Fairness Interventions as (Dis)Incentives for Strategic Manipulation

Xueru Zhang Mohammad Mahdi Khalili Kun Jin Parinaz Naghizadeh Mingyan Liu

Machine Learning for People

- ML has been increasingly used to help **make decisions about people**
  - College admission, Hiring, Lending, …

salesforce.org education cloud

Kira Talent

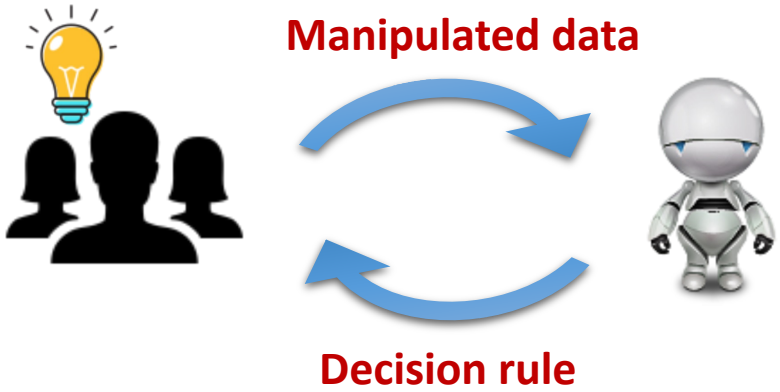
TurnKey Lender
Intelligent Lending Automation Made Simple

Kabbage

ZAML Zest Automated Machine Learning

Challenges

How to receive favorable decisions with lowest effort?

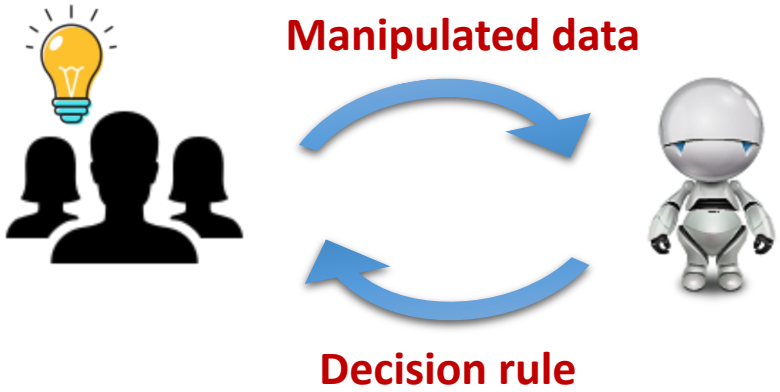


How to make accurate decisions?

- ML is vulnerable to strategic manipulation

Challenges

How to receive favorable decisions with lowest effort?

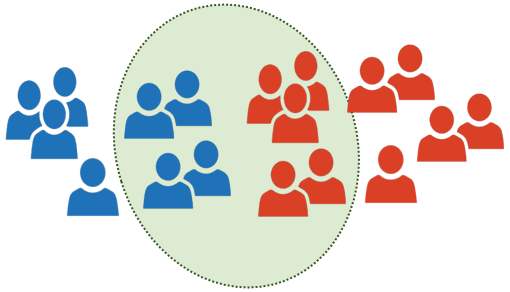


How to make accurate and **fair** decisions?

- ML is vulnerable to strategic manipulation
- ML can be biased against certain social groups

Existing Work

Fair machine learning



$$\min \quad Loss$$

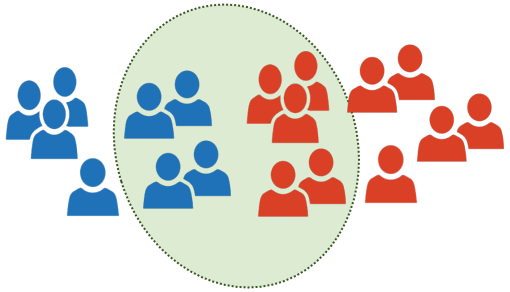
$$s.t., \quad \phi(\text{👤}) \approx \phi(\text{👤})$$

Fairness constraint

- **Demographic parity**: equal positive rate
- **Equal Opportunity**: equal true positive rate

Existing Work

Fair machine learning



$$\min \quad Loss$$
$$s.t., \quad \phi(\text{👤}) \approx \phi(\text{👤})$$

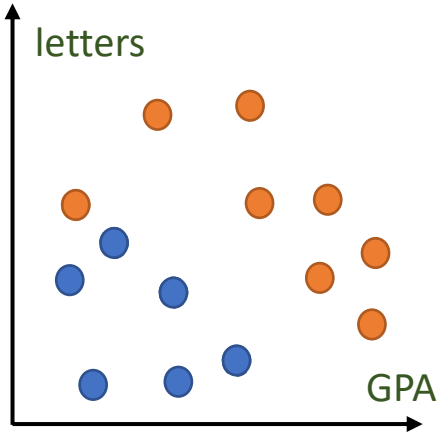
Fairness constraint

- **Demographic parity**: equal positive rate
- **Equal Opportunity**: equal true positive rate

Strategic classification

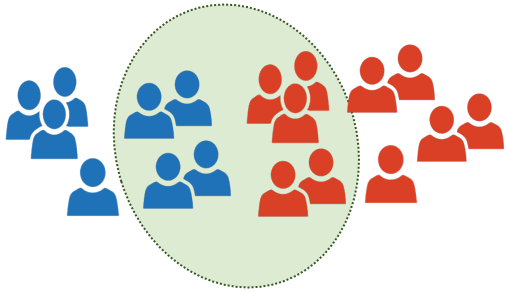
- Stackelberg game formulation

Hardt et al., 2016a; Dong et al. 2018; Milli et al., 2019; Hu et al., 2019; Braverman & Garg, 2020



Existing Work

Fair machine learning



$$\min \quad Loss$$
$$s.t., \quad \phi(\text{👤}) \approx \phi(\text{👥})$$

Fairness constraint

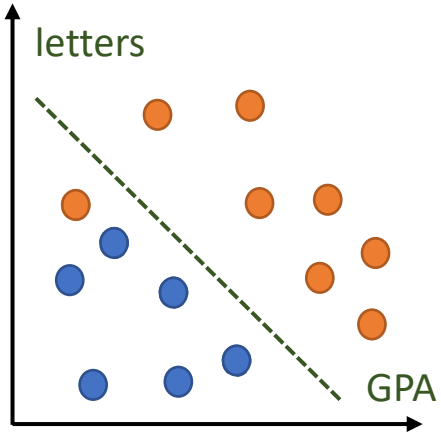
- **Demographic parity**: equal positive rate
- **Equal Opportunity**: equal true positive rate

Strategic classification

- Stackelberg game formulation

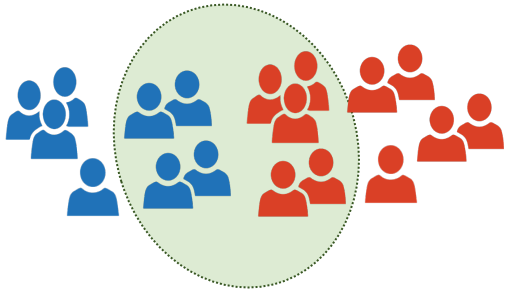
Hardt et al., 2016a; Dong et al. 2018; Milli et al., 2019; Hu et al., 2019; Braverman & Garg, 2020

Classifier f



Existing Work

Fair machine learning



$$\min \quad Loss$$
$$s.t., \quad \phi(\text{👤}) \approx \phi(\text{👤})$$

Fairness constraint

- **Demographic parity**: equal positive rate
- **Equal Opportunity**: equal true positive rate

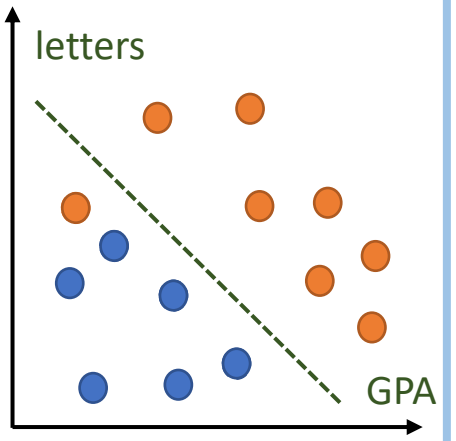
Strategic classification

- Stackelberg game formulation

Hardt et al., 2016a; Dong et al. 2018; Milli et al., 2019; Hu et al., 2019; Braverman & Garg, 2020

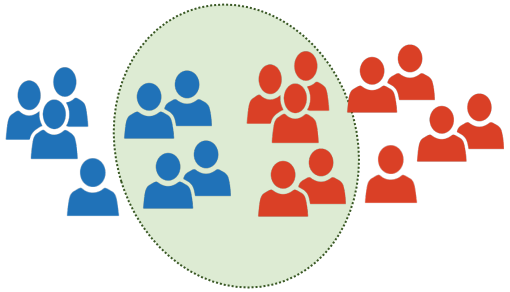
Classifier f

Initial data x



Existing Work

Fair machine learning



$$\min \quad Loss$$
$$s.t., \quad \phi(\text{👤}) \approx \phi(\text{👤})$$

Fairness constraint

- **Demographic parity**: equal positive rate
- **Equal Opportunity**: equal true positive rate

Strategic classification

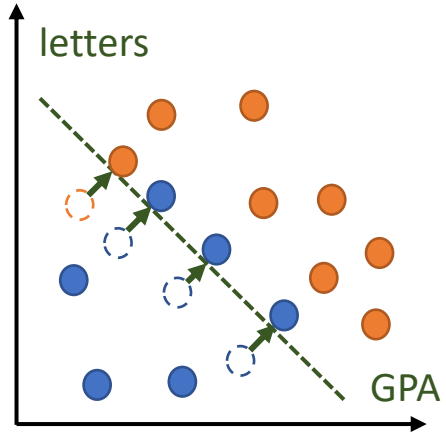
- Stackelberg game formulation

Hardt et al., 2016a; Dong et al. 2018; Milli et al., 2019; Hu et al., 2019; Braverman & Garg, 2020

Classifier f

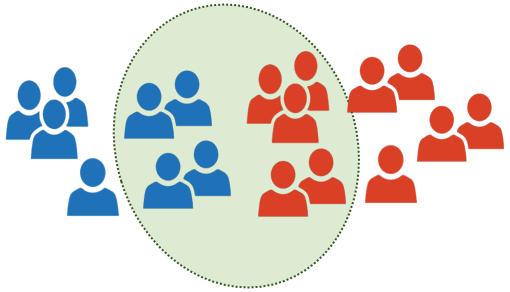
Initial data x

Manipulated data $\Delta(x)$



Existing Work

Fair machine learning



$$\min \quad Loss$$
$$s.t., \quad \phi(\text{👤}) \approx \phi(\text{👤})$$

Fairness constraint

- **Demographic parity**: equal positive rate
- **Equal Opportunity**: equal true positive rate

Strategic classification

- Stackelberg game formulation

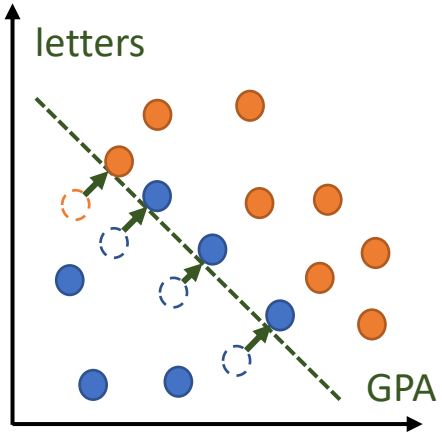
Hardt et al., 2016a; Dong et al. 2018; Milli et al., 2019; Hu et al., 2019; Braverman & Garg, 2020

Classifier f

Initial data x

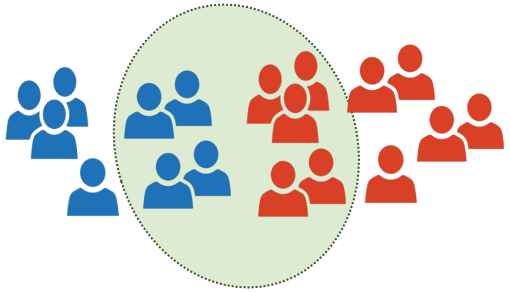
Manipulated data $\Delta(x)$

Cost $c(x, \Delta(x))$



Existing Work

Fair machine learning



$$\min \quad Loss$$
$$s.t., \quad \phi(\text{👤}) \approx \phi(\text{👤})$$

Fairness constraint

- **Demographic parity**: equal positive rate
- **Equal Opportunity**: equal true positive rate

Strategic classification

- Stackelberg game formulation

Hardt et al., 2016a; Dong et al. 2018; Milli et al., 2019; Hu et al., 2019; Braverman & Garg, 2020

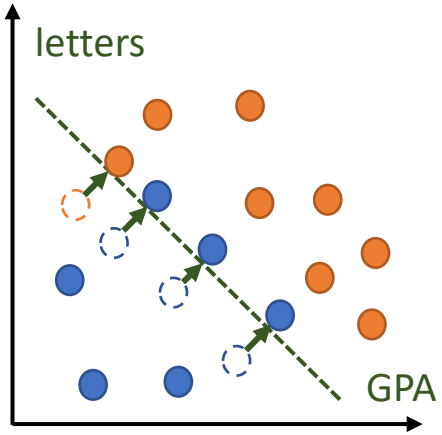
Classifier f

Initial data x

Manipulated data $\Delta(x)$

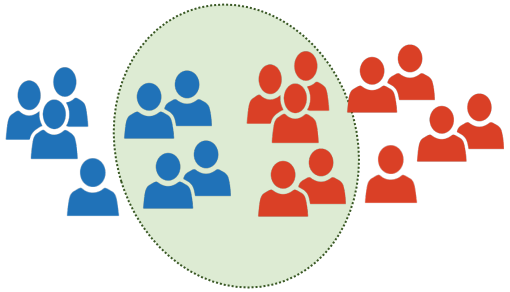
Cost $c(x, \Delta(x))$

$$\max f(\Delta(x)) - c(x, \Delta(x))$$



Existing Work

Fair machine learning



$$\min \quad Loss$$
$$s.t., \quad \phi(\text{👤}) \approx \phi(\text{👤})$$

Fairness constraint

- **Demographic parity**: equal positive rate
- **Equal Opportunity**: equal true positive rate

Strategic classification

- Stackelberg game formulation

Hardt et al., 2016a; Dong et al. 2018; Milli et al., 2019; Hu et al., 2019; Braverman & Garg, 2020

Classifier f

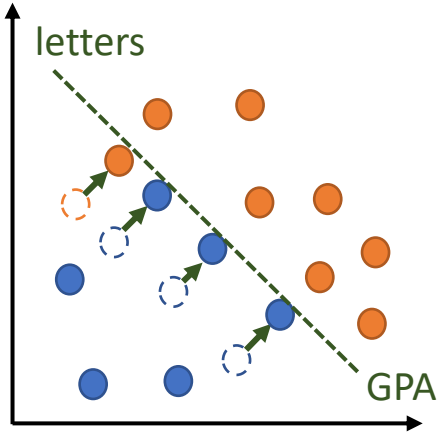
$$\max \Pr[h(x) = f(\Delta(x))]$$

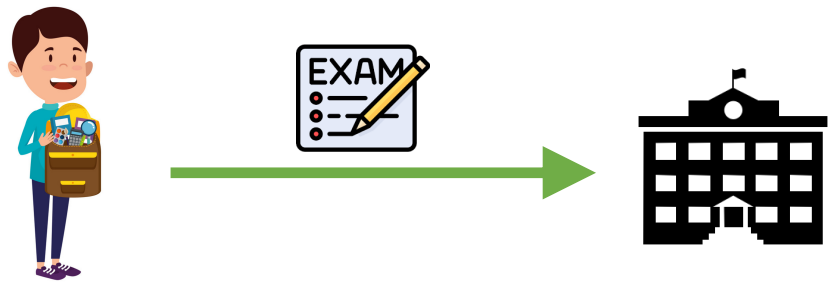
Initial data x

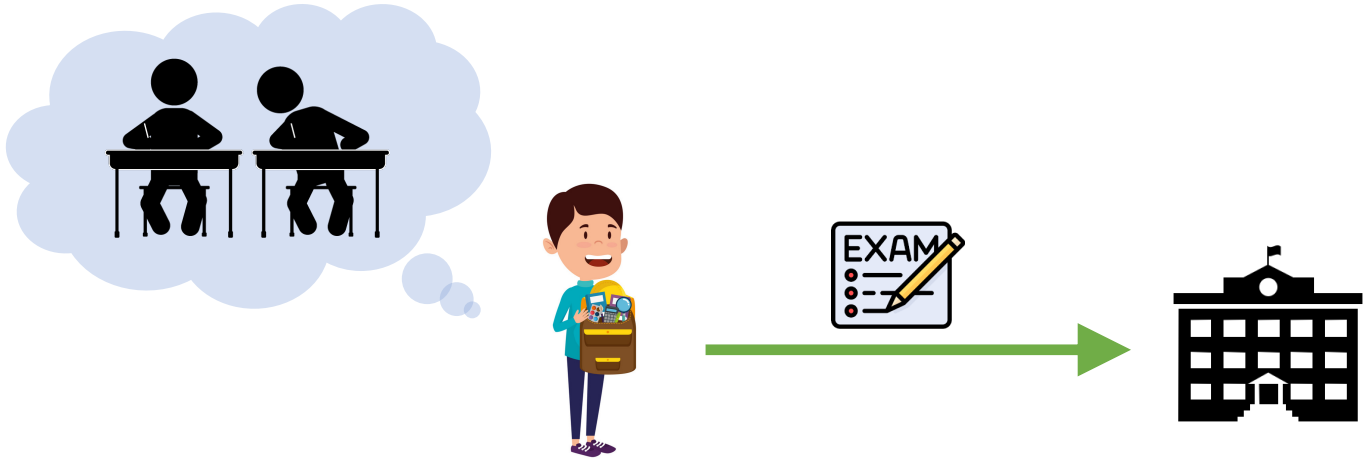
Manipulated data $\Delta(x)$

Cost $c(x, \Delta(x))$

$$\max f(\Delta(x)) - c(x, \Delta(x))$$




EXAM


EXAM

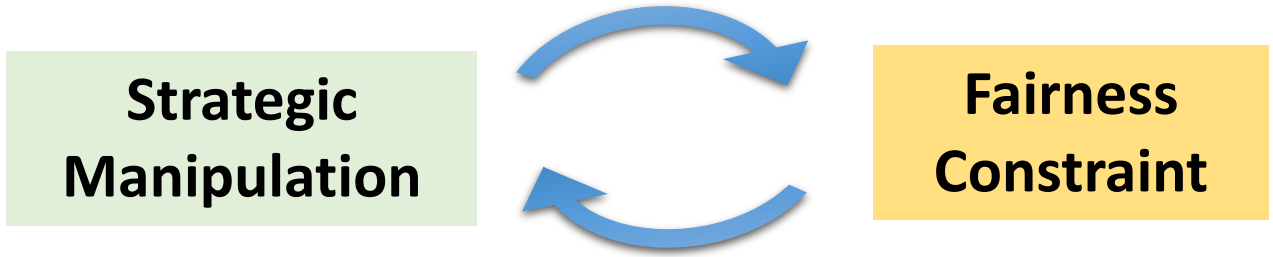

EXAM

This paper:

- A new Stackelberg game formulation that admits
  - **uncertain** manipulation outcomes

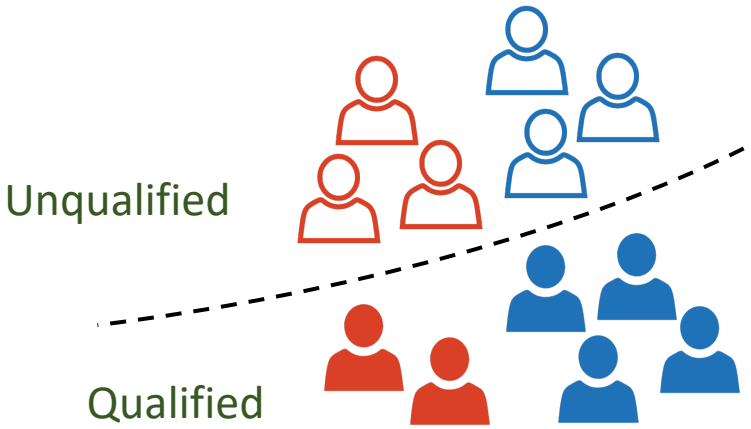
This paper:

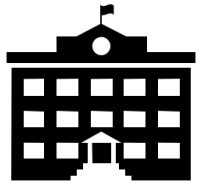
- A new Stackelberg game formulation that admits
  - **uncertain** manipulation outcomes
- How strategic manipulation and fairness intervention impact each other?



Model

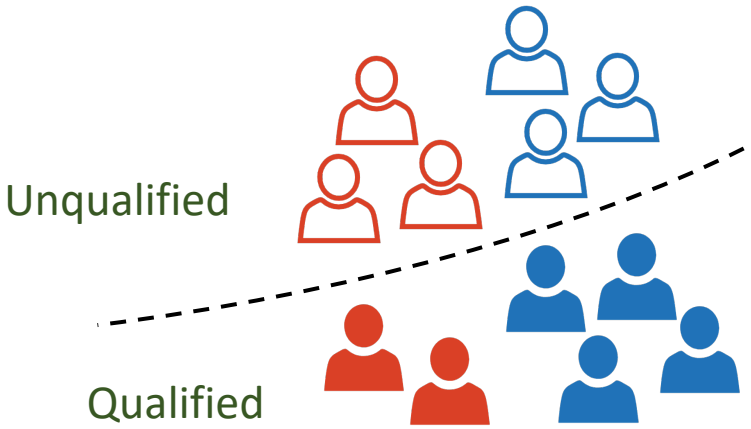
Two demographic groups

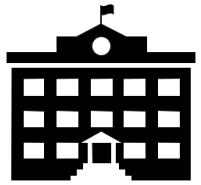




School

Model



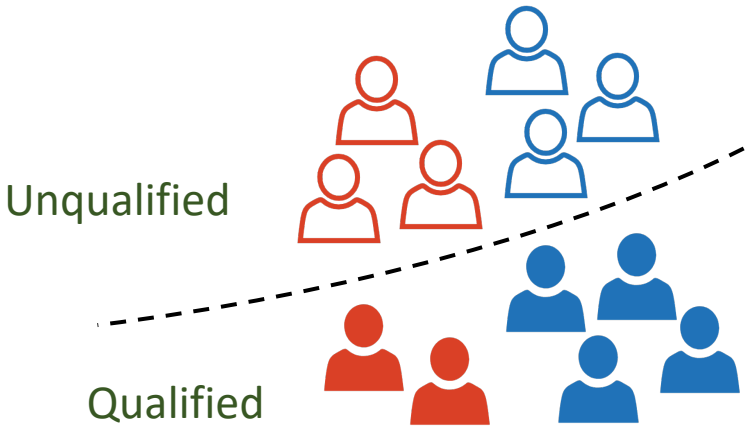


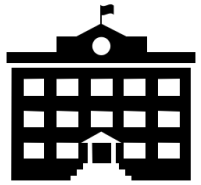
School

Two demographic groups

- Sensitive attribute $S \in \{a, b\}$ (race/gender)

Model



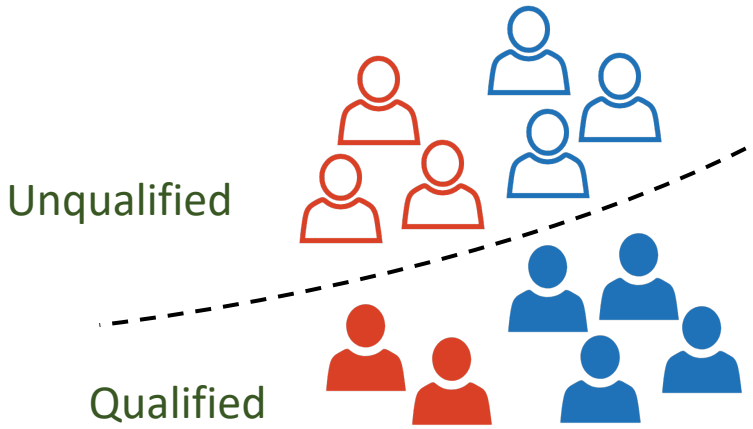


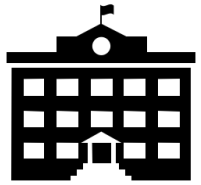
School

Two demographic groups

- Sensitive attribute $S \in \{a, b\}$ (race/gender)
- Feature X (exam score)

Model



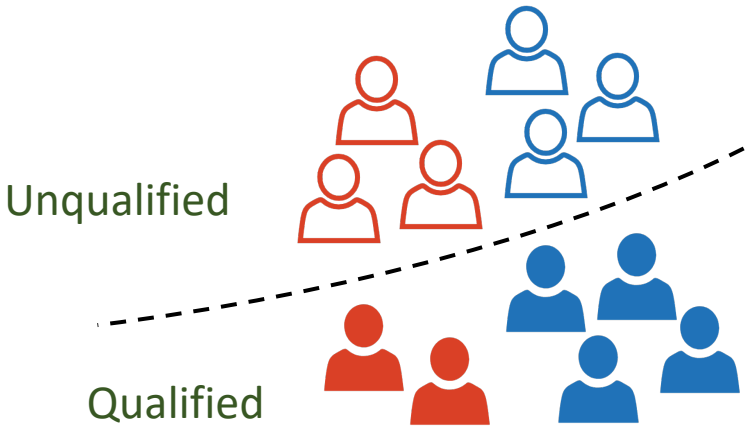


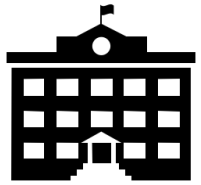
School

Two demographic groups

- Sensitive attribute $S \in \{a, b\}$ (race/gender)
- Feature X (exam score)
- Qualification $Y \in \{0,1\}$ (ability to graduate)

Model



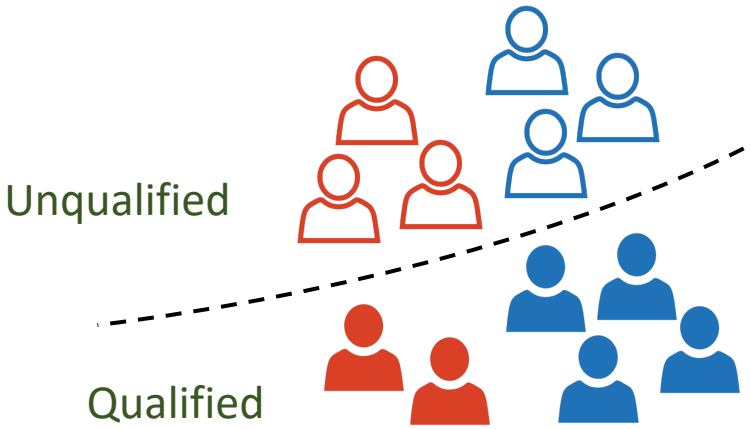


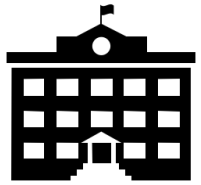
School

Two demographic groups

- Sensitive attribute $S \in \{a, b\}$ (race/gender)
- Feature X (exam score)
- Qualification $Y \in \{0,1\}$ (ability to graduate)
- Decision $D \in \{0,1\}$ (get admitted or not)

Model



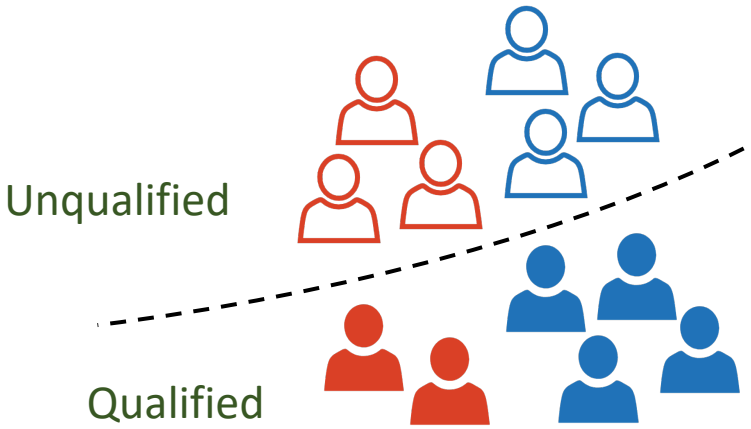


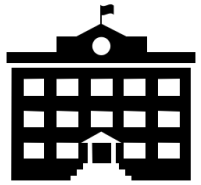
School

Two demographic groups

- Sensitive attribute $S \in \{a, b\}$ (race/gender)
- Feature X (exam score)
- Qualification $Y \in \{0,1\}$ (ability to graduate)
- Decision $D \in \{0,1\}$ (get admitted or not)
  - Decision-maker's policy $\pi_s(x) = P_{D|XS}(1|x, s)$

Model



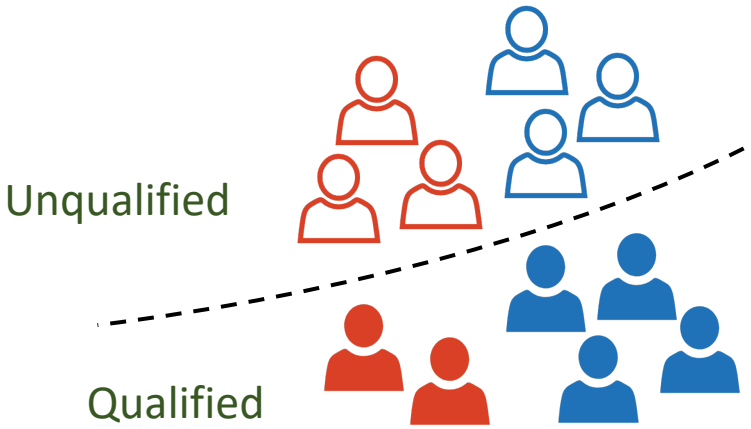


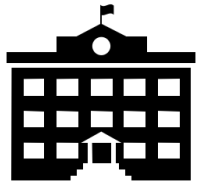
School

Two demographic groups

- Sensitive attribute $S \in \{a, b\}$ (race/gender)
- Feature X (exam score)
- Qualification $Y \in \{0,1\}$ (ability to graduate)
- Decision $D \in \{0,1\}$ (get admitted or not)
  - Decision-maker's policy $\pi_s(x) = P_{D|XS}(1|x, s)$
- Manipulation action $M \in \{0,1\}$ (whether to hire someone else to take the exam or not)

Model



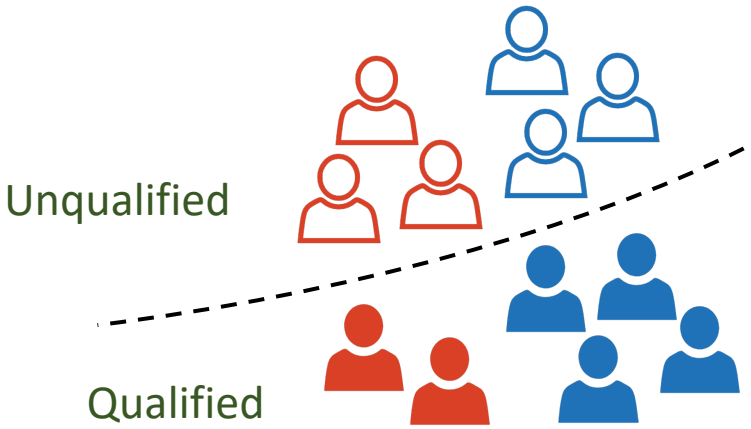


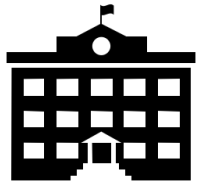
School

Two demographic groups

- Sensitive attribute $S \in \{a, b\}$ (race/gender)
- Feature X (exam score)
- Qualification $Y \in \{0,1\}$ (ability to graduate)
- Decision $D \in \{0,1\}$ (get admitted or not)
  - Decision-maker's policy $\pi_s(x) = P_{D|XS}(1|x, s)$
- Manipulation action $M \in \{0,1\}$ (whether to hire someone else to take the exam or not)
  - Manipulation doesn't affect qualification but results in a better feature distribution

Model



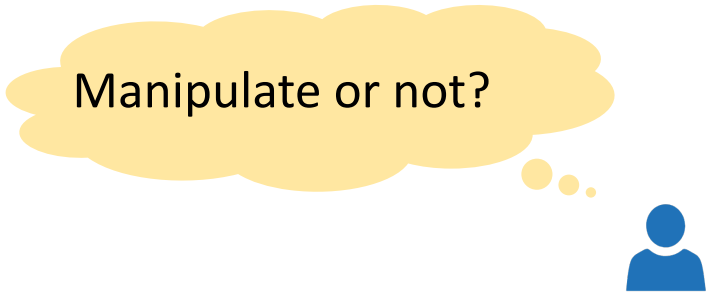


School

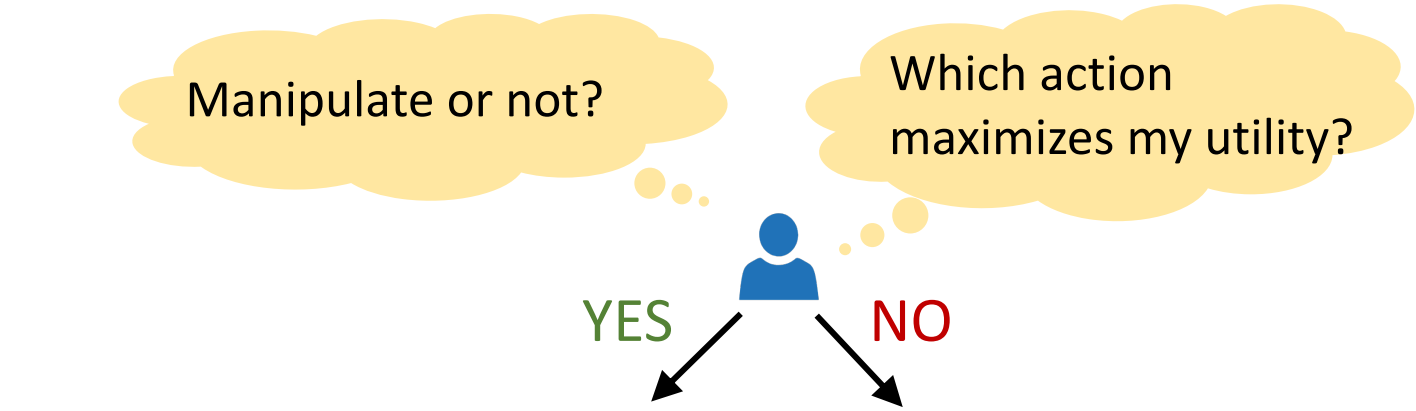
Two demographic groups

- Sensitive attribute $S \in \{a, b\}$ (race/gender)
- Feature X (exam score)
- Qualification $Y \in \{0,1\}$ (ability to graduate)
- Decision $D \in \{0,1\}$ (get admitted or not)
  - Decision-maker's policy $\pi_s(x) = P_{D|XS}(1|x, s)$
- Manipulation action $M \in \{0,1\}$ (whether to hire someone else to take the exam or not)
  - Manipulation doesn't affect qualification but results in a better feature distribution
  - Manipulation cost $C_s \geq 0$ (cost of hiring someone)

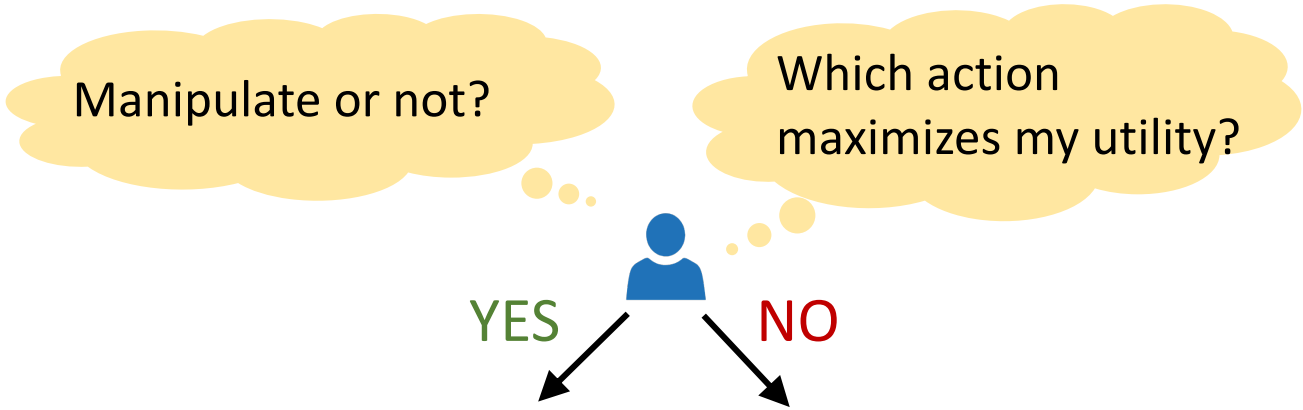
Model: individual best response



Model: individual best response

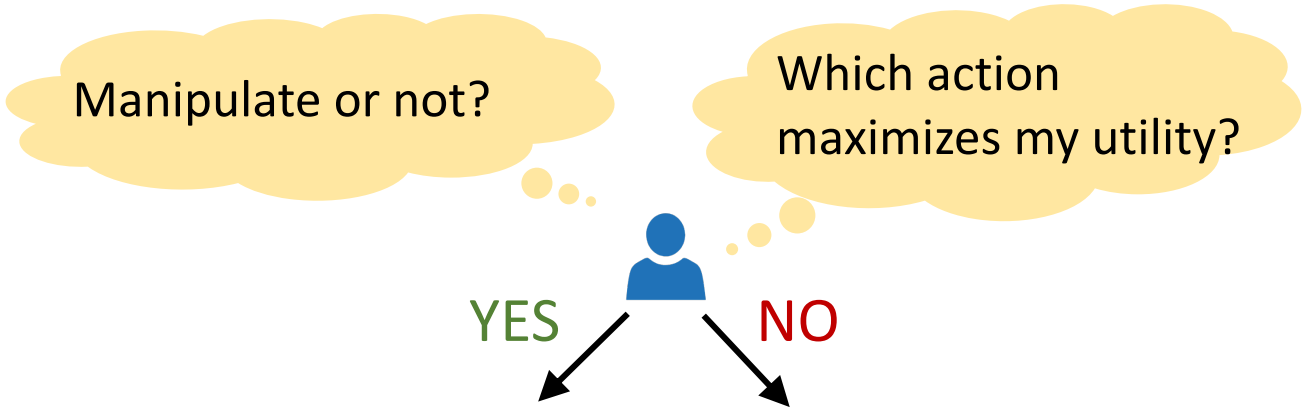


Model: individual best response



Manipulate ($M = 1$) if Benefit with manipulation – cost $P_{D|YMS}(1|y,1,s) - C_s$ $\geq$ Benefit without manipulation $P_{D|YMS}(1|y,0,s)$

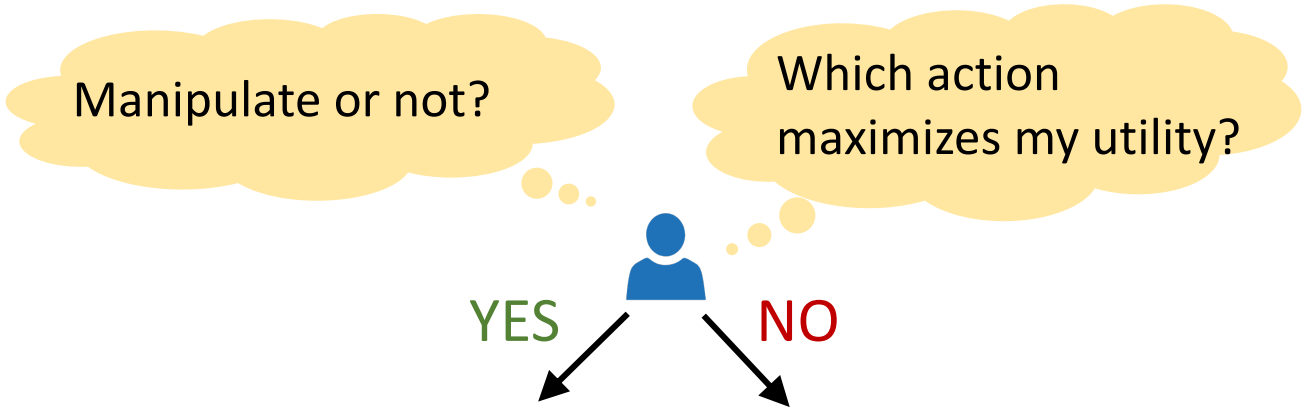
Model: individual best response



Manipulate ($M = 1$) if Benefit with manipulation – cost $P_{D|YMS}(1|y, 1, s) - C_s$ $\geq$ Benefit without manipulation $P_{D|YMS}(1|y, 0, s)$

- For an individual in group s with qualification y, given a policy π_s, he/she manipulates with probability:

Model: individual best response



Manipulate ($M = 1$) if Benefit with manipulation – cost $P_{D|YMS}(1|y,1,s) - C_s$ $\geq$ Benefit without manipulation $P_{D|YMS}(1|y,0,s)$

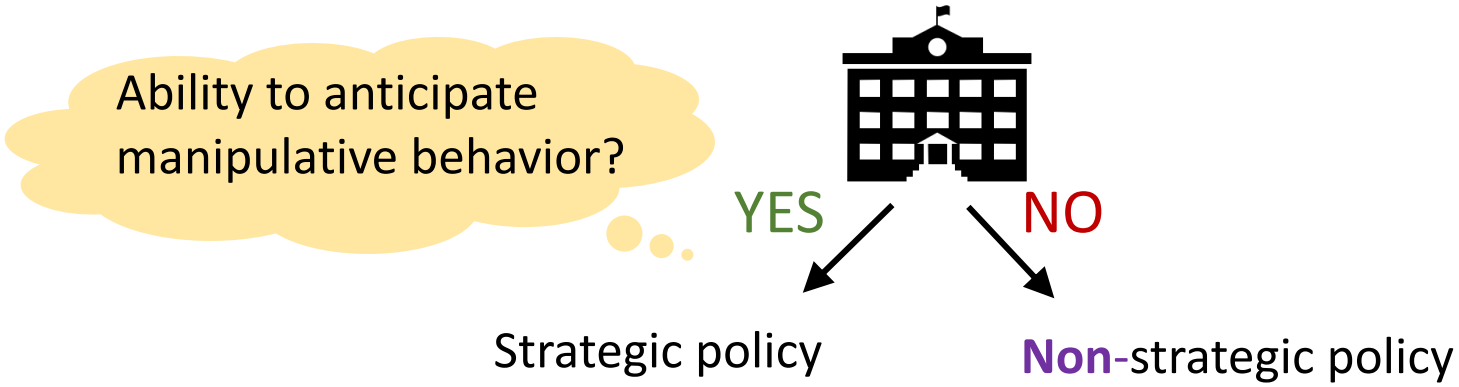
- For an individual in group s with qualification y, given a policy π_s, he/she manipulates with probability:

$$\Pr\left(C_s \leq P_{D|YMS}(1|y,1,s) - P_{D|YMS}(1|y,0,s)\right)$$

Model: decision-maker's optimal policies

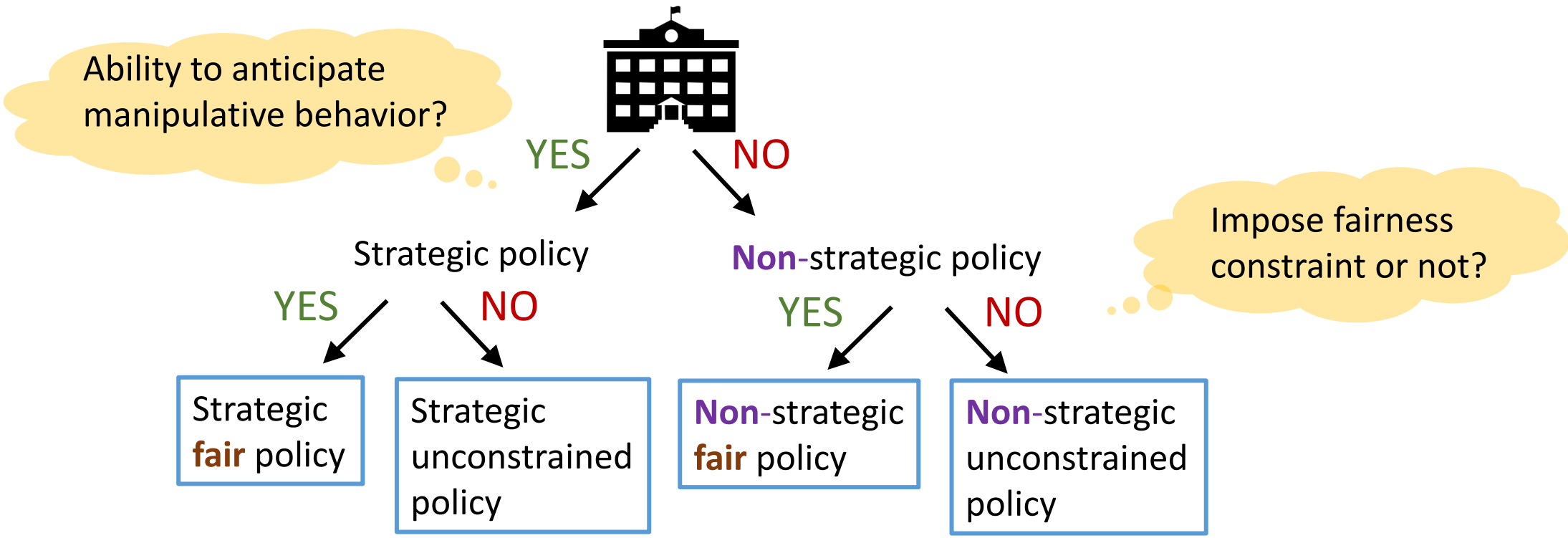
- Policy (π_a, π_b) that maximizes the expected utility $\mathbb{E}[R(Y, D)]$
  - True-positive benefit $R(1,1) = u_+$
  - False-positive penalty $R(0,1) = -u_-$

Model: decision-maker's optimal policies



- Policy (π_a, π_b) that maximizes the expected utility $\mathbb{E}[R(Y, D)]$
  - True-positive benefit $R(1,1) = u_+$
  - False-positive penalty $R(0,1) = -u_-$

Model: decision-maker's optimal policies



- Policy (π_a, π_b) that maximizes the expected utility $\mathbb{E}[R(Y, D)]$
  - True-positive benefit $R(1,1) = u_+$
  - False-positive penalty $R(0,1) = -u_-$

Results

- Characterize the equilibrium strategies of individuals & decision-maker

Strategic **fair** policy	Strategic unconstrained policy	**Non**-strategic **fair** policy	**Non**-strategic unconstrained policy

Results

- Characterize the equilibrium strategies of individuals & decision-maker

Strategic unconstrained policy	**vs.**	**Non**-strategic unconstrained policy

- Impact of decision-maker's anticipation of strategic manipulation
  - Conditions when strategic policy over/under accepts individuals
  - Conditions when strategic policy worsens/mitigates unfairness?

Results

- Characterize the equilibrium strategies of individuals & decision-maker

Non-strategic **fair** policy

vs.

Non-strategic unconstrained policy

- Impact of decision-maker's anticipation of strategic manipulation
  - Conditions when strategic policy over/under accepts individuals
  - Conditions when strategic policy worsens/mitigates unfairness?
- Impact of fairness constraint on non-strategic policies
  - Non-strategic decision-maker may benefit from fairness constraints

Results

- Characterize the equilibrium strategies of individuals & decision-maker

Strategic **fair** policy **vs.** Strategic unconstrained policy

- Impact of decision-maker's anticipation of strategic manipulation
  - Conditions when strategic policy over/under accepts individuals
  - Conditions when strategic policy worsens/mitigates unfairness?
- Impact of fairness constraint on non-strategic policies
  - Non-strategic decision-maker may benefit from fairness constraints
- Impact of fairness constraint on manipulative behavior
  - Fairness constraints can serve as incentives/disincentives for manipulation

Please stop by our poster (**Hall E #906**) to see more!