Fairness in Machine Learning: Delayed impact and other desiderata

Lydia T. Liu (Cornell University)

Simons Workshop on "Societal Considerations and Applications" | Nov 9, 2022

This talk

1. **Review the problem of "fairness" in machine learning in the context of algorithmic risk scores, and its formalizations as statistical criteria**
   [e.g. Calders et al 2009; Angwin et al 2016; Zafar et al, 2017; Hardt et al, 2016; Chouldechova, 2016; Kleinberg et al, 2017, **Liu** et al, 2019…]

2. ***Delayed Impact* model for characterizing downstream welfare implications of fairness criteria**
   [**Liu**, Sarah Dean, Esther Rolf, Max Simchowitz, Moritz Hardt, 2018]

3. **Follow-up work and broader impacts**
   [e.g. Mouzannar et al 2019; **Liu** et al 2020; Kannan et al 2019; Arunachaleswaran et al 2020; Dwork et al 2020; Morik et al 2020; Ge et al 2021; Nilforoshan et al 2022; D'Amour et al 2020; Holstein et al 2019; Fazelpour and Lipton 2020; Lee et al 2021…]

Many consequential decisions in society depend on algorithmic risk scores.

How Flawed Data Aggravates Inequality in Credit

AI offers new tools for calculating credit risk. But it can be tripped up by noisy data, leading to disadvantages for low-income and minority borrowers.

Aug 6, 2021 | Edmund L. Andrews

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Ofqual's A-level algorithm: why did it fail to make the grade?

There is a lot we can learn from the algebraic symbols used to determine results in England

- A university vice-chancellor's diary of A-level chaos

Promises

- Avoid arbitrariness of human decisions
  (e.g. judges affected by extraneous factors [Danziger et al., 2011])
- Better information
  (e.g. credit scoring led to increased credit access for high risk households [Edelberg 2006])

Problems

- Algorithms can make systematically “biased” assessments
- Algorithmic decision making can still lead to disparate impact
  (v.s. disparate treatment)

Source: [Angwin et al 2016]

Forms of “algorithmic bias” by group

Classification context. Individual has features X, binary decision D, based on score R(X). True outcome Y. Protected group attribute A.

e.g. X: credit history, R: credit score, D: loan approval, Y: on-time loan repayment, A: race

- Decision D (or score R) is group-dependent.
  - “Loan approval rate differs by group.”
    $\mathbb{E}[D] \neq \mathbb{E}[D \mid A]$
  - violates **Demographic Parity**

Forms of “algorithmic bias” by group

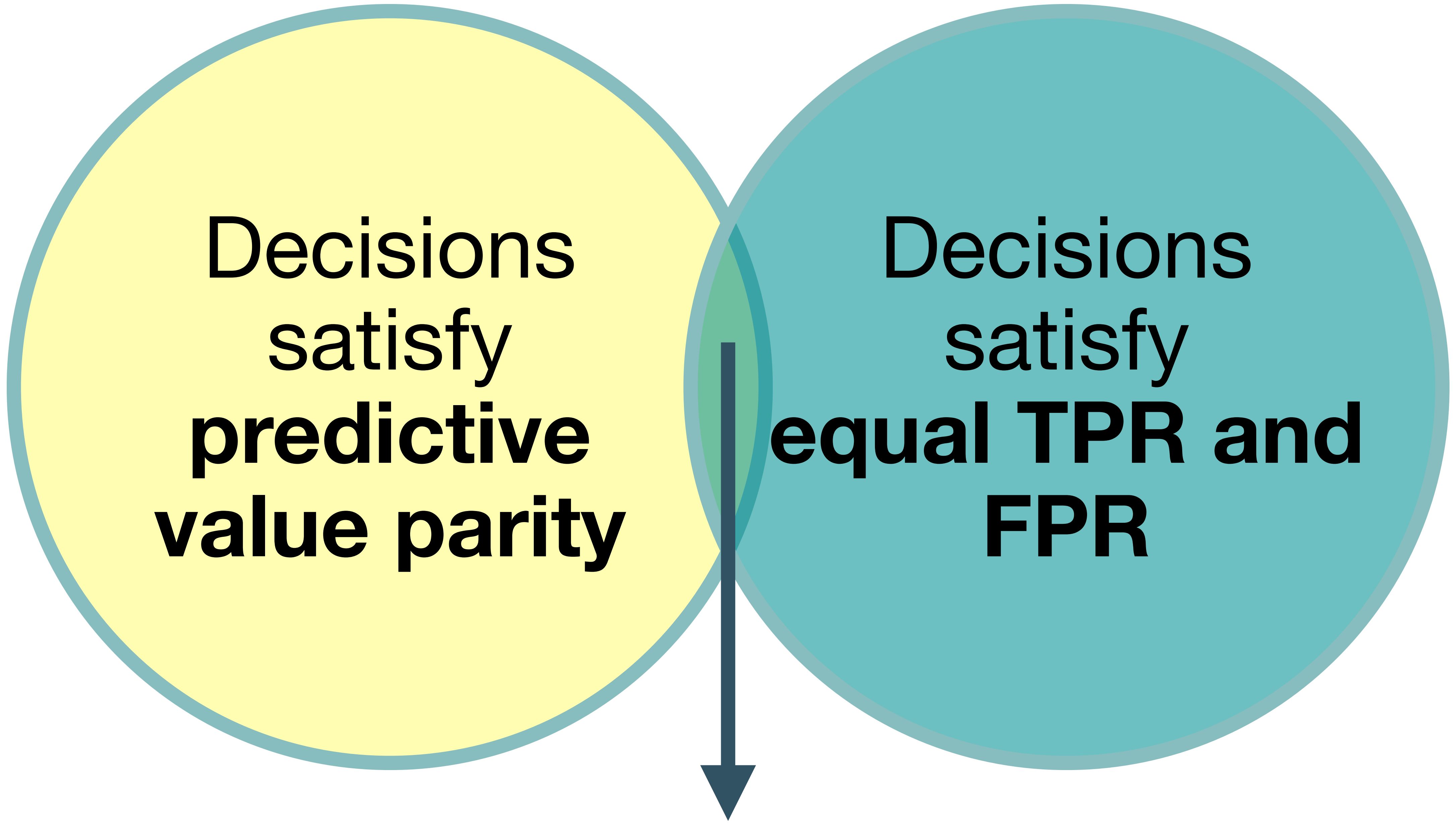
- False positive/negative rate of decisions is group-dependent.
  - e.g.

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Source: [Angwin et al 2016]

  - Decisions violate **equalized odds** (equal TPR and FPR). Scores violate **separation**, $\mathbb{E}[R \mid Y] \neq \mathbb{E}[R \mid Y, A]$
- Scores are not calibrated to probabilities of actual outcomes.
  - e.g. “For the same credit score, one group is more likely to repay than another.”
  - Decisions violate **predictive value parity** $\mathbb{E}[Y \mid D] \neq \mathbb{E}[Y \mid D, A]$. Scores violate **calibration**, $R \neq \mathbb{E}[Y \mid R, A]$

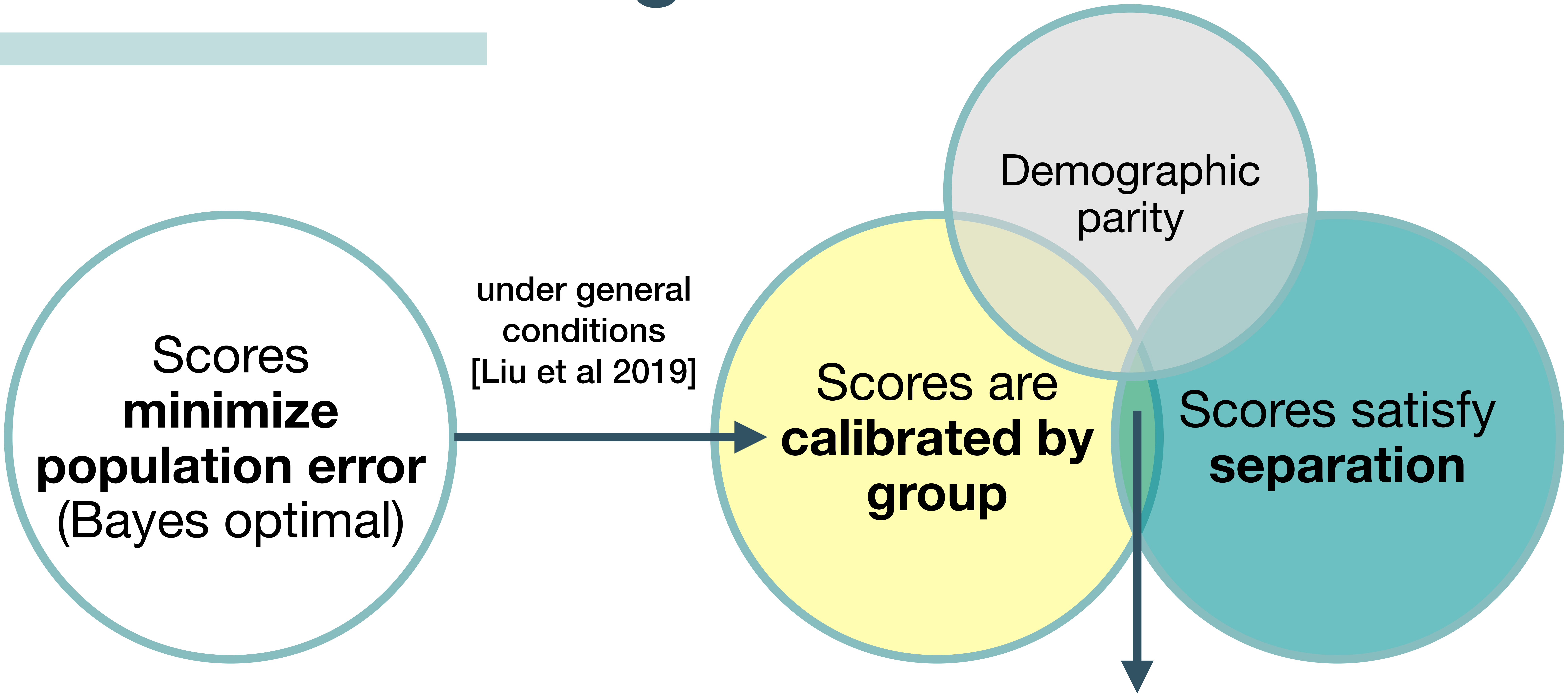
Reasonable disagreement on desiderata



- Error-free decisions $Y = D$
- Equal group base rates, $\mathbb{E}[Y] = \mathbb{E}[Y \mid A]$

[Chouldechova 2016]

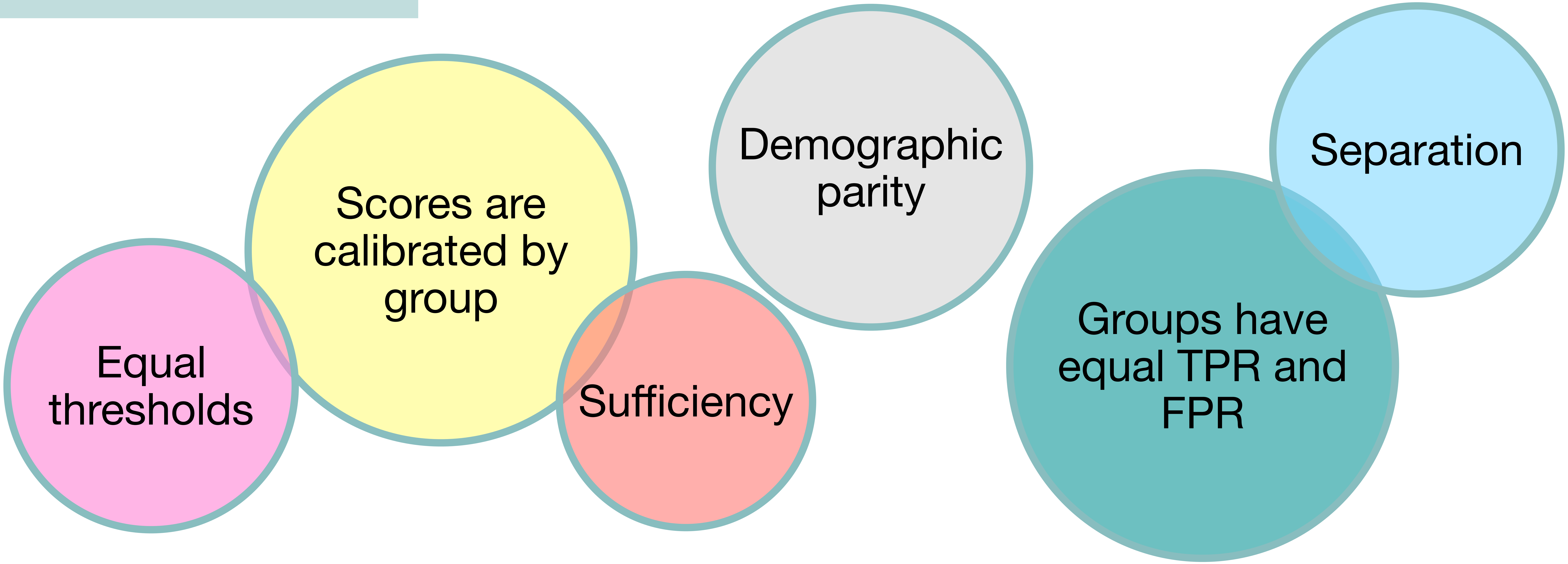
Reasonable disagreement on desiderata



- Error-free decisions $Y = D$
- Equal group base rates, $\mathbb{E}[Y] = \mathbb{E}[Y \mid A]$

[Chouldechova 2016; Kleinberg et al 2016]

From Algorithmic Bias…to Algorithmic Harm?

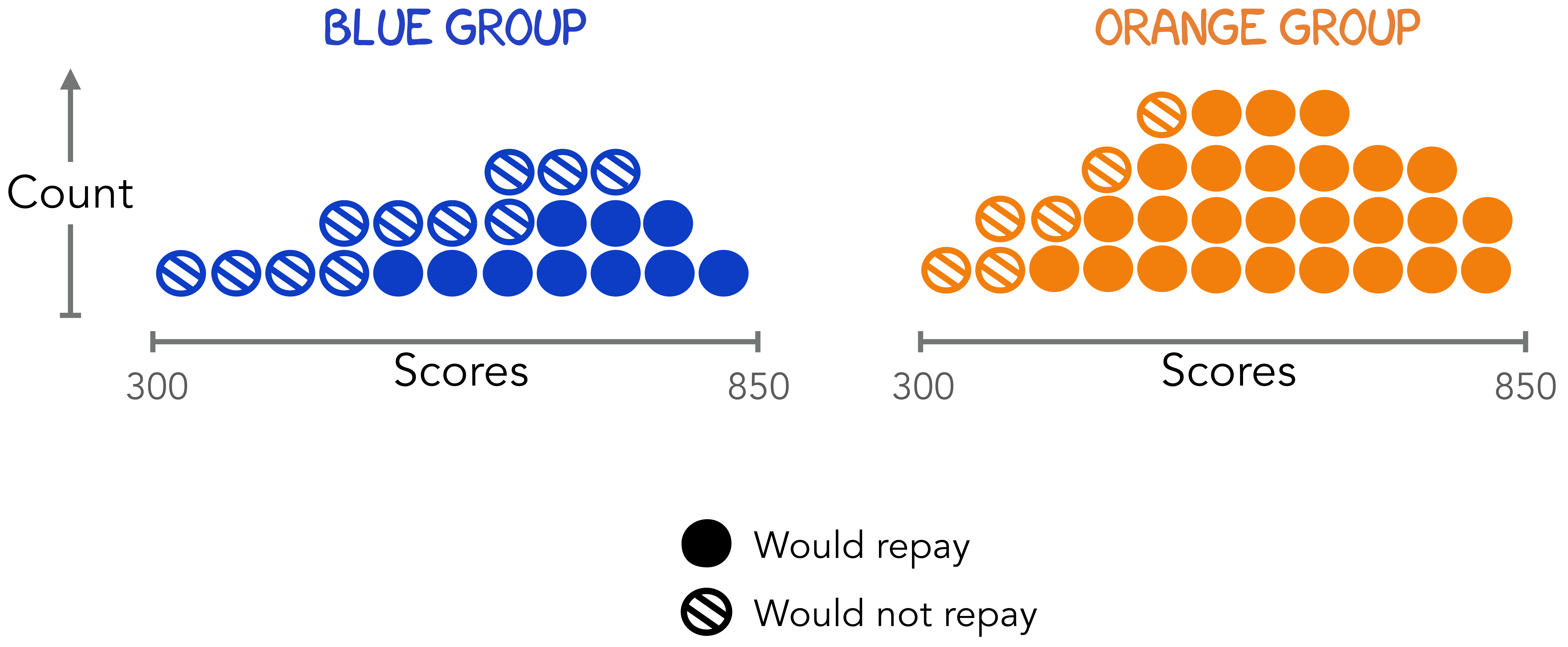


What are the downstream benefits or harms?

[Liu, Dean, Rolf, Simchowitz, Hardt. ICML 2018]

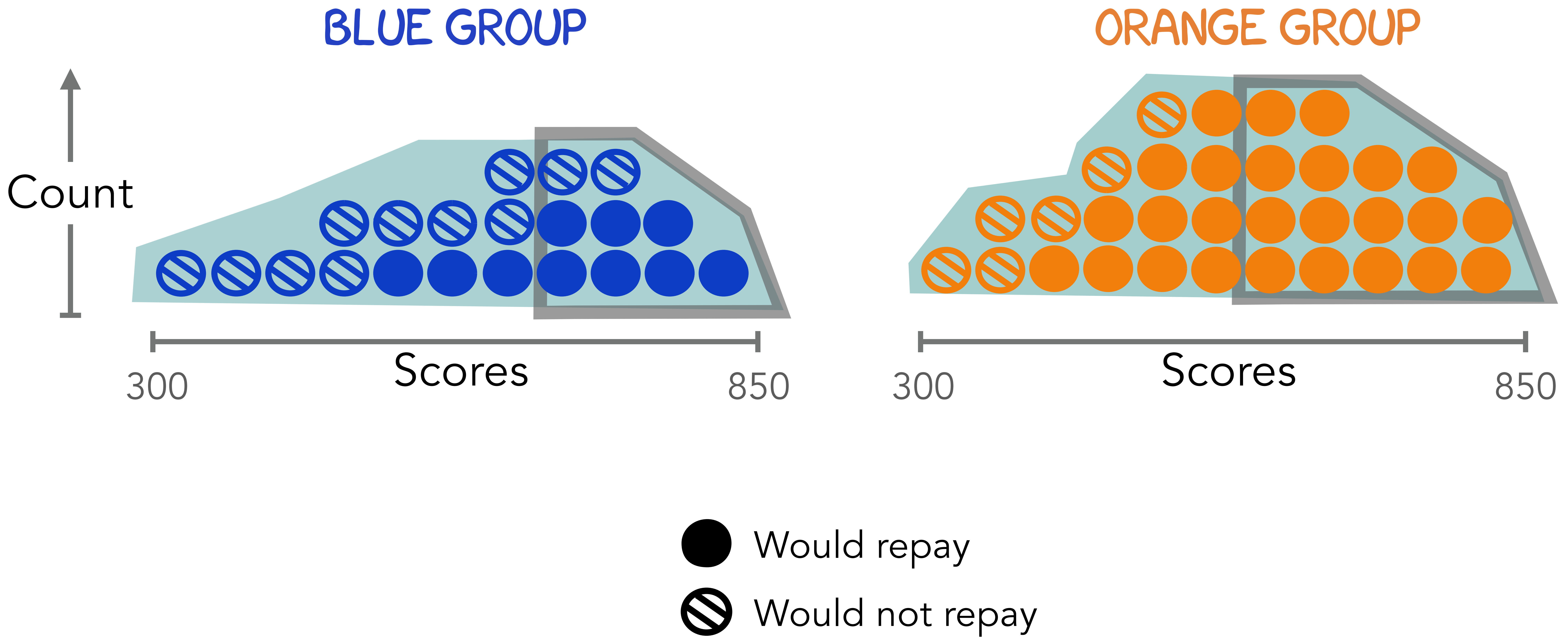
Example: Lending Decisions

Two groups with different *score* distributions (e.g. credit scores)



Algorithmic fairness equalizes loan approval rates.

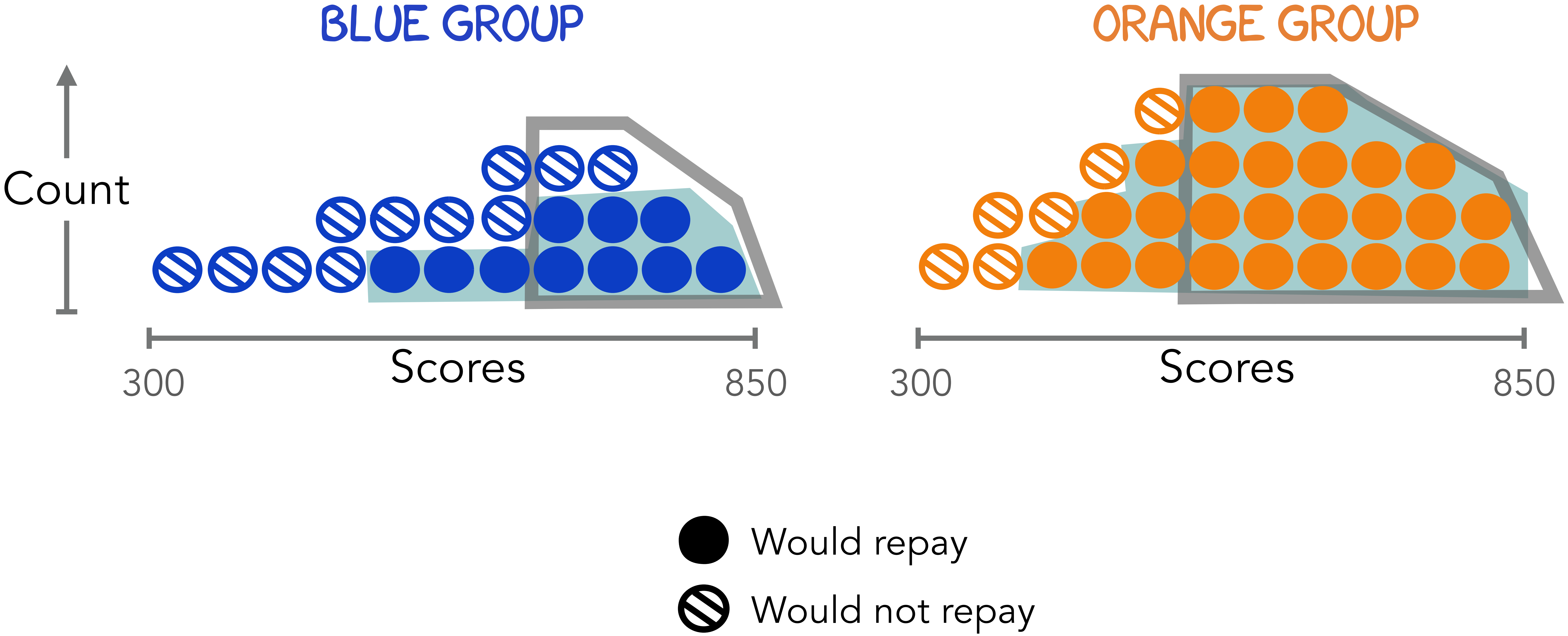
Demographic Parity [CKP09]: Same fraction of applicants accepted.



CONDITIONAL

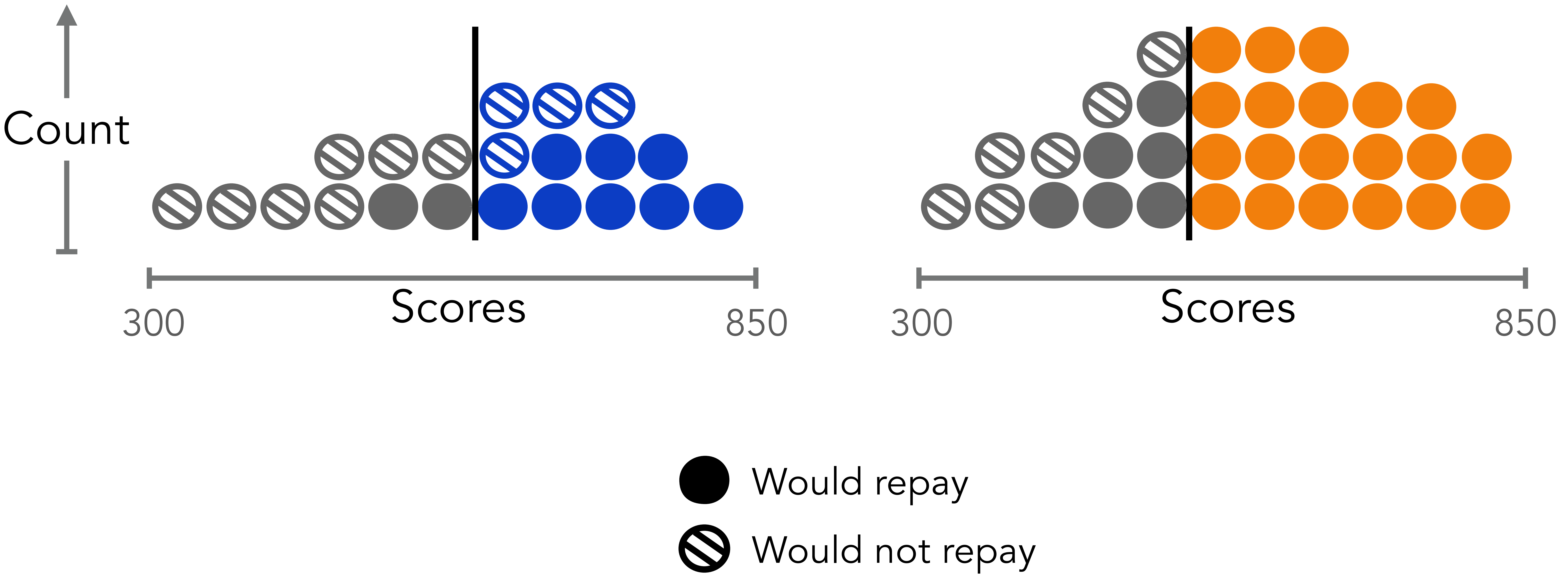
Algorithmic fairness equalizes loan approval rates.

Equal Opportunity [HPS2018]: Same fraction of repaying applicants accepted



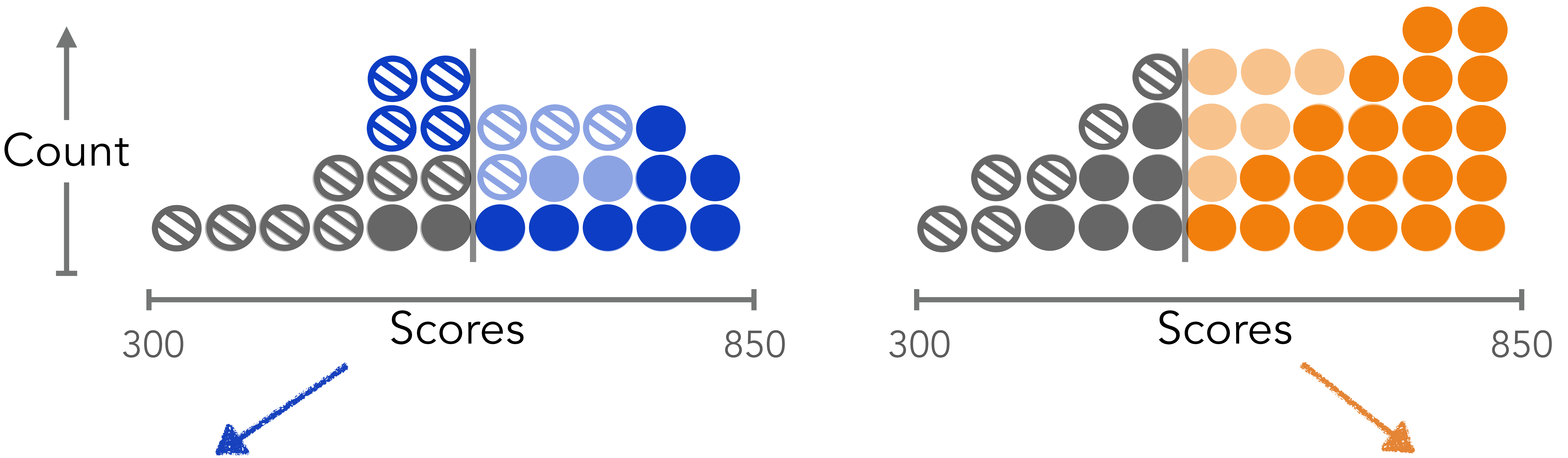
Lending Decisions

Policy: Accept applicants according to **DEMOGRAPHIC PARITY**.



Delayed Impact

Credit scores **change** with repayment (+) or default (-).



Scores got **worse** on average

Would repay

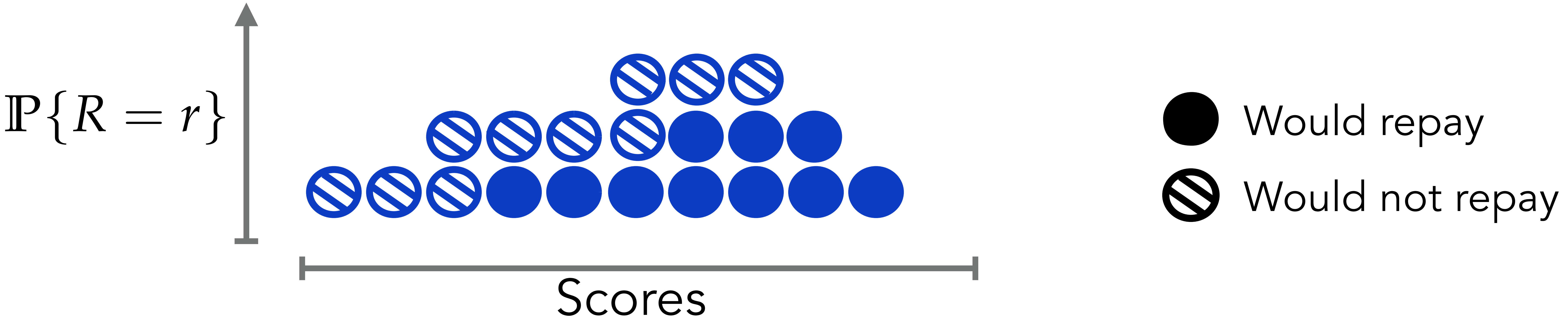
Would not repay

Scores got better on average

Average outcomes were more harmful for the lower-scoring group.

MODEL | SCORES

- A **score** $R(X)$ is a scalar random variable that is a function of an individual's features X
  - e.g. credit score is an integer from 300 to 850
- Any group of individuals has a particular **distribution** over scores:

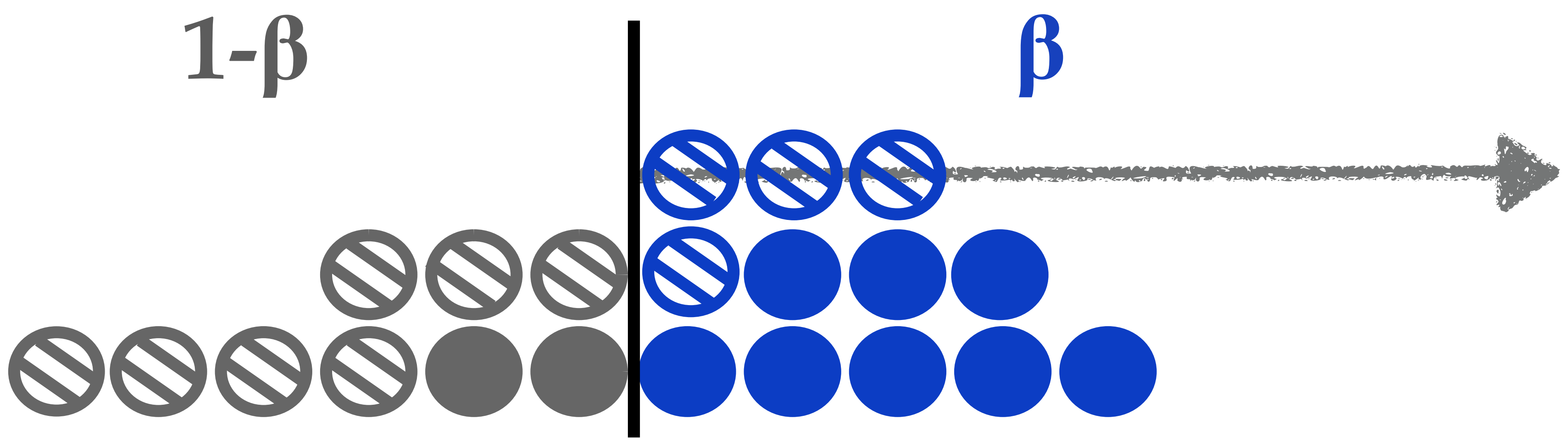


- Scores correspond to an individual's success probability (e.g. probability of repaying a loan) once accepted, $\rho(R)$, and are equally **calibrated** for each group.
- **<u>Monotonicity assumption for ρ</u>:** Higher scores implies **more likely** to repay.

MODEL | INSTITUTION CHOOSES ACCEPTANCE RATE

- Institution **accepts** individuals by choosing an **acceptance threshold** score T to maximize their expected **utility**:

$$\mathbb{E}[\text{utility}|T] = \mathbb{E}[\text{reward from repayments}|T] - \mathbb{E}[\text{loss from defaults}|T]$$



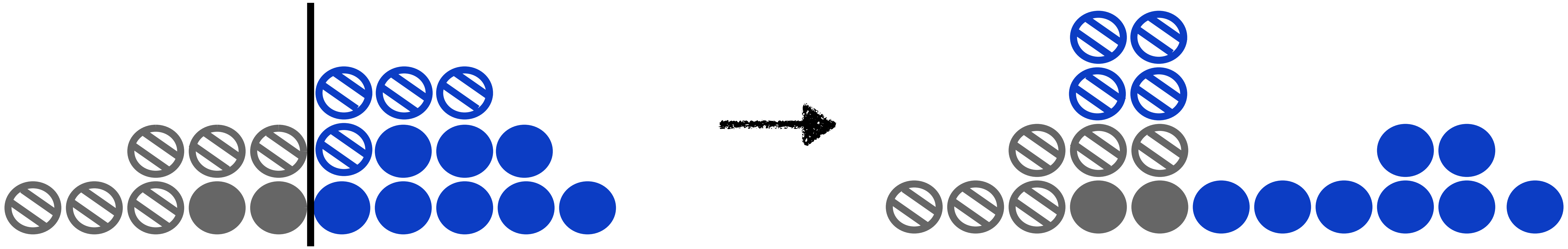
Threshold T corresponds to **acceptance rate β** for the group.

- When there are multiple groups, thresholds can be **group-dependent**.

MODEL | DELAYED IMPACT ON GROUPS

- Scores of accepted individuals change depending on their success.

$$R_{\text{new}} = \begin{cases} R_{\text{old}} + c_{+} & \text{if repaid} \\ R_{\text{old}} + c_{-} & \text{if defaulted} \end{cases}$$

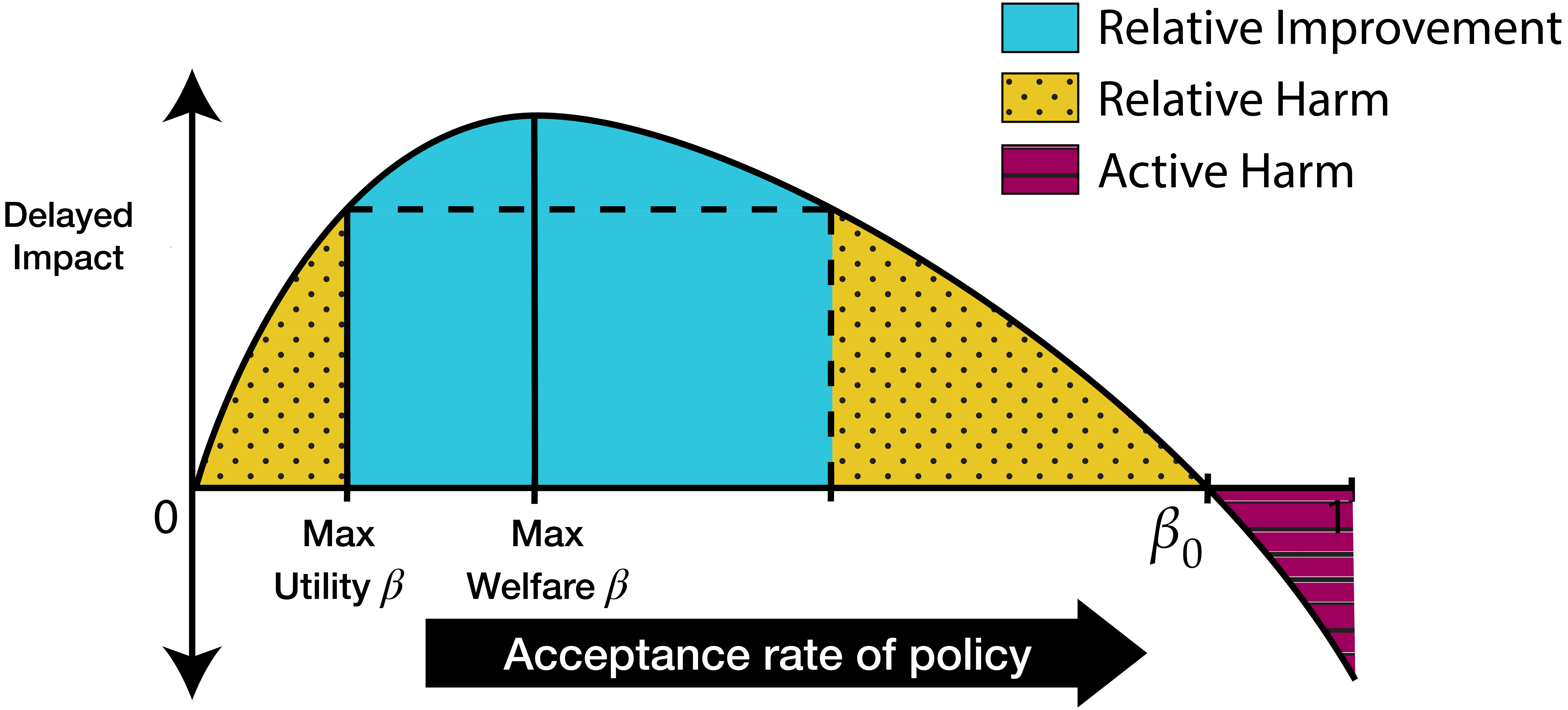


- The **average change in score** of each group is the **delayed impact**:

$$\Delta\mu = \mathbb{E}[R_{\text{new}} - R_{\text{old}}]$$

Outcome curve

Delayed impact is a **concave function** of **acceptance rate** β under mild assumptions.



Characterization of β under fairness constraint

- Assume two groups, A and B, with score quantile functions Q_A, Q_B, and population proportions g_A, g_B.
- The institution's expected utility as a function of score is $u(r)$.
- **Theorem 1 (Informal).** Under Demographic Parity, the acceptance rate β is completely determined by Q_A, Q_B, g_A, g_B, and u:

$$g_A u(Q_A(\beta)) + g_B(Q_B(\beta)) = 0$$

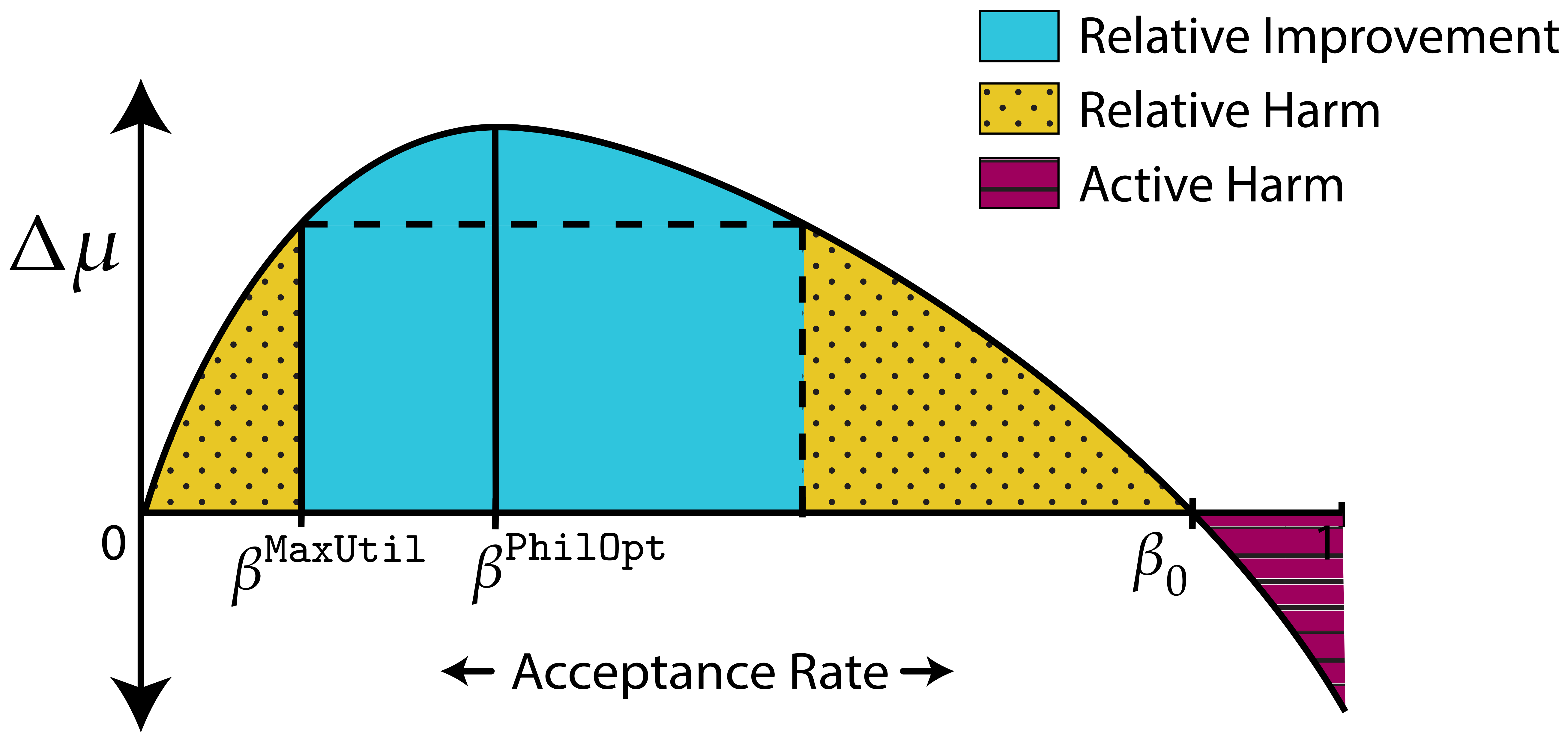
- **Theorem 2 (Informal).** Under Equal Opportunity, the acceptance rates β_A, β_B are completely determined by Q_A, Q_B, g_A, g_B, u, and ρ.

FAIRNESS CONSTRAINTS DO NOT GUARANTEE LONG-TERM WELFARE.

Corollary 1 [All outcome regimes are possible]

Equal opportunity and demographic parity may cause relative improvement, relative harm, or active harm.

- unconstrained utility maximization never causes active harm.



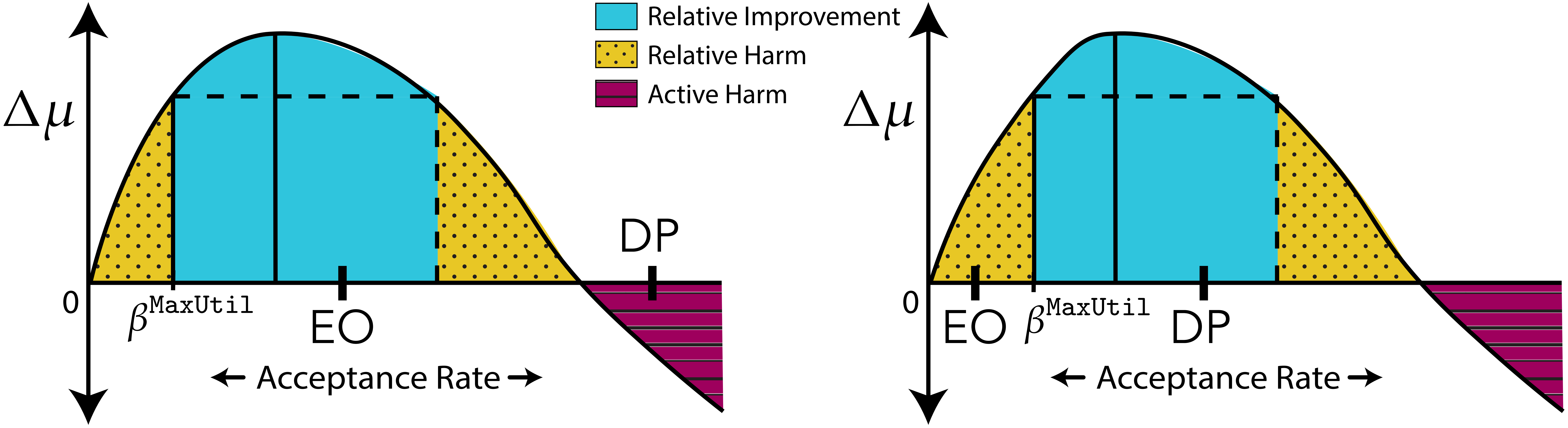
CHOICE OF FAIRNESS CRITERIA MATTERS.

Corollary 2

Demographic parity (DP) may cause **active** or **relative harm** by over-acceptance; equal opportunity (EO) does not.

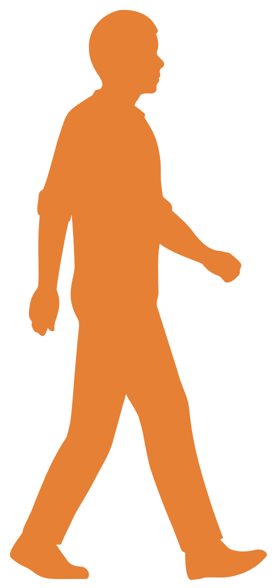
Corollary 3

Equal opportunity (EO) may cause **relative harm** by under-acceptance; demographic parity never under-accepts

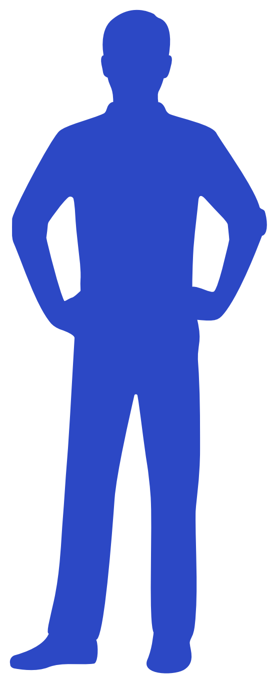


CALIBRATION ERRORS FOR ONE GROUP

- Suppose the bank systematically **underestimates** the repayment ability of the disadvantaged group



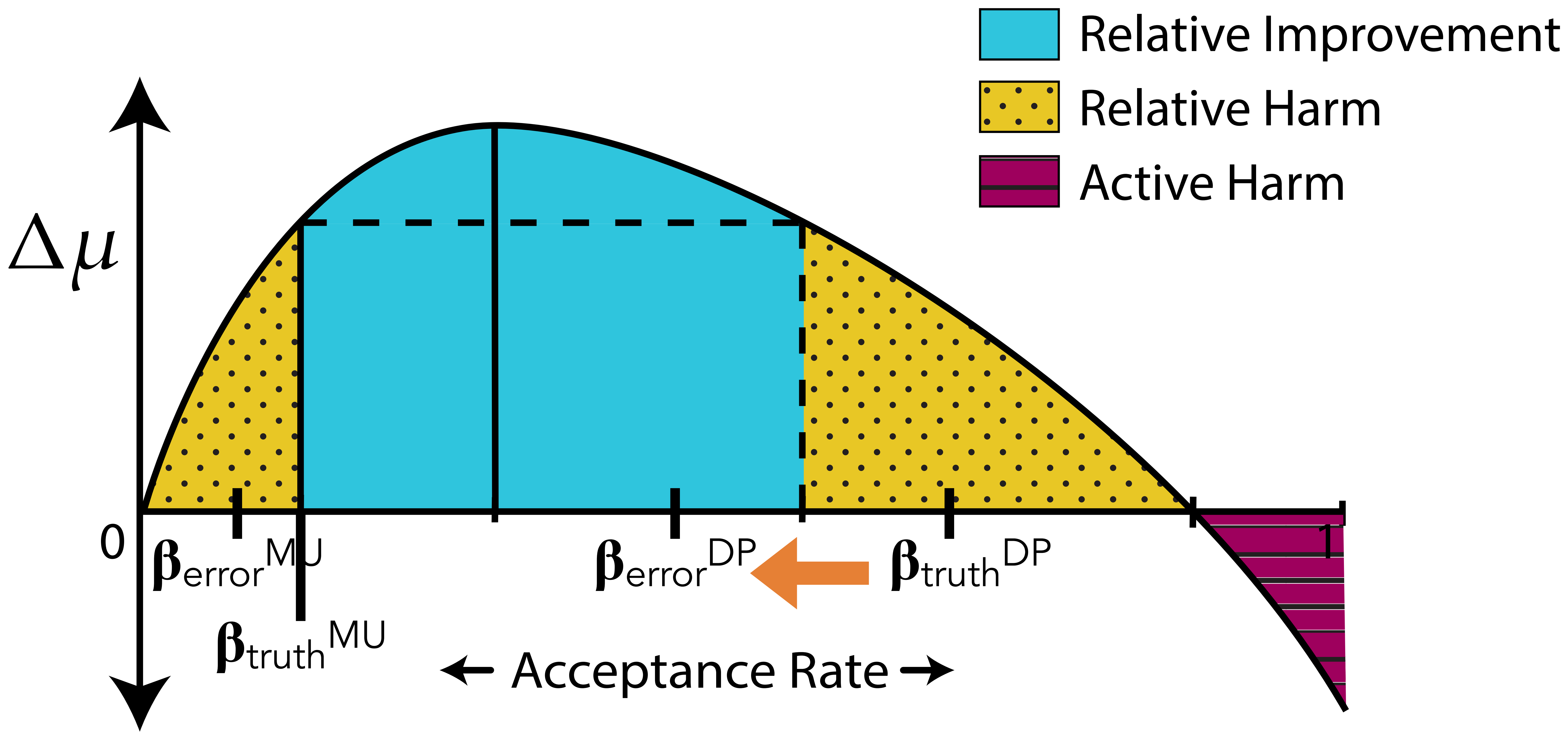
- orange group
- **0.8** probability of repaying loan
- assigned credit score of **700**



- blue group
- **0.8** probability of repaying loan
- but assigned credit score of **600** (*underestimated*)

UNDERESTIMATION CAUSES UNDERACCEPTANCE

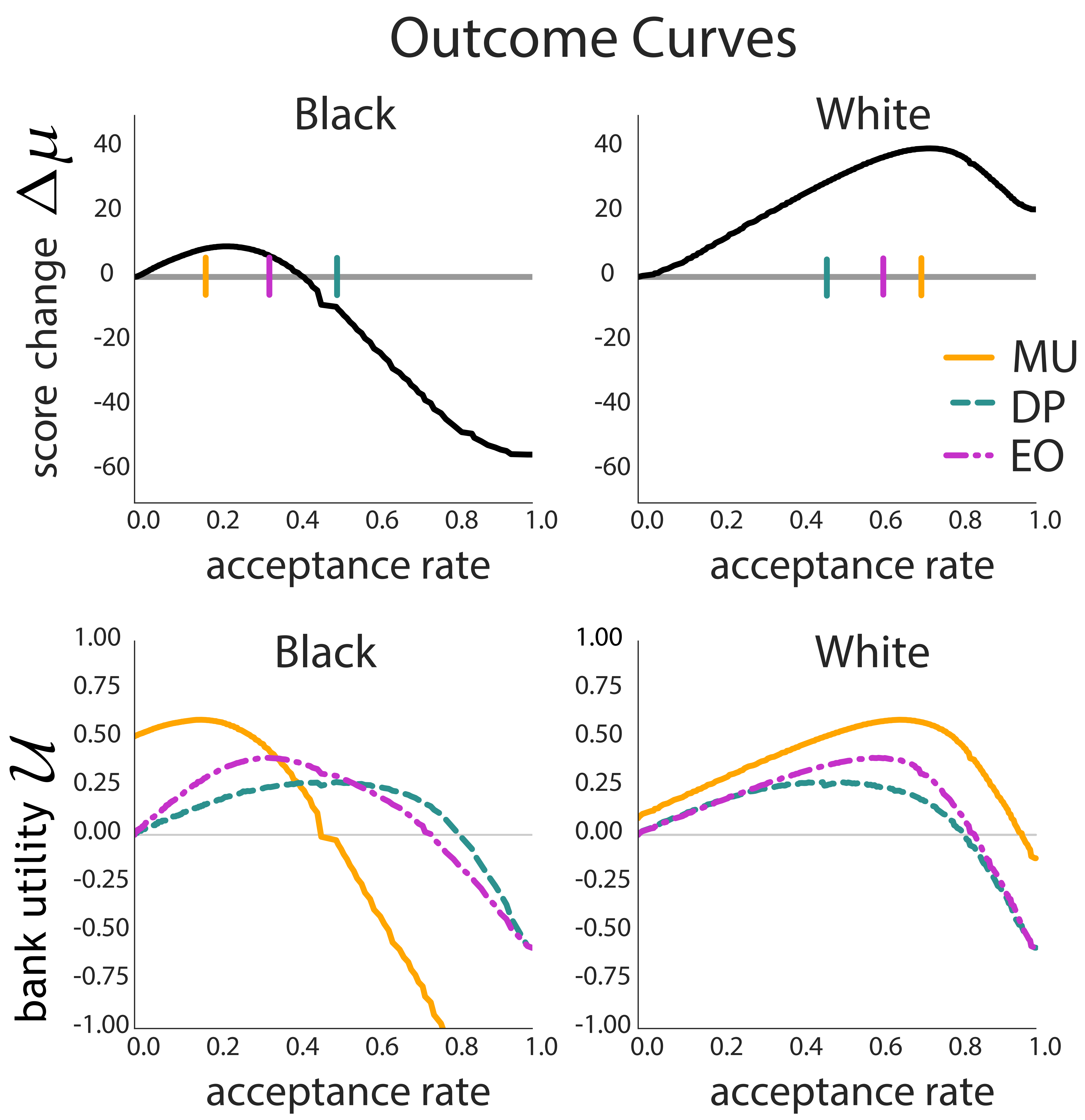
- *Corollary 4*: **Acceptance rate for group is lower** if their scores are systematically underestimated than when their scores reflect true probability of repayment.
- This holds for **unconstrained utility maximization, demographic parity,** as well as **equal opportunity*.**
- Example:
  If there's calibration error, **demographic parity** yields more favorable delayed impact by promoting a higher acceptance rate.



*under an additional condition (true TPR dominates estimated TPR).

ESTIMATING DELAYED IMPACT WITH FICO CREDIT SCORES

- 300,000+ TransUnion TransRisk scores from 2003
- Use data labeled by race to **estimate** group score distributions, repayment probabilities, and proportion.
- **Plug in** bank's profit/loss ratio, e.g. +1:-4, and the impact of repayment/default on credit score, e.g. +75/-150



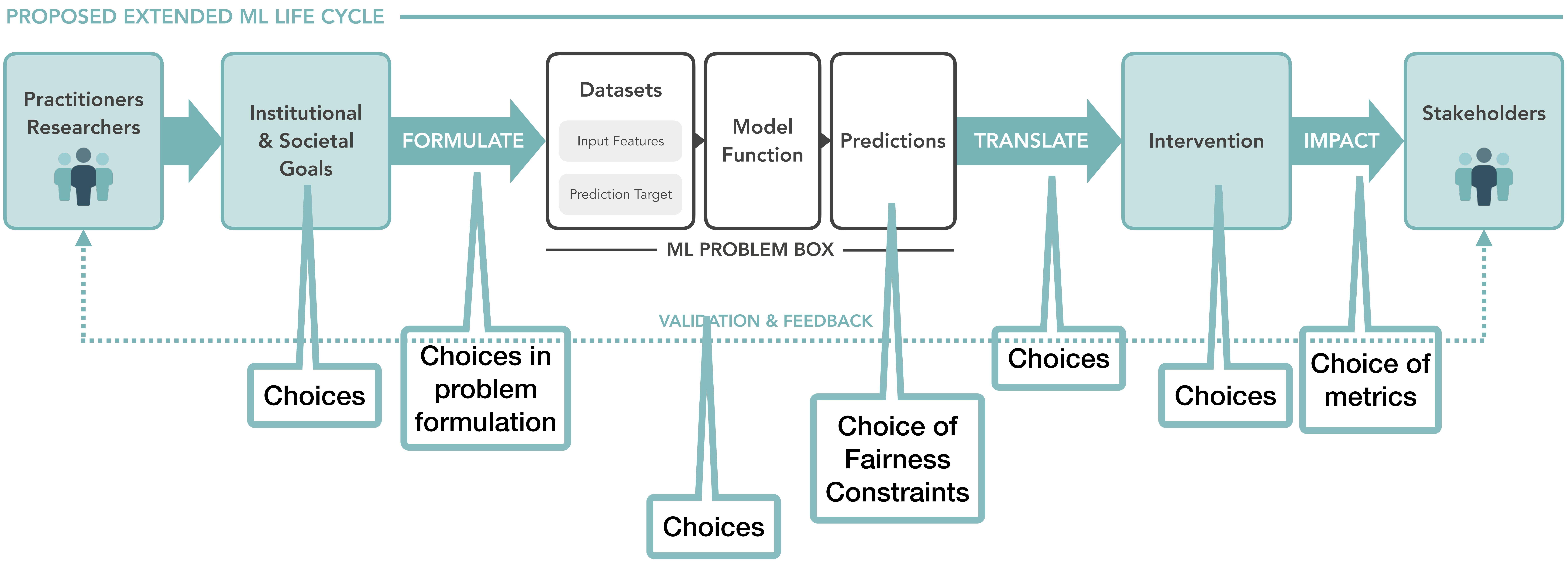
Selected related work and impact

* prior or contemporaneous work

- **Growing research area:** Long Term Dynamics and Societal Impact of Algorithmic Decisions
  - Feedback loops and populations [Ensign et al 2017*; Hashimoto et al 2018*; Mouzannar et al 2019; **Liu** et al 2020]
  - Fairness in pipelines [Hu and Chen 2017*; Kannan et al 2019; Arunachaleswaran et al 2020; Dwork et al 2020]
  - Fairness in recommendation systems [Morik et al 2020; Ge et al 2021]
  - Delayed Impact of Causal Fairness notions [Nilforoshan et al 2022]
- **Practical Impact:** Simulation toolkits for anticipating real world impact of ML systems
  - ML Fairness Gym [D'Amour et al 2020], Importance for industry practitioners [Holstein et al 2019]
  - Fairkit-learn [Johnson et al 2020]
- **Broader impact on AI ethics and normative discourse**
  - Non-ideal theory of algorithmic fairness, broader assessments [Fazelpour and Lipton 2020; Lee et al 2021]

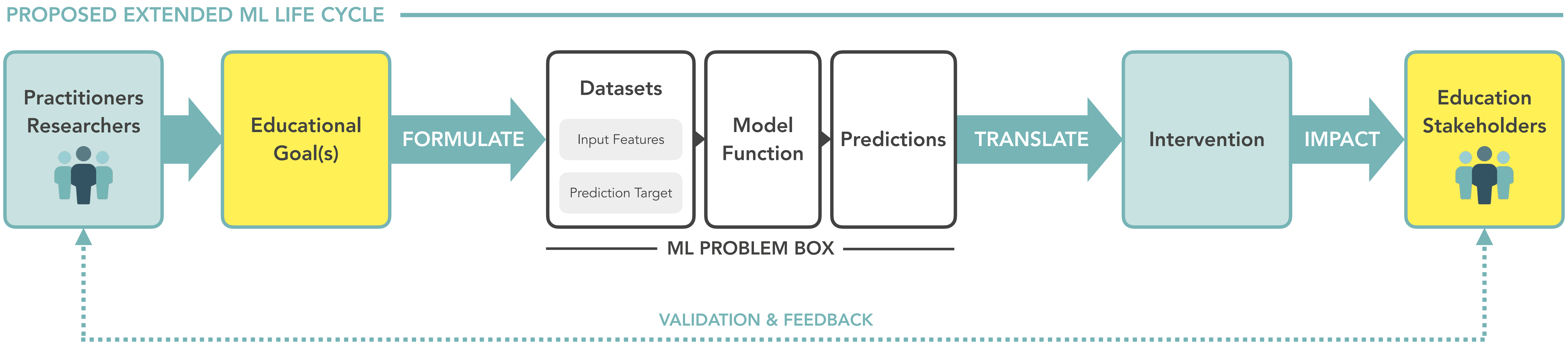
"Delayed Impact" in Practice

Improving downstream outcomes of ML and algorithmic decision making in consequential domains



"Delayed Impact" in Practice

Improving downstream outcomes of ML and algorithmic decision making in consequential domains



"Reimagining the Machine Learning Life Cycle to Improve Educational Outcomes of Students"

L. T. L., Serena Wang, Tolani Britton, Rediet Abebe. PNAS (forthcoming). 2022.

Thank you!

lydiatliu.com | lydiatliu (at) cornell (dot) edu