Information Discrepancy in Strategic Learning

Chara Podimata (UC Berkeley → MIT)

Joint work with Yahav Bechavod (Hebrew University), Steven Wu (CMU), Juba Ziani (Georgia Tech)
ML algorithms for decision-making are almost everywhere nowadays.

Is an Algorithm Less Racist Than a Loan Officer?
Digital mortgage platforms have the potential to reduce discrimination. But automated systems provide rich opportunities to perpetuate bias, too.

- increase # credit cards
- increase # bank accounts
- improve credit history

Your end-to-end hiring platform with video interview software, conversational AI, and assessments.

Build a faster, fairer, friendlier hiring process with HireVue's end-to-end hiring platform. Together, we can improve the way you discover, engage, and hire talent.

- dress a certain way
- hide piercings/tattoos
- change way you talk

Student tracking, secret scores: How college admissions offices rank prospects before they apply

- improve GPA
- retake GRE/pay for classes
- change schools
Problem

If ML algorithms ignore this “strategic”/“responsive” behavior, they risk making policy decisions that are incompatible with the original policy’s goal.
I study the effects of “strategic” behavior both to institutions and society as a whole and propose ways to adapt ML algorithms to it.
Incentive-Aware ML Stakeholders

**Institution**
- **Who?** mechanism/algorithm designers
- **Goal:** profit, justice, ...
- **Action:** learning task for accurate prediction

**Individual**
- **Who?** Person (data provider)
- **Goal:** get best outcomes for them
- **Action:** change their data

**Society**
- **Who?** All people as a whole
- **Goal:** fairness, robustness, welfare
- **Action:** regulate, public pressure
Incentive-Aware ML Stakeholders

Contributions

1) Algorithms robust to incentives. 
[CP, EC18 (best paper finalist)], [FP, ICML20], [CL, NeurIPS20]

2) Algorithms robust to irrationalities. 
[KL, STOC21 & OR22], [LP, COLT22]

Contributions

Societal effects of non-transparency. 
[BPWZ, ICML22]
Lots of Recent, Exciting Work


- **Fairness:** [Milli, Miller, Dragan, Hardt, *FAT*19], [Hu, Immorlica, Vaughan, *FAT*19], [Liu, Wilson, Haghtalab, Kalai, Borgs, Chayes, *FAT*19], [Braverman, Garg, *FORC20*]


Strategic/Incentive-Aware Learning

Mathematically:
- Learner commits to a decision rule \( w: \mathcal{X} \rightarrow [0,1] \)
- Agent with feature vector \( x \in \mathcal{X} \) and score \( y \in [0,1] \), observes \( w \) and best-responds by reporting
\[
\hat{x}(w) = \arg \max_{x' \in \mathcal{X}} u(x; w)
\]
- Learner’s rule = Stackelberg equilibrium. For example: \( w = \arg \min_{w'} (w', \hat{x}(w')) - y)^2 \)
Strategic/Incentive-Aware Learning

Mathematically:

- Learner commits to a decision rule $w: \mathcal{X} \rightarrow [0,1]$
- Agent with feature vector $x \in \mathcal{X}$ and score $y \in [0,1]$, observes $w$ and best-responds by reporting $\hat{x}(w) = \arg \max_{x' \in \mathcal{X}} u(x; w)$
- Learner’s rule = Stackelberg equilibrium. For example: $w = \arg \min_{w'} (w', \hat{x}(w')) - y)^2$
Strategic/Incentive-Aware Learning Revisited

In reality: institutions **rarely reveal** their decision rules (reasons: privacy, proprietary software etc)!

Instead: explanations or examples of past decisions
Our Model at a High Level

Decision-making rule (e.g., regression etc)

Learner

Policy

Strategically change features
Our Model Formally

Interaction Protocol

1. Nature decides the ground truth assessment: $\mathbf{w}^* \in \mathbb{R}^d$.
2. Learner deploys score rule $\mathbf{w} \in \mathbb{R}^d$ but does not reveal it to agents.
3. Agents (per subgroup $g$) draw their private feature vectors from space $\mathcal{X}$: $x_1 \sim \mathcal{D}_1$ and $x_2 \sim \mathcal{D}_2$.
4. Given peer dataset $S_g$, private feature vector $x_g$, & their utility $u(x_g, x_g'; g)$, the agents best-respond with feature vector: $\hat{x}_g = \arg \max_{x'} u(x_g, x'; g)$.

**Learner’s Goal**

Choose decision rule that maximizes the social welfare wrt the ground truth assessment

$$\mathbf{w} = \arg \max_{\mathbf{w}'} \left( \mathbb{E}_{x_1 \sim \mathcal{D}_1} [\langle \hat{x}_1, \mathbf{w}^* \rangle] + \mathbb{E}_{x_2 \sim \mathcal{D}_2} [\langle \hat{x}_2, \mathbf{w}^* \rangle] \right)$$

Why is $\mathbf{w} \neq \mathbf{w}^*$?
How do information discrepancies regarding the principal’s decision rule affect the ability of the agents to improve their outcomes?
Our Model Formally

Interaction Protocol

1. Nature decides the ground truth assessment: \( w^* \in \mathbb{R}^d \).
2. Learner deploys score/decision rule \( w \in \mathbb{R}^d \) but does not reveal it to agents.
3. Agents (per subgroup \( g \)) draw their private feature vectors from space \( \mathcal{X} \): \( x_1 \sim \mathcal{D}_1 \) and \( x_2 \sim \mathcal{D}_2 \).
4. Given peer dataset \( S_g \), private feature vector \( x_g \), & their utility \( u(x_g, x'_g; g) \), the agents best-respond with feature vector: \( \hat{x}_g = \arg \max_{x'} u(x_g, x'; g) \).

Subgroup Feature Vector Discrepancies

- \( S_1, S_2 \): subspaces of \( \mathcal{X} \) defined by the supports of \( \mathcal{D}_1, \mathcal{D}_2 \)
- \( \Pi_1, \Pi_2 \in \mathbb{R}^d \): orthogonal projection matrices onto \( S_1, S_2 \rightarrow x_g = \Pi_g x_g \) (feature discrepancy)

Subgroup Utilities

Score they get with their estimated decision rule
\[
u(x_g, x'; g) = \frac{\text{EstScore}(x')}{-\text{Cost}(x_g \rightarrow x')}
\]
\[
= \langle x', w_{\text{est}}(g) \rangle - \| A_g (x_g - x') \|_2
\]
How Do the Subgroups Estimate $w$

Each subgroup runs ERM on labeled examples to recover $w$. \( \Rightarrow \) Recovers: $w_{est}(g) = \Pi_g w$
**Principal’s Equilibrium Decision Rule**

- Agents’ best response: 
  \[ \hat{x}_g = \arg \max_{x'} u(x_g, x'; g) \]
  \[ \Rightarrow \hat{x}_g = x + A_g^{-1} \Pi_g w = x + \Delta_g(w) \]

- Principal’s rule optimizing SW: 
  \[ w_{SW} = \arg \max_w (\mathbb{E}_{x_1 \sim D_1} [(\hat{x}_1, w^*)] + \mathbb{E}_{x_2 \sim D_2} [(\hat{x}_2, w^*)]) \]
  \[ = \frac{\left( \Pi_1 A_1^{-1} + \Pi_2 A_2^{-1} \right) w^*}{\| (\Pi_1 A_1^{-1} + \Pi_2 A_2^{-1}) w^* \|} \]

**Answer:**

- Sometimes (e.g., \( A_1 = A_2 = I \) and \( \Pi_1 + \Pi_2 = \mathcal{X} \)).
- In general, not true.

  1. Disparities in feature modification costs
  2. Maybe worth incentivizing feature changes that benefit both groups

Example: \( w^* = \left( \frac{2}{3}, \frac{2}{3}, \frac{1}{3} \right) \) and \( \Pi_1 = (1, 0, 1), \Pi_2 = (0, 1, 1). \Delta(SW(w^*)) = 10/9. \)

For \( w = \frac{1}{\sqrt{3}} (1,1,1): \Delta(SW(w)) > 10/9 \)
Measures of Outcome Improvement in Equilibrium

Improvement for group $g$: $J_g(w) = \langle \hat{x}(w), w^* \rangle - \langle x, w^* \rangle$

1. Do-no-harm: “Are all individuals better off?”
2. Total improvement: “By how much?”
3. Per-unit improvement: “Is effort exerted optimally?”
Results

\[ I_g(w) = \langle \hat{x}(w), w^* \rangle - \langle x, w^* \rangle = \langle A_g^{-1} \Pi_g w, w^* \rangle \]

1. **Do-no-harm**: “Are all individuals better off?”
2. **Total improvement**: “By how much?”
3. **Per-unit improvement**: “Is effort exerted optimally?”

For general costs and projection matrices: \textbf{NO!}

→ “contentious” information from each group, but principal still maximizing the total social welfare

**Notable examples for guaranteeing no negative externality:**

(1) Proportional movement costs \( A_1 = c \cdot A_2 \)
(2) Non-interfering information: \( \Pi_1 \perp \Pi_2 \)
Results

Improvement for group $g$: $J_g(w) = \langle \hat{x}(w), w^* \rangle - \langle x, w^* \rangle = \langle A_g^{-1} \Pi_g w, w^* \rangle$

1. Do-no-harm: “Are all individuals better off?”
2. Total improvement: “By how much?”
3. Per-unit improvement: “Is effort exerted optimally?”

In general: $|J_1(w) - J_2(w)| \leq ||\Pi_1 w^* - \Pi_2 w^*||_2$

Equal outcome improvement iff: $A_1^{-1} \Pi_1 A_1^{-1} = A_2^{-1} \Pi_2 A_2^{-1}$
Results

Improvement for group $g$: $J_g(w) = \langle \hat{x}(w), w^* \rangle - \langle x, w^* \rangle = \langle A_g^{-1}\Pi_g w, w^* \rangle$

1. Do-no-harm: “Are all individuals better off?”
2. Total improvement: “By how much?”
3. Per-unit improvement: “Is effort exerted optimally?”

Properties
- Considers only the part of the decision rule that belongs in the relevant subspace for each group
- Measures how efficient the direction of this rule projected onto the relevant subspace is at inducing improvement for the group

Notable examples for optimal effort exertion:
1. Non-interfering information: $\Pi_1 \perp \Pi_2$
2. Proportional movement costs and $\Pi_1 = \Pi_2$. 
The Adult Dataset

- Publicly available at UCI repository: [https://archive.ics.uci.edu/ml/datasets/adult](https://archive.ics.uci.edu/ml/datasets/adult)
- ~50K datapoints
- 14 attributes including Age, Country, Workclass, Education, Race, etc.
- Label (annual income): <50K, >= 50K

**Our process:**

- 4 experiments separating subpopulations based on:

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Subpopulation 1</th>
<th>Subpopulation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>&lt;35 yrs old</td>
<td>&gt;=35 yrs old</td>
</tr>
<tr>
<td>Country</td>
<td>All others</td>
<td>Western countries</td>
</tr>
<tr>
<td>Education</td>
<td>All others</td>
<td>Above high school</td>
</tr>
<tr>
<td>Race</td>
<td>All others</td>
<td>White</td>
</tr>
</tbody>
</table>

- Predict income **improvement** (final income – original income) for each subpopulation.
Results Snapshot: Adult Dataset

1. One subpopulation may get worse off.

- Total income improvement currently subpopulation 1
- Total income improvement currently subpopulation 2

Subpopulations breakdown criteria

<table>
<thead>
<tr>
<th>Subpopulation</th>
<th>Race</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>All others</td>
</tr>
<tr>
<td>2</td>
<td>White</td>
</tr>
</tbody>
</table>
Total improvement may be very unequal across subpopulations.

- Total income improvement currently subpopulation 1
- Total income improvement currently subpopulation 2
Summary

When there exists information discrepancy regarding the decision-making rule among the subgroups:

1. **Do-no-harm**: “Are all individuals better off?”  
   Not in general! Yes, if (e.g.,) proportional movement costs or non-conflicting information between subgroups.

2. **Total improvement**: “By how much?”  
   Equal among subgroups if $A_1^{-1} \Pi_1 A_1^{-1} = A_2^{-1} \Pi_2 A_2^{-1}$

3. **Per-unit improvement**: “Is effort exerted optimally?”  
   Yes if (e.g.,) non-interfering information: $\Pi_1 \perp \Pi_2$ or proportional movement costs and $\Pi_1 = \Pi_2$. 
Summary

When there exists information discrepancy regarding the decision-making rule among the subgroups:

1. **Do-no-harm**: “Are all individuals better off?” Not in general! Yes, if (e.g.,) proportional movement costs or non-conflicting information between subgroups.

2. **Total improvement**: “By how much?” Equal among subgroups if $A_1^{-1} \Pi_1 A_3^{-1} = A_2^{-1} \Pi_2 A_3^{-1}$.

3. **Per-unit improvement**: “Is effort exerted optimally?” Yes if (e.g.,) non-interfering information: $\Pi_1 \perp \Pi_2$, or proportional movement costs and $\Pi_1 = \Pi_2$. 
When there exists information discrepancy regarding the decision-making rule among the subgroups:

1. **Do-no-harm**: “Are all individuals better off?” Not in general! Yes, if (e.g.,) proportional movement costs or non-conflicting information between subgroups.

2. **Total improvement**: “By how much?” Equal among subgroups if \( A_1^{-1} \Pi_1 A_1^{-1} = A_2^{-1} \Pi_2 A_2^{-1} \)

3. **Per-unit improvement**: “Is effort exerted optimally?” Yes if (e.g.,) non-interfering information: \( \Pi_1 \perp \Pi_2 \) or proportional movement costs and \( \Pi_1 = \Pi_2 \).
Summary

When there exists information discrepancy regarding the decision-making rule among the subgroups:

1. **Do-no-harm**: “Are all individuals better off?” Not in general! Yes, if (e.g.,) proportional movement costs or non-conflicting information between subgroups.

2. **Total improvement**: “By how much?” Equal among subgroups if $A_1^{-1} \Pi_1 A_1^{-1} = A_2^{-1} \Pi_2 A_2^{-1}$

3. **Per-unit improvement**: “Is effort exerted optimally?” Yes if (e.g.,) non-interfering information: $\Pi_1 \perp \Pi_2$ or proportional movement costs and $\Pi_1 = \Pi_2$.

Extensions included in the paper:

(1) Principal’s learning problem when $\Pi_g$’s, $A_g$’s, and $w^*$ are not known a priori.

(2) Generalization for $g \geq 3$.

(3) Principal that cares about a combination of accuracy and social welfare.
Interpretability and Incentives

"Obscure" ML algorithms
+ Stop strategic behavior
- Non-transparent

Policy

Rule is known to the agents.

Decision-making rule (e.g., classification/regression etc)

Strategically change features

Learner

Policy
Interpretability and Incentives

"Obscure" ML algorithms
+ Stop strategic behavior
- Non-transparent

Public ML algorithms
+ Incentivize efforts for outcome improvement.
- Prone to strategic behavior

Learner
Decision-making rule (e.g., classification/regression etc)

Rule is known to the agents.

Policy

Strategically change features
Interpretability and Incentives

"Obscure" ML algorithms
+ Stop strategic behavior
- Non-transparent

Interpretable ML Algorithms

Public ML algorithms
+ Incentivize efforts for outcome improvement.
- Prone to strategic behavior

Rule is interpretable by the agents.

Policy

Strategically change features

Decision-making rule (e.g., classification/regression etc)

Learner

[Diagram showing interactions and decision-making process]
Interpretability and Incentives

Current state of Incentive-Aware ML research

• Learner: **Non-linear rules** (e.g., coming from neural nets).
• Agent: **understand** rules fully + **best-respond**

Large **case studies** to move from theory to practice and drive change.

**Interpretable** ML rules that are **robust to strategizing** but **incentivize** honest outcome improvement.

Tutorial at FAccT21

Thank you!