflates her agency over the project

Bibliography

- **Evidence for Perception Manipulation**: Hiroto & Seligman (1975), Maier & Seligman (1976), Alloy & Abramson (1979), Quattrone & Tversky (1984), Sharot, Korn & Dolan (2011), Eil & Rao (2011), Von Hippel & Trivers (2011), Schacter (2012), Sweet (2019), Fan & Bolte (2022)

- **Theories on Motivated Belief**: Akerlof & Dickens (1982), Akerlof & Kranton (2000), Carrillo & Mariotti (2000), **Bénabou & Tirole** (2002, 2004, 2011, **2016**), **Brunnermeier & Parker (2005)**, Eyster, Li & Ridout (2021)

Bibliography

- **Misspecified Learning**
  - Original concepts: Nyarko (1991), **Esponda & Pouzo (2016)**
  - Techniques to analyze asymptotic properties: Fudenberg, Romanyuk & Strack (2017), Bohren & Hauser (2021), Esponda & Pouzo (2021), Heidhues, Kőszegi & Strack (2021), Esponda, Pouzo & Yamamoto (2021), Fudenberg, Lanzani & Strack (2021,2023), Frick, Iijima & Ishii (2023)
  - Implications of misspecified learning: Andreoni & Mylovanov (2012), Bohren (2016), **Heidhues, Kőszegi & Strack(2018**, 2023), Frick, Iijima & Ishii (2020), Ba & Gindin (2023), Gagnon-Bartsch & Bushong (2022), He (2022), Kőszegi, Loewenstein & Murooka (2022), Levy, Razin & Young (2022)
  - Model selection: He & Libgober (2020), Montiel Olea, Ortoleva, Pai & Prat (2022), Ba (2023), Fudenberg & Lanzani (2023)