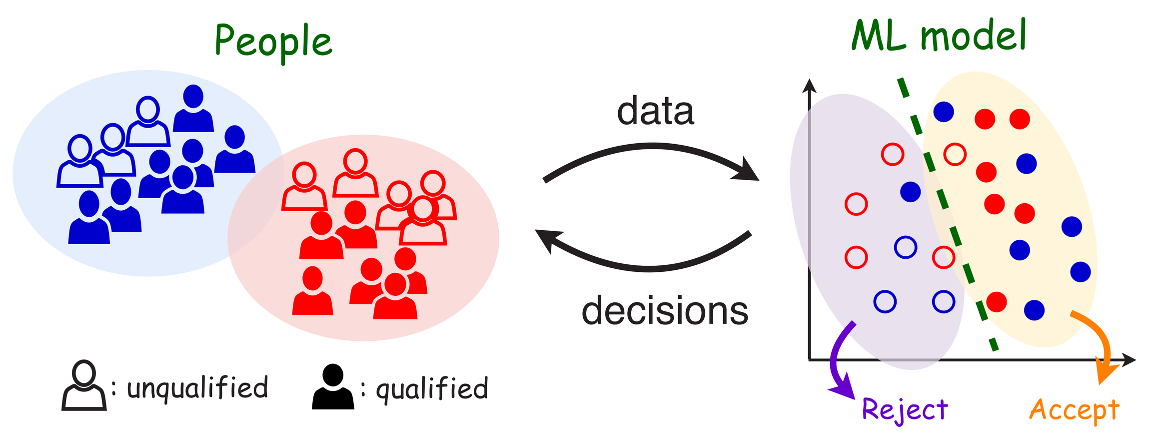
How Do Fair Decisions Fare in Long-term Qualification?

XUERU ZHANG*, RUIBO TU*, YANG LIU, MINGYAN LIU, HEDVIG KJELLSTRÖM, KUN ZHANG, CHENG ZHANG

NEURAL INFORMATION PROCESSING SYSTEMS

OBJECTIVES

- **Setting:** a decision-maker aims to select people from applicants that are qualified for tasks.
- Impose fairness constraint to make fair decisions (e.g., same acceptance rates across groups)
- Interplay between ML models and people
  - ML decisions affect people's behaviors
  - People generate data for training ML models



Goal: study the long-term impact of the fairness constraints on qualifications of different groups

MODEL

Two demographic groups $\mathcal{G}_a$, $\mathcal{G}_b$

- Sensitive attribute $S \in \{a, b\}$
- Time-varying feature $X_t \in \mathbb{R}^d$ and qualification state $Y_t \in \{0, 1\}$
  - ***Feature generation process:*** time-invariant $P_{X|Y,S}(x|y,s) = \mathbb{P}(X_t = x|Y_t = y, S = s)$
  - ***Transitions of qualification state:*** time-invariant $T^s_{yd} = \mathbb{P}(Y_{t+1} = 1|Y_t = y, D_t = d, S = s)$
- Qualification rate $\alpha^s_t = P_{Y_t|S}(1|s)$
- Inequality measure: disparity between α^a_t and α^b_t

Myopic decision-maker's optimal fair policies π^a_t, π^b_t

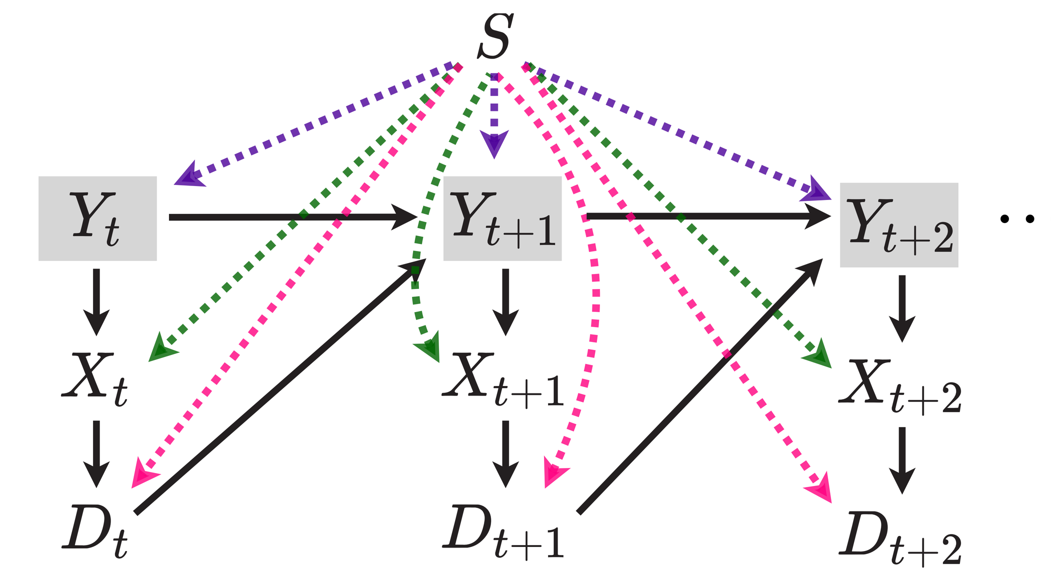
$$\max_{\pi^a,\pi^b} \quad U_t(\pi^a,\pi^b) = \mathbb{E}[R(D_t, Y_t)]$$

$$\text{s.t.} \quad \mathbb{E}_{X_t\sim\mathcal{P}^a_{\mathcal{C}}}[\pi^a(X_t)] = \mathbb{E}_{X_t\sim\mathcal{P}^b_{\mathcal{C}}}[\pi^b(X_t)]$$

- Unconstrained (UN)
- Demographic Parity (DP): $\mathcal{P}^s_{\text{DP}}(x) = P_{X|S}(x|s)$
- Equal Opportunity (EqOpt): $\mathcal{P}^s_{\text{EqOpt}}(x) = P_{X|Y,S}(x|1,s)$

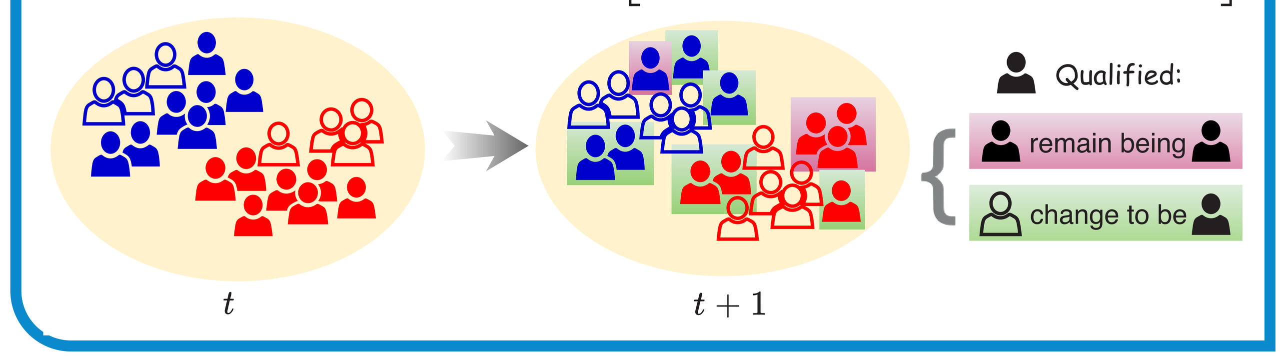
- Decision $D_t \in \{0, 1\}$ is based on $\pi^s_t(x) = \mathbb{P}(D_t = 1|X_t = x, S = s)$
- Utility function $R(1, 1) = u_+$, $R(1, 0) = -u_-$, $R(0, 1) = R(0, 0) = 0$

DYNAMICS



$$\alpha^s_{t+1} = g^{0s}(\alpha^a_t, \alpha^b_t) \cdot (1 - \lambda^s_t) + g^{1s}(\alpha^a_t, \alpha^b_t) \cdot \alpha^s_t$$

$$g^{ys}(\alpha^a_t, \alpha^b_t) = \mathbb{E}_{X_t|Y_t=y,S=s}\left[(1-\pi^s_t(X_t))T^s_{y0} + \pi^s_t(X_t)T^s_{y1}\right]$$



EQUILIBRIUM ANALYSIS

- **Optimal (fair) policies:** threshold policies are optimal.
- **Existence of equilibrium:** $\forall T^s_{dy} \in (0, 1)$, the dynamics have at least one equilibrium $(\widehat{\alpha}^a, \widehat{\alpha}^b)$.
- **Uniqueness of equilibrium:** sufficient conditions for the uniqueness of equilibrium under (A)(B).

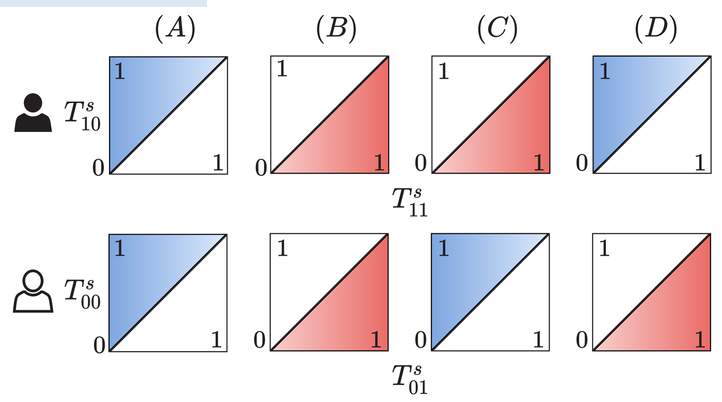
Two effects on people

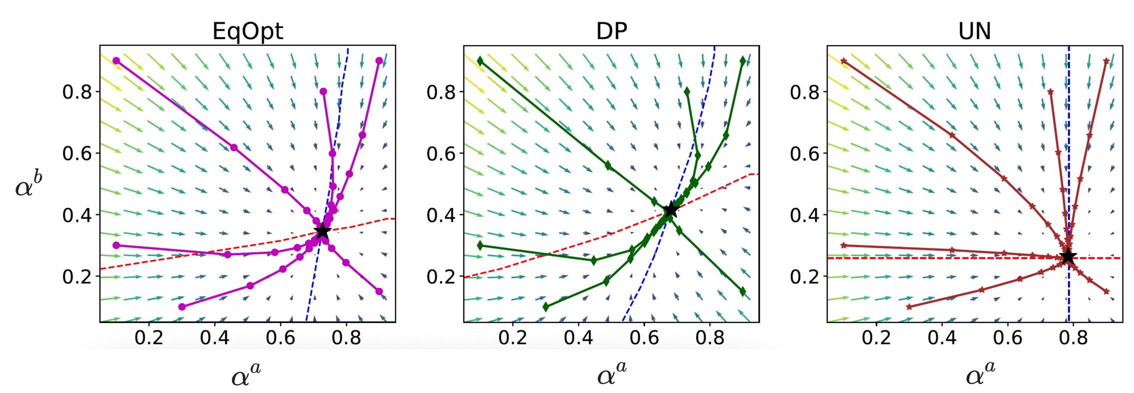
– "Lack of motivation"

$T^s_{y1} \leq T^s_{y0}$

– "Leg-up"

$T^s_{y1} \geq T^s_{y0}$



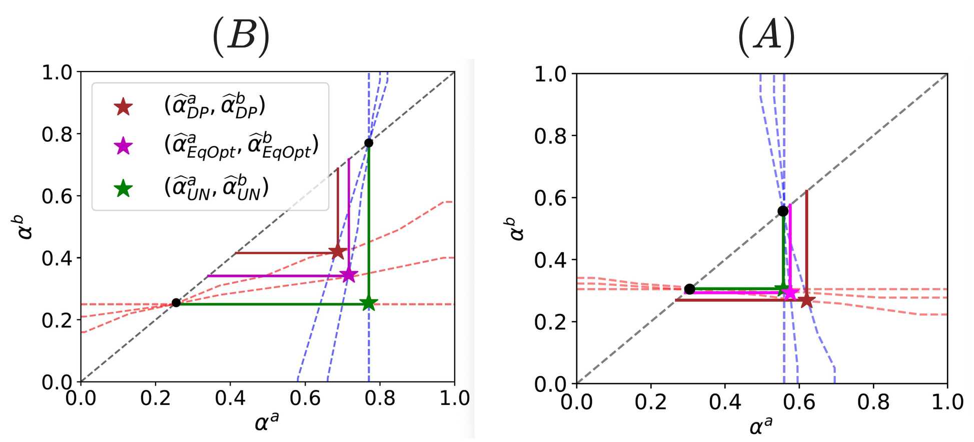


LONG-TERM IMPACT OF FAIRNESS CONSTRAINTS

- **Natural equality:** $\forall P_{X|Y,S}$ and $\forall \alpha \in (0, 1)$, $\exists$ transitions T^s_{yd} under (A) or (B) s.t. $\widehat{\alpha}^a_{\text{UN}} = \widehat{\alpha}^b_{\text{UN}} = \alpha$.
  - If $P_{X|Y,S=a} = P_{X|Y,S=b}$, then fairness $\mathcal{C}$ = DP or EqOpt maintains equality: $\widehat{\alpha}^a_{\mathcal{C}} = \widehat{\alpha}^b_{\mathcal{C}}$
  - If $P_{X|Y,S=a} \neq P_{X|Y,S=b}$, then fairness $\mathcal{C}$ = DP or EqOpt violates equality: $\widehat{\alpha}^a_{\mathcal{C}} \neq \widehat{\alpha}^b_{\mathcal{C}}$
- **Natural inequality ($\widehat{\alpha}^a_{\text{UN}} \neq \widehat{\alpha}^b_{\text{UN}}$):**

Case 1: due to different transitions

- Under (A), DP and EqOpt exacerbate inequality
- Under (B), DP and EqOpt mitigate inequality
- Disadvantaged group remains being disadvantaged



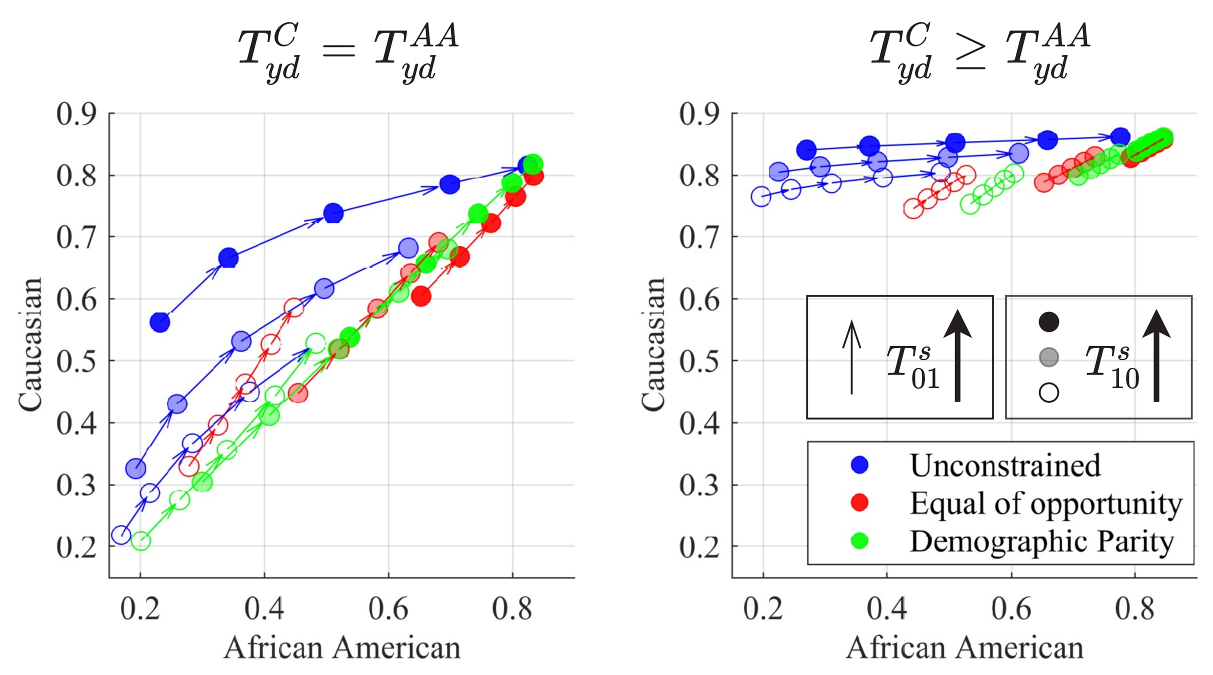
Case 2: due to different feature that generated

Under some conditions on $P_{X|Y,S}$, u_+, u_- and T^s_{yd} satisfying (B):

- EqOpt mitigates inequality and disadvantaged group remains being disadvantaged
- DP either mitigates inequality, or flips disadvantaged group

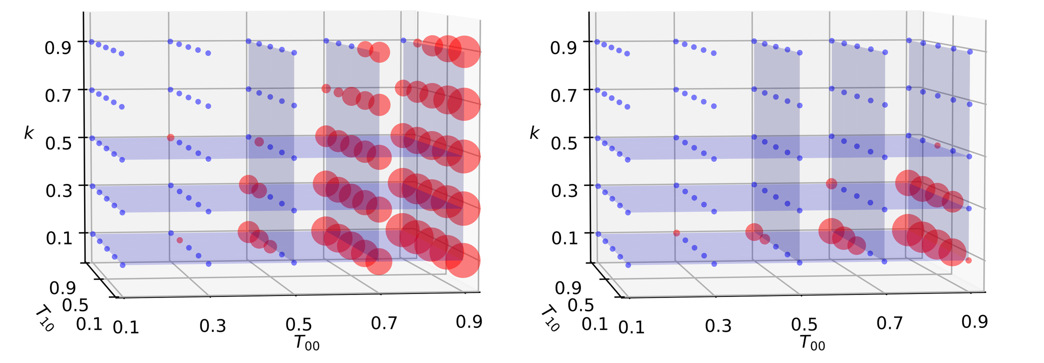
NUMERICAL RESULTS

- **FICO score dataset**
  - Effect of transition intervention



- **COMPAS dataset**
  - Oscillation may happen in the long-run

	$\widehat{\alpha}_{\theta_H} < \widehat{\alpha}^*$	$\widehat{\alpha}_{\theta_L} < \widehat{\alpha}^*$	osi*	osi$_H$	osi$_L$
A	0	1	0.29	0.12	0.36
B	0.99	0.01	0	0	0
C	0.37	0.28	0	0	0
D	0.79	0.63	0.06	0	0.13



EFFECTIVE INTERVENTION

- **Policy Intervention:**

– Sub-optimal fair policies can improve $(\widehat{\alpha}^a, \widehat{\alpha}^b)$

– $\exists$ threshold policies s.t. $\widehat{\alpha}^a = \widehat{\alpha}^b$ as long as T^a_{yd} and T^b_{yd} are not different significantly

- **Transition Intervention:**

– Increasing any T^s_{yd} increases $\widehat{\alpha}^s$

CONCLUSIONS

- Construct a POMDP framework for sequential decision-making and analyze its equilibrium.
- Imposing fairness constraints may or may **not** help in promoting long-term equality.
- Importance of understanding real-world dynamics in decision-making systems.