# Data and Incentives

Annie Liang[1] Erik Madsen[2]

[1]Northwestern

[2]NYU

# Big Data

Online platforms and data brokers track consumer activities, producing new measurements of everything from

# Big Data

Online platforms and data brokers track consumer activities, producing new measurements of everything from

> the size of your **social network**, to how often you **move**, to how many hours you spend playing **video games**

# Big Data

Online platforms and data brokers track consumer activities, producing new measurements of everything from

> the size of your **social network**, to how often you **move**, to how many hours you spend playing **video games**

This new data is increasingly available to firms and organizations

# Big Data

Online platforms and data brokers track consumer activities, producing new measurements of everything from

> the size of your **social network**, to how often you **move**, to how many hours you spend playing **video games**

This new data is increasingly available to firms and organizations, and regulators are concerned about

- banks using this data to predict whether an agent is creditworthy
- firms using this data to predict a worker's productivity at a new job

# Big Data

Online platforms and data brokers track consumer activities, producing new measurements of everything from

> the size of your **social network**, to how often you **move**, to how many hours you spend playing **video games**

This new data is increasingly available to firms and organizations, and regulators are concerned about

- banks using this data to predict whether an agent is creditworthy
- firms using this data to predict a worker's productivity at a new job

Recent EU regulation draft labels these as "high risk" — not just privacy concerns, but also possibility of the forecasts to **"distort behavior."**

# Reputational Incentives for Effort

To understand when and how to regulate new data, need to first understand the economic implications.

# Reputational Incentives for Effort

To understand when and how to regulate new data, need to first understand the economic implications.

Today focus on impact on markets with **moral hazard**.

# Reputational Incentives for Effort

To understand when and how to regulate new data, need to first understand the economic implications.

Today focus on impact on markets with **moral hazard**.

- Agents exert hidden effort (which may be productive, but is costly).
- Main incentive to do this is a reputational payoff: improve a market's perception of their quality.

# Reputational Incentives for Effort

To understand when and how to regulate new data, need to first understand the economic implications.

Today focus on impact on markets with **moral hazard**.

- Agents exert hidden effort (which may be productive, but is costly).
- Main incentive to do this is a reputational payoff: improve a market's perception of their quality.

Our question: How does information extracted from big data impact effort choice and welfare?

# Model

# Reputational Incentives for Effort (Holmstrom, 1999)

An agent participates in a market across two periods.

# Reputational Incentives for Effort (Holmstrom, 1999)

An agent participates in a market across two periods.

The agent has a type $\theta \sim F_\theta$ which is initially unknown.

# Reputational Incentives for Effort (Holmstrom, 1999)

An agent participates in a market across two periods.

The agent has a type $\theta \sim F_\theta$ which is initially unknown.

- In **period 1**:
  - The agent **privately** exerts effort $e \in \mathbb{R}_+$ at cost $C(e) = \frac{1}{2}e^2$. (Results generalize to other convex cost functions.)
  - An outcome
    $$Y = e + \theta + \varepsilon$$
    is observed, where $\varepsilon \sim F_\varepsilon$ is a shock independent of $\theta$

# Reputational Incentives for Effort (Holmstrom, 1999)

An agent participates in a market across two periods.

The agent has a type $\theta \sim F_\theta$ which is initially unknown.

- In **period 1**:
  - The agent **privately** exerts effort $e \in \mathbb{R}_+$ at cost $C(e) = \frac{1}{2}e^2$. (Results generalize to other convex cost functions.)
  - An outcome
    $$Y = e + \theta + \varepsilon$$
    is observed, where $\varepsilon \sim F_\varepsilon$ is a shock independent of $\theta$
- In **period 2**, the agent receives a reputational payoff $\mathbb{E}[\theta \mid Y]$

# Reputational Incentives for Effort (Holmstrom, 1999)

An agent participates in a market across two periods.

The agent has a type $\theta \sim F_\theta$ which is initially unknown.

- In **period 1**:
  - The agent **privately** exerts effort $e \in \mathbb{R}_+$ at cost $C(e) = \frac{1}{2}e^2$. (Results generalize to other convex cost functions.)
  - An outcome
    $$Y = e + \theta + \varepsilon$$
    is observed, where $\varepsilon \sim F_\varepsilon$ is a shock independent of $\theta$
- In **period 2**, the agent receives a reputational payoff $\mathbb{E}[\theta \mid Y]$

Total payoff: $\beta \cdot \mathbb{E}[\theta \mid Y] - (1 - \beta) \cdot C(e)$ for some $\beta \in (0, 1)$.

# Equilibrium Effort

The agent's expected payoff given conjectured effort $\hat{e}$ is:

$$U(e;\hat{e}) = -(1-\beta)\cdot\frac{e^2}{2} + \beta\cdot\mathbb{E}^{e}[\mathbb{E}^{\hat{e}}(\theta \mid Y)]$$

# Equilibrium Effort

The agent's expected payoff given conjectured effort $\hat{e}$ is:

$$U(e;\hat{e}) = \underbrace{-(1-\beta)\cdot\frac{e^2}{2}}_{\text{Cost of effort}} + \beta\cdot\mathbb{E}^{e}[\mathbb{E}^{\hat{e}}(\theta \mid Y)]$$

# Equilibrium Effort

The agent's expected payoff given conjectured effort $\hat{e}$ is:

$$U(e; \hat{e}) = -(1-\beta) \cdot \frac{e^2}{2} + \beta \cdot \mathbb{E}^e[\underbrace{\mathbb{E}^{\hat{e}}(\theta \mid Y)}]$$

Market's posterior expectation of $\theta$
(updating with respect to $Y = \hat{e} + \theta + \varepsilon$)

# Equilibrium Effort

The agent's expected payoff given conjectured effort $\hat{e}$ is:

$$U(e;\hat{e}) = -(1-\beta)\cdot\frac{e^2}{2} + \beta\cdot\underbrace{\mathbb{E}^{e}[\mathbb{E}^{\hat{e}}(\theta \mid Y)]}$$

Agent's expectation of market's expectation (updating with respect to $Y = e + \theta + \varepsilon$)

# Equilibrium Effort

The agent's expected payoff given conjectured effort $\hat{e}$ is:

$$U(e;\hat{e}) = -(1-\beta)\cdot\frac{e^2}{2} + \beta\cdot\mathbb{E}^{e}[\mathbb{E}^{\hat{e}}(\theta \mid Y)]$$

Equilibrium effort $e^*$ is uniquely pinned down by the first-order condition:

$$\underbrace{\frac{\beta}{1-\beta}\cdot\left.\frac{\partial\mathbb{E}^{e}[\mathbb{E}^{e^*}(\theta \mid Y)]}{\partial e}\right|_{e=e^*}}_{\text{Marginal value}} = \underbrace{e^*}_{\text{Marginal cost}}$$

# Our Model:

We suppose that $\theta$ and $\varepsilon$ are predictable from primitive covariates:

$$\theta = \theta_1 + \cdots + \theta_J$$
$$\varepsilon = \varepsilon_1 + \cdots + \varepsilon_K$$

where

- $\theta_1, \ldots, \theta_J$ are called **attributes**
- $\varepsilon_1, \ldots, \varepsilon_K$ are called **circumstances**

# Our Model:

We suppose that $\theta$ and $\varepsilon$ are predictable from primitive covariates:

$$\theta = \theta_1 + \cdots + \theta_J$$
$$\varepsilon = \varepsilon_1 + \cdots + \varepsilon_K$$

where

- $\theta_1, \ldots, \theta_J$ are called **attributes**
- $\varepsilon_1, \ldots, \varepsilon_K$ are called **circumstances**

Examples:

- Labor market: $\theta$ is productivity, $Y$ is output. "Reliability" is an attribute; "industry shock" is a circumstance.
- College admissions: $\theta$ is ability, $Y$ is GPA. "Attention span" is an attribute; "illness or injury" are circumstances.

# Expansion of Measured Covariates

Some covariates are **measured**, revealing their values for all agents.

- Measured attributes: $\mathcal{J} \subseteq \{1, \ldots, J\}$
- Measured circumstances: $\mathcal{K} \subseteq \{1, \ldots, K\}$

# Expansion of Measured Covariates

Some covariates are **measured**, revealing their values for all agents.

- Measured attributes: $\mathcal{J} \subseteq \{1, \ldots, J\}$
- Measured circumstances: $\mathcal{K} \subseteq \{1, \ldots, K\}$

Timeline:

$t = 0$ : Agent and market observe agent's realization $(\theta_{\mathcal{J}}, \varepsilon_{\mathcal{K}})$

$t = 1$ : Agent chooses effort $e$ and incurs cost of effort.

$t = 2$ : The outcome $Y = e + \theta + \varepsilon$ is realized, and the agent receives the market's forecast $\mathbb{E}[\theta \mid Y, \theta_{\mathcal{J}}, \varepsilon_{\mathcal{K}}]$.

Distribution of effort depends on family $(\mathcal{J}, \mathcal{K})$

# Expansion of Measured Covariates

Some covariates are **measured**, revealing their values for all agents.

- Measured attributes: $\mathcal{J} \subseteq \{1, \ldots, J\}$
- Measured circumstances: $\mathcal{K} \subseteq \{1, \ldots, K\}$

Timeline:

$t = 0$ : Agent and market observe agent's realization $(\theta_{\mathcal{J}}, \varepsilon_{\mathcal{K}})$

$t = 1$ : Agent chooses effort $e$ and incurs cost of effort.

$t = 2$ : The outcome $Y = e + \theta + \varepsilon$ is realized, and the agent receives the market's forecast $\mathbb{E}[\theta \mid Y, \theta_{\mathcal{J}}, \varepsilon_{\mathcal{K}}]$.

Distribution of effort depends on family $(\mathcal{J}, \mathcal{K})$

**Big data:** Measured covariates expand from $(\mathcal{J}, \mathcal{K})$ to $(\mathcal{J}', \mathcal{K}')$.

# Some Comments on Knowledge Assumptions

Should the agent know more?

Should the agent know less?

# Some Comments on Knowledge Assumptions

Should the agent know more?

- Each $\theta_j$ is really the **effect** of a given covariate
- Agents may know, e.g., # hours spent playing video games, but not how that translates into productivity

Should the agent know less?

# Some Comments on Knowledge Assumptions

Should the agent know more?

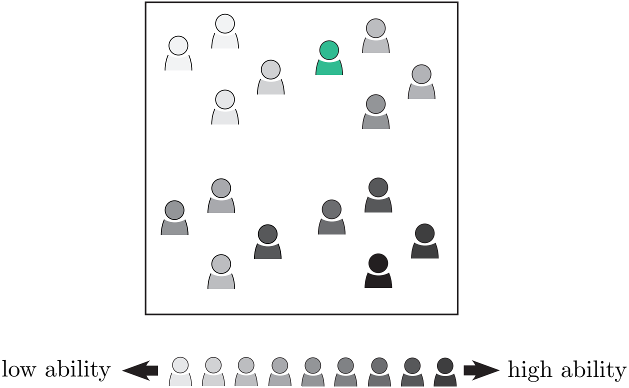
- Each $\theta_j$ is really the **effect** of a given covariate
- Agents may know, e.g., # hours spent playing video games, but not how that translates into productivity

Should the agent know less?

- Agents only need to understand the value of their effort, which may be learned by experience in the market.
- Our results generalize under model uncertainty where the agent is uncertain about the market's understanding of $(\theta, \varepsilon)$.

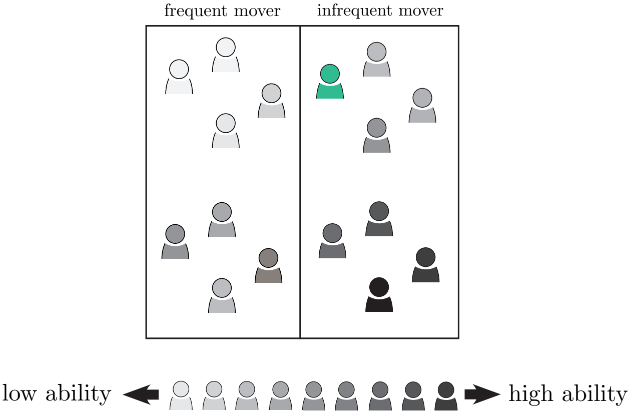
# What We Ask

How does expansion of measured covariates from $(\mathcal{J}, \mathcal{K})$ to some larger $(\mathcal{J}', \mathcal{K}')$ affect the distribution of effort?



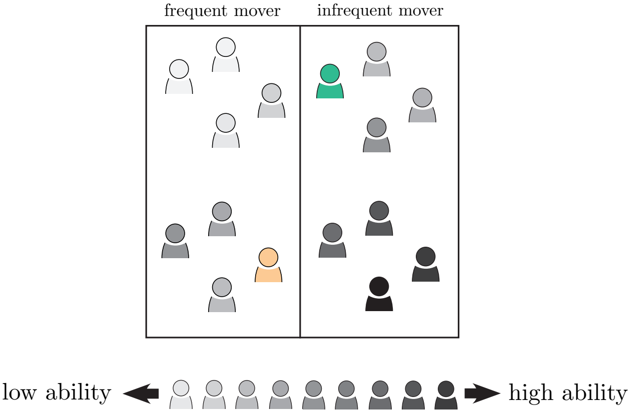
# What We Ask

How does expansion of measured covariates from $(\mathcal{J}, \mathcal{K})$ to some larger $(\mathcal{J}', \mathcal{K}')$ affect the distribution of effort?



# What We Ask

How does expansion of measured covariates from $(\mathcal{J}, \mathcal{K})$ to some larger $(\mathcal{J}', \mathcal{K}')$ affect the distribution of effort?



# Related Literature

Growing literature about the economic consequences of big data.

- see e.g. Bergemann, Bonatti, and Smolin (2018); Ichihashi (2019); Bergemann, Bonatti, and Gan (2020); Hidir and Vellodi (2021); Elliot et al. (2021).

Specifically related to us, papers about impact on consumer effort:

- incentives for "gaming" forecasts (Eliaz and Spiegler, 2019; Frankel and Kartik, 2020; Ball, 2020)
- to improve own characteristics (Haghtalab et al., 2020)

These papers treat the data environment as fixed. We vary it.

Methodologically, we build on the career concerns literature.

- Closest paper: Dewatripont et al. (1999). Effort chosen prior to information realizations.

# Preview of Main Results

- The change in effort **on average** is determined completely by whether the covariate is an attribute or circumstance:
  - New attributes reduce average effort
  - New circumstances increase average effort

# Preview of Main Results

- The change in effort **on average** is determined completely by whether the covariate is an attribute or circumstance:
  - New attributes reduce average effort
  - New circumstances increase average effort
- But the change in effort may differ across agents not only in magnitude but also in sign — we call this **disparate impact**

# Preview of Main Results

- The change in effort **on average** is determined completely by whether the covariate is an attribute or circumstance:
  - New attributes reduce average effort
  - New circumstances increase average effort
- But the change in effort may differ across agents not only in magnitude but also in sign — we call this **disparate impact**
- Demonstrate a statistical condition which guarantees that disparate impact does not emerge.

# Preview of Main Results

- The change in effort **on average** is determined completely by whether the covariate is an attribute or circumstance:
  - New attributes reduce average effort
  - New circumstances increase average effort
- But the change in effort may differ across agents not only in magnitude but also in sign — we call this **disparate impact**
- Demonstrate a statistical condition which guarantees that disparate impact does not emerge.
- Use these results to determine when measurement of a new covariate improves welfare.

# Illustrative Example

# Example: Labor Market

Worker output is

$$Y = e + \theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 100)$$

# Example: Labor Market

Worker output is

$$Y = e + \theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 100)$$

Worker productivity $\theta$ decomposes into a **residential stability** $\theta_2$ and a **worker reliability** $\theta_2$.

# Example: Labor Market

Worker output is

$$Y = e + \theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 100)$$

Worker productivity $\theta$ decomposes into a **residential stability** $\theta_2$ and a **worker reliability** $\theta_2$.

The two attributes are correlated: $\theta_1 \sim U([0,1])$ and

$$\theta_2|\theta_1 \sim \begin{cases} U([90,100]), & \theta_1 \geq 0.05 \\ U([0,100]), & \theta_1 < 0.05 \end{cases}$$

# Example: Labor Market

Worker output is

$$Y = e + \theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 100)$$

Worker productivity $\theta$ decomposes into a **residential stability** $\theta_2$ and a **worker reliability** $\theta_2$.

The two attributes are correlated: $\theta_1 \sim U\left([0,1]\right)$ and

$$\theta_2 | \theta_1 \sim \begin{cases} U([90, 100]), & \theta_1 \geq 0.05 \\ U([0, 100]), & \theta_1 < 0.05 \end{cases}$$

Workers with very low residential stability are less reliable on average and also very heterogeneous

- includes individuals who move frequently because of evictions
- Paul Erdős famously had no permanent residence

# Comparison Across Two Data Environments

Suppose no attributes are measured.

- All agents exert effort $e^* \approx 0.16$.

# Comparison Across Two Data Environments

Suppose no attributes are measured.

- All agents exert effort $e^* \approx 0.16$.

Suppose worker reliability remains unmeasured, but big data makes it possible to observe a worker's residential stability.

# Comparison Across Two Data Environments

Suppose no attributes are measured.

- All agents exert effort $e^* \approx 0.16$.

Suppose worker reliability remains unmeasured, but big data makes it possible to observe a worker's residential stability.

- Average effort **declines** to 0.11.

# Comparison Across Two Data Environments

Suppose no attributes are measured.

- All agents exert effort $e^* \approx 0.16$.

Suppose worker reliability remains unmeasured, but big data makes it possible to observe a worker's residential stability.

- Average effort **declines** to 0.11.
- Effort for infrequent movers **declines** to $e^{**} \approx 0.08$

# Comparison Across Two Data Environments

Suppose no attributes are measured.

- All agents exert effort $e^* \approx 0.16$.

Suppose worker reliability remains unmeasured, but big data makes it possible to observe a worker's residential stability.

- Average effort **declines** to 0.11.
- Effort for infrequent movers **declines** to $e^{**} \approx 0.08$
- Effort for frequent movers **rises** to $\tilde{e}^{**} \approx 0.82$

# Comparison Across Two Data Environments

Suppose no attributes are measured.

- All agents exert effort $e^* \approx 0.16$.

Suppose worker reliability remains unmeasured, but big data makes it possible to observe a worker's residential stability.

- Average effort **declines** to 0.11.
- Effort for infrequent movers **declines** to $e^{**} \approx 0.08$
  $\longrightarrow$ equilibrium payoffs **increase**
- Effort for frequent movers **rises** to $\tilde{e}^{**} \approx 0.82$
  $\longrightarrow$ equilibrium payoffs **decrease**

# Comparison Across Two Data Environments

Suppose no attributes are measured.

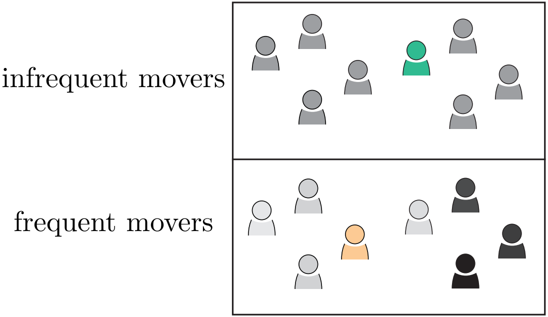
- All agents exert effort $e^* \approx 0.16$.

Suppose worker reliability remains unmeasured, but big data makes it possible to observe a worker's residential stability.

- Average effort **declines** to 0.11.
- Effort for infrequent movers **declines** to $e^{**} \approx 0.08$
  $\longrightarrow$ equilibrium payoffs **increase**
- Effort for frequent movers **rises** to $\tilde{e}^{**} \approx 0.82$
  $\longrightarrow$ equilibrium payoffs **decrease**

Measuring residential stability leads to **disparate impact**.

# Why Does Disparate Impact Occur?




Measurement of $\theta_1$ **redistributes uncertainty** across agents, with uncertainty about $\theta$ going down for one group and up for another.

- Lower uncertainty about $\theta$ reduces the value to effort
- Higher uncertainty about $\theta$ improves the value to effort

# Main Results

# Data Environments

To simplify notation, suppose in talk that:

$$\theta = \theta_1 + \theta_2$$

$$\varepsilon = \varepsilon_1 + \varepsilon_2$$

Compare the distribution of effort across two data environments:

- No covariates measured vs. $\theta_1$ measured
- No covariates measured vs. $\varepsilon_1$ measured

# Affiliated Covariates

$$\theta = \theta_1 + \theta_2 \qquad \varepsilon = \varepsilon_1 + \varepsilon_2$$

# Affiliated Covariates

$$\theta = \theta_1 + \theta_2 \qquad\qquad \varepsilon = \varepsilon_1 + \varepsilon_2$$

**Definition.** Say that attribute 1 is Affiliated if $(\theta_1, \theta_2)$ are affiliated, and circumstance 1 is Affiliated if $(\varepsilon_1, \varepsilon_2)$ are affiliated.

- Higher realizations for the measured $\theta_1$ imply higher realizations of the unmeasured $\theta_2$.

# Affiliated Covariates

$$\theta = \theta_1 + \theta_2 \qquad \varepsilon = \varepsilon_1 + \varepsilon_2$$

**Definition.** Say that attribute 1 is Affiliated if $(\theta_1, \theta_2)$ are affiliated, and circumstance 1 is Affiliated if $(\varepsilon_1, \varepsilon_2)$ are affiliated.

- Higher realizations for the measured $\theta_1$ imply higher realizations of the unmeasured $\theta_2$.

**Theorem**

(a) If attribute 1 is Affiliated, then measuring the attribute **reduces** average effort.

(b) If circumstance 1 is Affiliated, then measuring the circumstance **increases** average effort.

# General Intuition

Returns to effort depend on the sensitivity of $\mathbb{E}[\theta \mid Y, \theta_{\mathcal{J}}, \varepsilon_{\mathcal{K}}]$ to $Y$.

$$Y = e + \theta + \varepsilon$$

# General Intuition

Returns to effort depend on the sensitivity of $\mathbb{E}[\theta \mid Y, \theta_{\mathcal{J}}, \varepsilon_{\mathcal{K}}]$ to $Y$.

$$Y = e + \theta + \varepsilon$$

Lower uncertainty about $\theta$

# General Intuition

Returns to effort depend on the sensitivity of $\mathbb{E}[\theta \mid Y, \theta_{\mathcal{J}}, \varepsilon_{\mathcal{K}}]$ to $Y$.

$$Y = e + \theta + \varepsilon$$

Lower uncertainty about $\theta$

$\longrightarrow$ Variations in $Y$ are attributed to shock rather than type

# General Intuition

Returns to effort depend on the sensitivity of $\mathbb{E}[\theta \mid Y, \theta_{\mathcal{J}}, \varepsilon_{\mathcal{K}}]$ to $Y$.

$$Y = e + \theta + \varepsilon$$

Lower uncertainty about $\theta$

$\longrightarrow$ Variations in $Y$ are attributed to shock rather than type

$\longrightarrow$ **Lower** equilibrium effort.

# General Intuition

Returns to effort depend on the sensitivity of $\mathbb{E}[\theta \mid Y, \theta_{\mathcal{J}}, \varepsilon_{\mathcal{K}}]$ to $Y$.

$$Y = e + \theta + \varepsilon$$

Lower uncertainty about $\theta$

$\longrightarrow$ Variations in $Y$ are attributed to shock rather than type

$\longrightarrow$ **Lower** equilibrium effort.

Lower uncertainty about $\varepsilon$

# General Intuition

Returns to effort depend on the sensitivity of $\mathbb{E}[\theta \mid Y, \theta_{\mathcal{J}}, \varepsilon_{\mathcal{K}}]$ to $Y$.

$$Y = e + \theta + \varepsilon$$

Lower uncertainty about $\theta$

$\longrightarrow$ Variations in $Y$ are attributed to shock rather than type

$\longrightarrow$ **Lower** equilibrium effort.

Lower uncertainty about $\varepsilon$

$\longrightarrow$ Variations in $Y$ are attributed to type rather than shock

# General Intuition

Returns to effort depend on the sensitivity of $\mathbb{E}[\theta \mid Y, \theta_{\mathcal{J}}, \varepsilon_{\mathcal{K}}]$ to $Y$.

$$Y = e + \theta + \varepsilon$$

Lower uncertainty about $\theta$

$\longrightarrow$ Variations in $Y$ are attributed to shock rather than type

$\longrightarrow$ **Lower** equilibrium effort.

Lower uncertainty about $\varepsilon$

$\longrightarrow$ Variations in $Y$ are attributed to type rather than shock

$\longrightarrow$ **Higher** equilibrium effort.

# General Intuition

Returns to effort depend on the sensitivity of $\mathbb{E}[\theta \mid Y, \theta_{\mathcal{J}}, \varepsilon_{\mathcal{K}}]$ to $Y$.

$$Y = e + \theta + \varepsilon$$

Lower uncertainty about $\theta$

$\longrightarrow$ Variations in $Y$ are attributed to shock rather than type

$\longrightarrow$ **Lower** equilibrium effort.

Lower uncertainty about $\varepsilon$

$\longrightarrow$ Variations in $Y$ are attributed to type rather than shock

$\longrightarrow$ **Higher** equilibrium effort.

But what measurement of a new covariate implies for **ex post** uncertainty about $\theta$ and $\varepsilon$ is not straightforward.

# Two Effects

Consider measurement of $\theta_1$ (where recall that $\theta = \theta_1 + \theta_2$).

# Two Effects

Consider measurement of $\theta_1$ (where recall that $\theta = \theta_1 + \theta_2$).

**Direct** effect:

- $\theta_1$ is known, which taken alone implies lower uncertainty about $\theta$.

# Two Effects

Consider measurement of $\theta_1$ (where recall that $\theta = \theta_1 + \theta_2$).

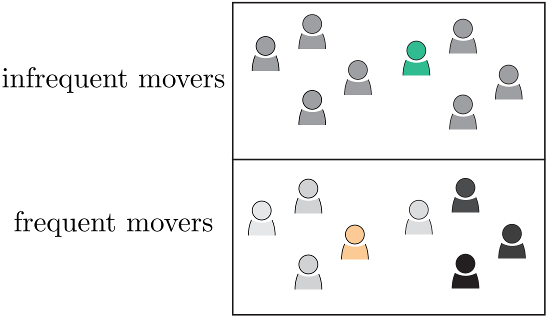
**Direct** effect:

- $\theta_1$ is known, which taken alone implies lower uncertainty about $\theta$.

**Indirect** effect (redistribution of uncertainty):

- Conditional uncertainty about $\theta_2$ (given $\theta_1$) may be large, leading uncertainty about $\theta$ to **increase** at some realizations of $\theta_1$
  $\longrightarrow$ higher value of effort at those realizations of $\theta_1$

# Redistribution of Uncertainty



# Two Effects

Consider measurement of $\theta_1$.

**Direct** effect:

- $\theta_1$ is known, which taken alone implies lower uncertainty about $\theta$.

**Indirect** effect (redistribution of uncertainty):

- Conditional uncertainty about $\theta_2$ (given $\theta_1$) may be large, leading uncertainty about $\theta$ to **increase** at some realizations of $\theta_1$
  $\longrightarrow$ higher value of effort at those realizations of $\theta_1$

# Two Effects

Consider measurement of $\theta_1$.

**Direct** effect:

- $\theta_1$ is known, which taken alone implies lower uncertainty about $\theta$.

**Indirect** effect (redistribution of uncertainty):

- Conditional uncertainty about $\theta_2$ (given $\theta_1$) may be large, leading uncertainty about $\theta$ to **increase** at some realizations of $\theta_1$
  $\longrightarrow$ higher value of effort at those realizations of $\theta_1$

Previous result says that when a covariate is Affiliated, then the average change in effort is in the same direction as the direct effect.

# Two Effects

Consider measurement of $\theta_1$.

**Direct** effect:

- $\theta_1$ is known, which taken alone implies lower uncertainty about $\theta$.

**Indirect** effect (redistribution of uncertainty):

- Conditional uncertainty about $\theta_2$ (given $\theta_1$) may be large, leading uncertainty about $\theta$ to **increase** at some realizations of $\theta_1$
  $\longrightarrow$ higher value of effort at those realizations of $\theta_1$

Previous result says that when a covariate is Affiliated, then the average change in effort is in the same direction as the direct effect.

Next result will further clarify these two forces by shutting down redistribution of uncertainty.

# Strong Homoskedasticity

$$\theta = \theta_1 + \theta_2 \qquad \varepsilon = \varepsilon_1 + \varepsilon_2$$

# Strong Homoskedasticity

$$\theta = \theta_1 + \theta_2 \qquad\qquad \varepsilon = \varepsilon_1 + \varepsilon_2$$

**Definition.** Say that attribute 1 satisfies **Strong Homoskedasticity**

$$\underbrace{\theta_2 - \mathbb{E}(\theta_2 \mid \theta_1)}_{\text{"de-meaned residual"}} \perp\!\!\!\perp \theta_1$$

and say that circumstance 1 satisfies **Strong Homoskedasticity** if

$$\varepsilon_2 - \mathbb{E}(\varepsilon_2 \mid \varepsilon_1) \perp\!\!\!\perp \varepsilon_1$$

# Strong Homoskedasticity

$$\theta = \theta_1 + \theta_2 \qquad\qquad \varepsilon = \varepsilon_1 + \varepsilon_2$$

**Definition.** Say that attribute 1 satisfies **Strong Homoskedasticity**

$$\underbrace{\theta_2 - \mathbb{E}(\theta_2 \mid \theta_1)}_{\text{“de-meaned residual”}} \perp\!\!\!\perp \theta_1$$

and say that circumstance 1 satisfies **Strong Homoskedasticity** if

$$\varepsilon_2 - \mathbb{E}(\varepsilon_2 \mid \varepsilon_1) \perp\!\!\!\perp \varepsilon_1$$

**Interpretation:** Distribution of the unmeasured term does not depend on the measured covariate, except possibly in mean.

# Strong Homoskedasticity

$$\theta = \theta_1 + \theta_2 \qquad\qquad \varepsilon = \varepsilon_1 + \varepsilon_2$$

**Definition.** Say that attribute 1 satisfies **Strong Homoskedasticity**

$$\underbrace{\theta_2 - \mathbb{E}(\theta_2 \mid \theta_1)}_{\text{"de-meaned residual"}} \perp\!\!\!\perp \theta_1$$

and say that circumstance 1 satisfies **Strong Homoskedasticity** if

$$\varepsilon_2 - \mathbb{E}(\varepsilon_2 \mid \varepsilon_1) \perp\!\!\!\perp \varepsilon_1$$

**Interpretation:** Distribution of the unmeasured term does not depend on the measured covariate, except possibly in mean.

**Examples**:

- Multivariate normal covariates (with any covariance matrix)
- Additive shifts, e.g. $\theta_2 = X + \theta_1$ for any $X \perp\!\!\!\perp \theta_1$

# Strongly Homoskedastic Covariates

Previous directional change in average effort manifests as a stronger uniform effect:

**Theorem**

(a) If attribute 1 satisfies Strong Homoskedasticity, then measuring the attribute reduces **every agent's** effort.

(b) If circumstance 1 satisfies Strong Homoskedasticity, then measuring the circumstance increases **every agent's** effort.

Moreover the size of change is the same for every agent.

- No disparate impact
- Can interpret as a limiting case where redistribution of uncertainty becomes small.

# Welfare and Regulation

# Which Covariates Should be Permitted?

Suppose measurement of a new covariate becomes available for forecasting

When should the social planner permit the market access to this covariate?

# Measuring Welfare

Welfare = expected social surplus generated by the agent's effort:

$$w(\theta, e) \equiv \mathbb{E}^{e}(Y \mid \theta) - C(e)$$

# Measuring Welfare

Welfare = expected social surplus generated by the agent's effort:

$$w(\theta, e) \equiv \mathbb{E}^e(Y \mid \theta) - C(e) = \theta + e - \frac{1}{2}e^2.$$

Strictly concave in $e$ and maximized at $e^{FB} = 1$ for every $\theta$.

# Measuring Welfare

Welfare = expected social surplus generated by the agent's effort:

$$w(\theta, e) \equiv \mathbb{E}^e(Y \mid \theta) - C(e) = \theta + e - \frac{1}{2}e^2.$$

Strictly concave in $e$ and maximized at $e^{FB} = 1$ for every $\theta$.

For any family of measured covariates $(\mathcal{J}, \mathcal{K})$, let $e^*_{\mathcal{J},\mathcal{K}}$ denote the (random) equilibrium effort.

# Measuring Welfare

Welfare = expected social surplus generated by the agent's effort:

$$w(\theta, e) \equiv \mathbb{E}^e(Y \mid \theta) - C(e) = \theta + e - \frac{1}{2}e^2.$$

Strictly concave in $e$ and maximized at $e^{FB} = 1$ for every $\theta$.

For any family of measured covariates $(\mathcal{J}, \mathcal{K})$, let $e^*_{\mathcal{J},\mathcal{K}}$ denote the (random) equilibrium effort.

Aggregate welfare is

$$W(\mathcal{J}, \mathcal{K}) \equiv \mathbb{E}\left(w(\theta, e^*_{\mathcal{J},\mathcal{K}})\right) = \mu + \mathbb{E}\left(e^*_{\mathcal{J},\mathcal{K}} - \frac{1}{2}{e^*_{\mathcal{J},\mathcal{K}}}^2\right)$$

# Measuring Welfare

Welfare = expected social surplus generated by the agent's effort:

$$w(\theta, e) \equiv \mathbb{E}^e(Y \mid \theta) - C(e) = \theta + e - \frac{1}{2}e^2.$$

Strictly concave in $e$ and maximized at $e^{FB} = 1$ for every $\theta$.

For any family of measured covariates $(\mathcal{J}, \mathcal{K})$, let $e^*_{\mathcal{J},\mathcal{K}}$ denote the (random) equilibrium effort.

Aggregate welfare is

$$W(\mathcal{J}, \mathcal{K}) \equiv \mathbb{E}\left(w(\theta, e^*_{\mathcal{J},\mathcal{K}})\right) = \mu + \mathbb{E}\left(e^*_{\mathcal{J},\mathcal{K}} - \frac{1}{2}{e^*_{\mathcal{J},\mathcal{K}}}^2\right)$$

where $\mu \equiv \mathbb{E}(\theta)$ denotes the average type in the population.

# Inefficiency of Equilibrium Effort

Given any set of measured covariates $(\mathcal{J}, \mathcal{K})$, the (random) **marginal value of effort** is

$$MV_{\mathcal{J},\mathcal{K}} = \frac{\partial}{\partial e}\mathbb{E}^{e}\left(\mathbb{E}^{e^*}(\theta \mid Y, a_{\mathcal{J}}, c_{\mathcal{K}}) \mid a_{\mathcal{J}}, c_{\mathcal{K}}\right)\Big|_{e=e^*}$$

for any effort level $e^*$.

# Inefficiency of Equilibrium Effort

Given any set of measured covariates $(\mathcal{J}, \mathcal{K})$, the (random) **marginal value of effort** is

$$MV_{\mathcal{J},\mathcal{K}} = \frac{\partial}{\partial e}\mathbb{E}^{e}\left(\mathbb{E}^{e^*}(\theta \mid Y, a_{\mathcal{J}}, c_{\mathcal{K}}) \mid a_{\mathcal{J}}, c_{\mathcal{K}}\right)\Big|_{e=e^*}$$

for any effort level $e^*$.

Equilibrium effort $e^*_{\mathcal{J},\mathcal{K}} = \beta \cdot MV_{\mathcal{J},\mathcal{K}}$ is FOSD increasing in $\beta$.

# Inefficiency of Equilibrium Effort

Given any set of measured covariates $(\mathcal{J}, \mathcal{K})$, the (random) **marginal value of effort** is

$$MV_{\mathcal{J},\mathcal{K}} = \frac{\partial}{\partial e}\mathbb{E}^{e}\left(\mathbb{E}^{e^*}(\theta \mid Y, a_{\mathcal{J}}, c_{\mathcal{K}}) \mid a_{\mathcal{J}}, c_{\mathcal{K}}\right)\Big|_{e=e^*}$$

for any effort level $e^*$.

Equilibrium effort $e^*_{\mathcal{J},\mathcal{K}} = \beta \cdot MV_{\mathcal{J},\mathcal{K}}$ is FOSD increasing in $\beta$.

- (Recall that agent's total payoff is $(1-\beta)U_1 + \beta U_2$ where $U_1$ and $U_2$ are the payoffs from the two periods.)
- Effort can be inefficiently high or low depending on size of reputation weight $\beta$.

# Regularity

**Definition**

Fix a baseline set of measured covariates $(\mathcal{J}, \mathcal{K})$. Then:

- An attribute $j' \notin \mathcal{J}$ is **regular** if measuring $j'$ decreases average effort.
- A circumstance $k' \notin \mathcal{K}$ is **regular** if measuring $k'$ increases average effort.

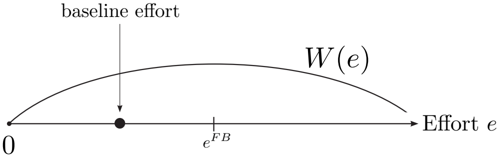
All Strongly Homoskedastic and Affiliated covariates are regular (by previous two results).

# Basic Forces

Suppose for simplicity that effort in the baseline and expanded environments deterministic (e.g. if population size $= 1$).

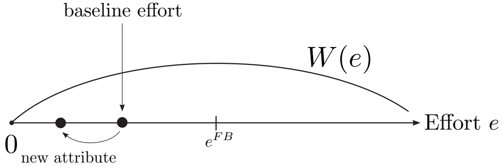
# Basic Forces

Suppose for simplicity that effort in the baseline and expanded environments deterministic (e.g. if population size = 1).



# Basic Forces

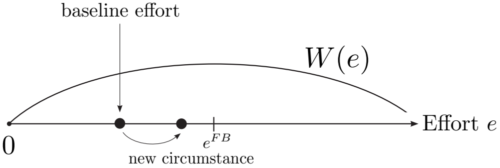
Suppose for simplicity that effort in the baseline and expanded environments deterministic (e.g. if population size = 1).




When baseline effort is below first-best, a new attribute always reduces welfare.

# Basic Forces

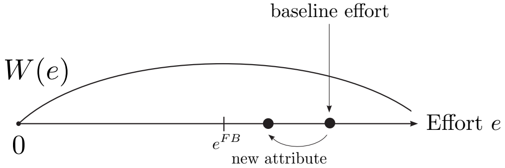
Suppose for simplicity that effort in the baseline and expanded environments deterministic (e.g. if population size = 1).




When baseline effort is below first-best, a new circumstance may improve welfare.

# Basic Forces

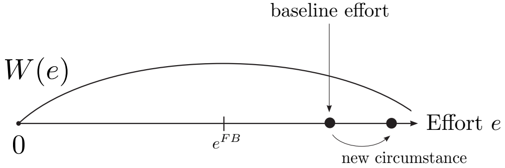
Suppose for simplicity that effort in the baseline and expanded environments deterministic (e.g. if population size = 1).




When baseline effort exceeds first-best, a new attribute may improve welfare.

# Basic Forces

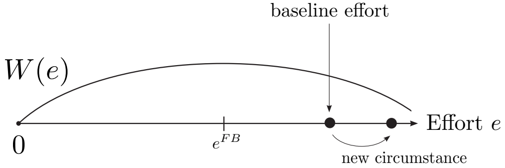
Suppose for simplicity that effort in the baseline and expanded environments deterministic (e.g. if population size = 1).




When baseline effort exceeds first-best, a new circumstance always harm welfare.

# Basic Forces

Suppose for simplicity that effort in the baseline and expanded environments deterministic (e.g. if population size = 1).




When baseline effort exceeds first-best, a new circumstance always harm welfare.

But more generally, effort is random both in the baseline and after measurement of the new covariate.

# Welfare Impact of New Covariates

**Proposition**

Fix any baseline family of measured covariates $(\mathcal{J}, \mathcal{K})$.

(a) For every regular attribute $j' \notin \mathcal{J}$, there is a $\beta^* \in (0, \infty]$ such that measuring $j'$ improves welfare iff $\beta \geq \beta^*$.

# Welfare Impact of New Covariates

**Proposition**

Fix any baseline family of measured covariates $(\mathcal{J}, \mathcal{K})$.

(a) For every regular attribute $j' \notin \mathcal{J}$, there is a $\beta^* \in (0, \infty]$ such that measuring $j'$ improves welfare iff $\beta \geq \beta^*$.

(b) For every regular circumstance $k' \notin \mathcal{K}$, there is a $\beta_* \in (0, \infty)$ such that measuring $k'$ improves welfare iff $\beta \leq \beta_*$.

# Welfare Impact of New Covariates

**Proposition**

Fix any baseline family of measured covariates $(\mathcal{J}, \mathcal{K})$.

(a) For every regular attribute $j' \notin \mathcal{J}$, there is a $\beta^* \in (0, \infty]$ such that measuring $j'$ improves welfare iff $\beta \geq \beta^*$.

(b) For every regular circumstance $k' \notin \mathcal{K}$, there is a $\beta_* \in (0, \infty)$ such that measuring $k'$ improves welfare iff $\beta \leq \beta_*$.

- Since $\beta_* > 0$, there is always a nondegenerate interval of reputation weights where the circumstance improves welfare.

# Welfare Impact of New Covariates

**Proposition**

Fix any baseline family of measured covariates $(\mathcal{J}, \mathcal{K})$.

(a) For every regular attribute $j' \notin \mathcal{J}$, there is a $\beta^* \in (0, \infty]$ such that measuring $j'$ improves welfare iff $\beta \geq \beta^*$.

(b) For every regular circumstance $k' \notin \mathcal{K}$, there is a $\beta_* \in (0, \infty)$ such that measuring $k'$ improves welfare iff $\beta \leq \beta_*$.

- Since $\beta_* > 0$, there is always a nondegenerate interval of reputation weights where the circumstance improves welfare.
- But $\beta^* = \infty$ is possible, so the same is not true for attributes.

# Disparate Impact and Welfare Improvement

**Proposition**

Fix a baseline set of measured covariates $(\mathcal{J}, \mathcal{K})$ and a new regular attribute $j' \notin \mathcal{J}$. Then $\beta^* < \infty$ if and only if

$$\mathbb{E}\left(MV^2_{\mathcal{J}\cup\{j'\},\mathcal{K}}\right) < \mathbb{E}\left(MV^2_{\mathcal{J},\mathcal{K}}\right)$$

This inequality is satisfied if $j'$ does not produce disparate impact.

# Disparate Impact and Welfare Improvement

## Proposition

Fix a baseline set of measured covariates $(\mathcal{J}, \mathcal{K})$ and a new regular attribute $j' \notin \mathcal{J}$. Then $\beta^* < \infty$ if and only if

$$\mathbb{E}\left(MV^2_{\mathcal{J}\cup\{j'\},\mathcal{K}}\right) < \mathbb{E}\left(MV^2_{\mathcal{J},\mathcal{K}}\right)$$

This inequality is satisfied if $j'$ does not produce disparate impact.

## Corollary

In our previous labor market example,

$$\mathbb{E}\left(MV^2_{\{1\},\emptyset}\right) \approx 0.039 > MV^2_{\emptyset,\emptyset} \approx 0.026.$$

Therefore aggregate welfare decreases upon observing attribute 1 for any reputation weight $\beta$.

# Regulation of Covariates: Takeaways

When reputation weights $\beta$ are too low, then **allow circumstances and ban attributes**.

When reputation weights $\beta$ are too high, then **allow attributes and ban circumstances**.

# Regulation of Covariates: Takeaways

When reputation weights $\beta$ are too low, then **allow circumstances and ban attributes**.

When reputation weights $\beta$ are too high, then **allow attributes and ban circumstances**.

In either case, disparate impact is harmful—prefer covariates that divide the population up into "similarly well understood" groups.

# Extensions

Model uncertainty:

- Agent is uncertain about observed covariates $(\mathcal{J}, \mathcal{K})$ or subpopulation distributions
- **All main results extend** as long as required statistical assumptions hold "model by model"

# Extensions

Model uncertainty:

- Agent is uncertain about observed covariates $(\mathcal{J}, \mathcal{K})$ or subpopulation distributions
- **All main results extend** as long as required statistical assumptions hold "model by model"

General nonlinear model:

$$\theta = \mu + f(\boldsymbol{a}) + u_\theta$$
$$\varepsilon = g(\boldsymbol{c}) + u_\varepsilon$$

# Extensions

Model uncertainty:

- Agent is uncertain about observed covariates $(\mathcal{J}, \mathcal{K})$ or subpopulation distributions
- **All main results extend** as long as required statistical assumptions hold "model by model"

General nonlinear model:

$$\theta = \mu + f(\boldsymbol{a}) + u_\theta$$
$$\varepsilon = g(\boldsymbol{c}) + u_\varepsilon$$

- **All main results have analogues**

# Conclusion

# Conclusion

Big data has the potential to reshape reputational incentives for socially relevant effort.

# Conclusion

Big data has the potential to reshape reputational incentives for socially relevant effort.

Our paper tackles two related questions:

- What is the effect of access to new data on reputational incentives for effort?
- When should a regulator permit/forbid use of a new covariate in forecasting?

# Conclusion

Big data has the potential to reshape reputational incentives for socially relevant effort.

Our paper tackles two related questions:

- What is the effect of access to new data on reputational incentives for effort?
- When should a regulator permit/forbid use of a new covariate in forecasting?

Future work:

- What data gets collected in a competitive market?
- Endogenous covariates