Information Discrepancy in Strategic Learning

Chara Podimata (UC Berkeley → MIT)

Joint work with Yahav Bechavod (Hebrew University), Steven Wu (CMU), Juba Ziani (Georgia Tech)

TOC4Fairness, October 2022

ML algorithms for **decision-making** are almost everywhere nowadays.

The New York Times

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

The Washington Post
Democracy Dies in Darkness

Business

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

Before many schools even look at an application, they comb through prospective students' personal data, such as web-browsing habits and financial history

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

HireVue Platform Why HireVue Hiring Resources

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.

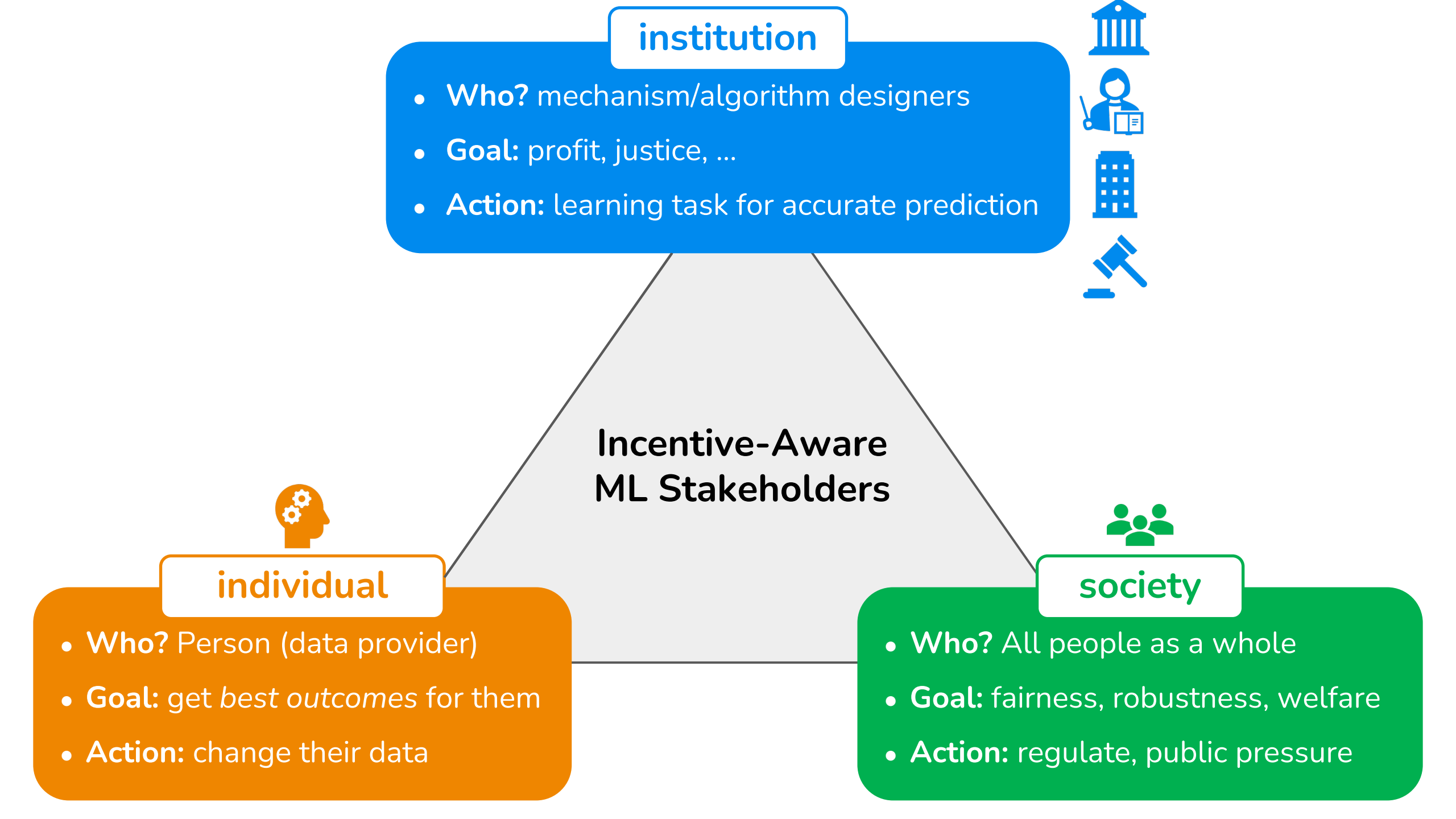
- dress a certain way
- hide piercings / tattoos
- change way you talk

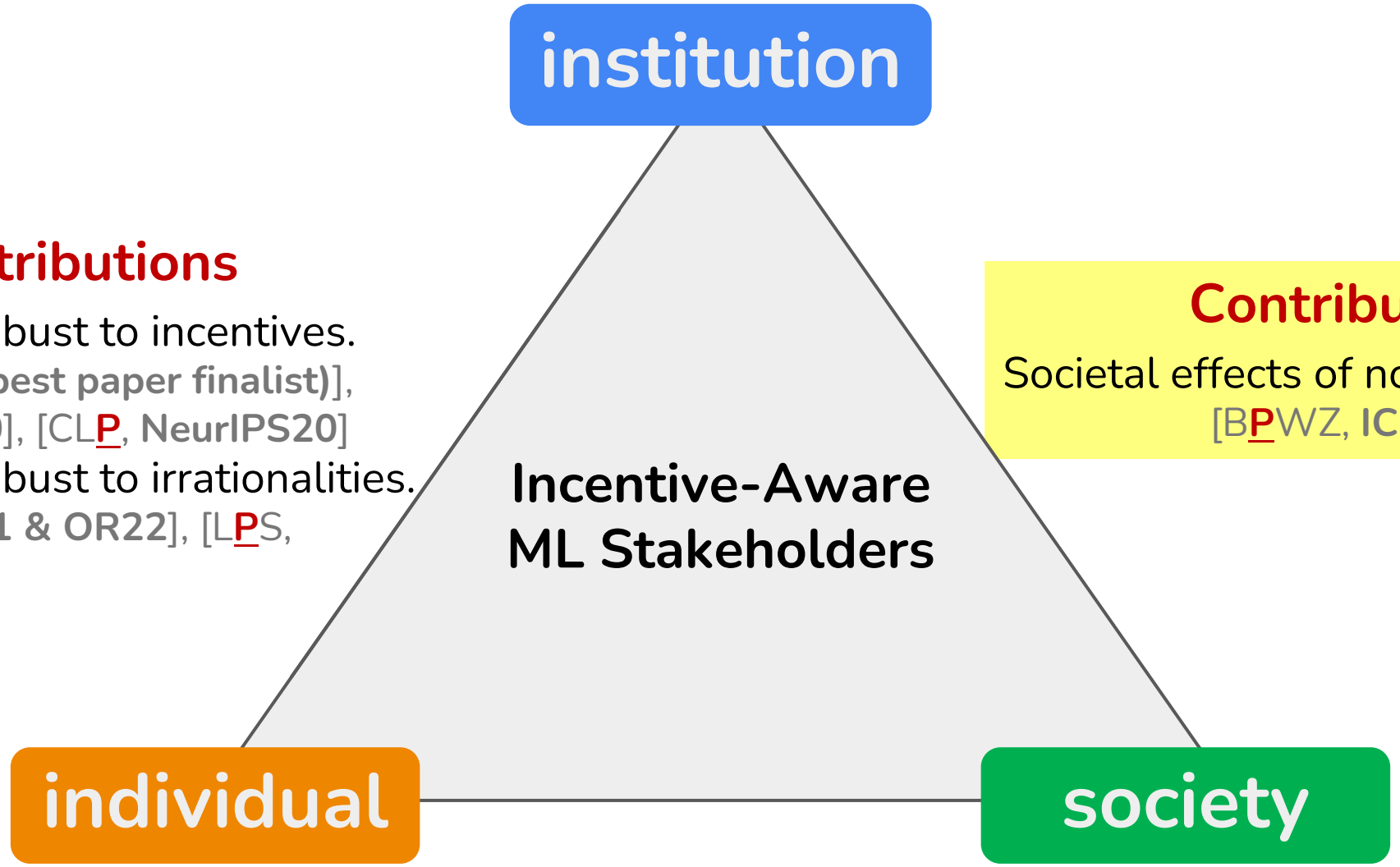
Problem

If ML algorithms **ignore** this **“strategic”/“responsive” behavior**, they risk making **policy decisions** that are **incompatible with the original policy’s goal**.

My Research Agenda: Incentive-Aware ML

I study the **effects of "strategic" behavior** both to institutions and society as a whole and propose **ways to adapt ML** algorithms to it.


institution
Who? mechanism/algorithm designers
Goal: profit, justice, ...
Action: learning task for accurate prediction
Incentive-Aware
ML Stakeholders
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



Contributions

1) Algorithms robust to incentives.
   [CPPS, **EC18 (best paper finalist)**],
   [FPPV, **ICML20**], [CLP, **NeurIPS20**]
2) Algorithms robust to irrationalities.
   [KLPS, **STOC21 & OR22**], [LPS,
   COLT22]

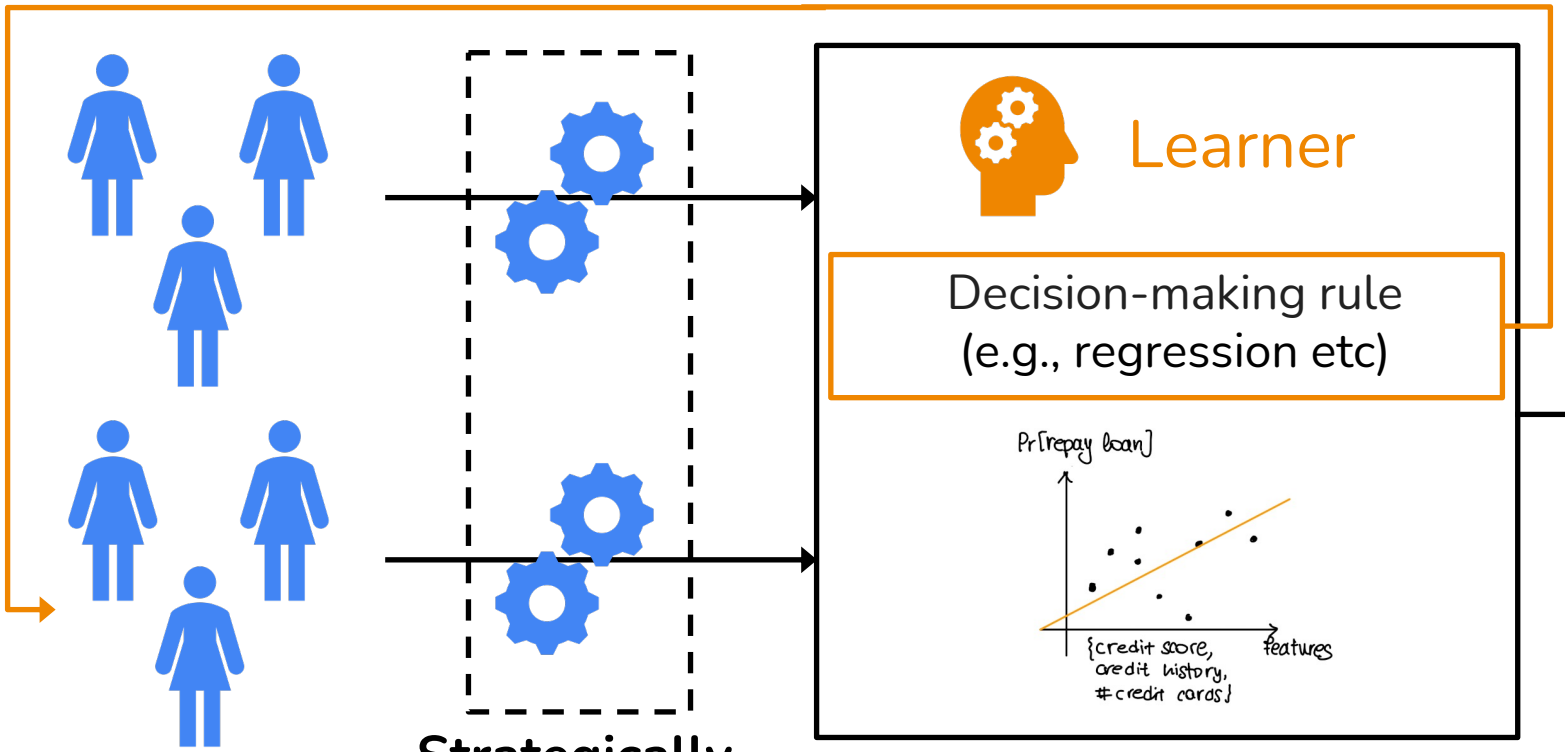
Contributions

Societal effects of non-transparency.
[BPWZ, **ICML22**]

Lots of Recent, Exciting Work

- **Robustness**: [Hardt, Megiddo, Papadimitriou, Wooters, **ITCS16**], [Dong, Roth, Schutzman, Waggoner, Wu, **EC18**], [Chen, Liu, **P.**, **NeurIPS20**], [Ahmadi, Beyhaghi, Blum, Naggita, **EC21**], [Sundaraman, Vullikanti, Xu, Yao, **ICML21**], [Ghalme , Nair, Eilat, Talgam-Cohen, Rosenfeld, **ICML21**], [Zrnic, Mazumdar, Sastry, Jordan, **NeurIPS21**], [Jagadeesan, Mendler-Dünner, Hardt, **ICML21**]
- **Fairness:** [Milli, Miller, Dragan, Hardt, **FAT*19**], [Hu, Immorlica, Vaughan, **FAT*19**], [Liu, Wilson, Haghtalab, Kalai, Borgs, Chayes, **FAT*19**], [Braverman, Garg, **FORC20**]
- **Recourse/Incentivizing Effort:** [Ustun, Spangher, Liu, **FAT*19**], [Kleinberg and Raghavan, **EC19**], [Khajehnejad, Tabibian, Scholkopf, Singla, Gomez-Rodriguez, arXiv19], [Gupta, Nokhiz, Roy, Venkatasubramanian, **arXiv19**], [Chen, Wang, Liu, **arXiv20**], [Tsirtsis, Gomez-Rodriguez, **NeurIPS20**], [Haghtalab, Immorlica, Lucier, Wang, **IJCAI20**], [Bechavod, **P.**, Wu, Ziani, **ICML22**]
- **Causality:** [Miller, Milli, Hardt, **FAT*19**], [Shavit, Edelman, Axelrod, **ICML20**], [Bechavod, Ligett, Wu, Ziani, **AISTATS21**]
- **Performative Prediction:** [Perdomo, Zrnic, Mendler-Dünner, Hardt, **ICML20**], [Mendler-Dünner, Perdomo, Zrnic, Hardt, **NeurIPS20**], [Miller, Perdomo, Zrnic, **ICML21**] [Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner, **ICML22**].

Strategic/Incentive-Aware Learning



Standard Assumption:
Rule is known to the agents.

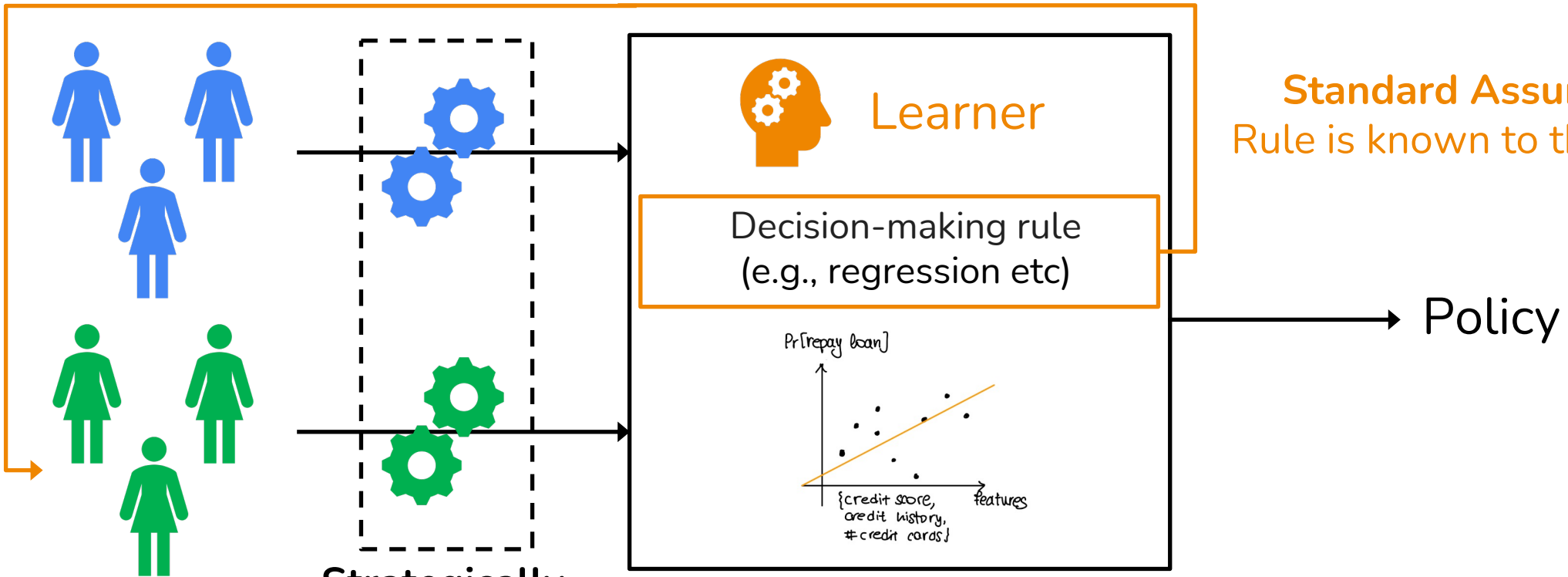
Mathematically:

- Learner commits to a decision rule $\boldsymbol{w}: \mathcal{X} \to [0,1]$
- Agent with feature vector $\boldsymbol{x} \in \mathcal{X}$ and score $y \in [0,1]$, observes $\boldsymbol{w}$ and best-responds by reporting

$$\hat{\boldsymbol{x}}(\boldsymbol{w}) = \arg\max_{x' \in \mathcal{X}} u(\boldsymbol{x}; \boldsymbol{w})$$

- Learner's rule = Stackelberg equilibrium. For example: $\boldsymbol{w} = \arg\min_{\boldsymbol{w}'} (\langle \boldsymbol{w}', \hat{\boldsymbol{x}}(\boldsymbol{w}') \rangle - y)^2$

Strategic/Incentive-Aware Learning



Standard Assumption:
Rule is known to the agents.

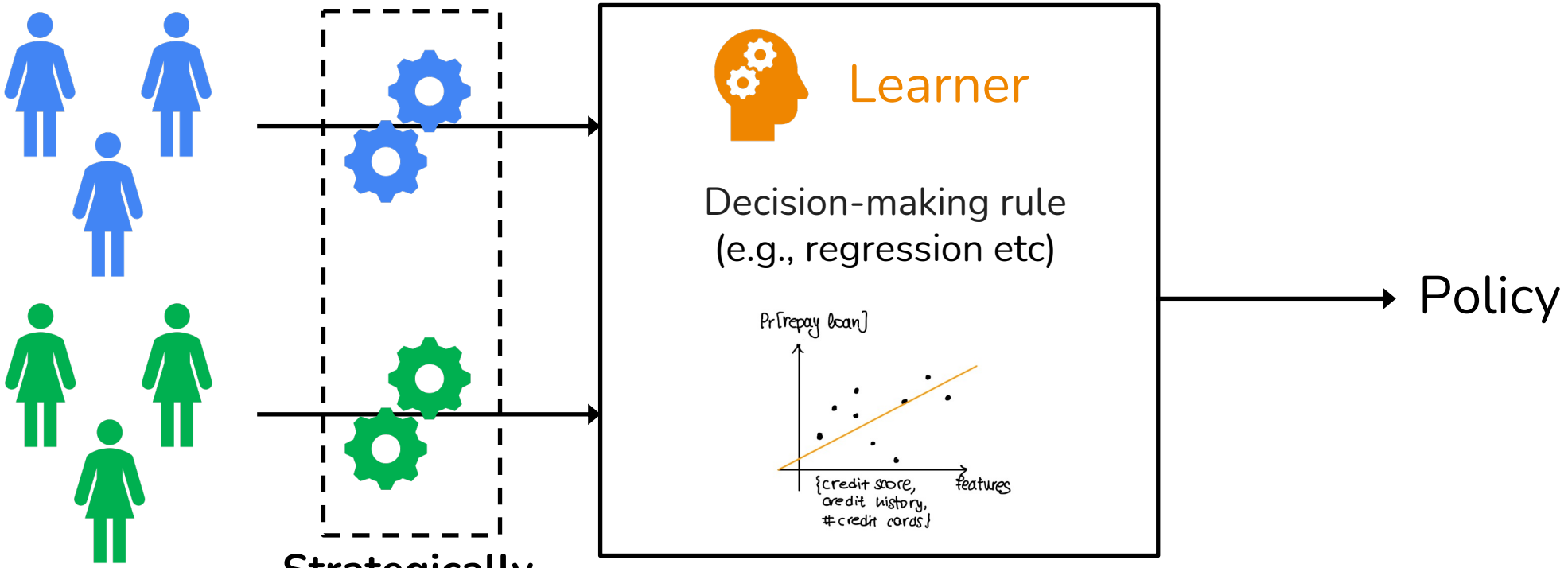
Mathematically:

- Learner commits to a decision rule $\boldsymbol{w}: \mathcal{X} \to [0,1]$
- Agent with feature vector $\boldsymbol{x} \in \mathcal{X}$ and score $y \in [0,1]$, observes w and best-responds by reporting

$$\hat{\boldsymbol{x}}(\boldsymbol{w}) = \arg\max_{x' \in \mathcal{X}} u(\boldsymbol{x}; \boldsymbol{w})$$

- Learner's rule = Stackelberg equilibrium. For example: $\boldsymbol{w} = \arg\min_{\boldsymbol{w}'}(\langle \boldsymbol{w}', \hat{\boldsymbol{x}}(\boldsymbol{w}') \rangle - y)^2$

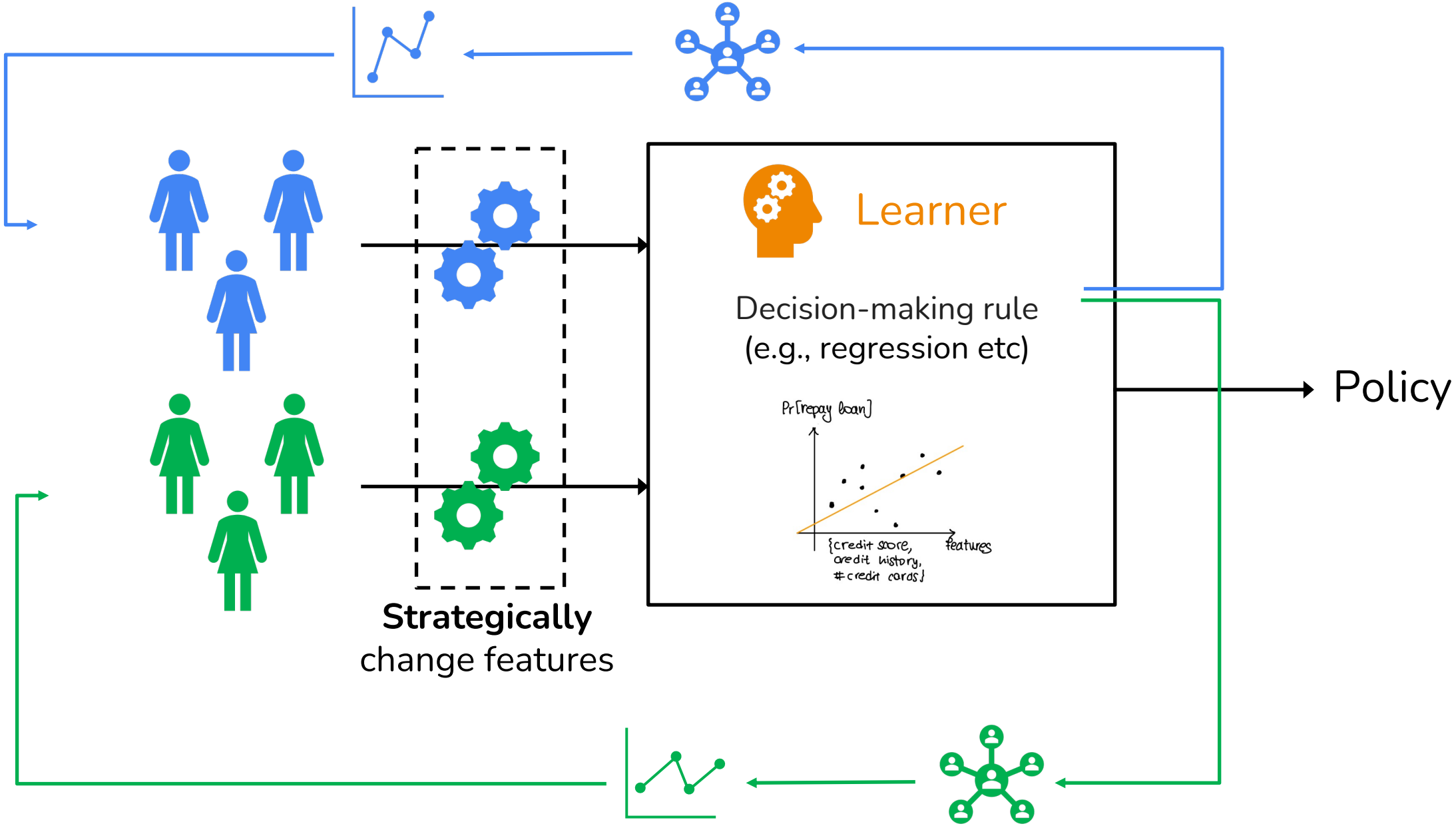
Strategic/Incentive-Aware Learning Revisited



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



Our Model Formally

Interaction Protocol

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

Learner's Goal

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

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

Why is $\boldsymbol{w} \neq \boldsymbol{w}^\star$?

Main Question

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

Subgroup Feature Vector Discrepancies

- $\mathcal{S}_1, \mathcal{S}_2$: 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 $\mathcal{S}_1, \mathcal{S}_2$ → $\boldsymbol{x}_g = \Pi_g \boldsymbol{x}_g$ (feature discrepancy)

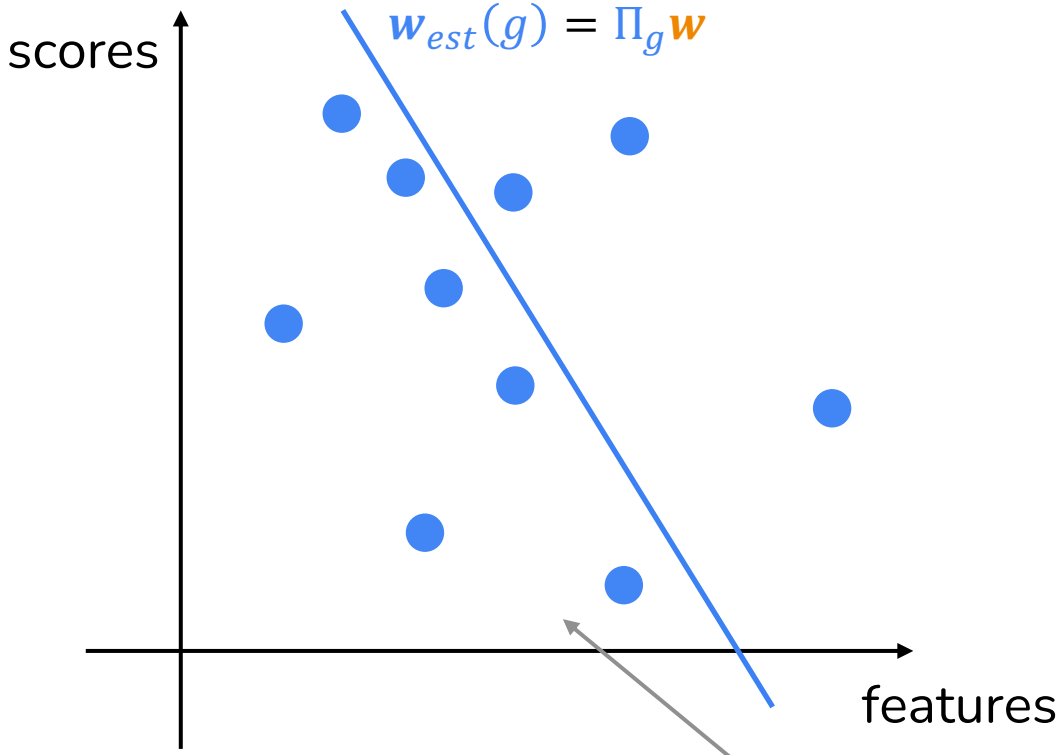
Subgroup Utilities

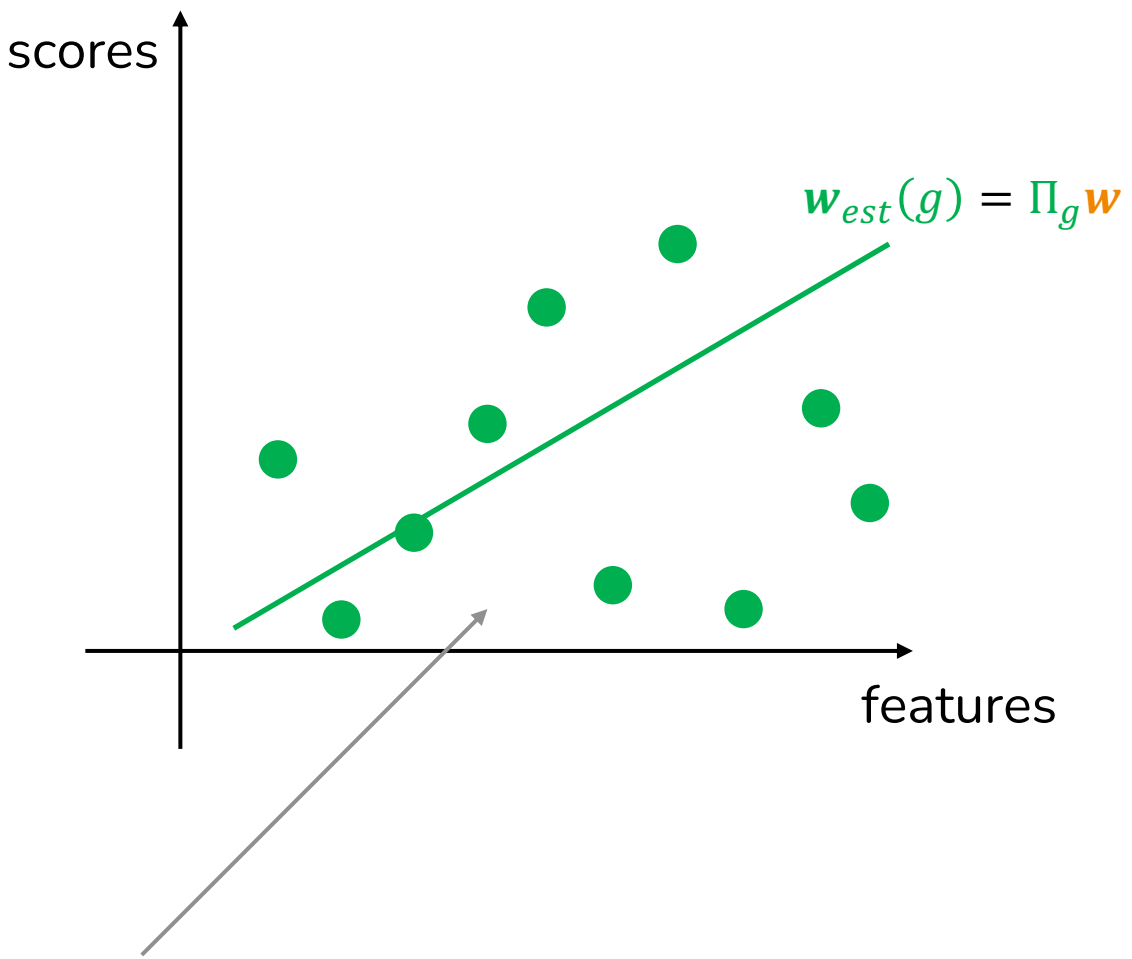
Score they get with their **estimated** decision rule

$$u(\boldsymbol{x}_g, \boldsymbol{x}'; g) = \overbrace{EstScore(\boldsymbol{x}')} - Cost(\boldsymbol{x}_g \to x')$$

$$= \langle \boldsymbol{x}', \boldsymbol{w}_{est}(g) \rangle - ||A_g(\boldsymbol{x}_g - \boldsymbol{x}')||_2$$

How Do the Subgroups Estimate $\boldsymbol{w}$





Labeled examples from a peer subgroup.

Each subgroup runs **ERM** on labeled examples to recover $\boldsymbol{w}$. → Recovers: $\boldsymbol{w}_{est}(g) = \Pi_g \boldsymbol{w}$

Principal's Equilibrium Decision Rule

estimated score manipulation cost

- Agents' best response: $\hat{\boldsymbol{x}}_g = \arg\max_{\boldsymbol{x}'} u(\boldsymbol{x}_g, \boldsymbol{x}'; g)$

$$\rightarrow \hat{\boldsymbol{x}}_g = \boldsymbol{x} + A_g^{-1}\Pi_g \boldsymbol{w} = \boldsymbol{x} + \Delta_g(\boldsymbol{w})$$

- Principal's rule optimizing SW: $\boldsymbol{w}_{SW} = \arg\max_{\boldsymbol{w}'} \left(\mathbb{E}_{\boldsymbol{x}_1 \sim \mathcal{D}_1}[\langle \hat{\boldsymbol{x}}_1, \boldsymbol{w}^\star \rangle] + \mathbb{E}_{\boldsymbol{x}_2 \sim \mathcal{D}_2}[\langle \hat{\boldsymbol{x}}_2, \boldsymbol{w}^\star \rangle]\right)$

$$= \frac{\left(\Pi_1 A_1^{-1} + \Pi_2 A_2^{-1}\right)\boldsymbol{w}^\star}{\left\|\left(\Pi_1 A_1^{-1} + \Pi_2 A_2^{-1}\right)\boldsymbol{w}^\star\right\|}$$

Is it true that $\boldsymbol{w} = \boldsymbol{w}^\star$?

Answer:

- Sometimes (e.g., $A_1 = A_2 = \mathbb{I}$ 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: $\boldsymbol{w}^\star = \left(\frac{2}{3}, \frac{2}{3}, \frac{1}{3}\right)$ and $\Pi_1 = (1,0,1), \Pi_2 = (0,1,1)$. $\Delta\left(SW(\boldsymbol{w}^\star)\right) = 10/9$.

    For $\boldsymbol{w} = \frac{1}{\sqrt{3}}(1,1,1)$: $\Delta\left(SW(\boldsymbol{w})\right) > 10/9$

Measures of Outcome Improvement in Equilibrium

Improvement for group g: $\mathcal{I}_g(\boldsymbol{w}) = \langle \hat{x}(\boldsymbol{w}), \boldsymbol{w}^\star \rangle - \langle x, \boldsymbol{w}^\star \rangle$

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

Results

Improvement for group g: $\mathcal{I}_g(\boldsymbol{w}) = \langle \hat{x}(\boldsymbol{w}), \boldsymbol{w}^\star \rangle - \langle x, \boldsymbol{w}^\star \rangle = \langle A_g^{-1} \Pi_g \boldsymbol{w}, \boldsymbol{w}^\star \rangle$

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: **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: $\mathcal{I}_g(\boldsymbol{w}) = \langle \hat{x}(\boldsymbol{w}), \boldsymbol{w}^\star \rangle - \langle x, \boldsymbol{w}^\star \rangle = \langle A_g^{-1} \Pi_g \boldsymbol{w}, \boldsymbol{w}^\star \rangle$

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: $|\mathcal{I}_1(\boldsymbol{w}) - \mathcal{I}_2(\boldsymbol{w})| \leq ||\Pi_1 \boldsymbol{w}^\star - \Pi_2 \boldsymbol{w}^\star||_2$

information overlap proxy

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: $\mathcal{I}_g(\boldsymbol{w}) = \langle \hat{x}(\boldsymbol{w}), \boldsymbol{w}^\star \rangle - \langle x, \boldsymbol{w}^\star \rangle = \langle A_g^{-1} \Pi_g \boldsymbol{w}, \boldsymbol{w}^\star \rangle$

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

Per-unit Improvement for group g: $\mathcal{I}_g \left(\frac{\Pi_g \boldsymbol{w}}{||\Pi_g \boldsymbol{w}||_2} \right)$

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
- ~50K datapoints
- 14 attributes including Age, Country, Workclass, Education, Race, etc.
- Label (annual income): <50K, >= 50K

Our process:

- 4 experiments separating subpopulations based on:

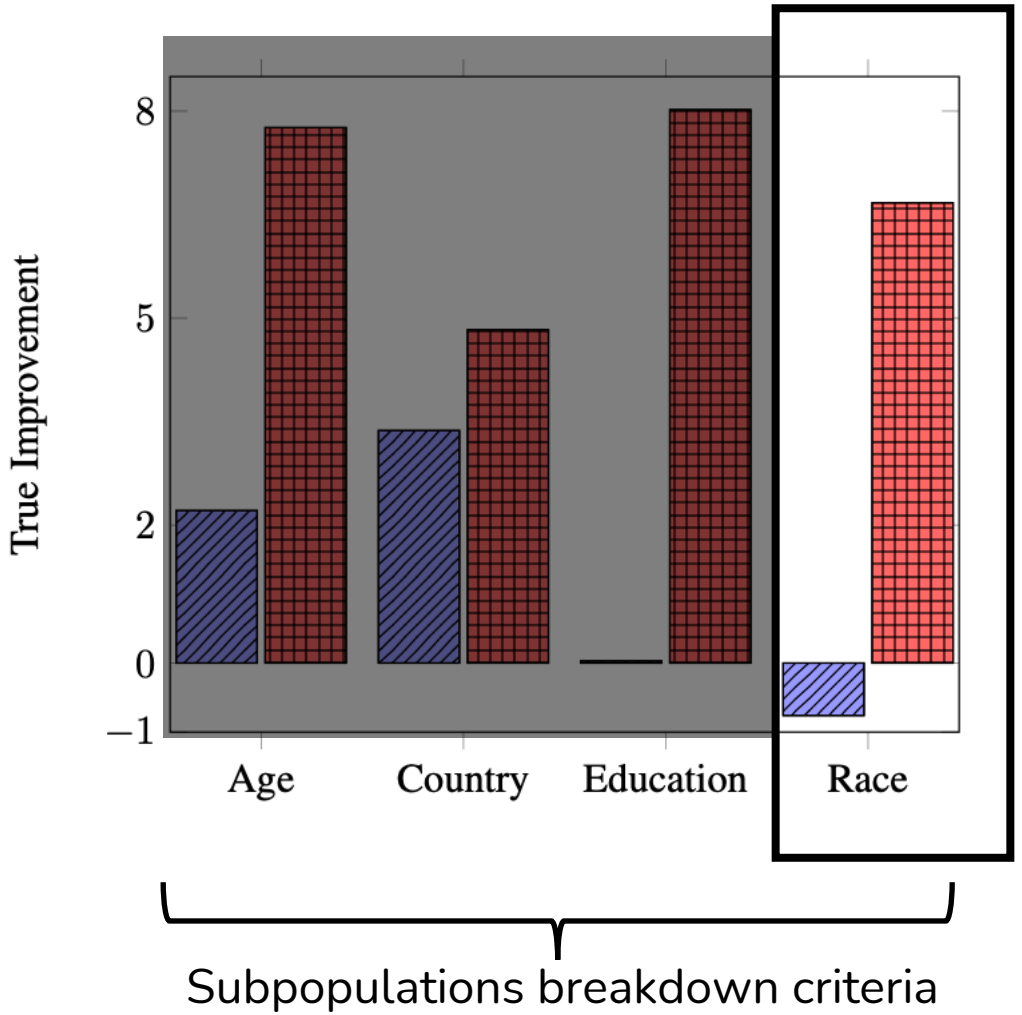
Characteristic	Subpopulation 1	Subpopulation 2
Age	<35 yrs old	>=35 yrs old
Country	All others	Western countries
Education	All others	Above high school
Race	All others	White

- 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

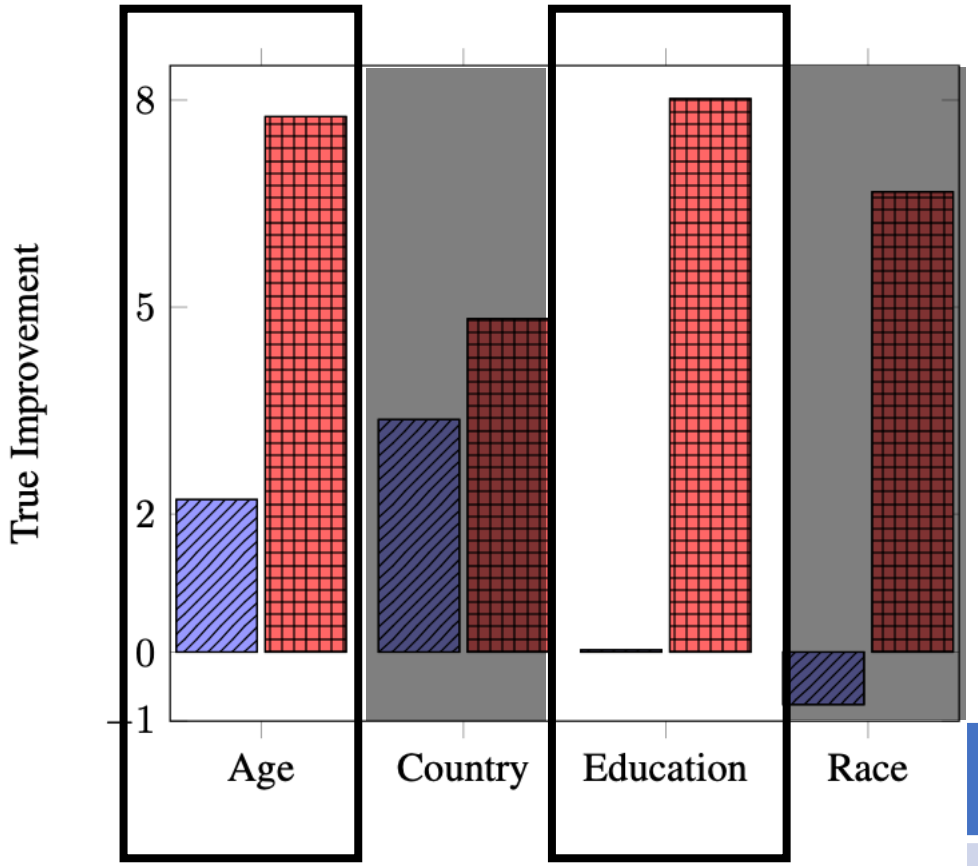


	Subpopulation 1	Subpopulation 2
Race	All others	White

Results Snapshot: Adult Dataset

2 Total improvement may be **very unequal** across subpopulations.

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



	Subpop. 1	Subpop. 2
Age	<35 yrs old	>=35 yrs old
Education	All others	Above HS

Summary

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

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

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 $A_1^{-1}\Pi_1 A_1^{-1} = A_2^{-1}\Pi_2 A_2^{-1}$
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 Π_g's, A_g's, and $\boldsymbol{w}^\star$ are not known a priori.

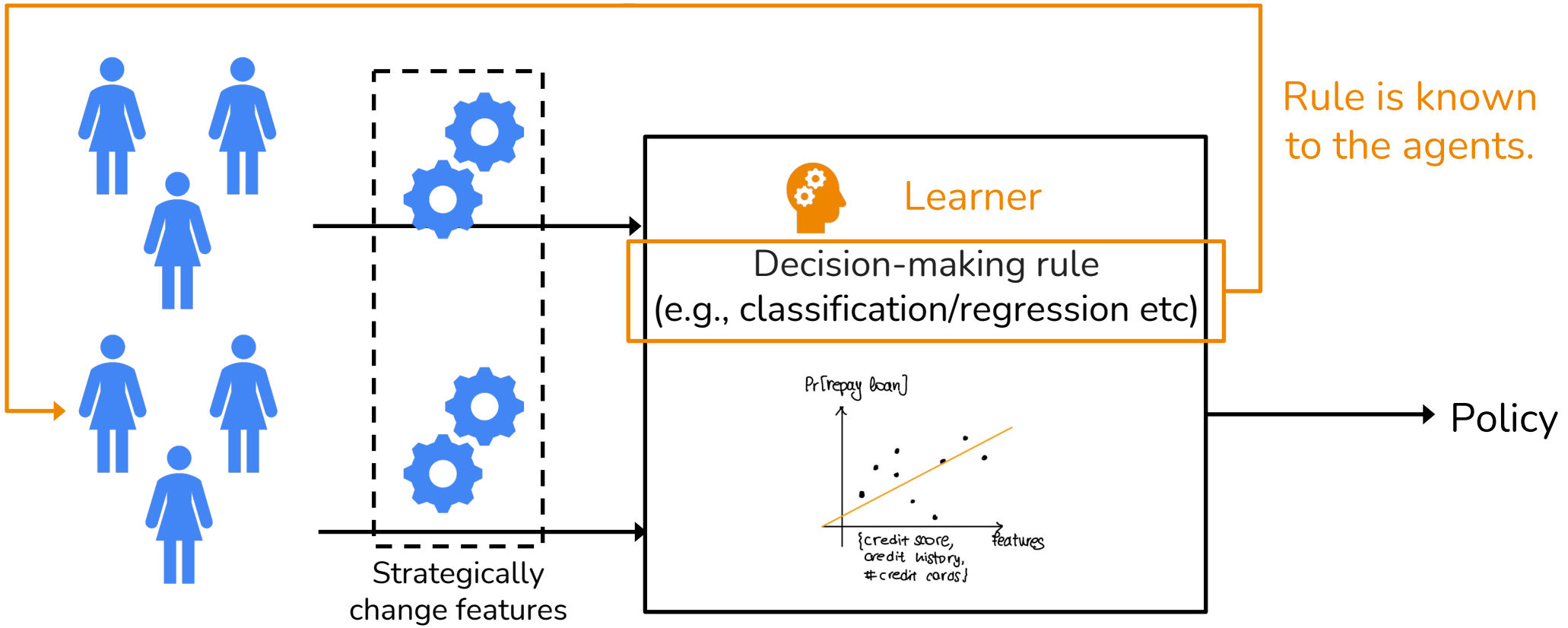
(2) Generalization for $g \geq 3$.

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

Interpretability and Incentives

"Obscure" ML algorithms

+ Stop strategic behavior
- Non-transparent



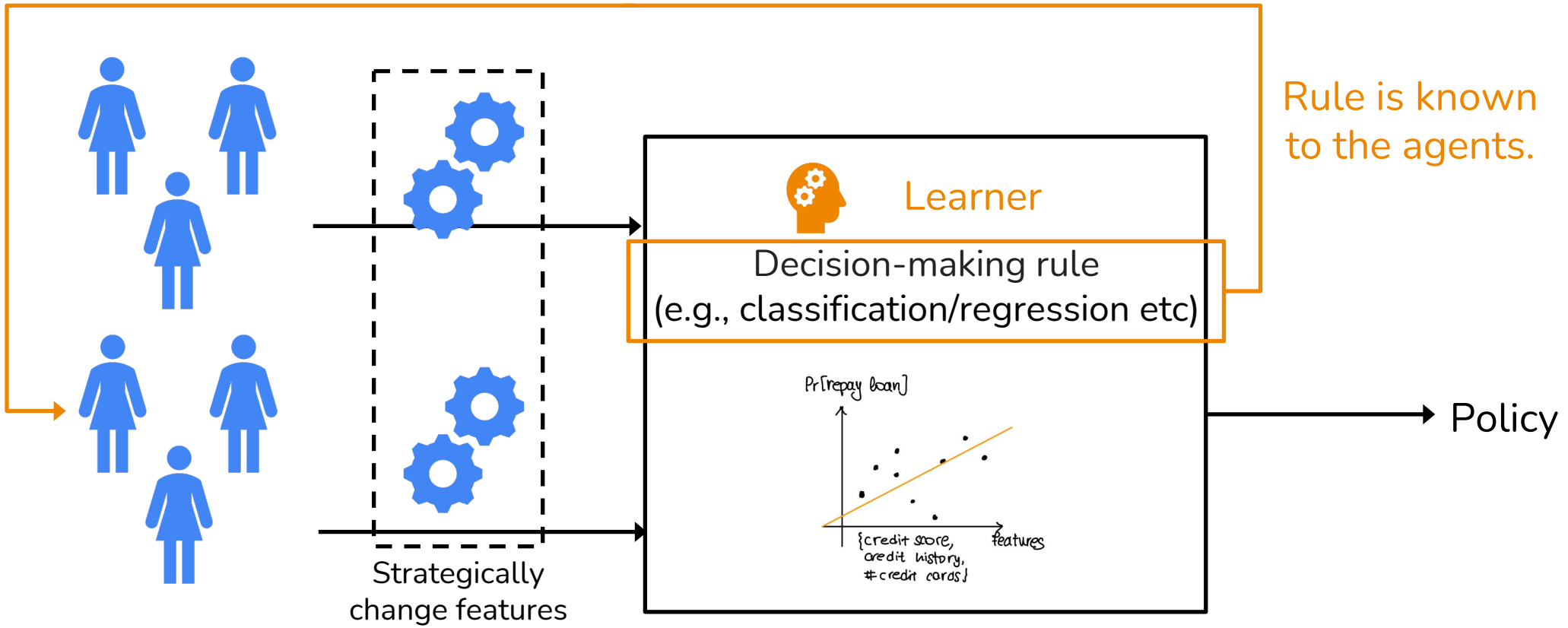
Interpretability and Incentives

"Obscure" ML algorithms

+ Stop strategic behavior
- Non-transparent

Public ML algorithms

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

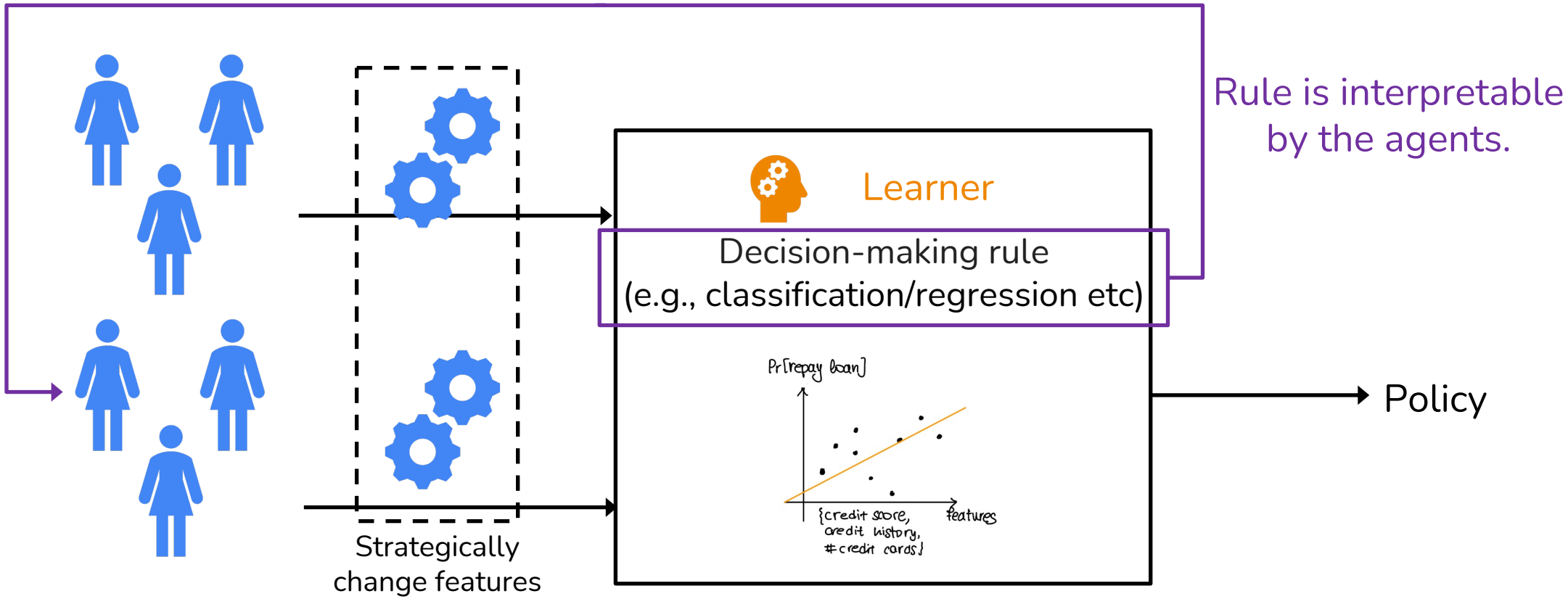


Interpretability and Incentives

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

Interpretable ML Algorithms

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



Interpretability and Incentives

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

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

Current state of Incentive-Aware ML research

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

Tutorial at FAccT21