Strategic Classification with Random Manipulation Outcomes

Xueru Zhang, Computer Science and Engineering, The Ohio State University

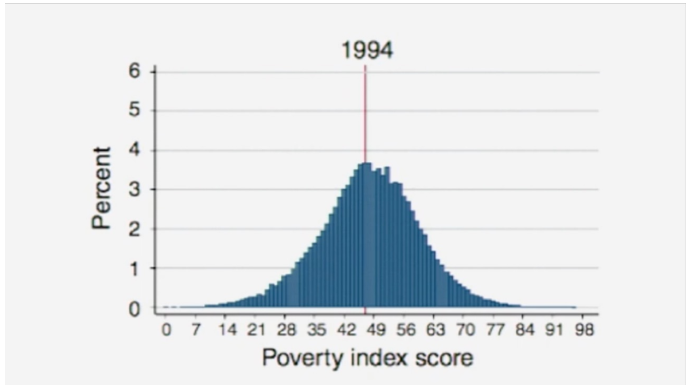
THE OHIO STATE UNIVERSITY

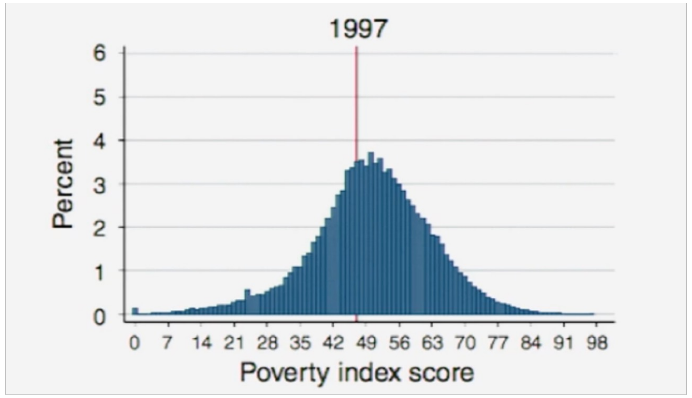
Machine Learning for People

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

Responsive and Interactive Distribution Shifts

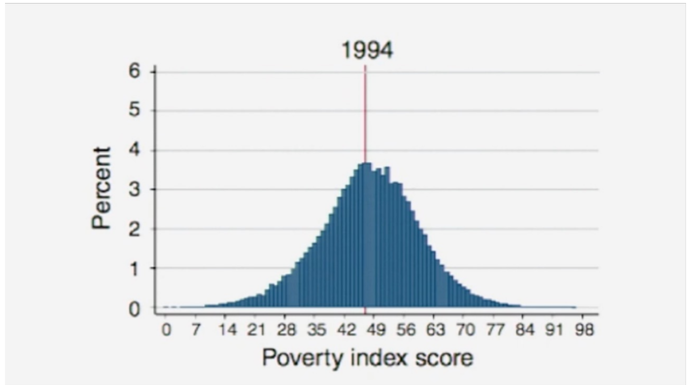
- Manipulation of social program eligibility *(Camacho et al., 2011)*

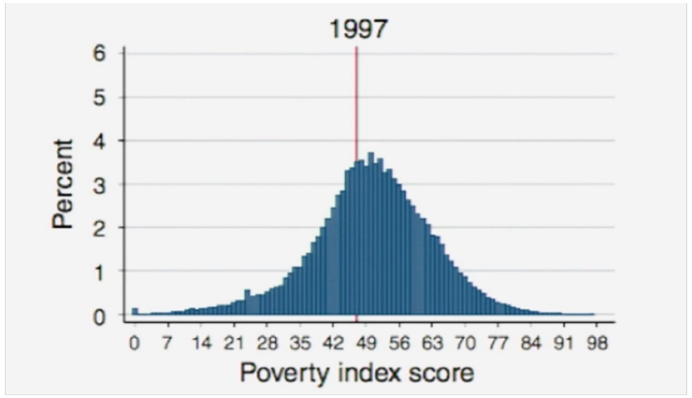




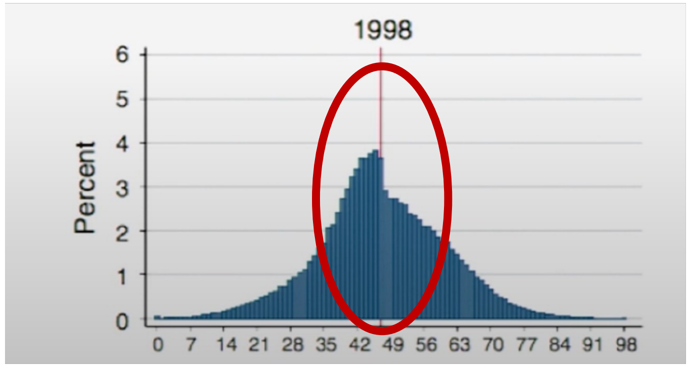
Responsive and Interactive Distribution Shifts

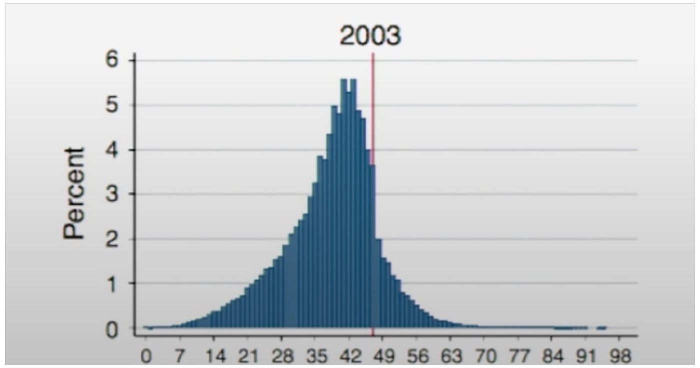
- Manipulation of social program eligibility *(Camacho et al., 2011)*





Government reveals some information about how the threshold is built



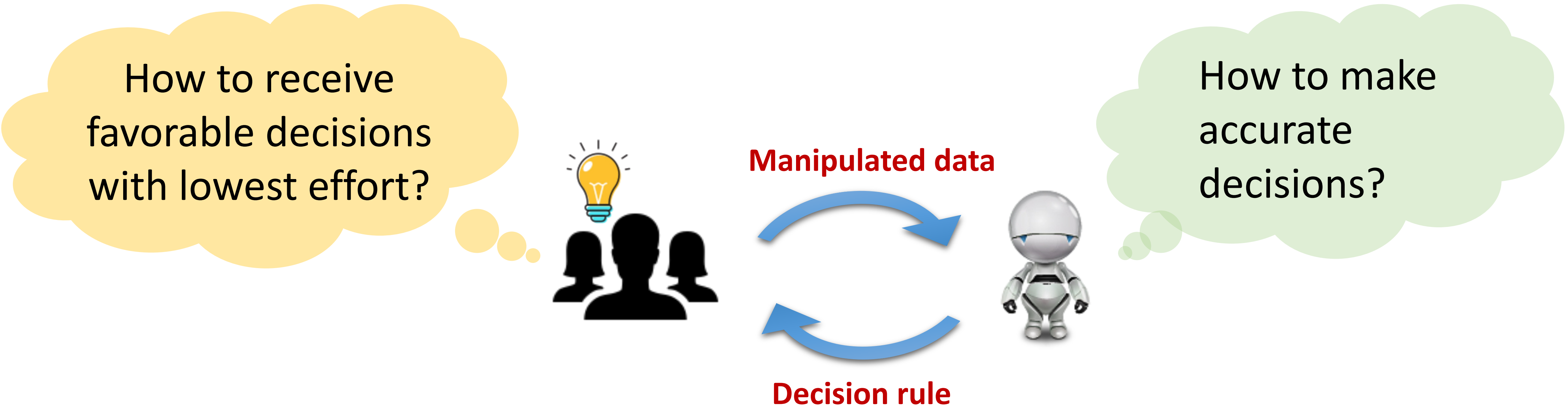


Responsive and Interactive Distribution Shifts

- Loan applicants apply for more credit cards to increase credit scores
- Job applicants manipulate the resumes to pass resume screening
- College applicants prepare application packages in a way that increase their chance of getting admitted

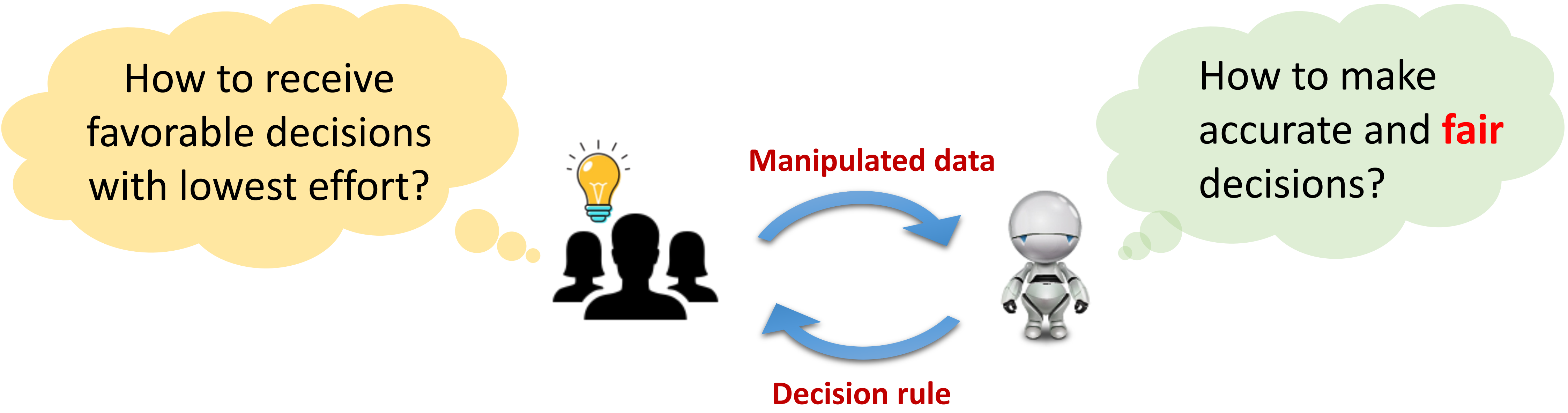
…

Challenge: ML under Strategic Behavior



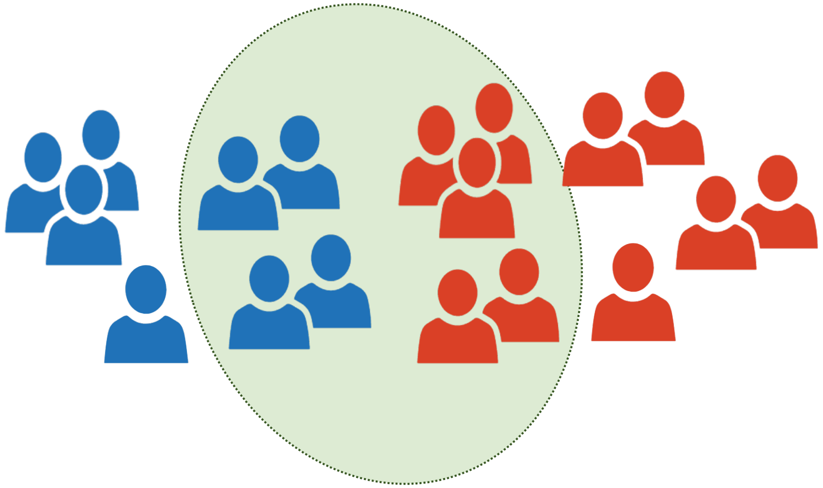
- ML is vulnerable to strategic manipulation

Another Challenge: Biases in ML



- ML can be biased against certain social groups

Existing Work: Fair Machine Learning



$$\min \quad Loss$$

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

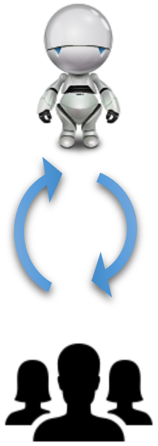
Fairness constraint

- **Demographic parity (DP)**: equal positive rate
- **Equal opportunity (EqOpt)**: equal true positive rate

…

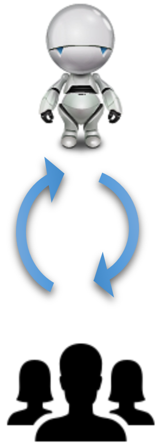
Existing Work: Strategic Classification

- Stackelberg game formulation *(Hardt et al., 2016a; Dong et al. 2018; Milli et al., 2019; Hu et al., 2019; Braverman & Garg, 2020)*

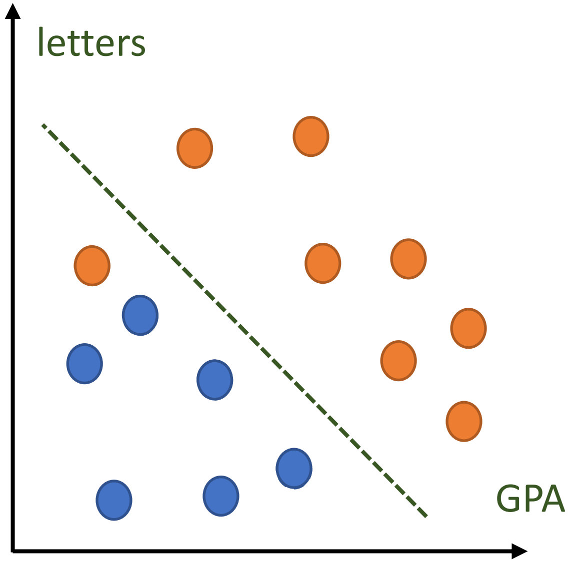


Existing Work: Strategic Classification

- Stackelberg game formulation *(Hardt et al., 2016a; Dong et al. 2018; Milli et al., 2019; Hu et al., 2019; Braverman & Garg, 2020)*

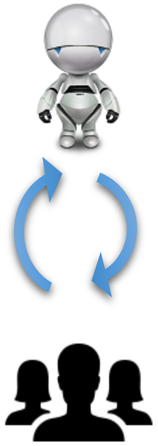


Classifier f



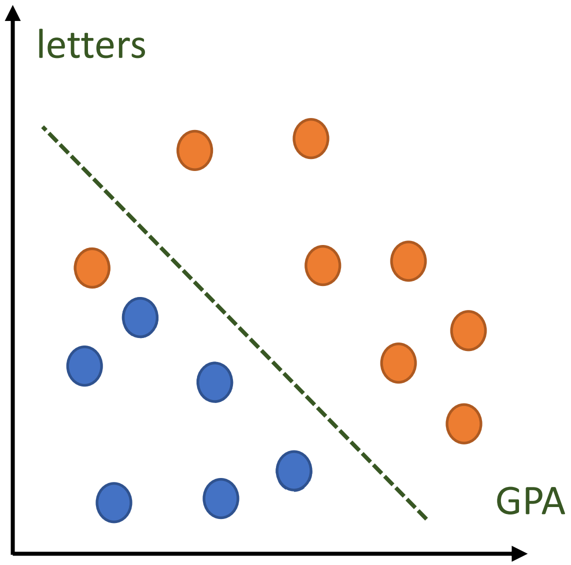
Existing Work: Strategic Classification

- Stackelberg game formulation *(Hardt et al., 2016a; Dong et al. 2018; Milli et al., 2019; Hu et al., 2019; Braverman & Garg, 2020)*



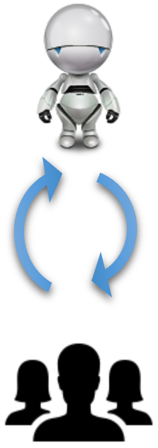
Classifier f

Initial data x



Existing Work: Strategic Classification

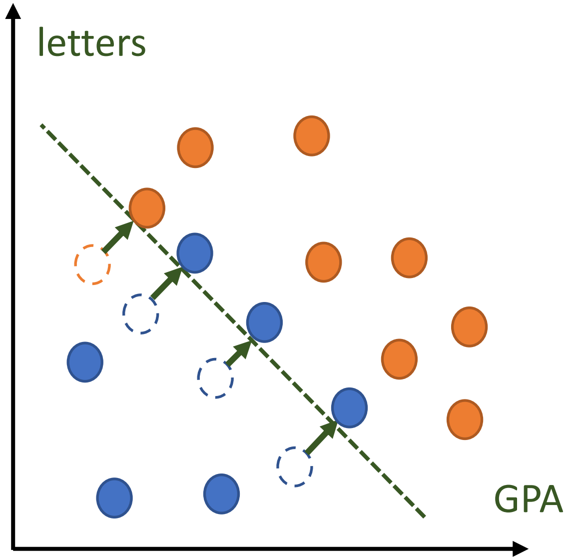
- Stackelberg game formulation *(Hardt et al., 2016a; Dong et al. 2018; Milli et al., 2019; Hu et al., 2019; Braverman & Garg, 2020)*



Classifier f

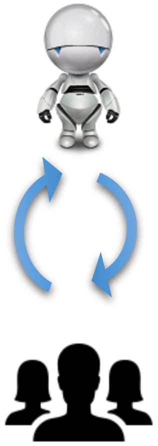
Initial data x

Manipulated data $\Delta(x)$



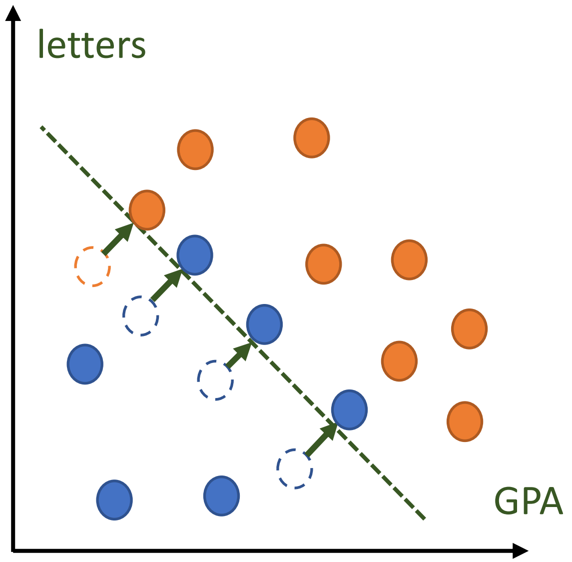
Existing Work: Strategic Classification

- Stackelberg game formulation *(Hardt et al., 2016a; Dong et al. 2018; Milli et al., 2019; Hu et al., 2019; Braverman & Garg, 2020)*



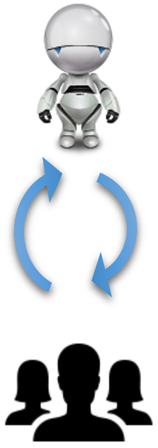
Classifier f

Initial data x
Manipulated data $\Delta(x)$
Cost $c(x, \Delta(x))$



Existing Work: Strategic Classification

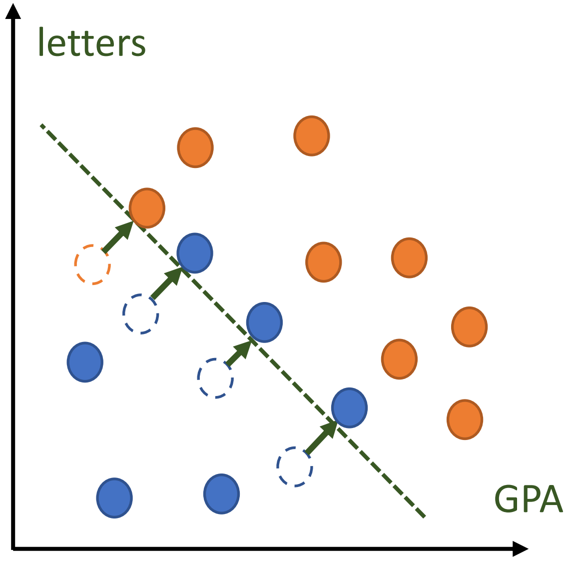
- Stackelberg game formulation *(Hardt et al., 2016a; Dong et al. 2018; Milli et al., 2019; Hu et al., 2019; Braverman & Garg, 2020)*



Classifier f

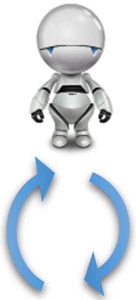
Initial data x
Manipulated data $\Delta(x)$
Cost $c(x, \Delta(x))$

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



Existing Work: 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

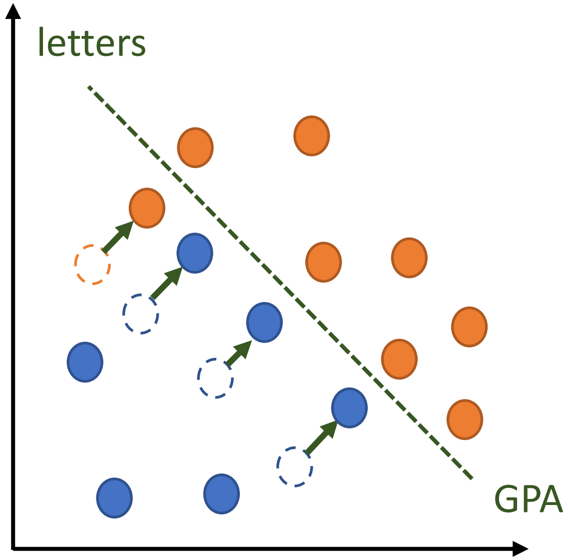
$$\max \Pr\left[h(x) = f\left(\Delta(x)\right)\right]$$

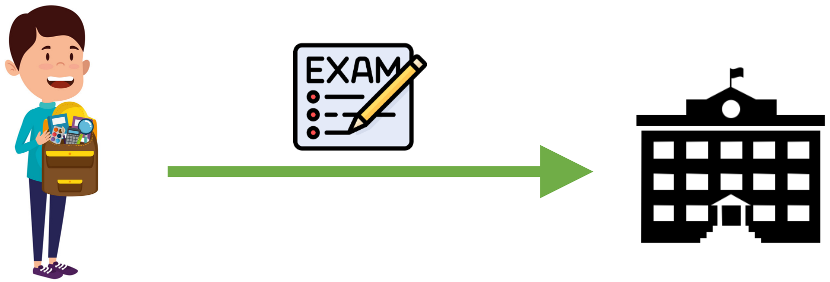
Initial data x

Manipulated data $\Delta(x)$

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

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




EXAM


EXAM


EXAM

$$\max f(\Delta(x)) - c(x, \cancel{\Delta}(x))$$

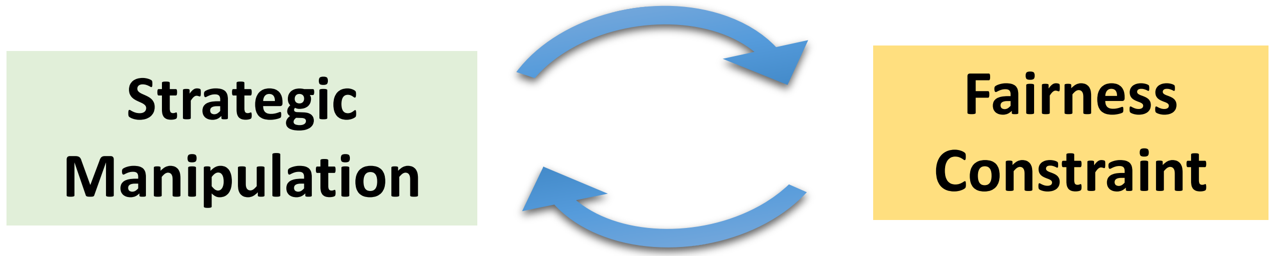
$$\max \Pr\left[h(x) = f\big(\Delta(x)\big)\right]$$

- **Random** manipulation outcomes
  - Unknown realizations before/after manipulation
- Cannot compute manipulation cost precisely
  - Random manipulation cost

Existing Stackelberg game formulation does not fit!

This talk:

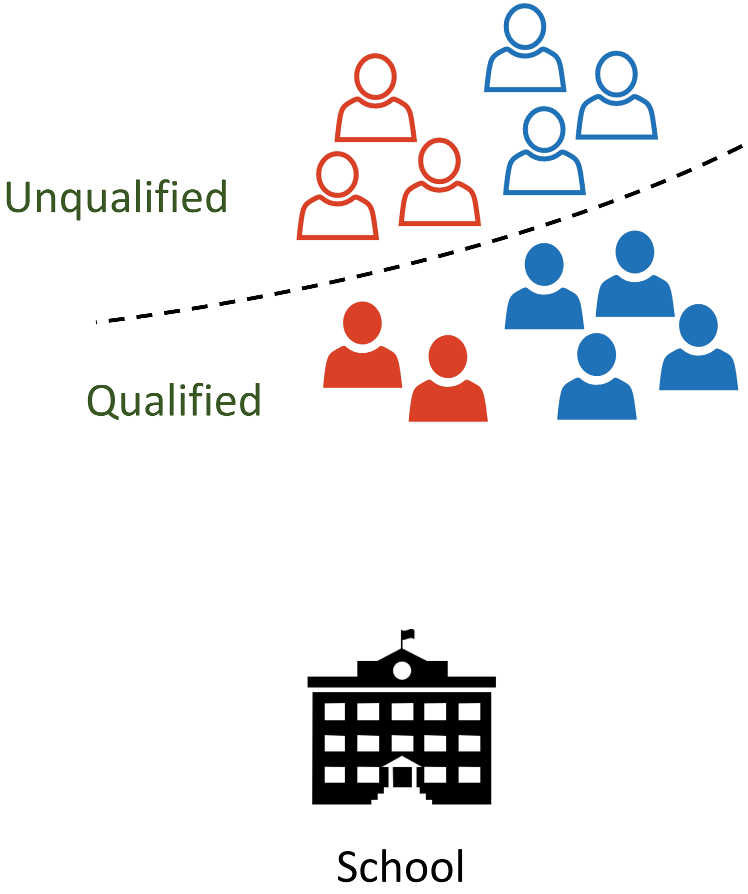
- A new Stackelberg game formulation that admits
  - **Random** manipulation outcomes & costs
- How strategic manipulation and fairness intervention impact each other?



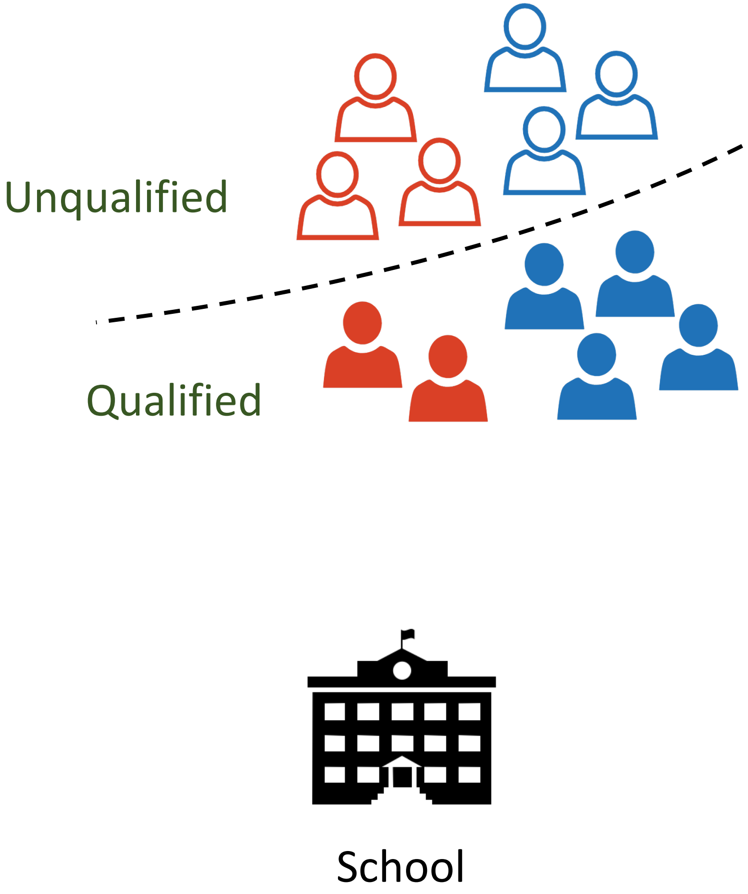
Zhang, X., Khalili, M. M., Jin, K., Naghizadeh, P., & Liu, M. (2022, June). Fairness Interventions as (Dis) Incentives for Strategic Manipulation. In *International Conference on Machine Learning* (ICML).

Model

Two demographic groups



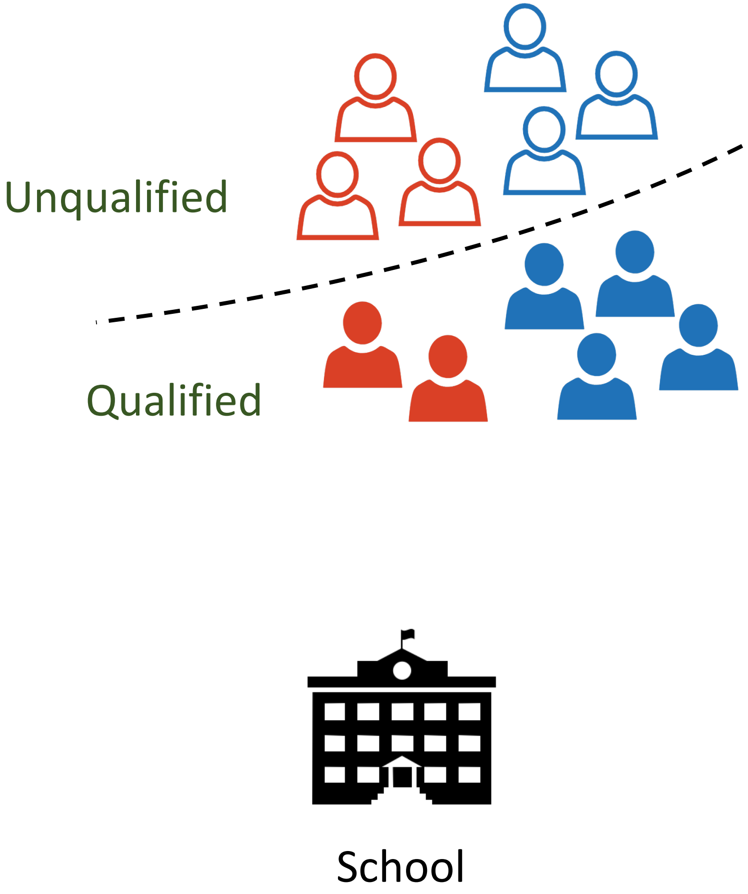
Model



Two demographic groups

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

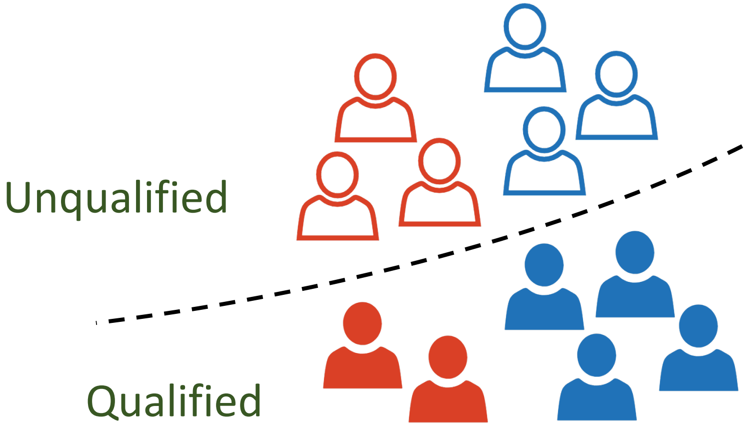
Model

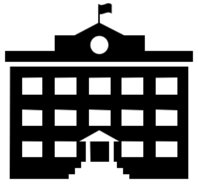


Two demographic groups

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

Model



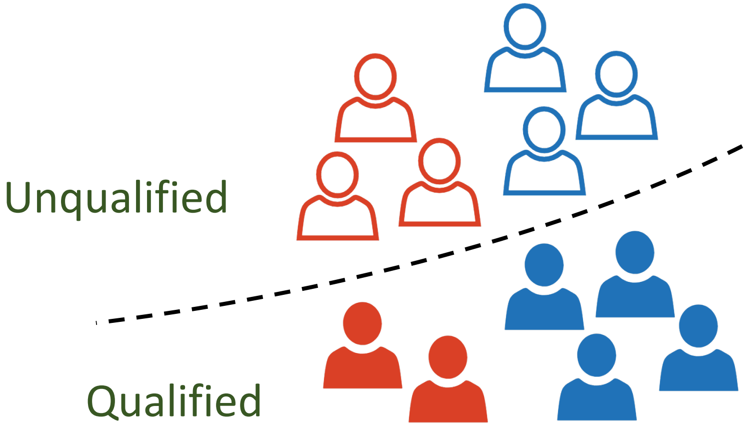


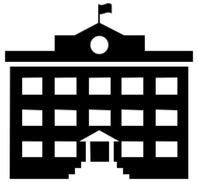
Two demographic groups

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

$$P_{X|YS}(x|y, s)$$

Model





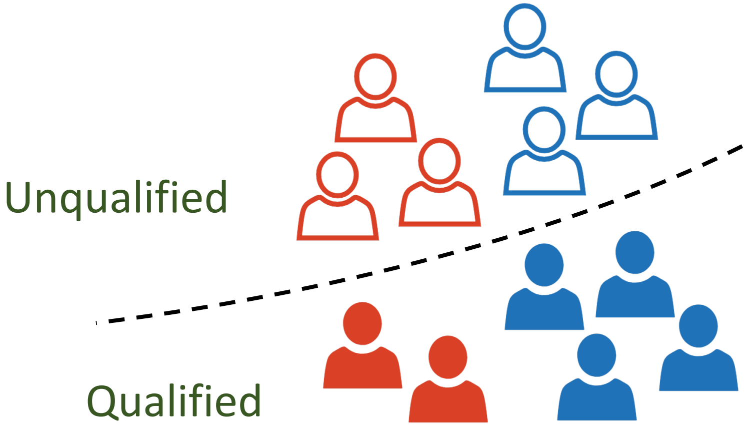
Two demographic groups

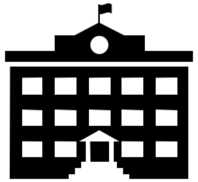
- Sensitive attribute $S \in \{a, b\}$ (race/gender)
- Feature X (exam score)
- Qualification $Y \in \{0,1\}$ (ability to graduate)

$$P_{X|YS}(x|y, s)$$

- Decision $D \in \{0,1\}$ (get admitted or not)

Model





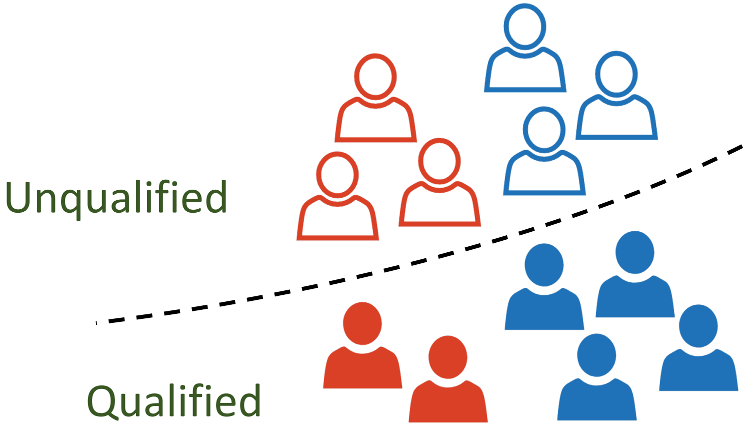
Two demographic groups

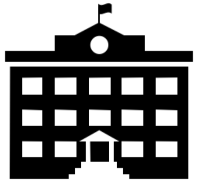
- Sensitive attribute $S \in \{a, b\}$ (race/gender)
- Feature X (exam score)
- Qualification $Y \in \{0,1\}$ (ability to graduate)

$$P_{X|YS}(x|y, s)$$

- Decision $D \in \{0,1\}$ (get admitted or not)
  - Decision-maker's policy $\pi_s(x) = P_{D|XS}(1|x, s)$

Model





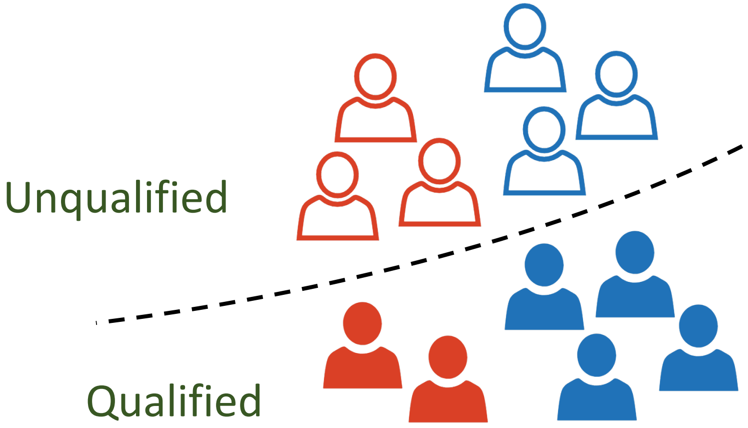
Two demographic groups

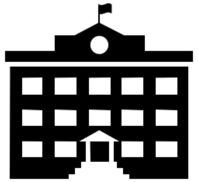
- Sensitive attribute $S \in \{a, b\}$ (race/gender)
- Feature X (exam score)
- Qualification $Y \in \{0,1\}$ (ability to graduate)

$$P_{X|YS}(x|y, s)$$

- 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

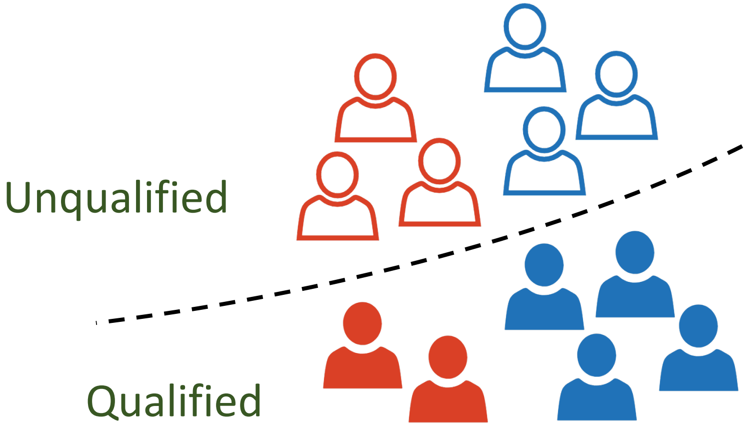
Two demographic groups

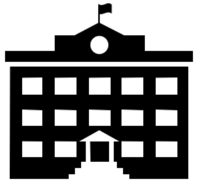
- Sensitive attribute $S \in \{a, b\}$ (race/gender)
- Feature X (exam score)
- Qualification $Y \in \{0,1\}$ (ability to graduate)

$$P_{X|YS}(x|y,s)$$

- 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

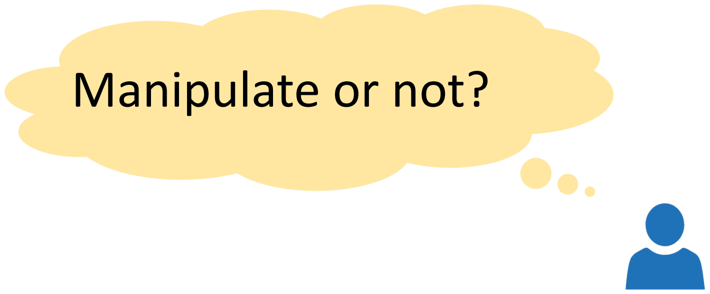
Two demographic groups

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

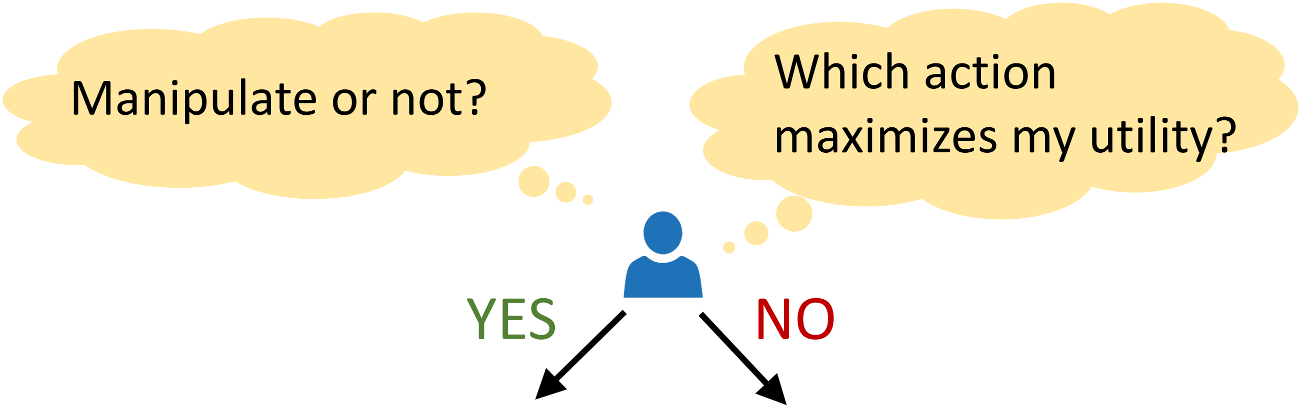
$$P_{X|YS}(x|y, s)$$

- 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



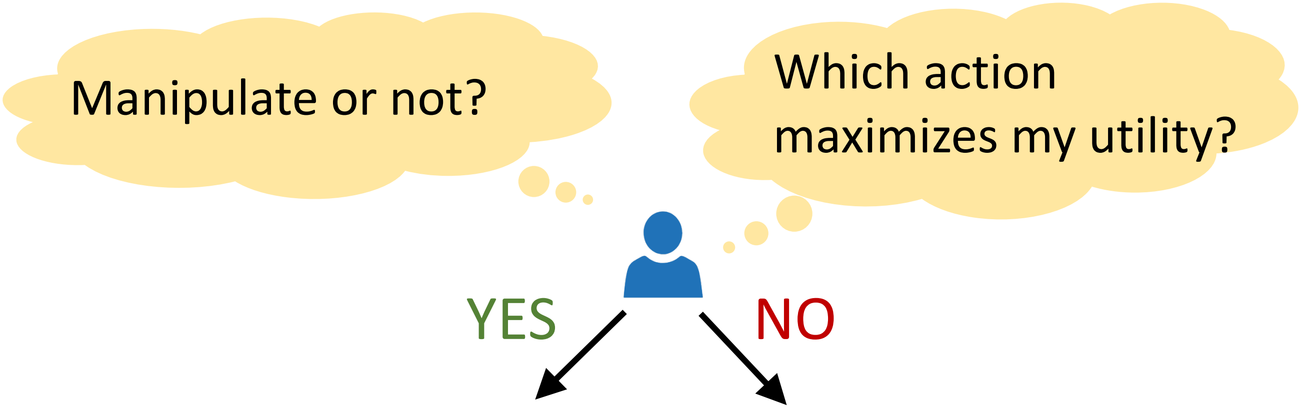
$$P_{D|YMS}(1|y, 1, s) - C_s$$

Benefit with manipulation – cost

$$P_{D|YMS}(1|y, 0, s)$$

Benefit without manipulation

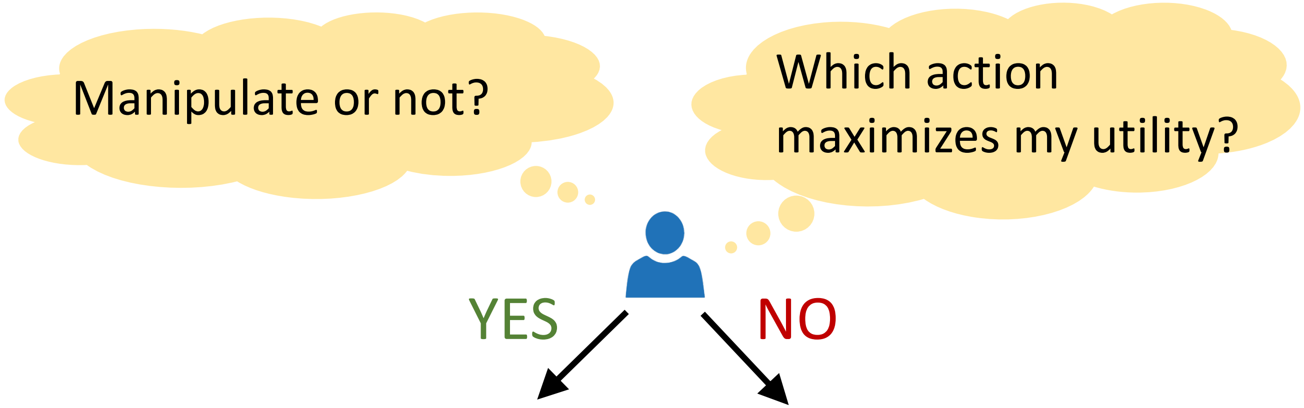
Model: individual best response



$$P_{D|YMS}(1|y,1,s) - C_s \qquad\qquad P_{D|YMS}(1|y,0,s)$$

Manipulate ($M = 1$) if $\boxed{\text{Benefit with manipulation – cost}} \geq \boxed{\text{Benefit without manipulation}}$

Model: individual best response

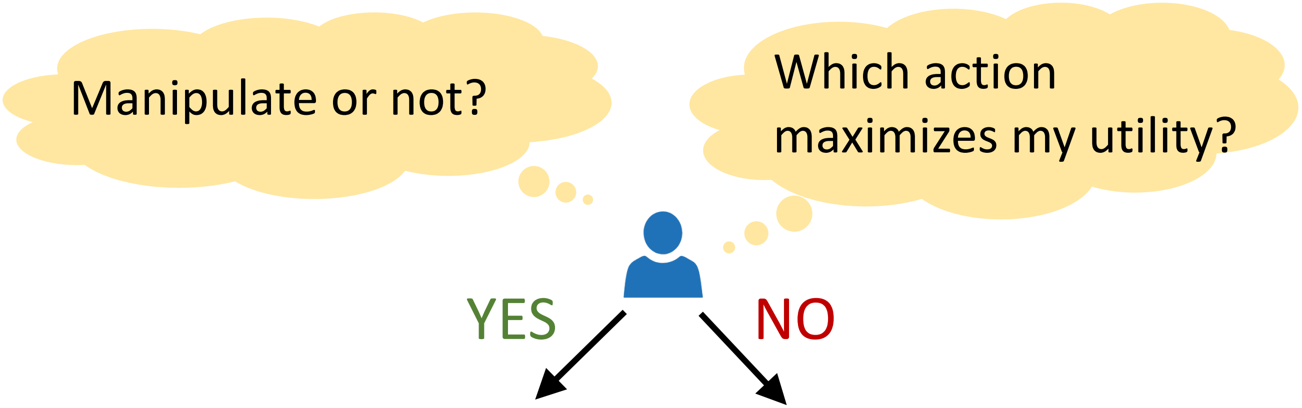


$$P_{D|YMS}(1|y,1,s) - C_s \qquad\qquad P_{D|YMS}(1|y,0,s)$$

Manipulate ($M = 1$) if $\boxed{\text{Benefit with manipulation – cost}} \geq \boxed{\text{Benefit without manipulation}}$

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

Model: individual best response



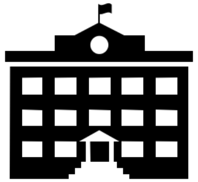
$$P_{D|YMS}(1|y,1,s) - C_S \qquad\qquad P_{D|YMS}(1|y,0,s)$$

Manipulate ($M = 1$) if $\boxed{\text{Benefit with manipulation – cost}} \geq \boxed{\text{Benefit without manipulation}}$

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

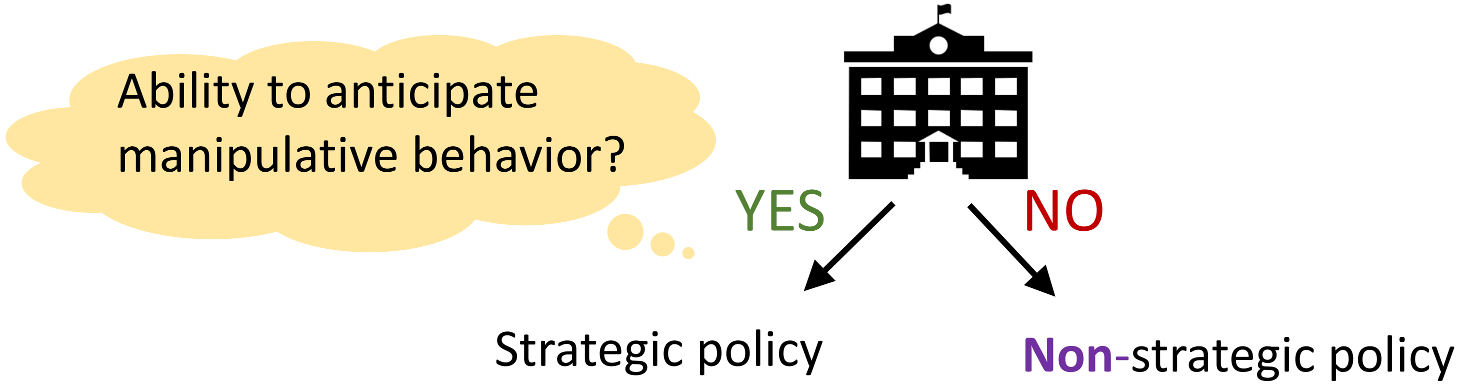
$$\Pr\left(C_S \leq P_{D|YMS}(1|y,1,s) - P_{D|YMS}(1|y,0,s)\right)$$

Model: decision-maker's optimal policies



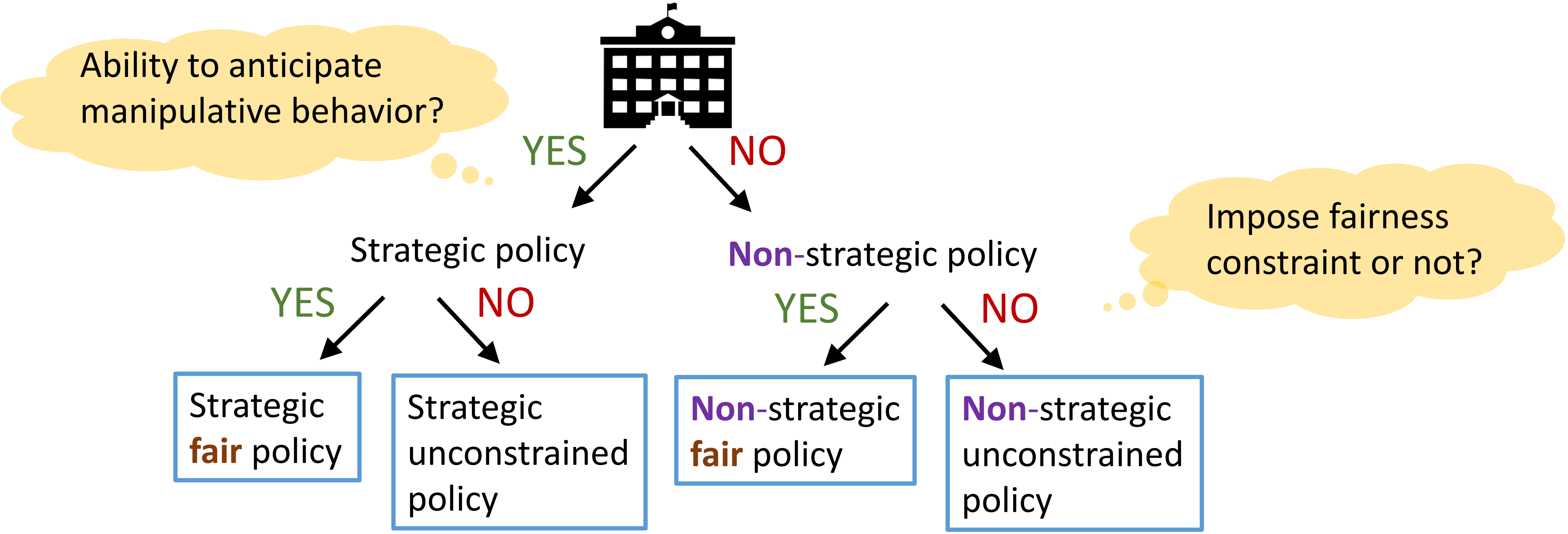
- Policy (π_a, π_b) that maximizes the expected utility $\mathbb{E}[R(Y, D)]$
  - True-positive benefit $R(1,1) = u_+$
  - False-positive penalty $R(0,1) = -u_-$

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

Special case:

- Individuals manipulate by imitating the features of qualified people
  - Only unqualified individuals have incentives to manipulate
- Threshold decision policy: $\pi_s(x) = P_{D|XS}(1|x,s) = \mathbf{1}(x \geq \theta_s)$
- Monotone likelihood ratio property: $\frac{P_{X|YS}(x|1,s)}{P_{X|YS}(x|0,s)}$ is increasing in $x \in \mathbb{R}$

Results

Strategic **fair** policy

Strategic unconstrained policy

Non-strategic **fair** policy

Non-strategic unconstrained policy

- Characterize the equilibrium strategies of individuals & decision-maker

Goal:

1. *How can policies (and fairness property) be affected when decision-maker has ability to anticipate strategic behavior?*
2. *What are the impacts of fairness interventions on policies and resulting manipulative behavior?*

Results

Strategic unconstrained policy

vs.

Non-strategic unconstrained policy

- Impacts on acceptance threshold

Let $\delta = \frac{u_-}{u_- + u_+}$, compared to non-strategic policy,

1. Strategic policy is the same if $P_{Y|S}(1|s) = \delta$
2. Strategic policy over accepts individuals if $P_{Y|S}(1|s) > \delta$ Majority-qualified
3. Strategic policy under accepts individuals if $P_{Y|S}(1|s) < \delta$ Majority-unqualified

Results

Strategic unconstrained policy

vs.

Non-strategic unconstrained policy

- Impacts on unfairness (DP/EqOpt): $(T)PR_a - (T)PR_b$
- Let disadvantaged group be the group with smaller (true) positive rate

If $P_{Y|S}(1|a) > \delta > P_{Y|S}(1|b)$ and group b is disadvantaged under non-strategic policy, then under strategic policy:

1. Unfairness get worse; **and**
2. Group b is still disadvantaged

Results

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

- Impacts on unfairness (DP/EqOpt): $(T)PR_a - (T)PR_b$
- Let disadvantaged group be the group with smaller (true) positive rate

If $\delta > P_{Y|S}(1|s)$ for both groups and group b is disadvantaged under non-strategic policy, then always there exists C_a for group a such that:

1. Strategic policy mitigates unfairness; **or**
2. The disadvantaged group is flipped from group b to group a

Results

Non-strategic fair policy

vs.

Non-strategic unconstrained policy

- Impacts of fairness constraint on non-strategic policy

Under certain scenarios, a non-strategic decision-maker can benefit from fairness interventions by receiving higher utility from both groups

Results

Strategic **fair** policy

vs.

Strategic unconstrained policy

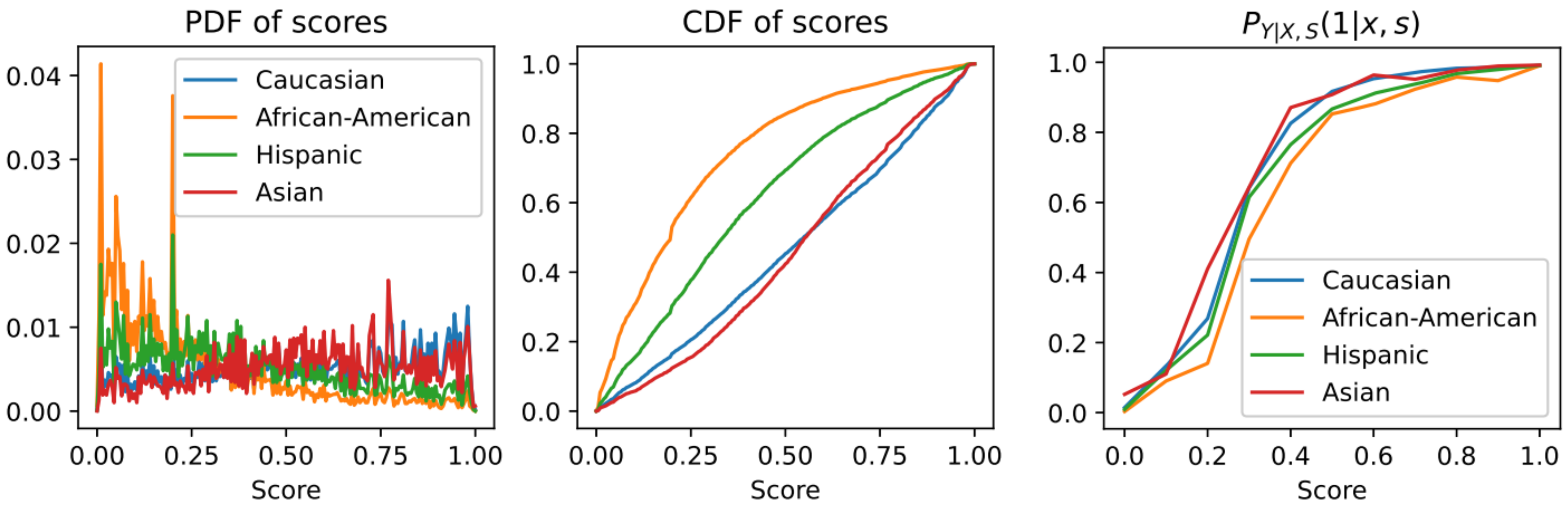
- Impacts of fairness constraint on strategic policy and individual behavior
- More complicated relations

Fairness interventions can serve as incentives and/or disincentives for strategic manipulation. We identified scenarios under which:

1. Both groups are more/less likely to manipulate under fair policy
2. One group is more likely to manipulate while the other is less likely to manipulate under fair policy.

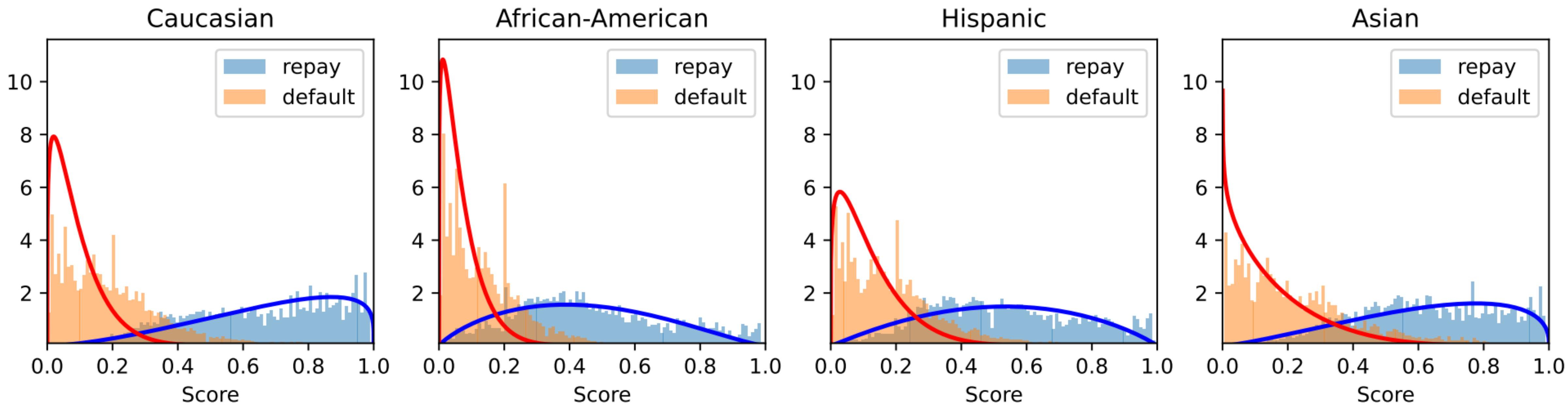
Experiments: FICO credit scores

- Scores are normalized from [300, 850] to [0, 1]
- Use empirical data to estimate:
  - Qualification (repayment) rates $P_{Y|S}(1|s)$, group proportion $P_S(s)$



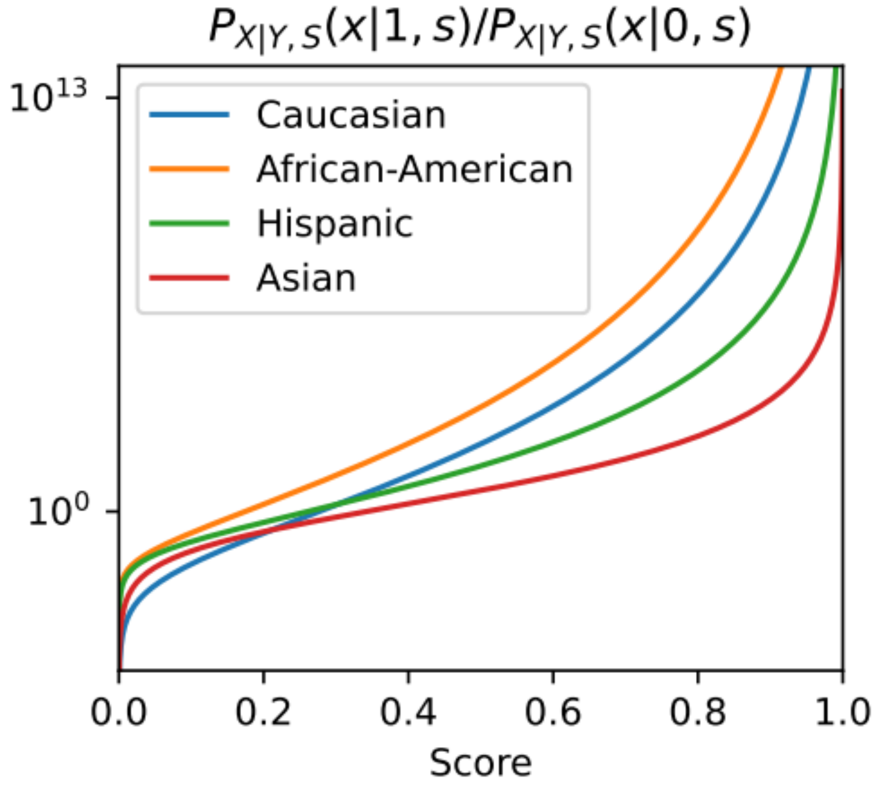
Experiments: FICO credit scores

- Fit Beta distribution to get $P_{X|YS}(x|y,s)$



Experiments: FICO credit scores

- Monotone likelihood ratio property



- Manipulation costs:
  - Uniform: $C_s \sim U[0, \bar{c}]$
  - Beta: $C_s \sim Beta[v, w]$

Experiments: FICO credit scores

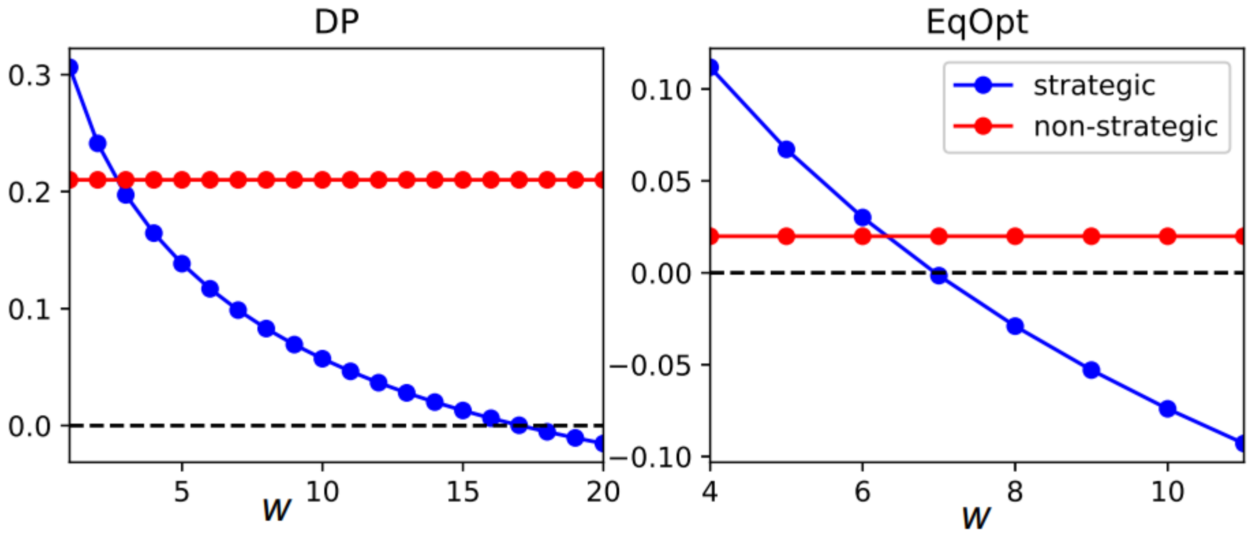
- Impacts of using strategic policy on unfairness: $(T)PR_a - (T)PR_b$
  - Group b: African-American
  - Group a: Caucasian/Hispanic/Asian
  - $u_- = u_+$, we have $P_{Y|S}(1|a) > \delta > P_{Y|S}(1|b)$
  - When $C_a \neq C_b$, it is less costly for group b to manipulate

	$\mathcal{G}_a$	strategic $C_a = C_b$	strategic $C_a \neq C_b$	non-strategic
EqOpt	Caucasian	0.355	0.556	0.136
	Hispanic	0.292	0.493	0.034
	Asian	0.333	0.533	0.123
DP	Caucasian	0.611	0.680	0.449
	Hispanic	0.421	0.490	0.242
	Asian	0.634	0.703	0.522

Unfairness get worse;
Group b is still disadvantaged

Experiments: FICO credit scores

- Impacts of using strategic policy on unfairness: $(T)PR_a - (T)PR_b$
  - Group b: African-American
  - Group a: Hispanic
  - $u_- = 2u_+$, we have $\delta > P_{Y|S}(1|s)$ for both groups
  - Fix group b and decrease the manipulation cost of group a



Unfairness can get mitigated;
The disadvantaged group may be flipped

Experiments: FICO credit scores

- Impacts of fairness constraint on non-strategic decision-maker
  - Group b: Caucasian
  - Group a: Asian

With fairness intervention

C_a	$U_a(\widehat{\theta}_a^{\mathrm{UN}})$	$U_a(\widehat{\theta}_a^{\mathcal{C}})$	$U_b(\widehat{\theta}_b^{\mathrm{UN}})$	$U_b(\widehat{\theta}_b^{\mathcal{C}})$
Beta(10, 10)	-0.190	-0.189	0.024	0.034
Beta(10, 1)	0.396	0.397	0.181	0.201

Under scenarios we identified, non-strategic decision-maker can benefit from fairness constraint by receiving higher utilities from both groups

Conclusion

- A new Stackelberg game formulation that admits
  - **Random** manipulation outcomes & costs
- Equilibrium strategies of both individuals & decision-maker

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

- What happens if decision-maker can (not) anticipate manipulative behavior?
- How is the population and decision-maker affected by fairness intervention?