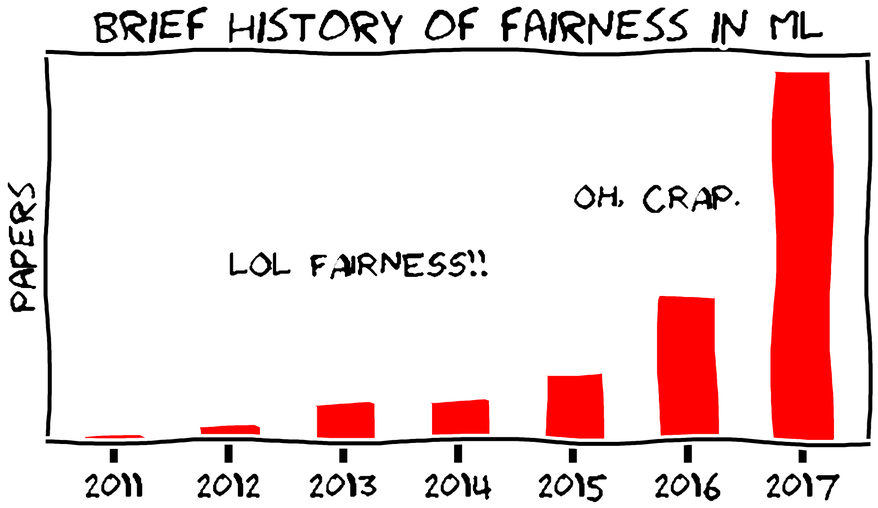
DELAYED IMPACT OF FAIR MACHINE LEARNING

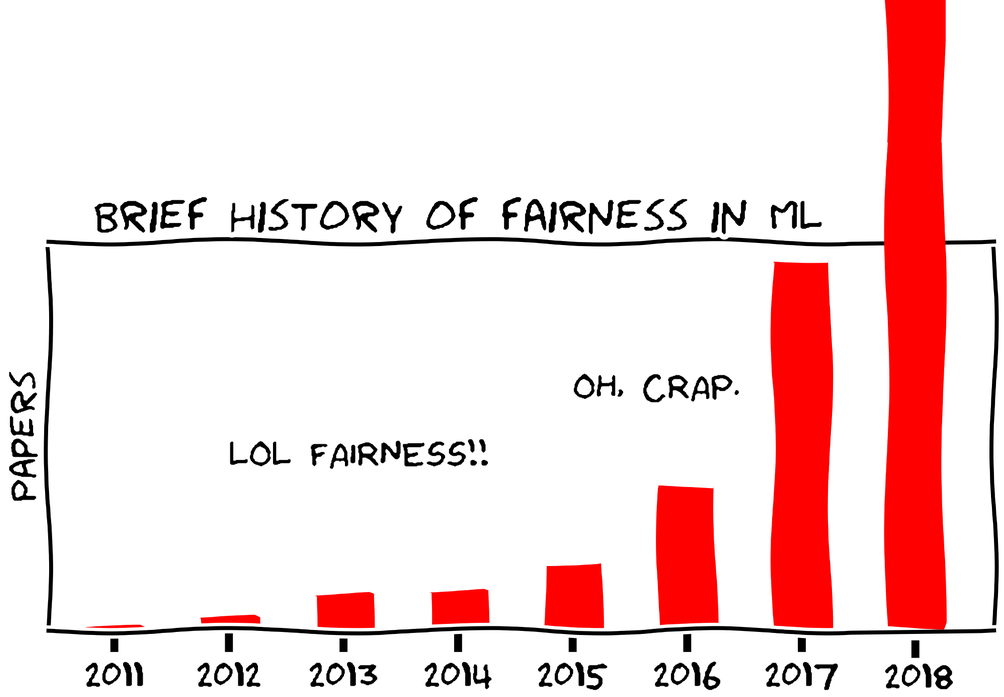
Lydia T. Liu (UC Berkeley)



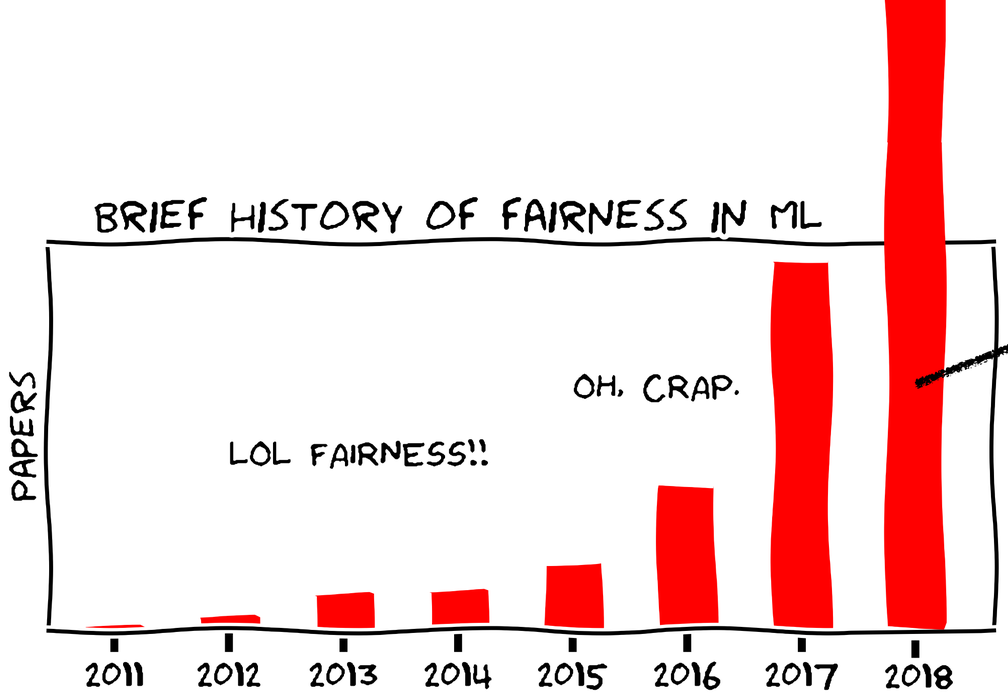

Joint work with **Sarah Dean, Esther Rolf, Max Simchowitz, Moritz Hardt**



source: mrtz.org



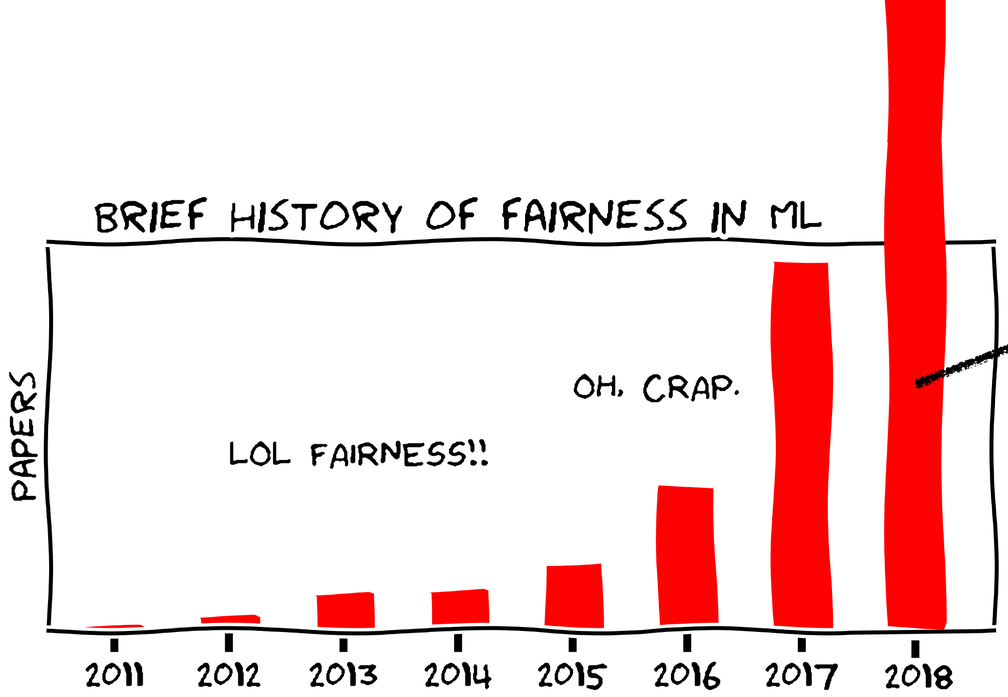
source: mrtz.org



source: mrtz.org

"21 DEFINITIONS OF FAIRNESS" [Narayanan 2018]

1. DEMOGRAPHIC PARITY
2. EQUALITY OF OPPORTUNITY
3. PREDICTIVE VALUE PARITY
4. GROUP CALIBRATION



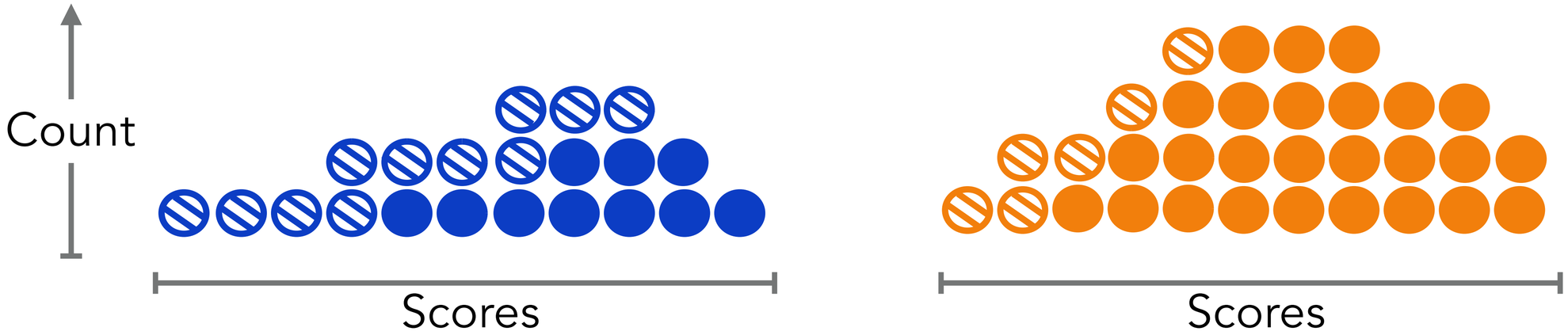
source: mrtz.org

"21 DEFINITIONS OF FAIRNESS" [Narayanan 2018]

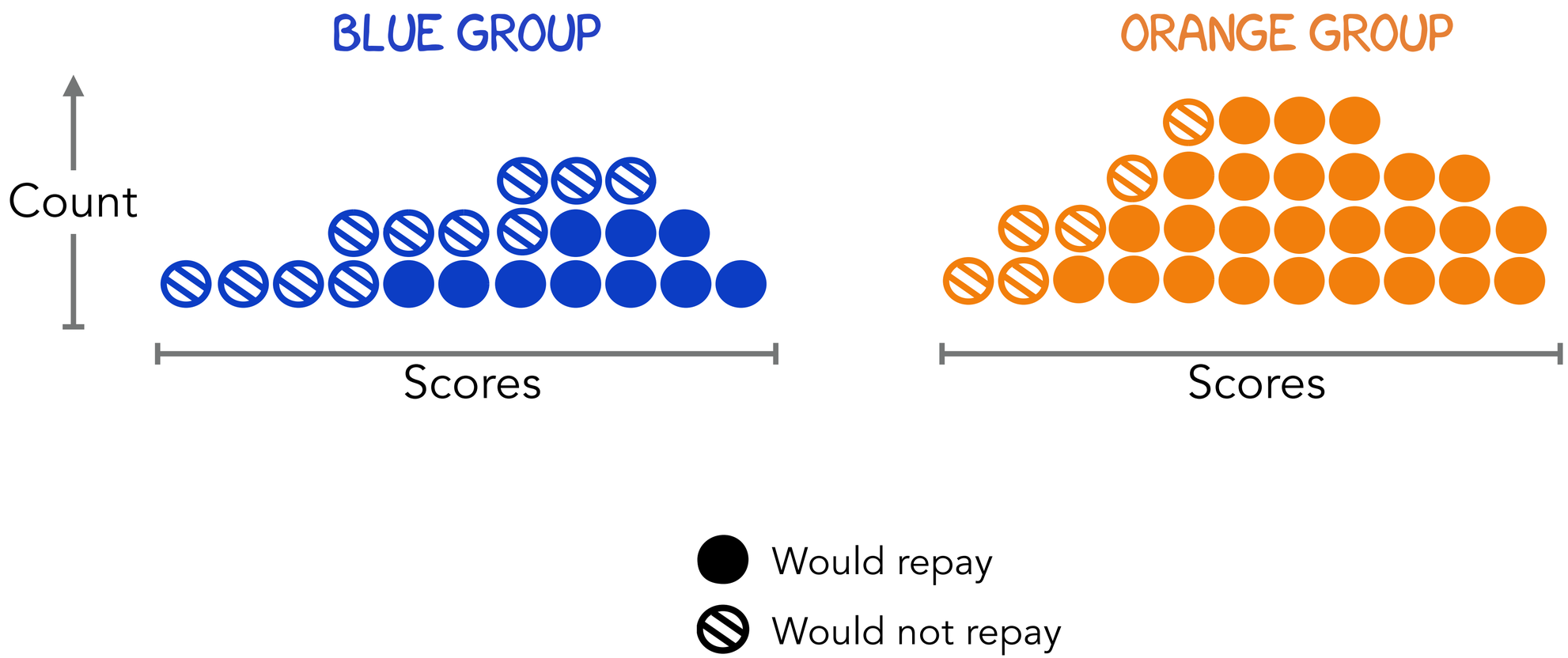
1. DEMOGRAPHIC PARITY
2. EQUALITY OF OPPORTUNITY
3. PREDICTIVE VALUE PARITY
4. GROUP CALIBRATION

Machine learning systems are "fair" ? Protected groups are "better off"

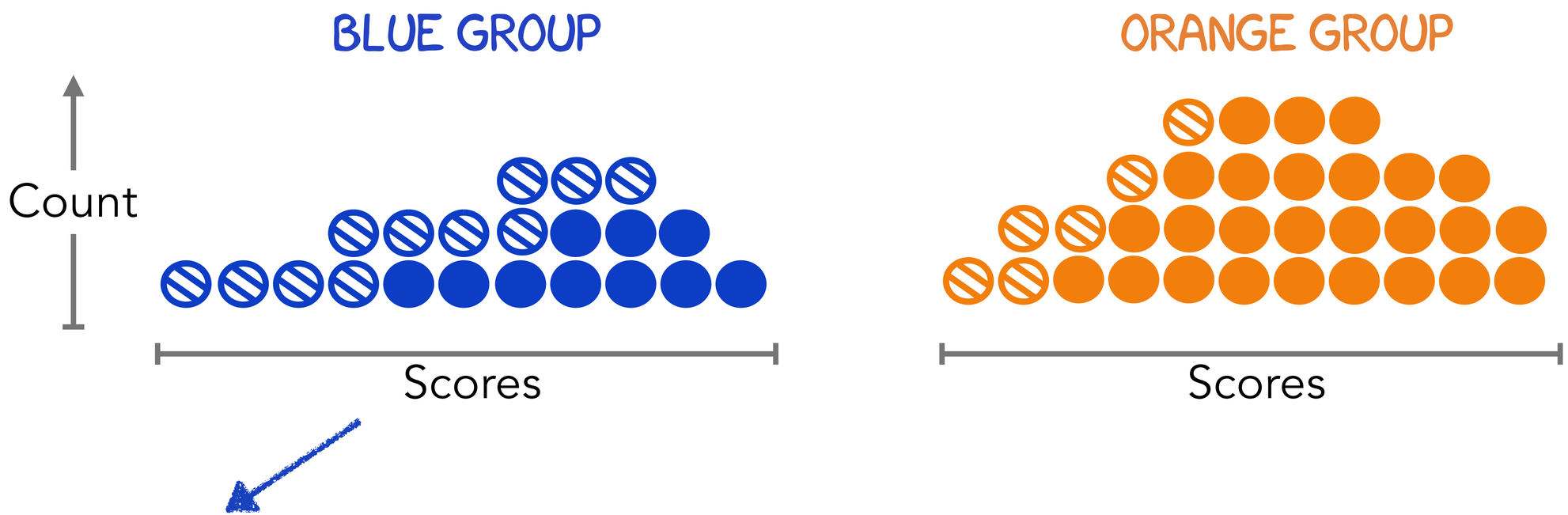
Two groups with different *score* distributions (e.g. credit scores)



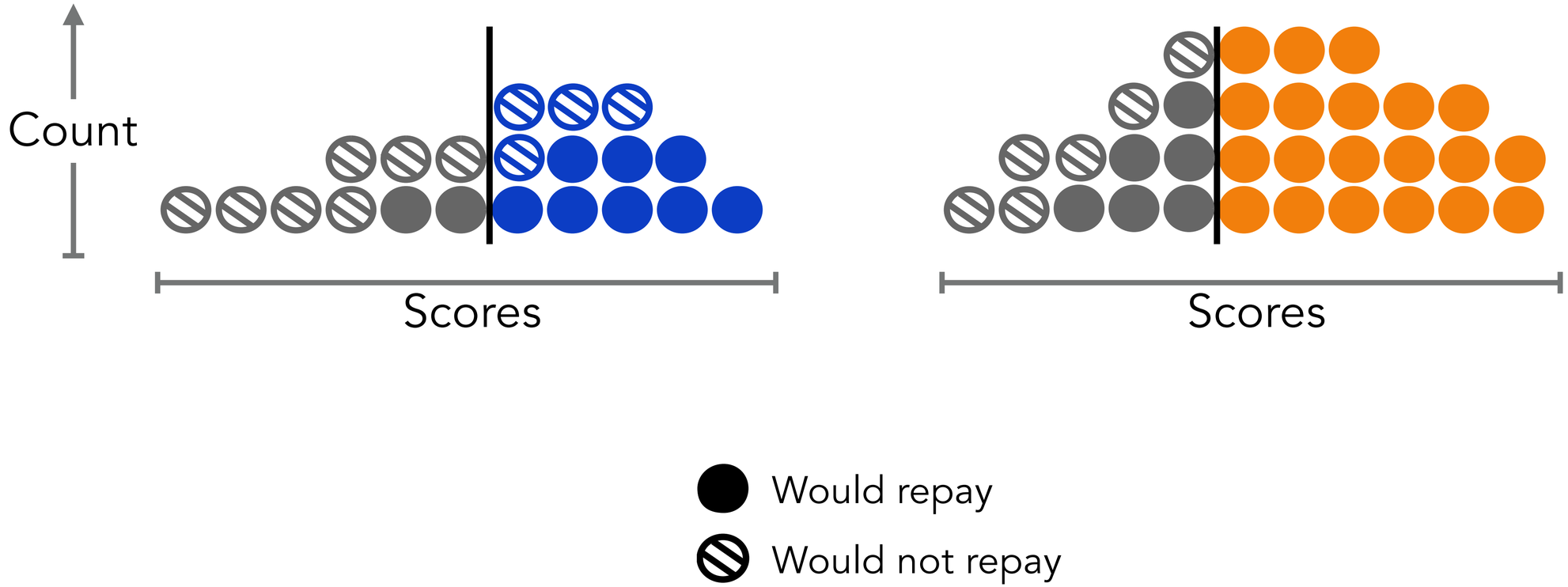
Two groups with different *score* distributions (e.g. credit scores)



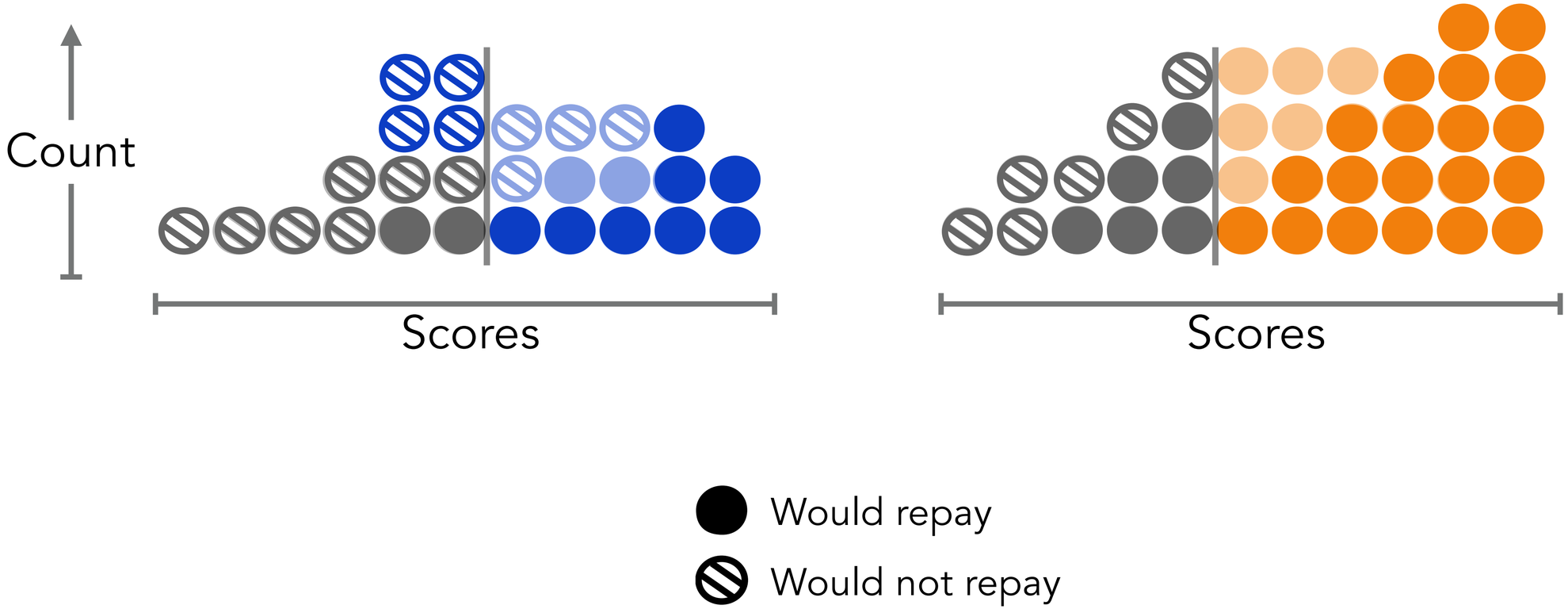
Two groups with different *score* distributions (e.g. credit scores)



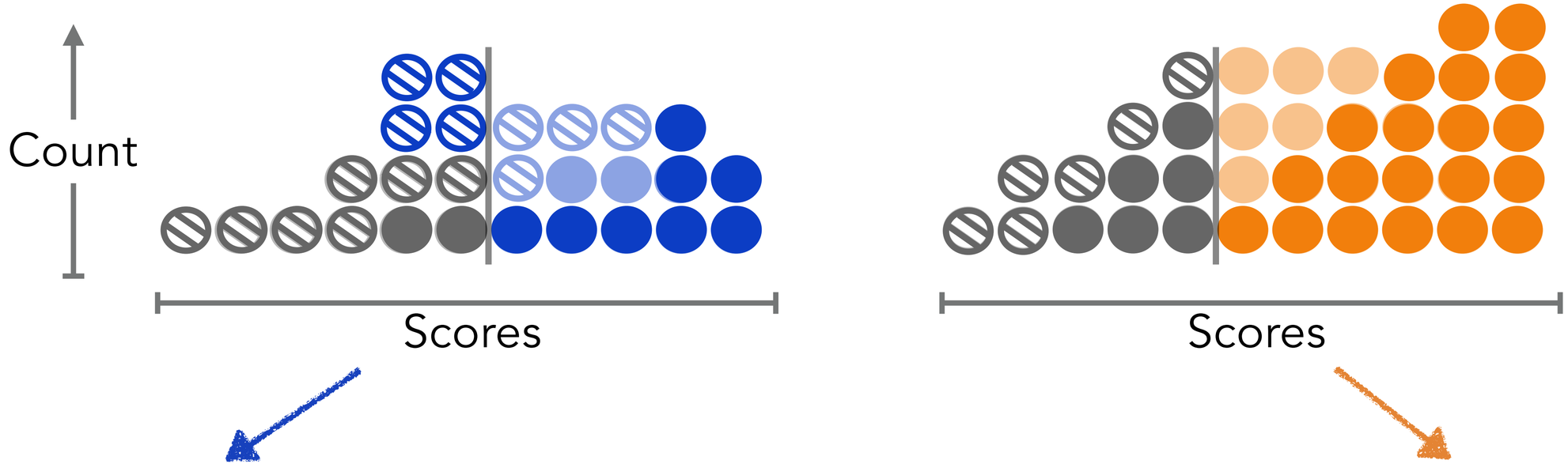
Approve loans according to **DEMOGRAPHIC PARITY**.



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



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



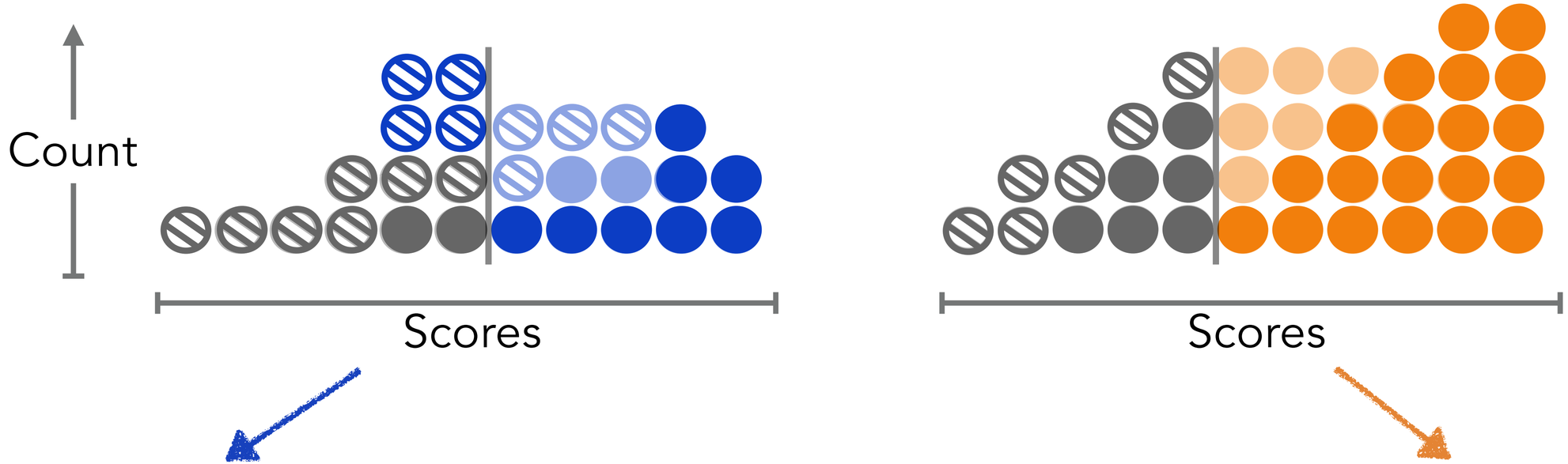
Scores got worse on average

Would repay

Would not repay

Scores got better on average

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



Scores got worse on average

HARM!

Would repay

Would not repay

Scores got better on average

WHAT HAPPENED?

WHAT HAPPENED?

Fairness criteria didn't seem to *help* the protected group,

once we considered the *impact* of loans on scores.

OUR WORK

OUR WORK

1. Introduce the "**outcome curve**", a tool for comparing the delayed impact of fairness criteria

OUR WORK

1. Introduce the "**outcome curve**", a tool for comparing the delayed impact of fairness criteria
2. Provide a **complete characterization** of the delayed impact of 3 different fairness criteria

OUR WORK

1. Introduce the "**outcome curve**", a tool for comparing the delayed impact of fairness criteria
2. Provide a **complete characterization** of the delayed impact of 3 different fairness criteria
3. Show that fairness constraints **may cause harm** to groups they intended to protect

SCORE DISTRIBUTION WITHIN A GROUP

SCORE DISTRIBUTION WITHIN A GROUP

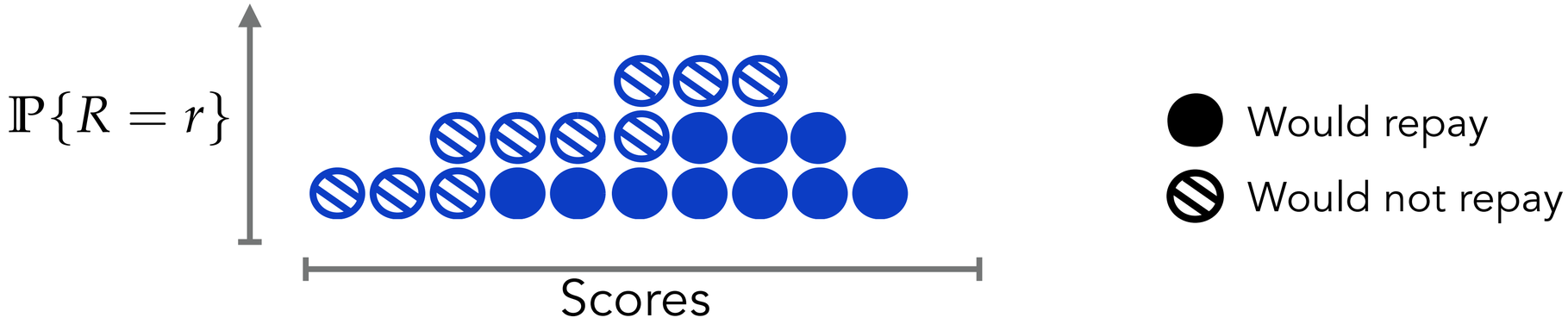
- A **score** $R(X)$ is a scalar random variable that is a function of an individual's features X
  - e.g. credit score is an integer from 300 to 850

SCORE DISTRIBUTION WITHIN A GROUP

- 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:

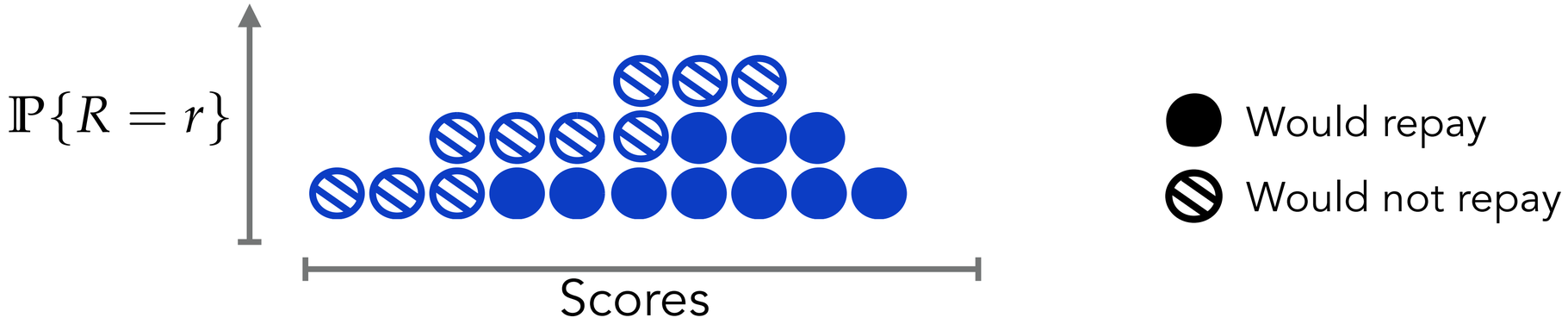
SCORE DISTRIBUTION WITHIN A GROUP

- 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:



SCORE DISTRIBUTION WITHIN A GROUP

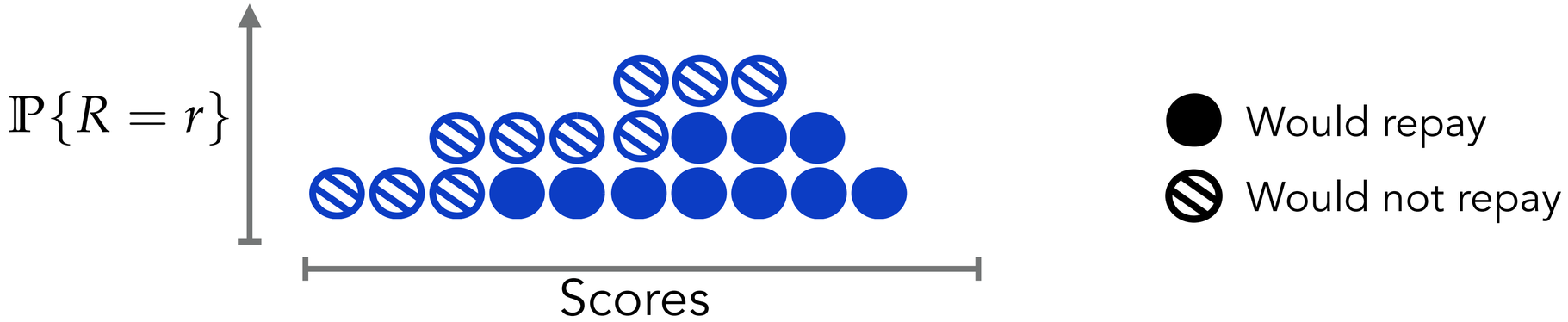
- A **score** $R(X)$ is a scalar random variable that is a function of an individual's features X
  - e.g. credit score is an integer from 300 to 850
- Any group of individuals has a particular **distribution** over scores:



- Each score corresponds to an individual's success probability (e.g. probability of repaying a loan) once accepted

SCORE DISTRIBUTION WITHIN A GROUP

- 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:



- Each score corresponds to an individual's success probability (e.g. probability of repaying a loan) once accepted
- **Monotonicity assumption:** Higher scores implies **more likely** to repay

CLASSIFIERS ARE THRESHOLDS

CLASSIFIERS ARE THRESHOLDS

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

CLASSIFIERS ARE THRESHOLDS

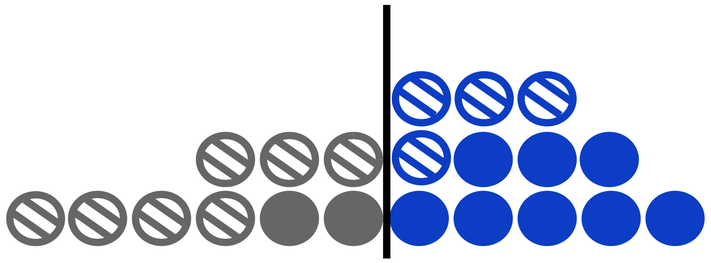
- Institution **classifies** individuals by choosing an **acceptance threshold** score T to maximize their expected **utility**:

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

CLASSIFIERS ARE THRESHOLDS

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

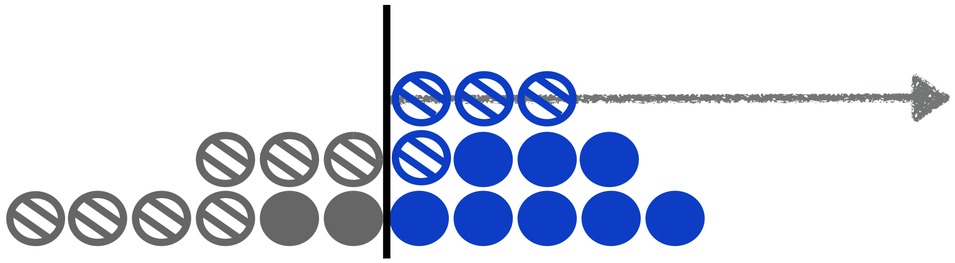
$$\mathbb{E}[\text{utility}|T] = \mathbb{E}[\text{reward from repayments}|T] - \mathbb{E}[\text{loss from defaults}|T]$$



CLASSIFIERS ARE THRESHOLDS

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

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

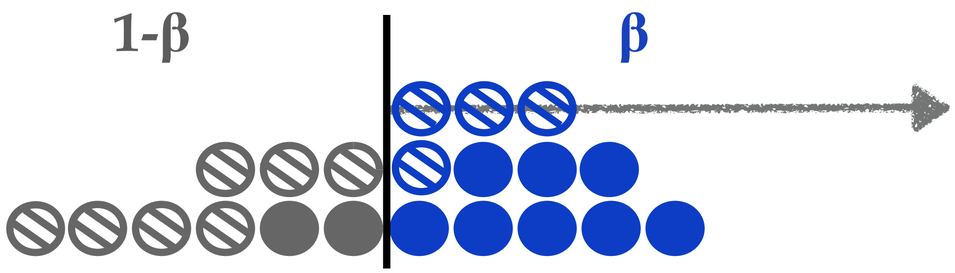


Threshold T corresponds to **acceptance rate β** for the group.

CLASSIFIERS ARE THRESHOLDS

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

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

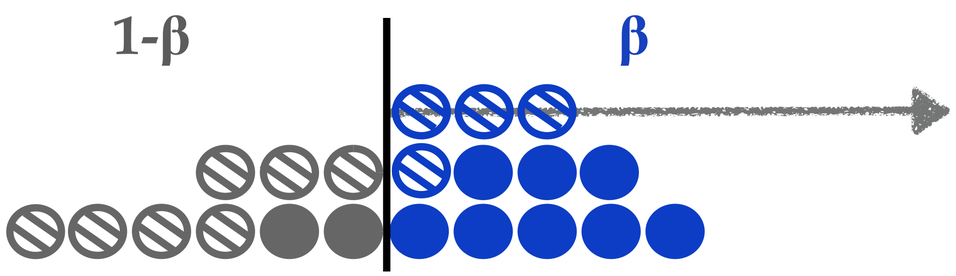


Threshold T corresponds to **acceptance rate β** for the group.

CLASSIFIERS ARE THRESHOLDS

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

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



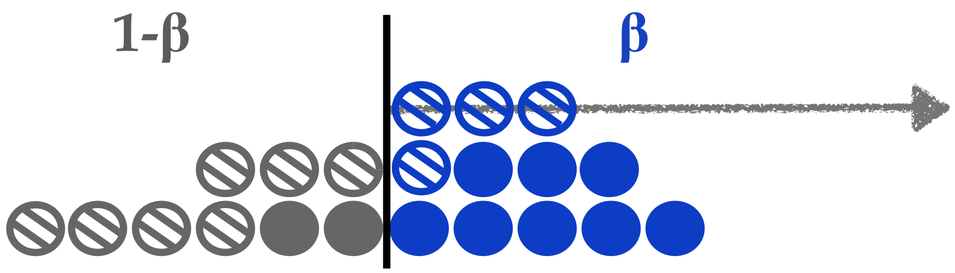
Threshold T corresponds to **acceptance rate β** for the group.

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

CLASSIFIERS ARE THRESHOLDS

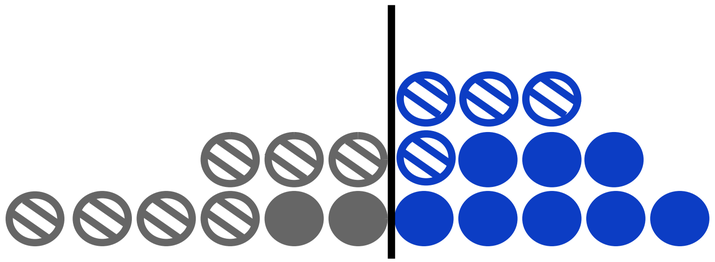
- Institution **classifies** individuals by choosing an **acceptance threshold** score T to maximize their expected **utility**:

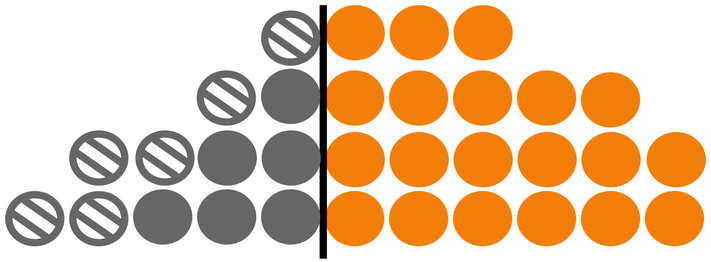
$$\mathbb{E}[\text{utility}|T] = \mathbb{E}[\text{reward from repayments}|T] - \mathbb{E}[\text{loss from defaults}|T]$$



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

MODEL DELAYED IMPACT ON GROUPS

- Scores of accepted individuals change depending on their success.

MODEL DELAYED IMPACT ON GROUPS

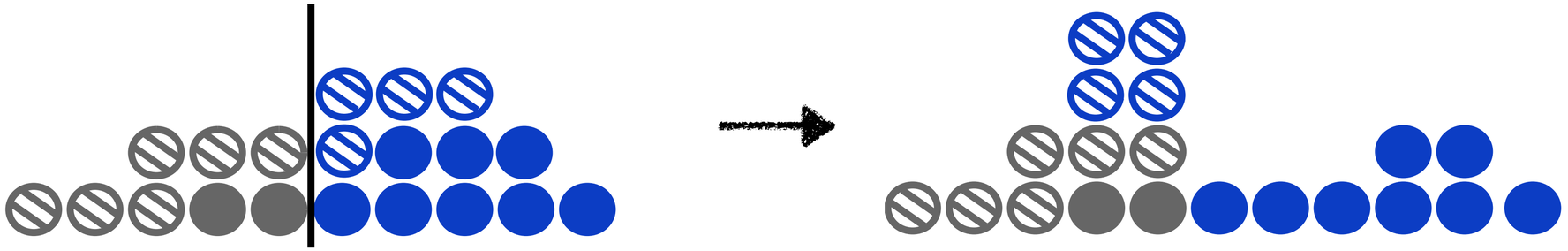
- Scores of accepted individuals change depending on their success.

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

MODEL DELAYED IMPACT ON GROUPS

- Scores of accepted individuals change depending on their success.

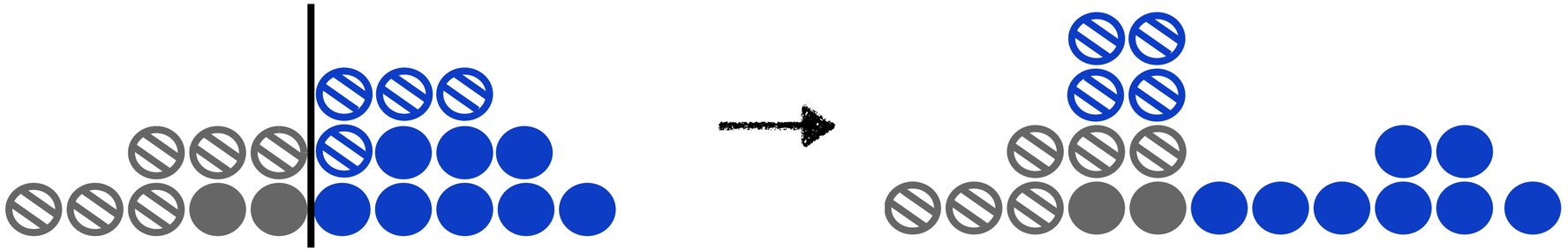
$$R_{\text{new}} = \begin{cases} R_{\text{old}} + c_+ & \text{if repaid} \\ R_{\text{old}} + c_- & \text{if defaulted} \end{cases}$$



MODEL DELAYED IMPACT ON GROUPS

- Scores of accepted individuals change depending on their success.

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

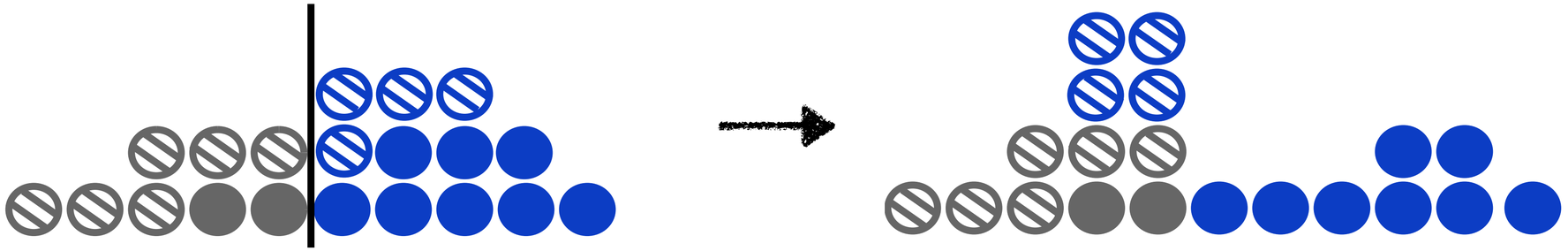


- The average change in score of each group is the **delayed impact**:

MODEL DELAYED IMPACT ON GROUPS

- Scores of accepted individuals change depending on their success.

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



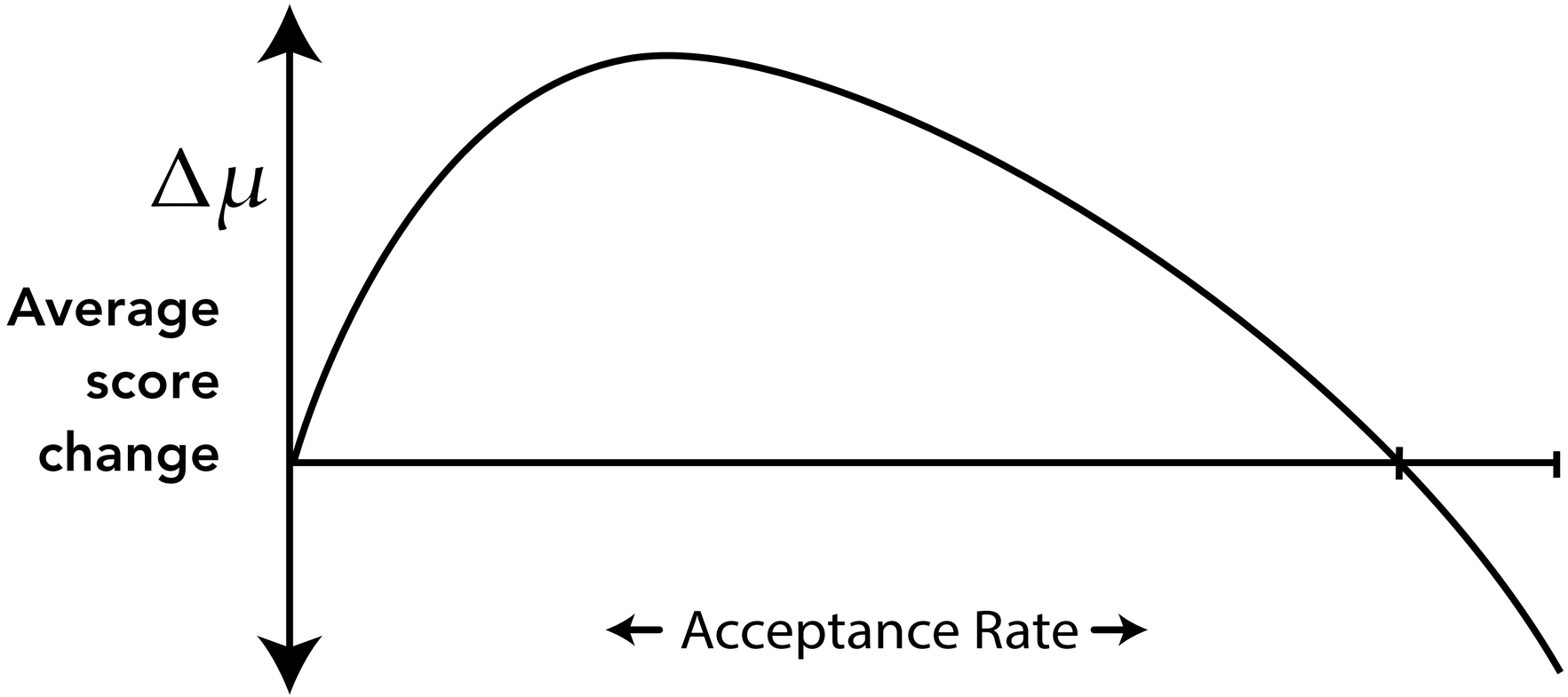
- The average change in score of each group is the **delayed impact**:

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

OUTCOME CURVE

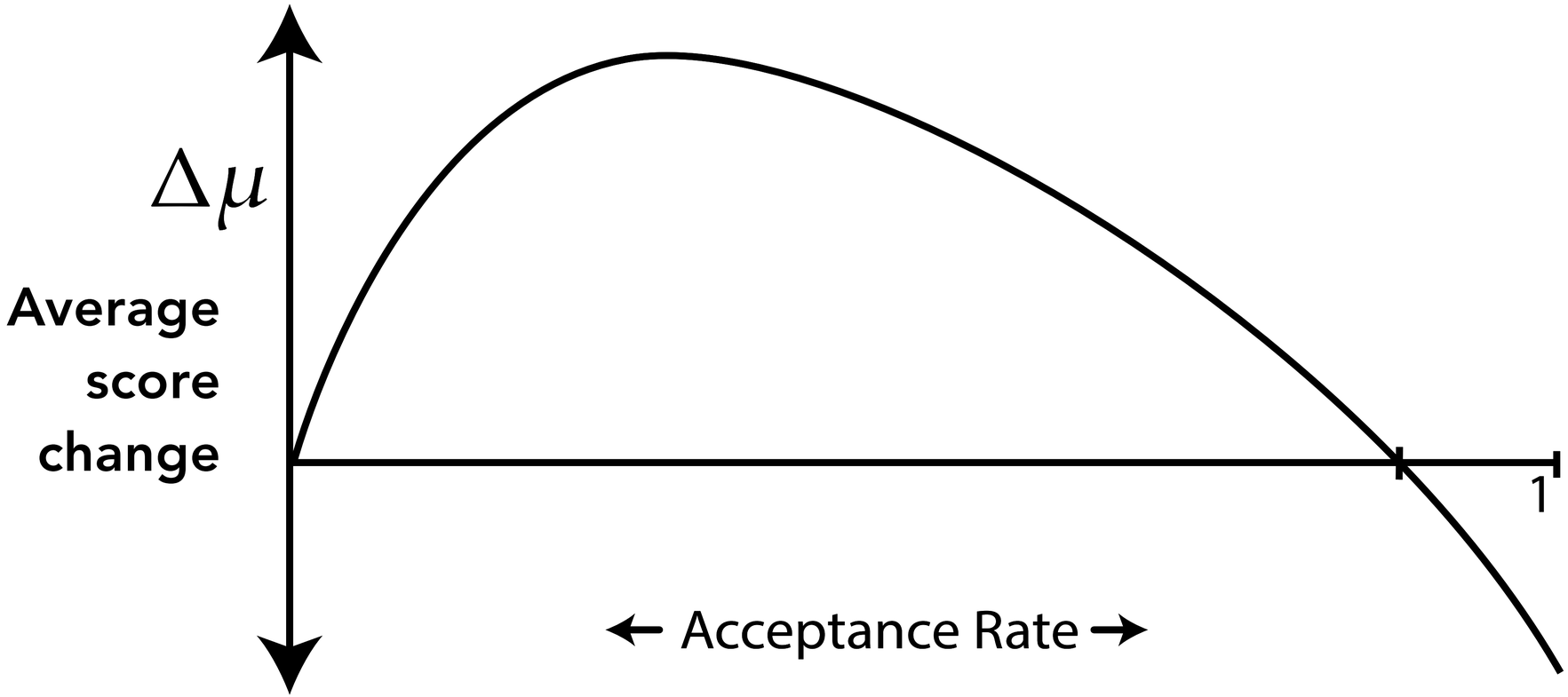
OUTCOME CURVE

Lemma: $\Delta\mu$ is a **concave** function of acceptance rate $\boldsymbol{\beta}$ under mild assumptions.



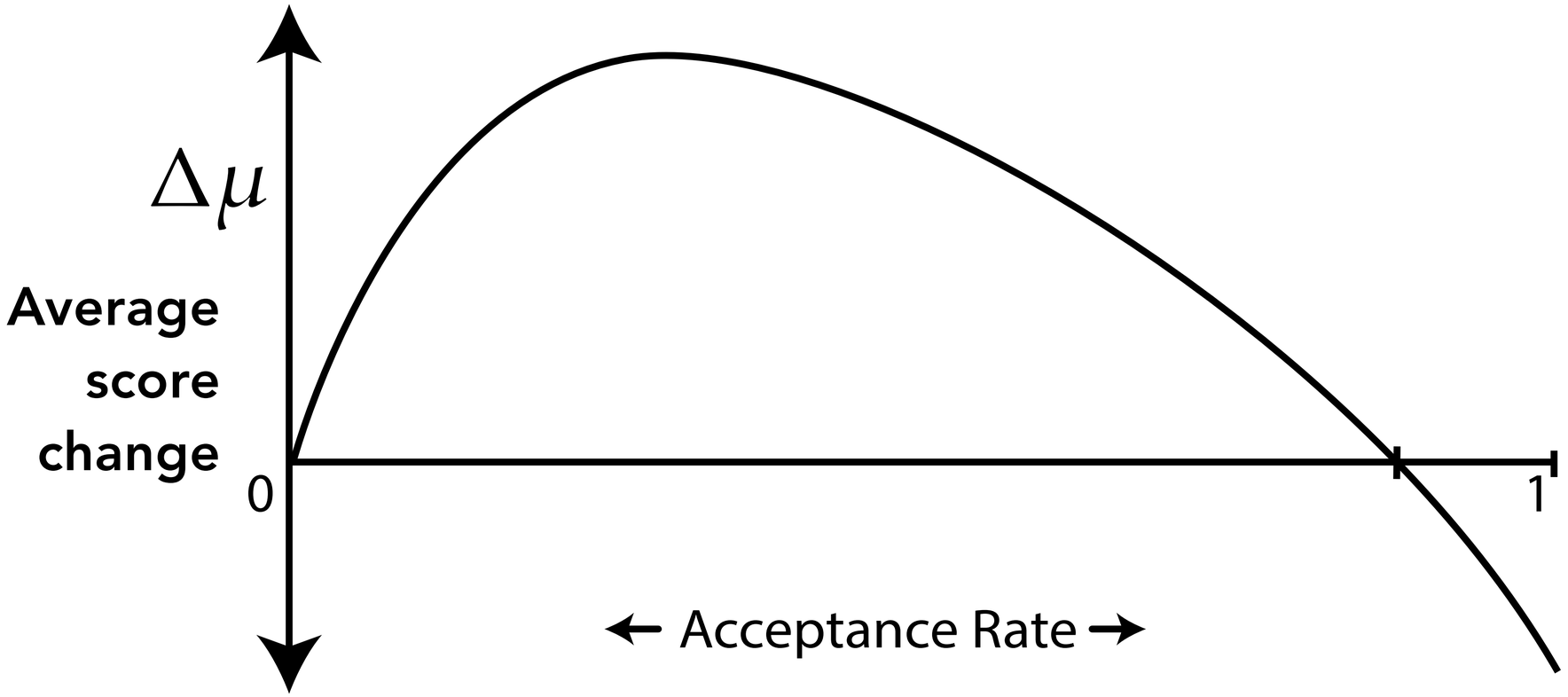
OUTCOME CURVE

Lemma: $\Delta\mu$ is a **concave** function of acceptance rate $\boldsymbol{\beta}$ under mild assumptions.



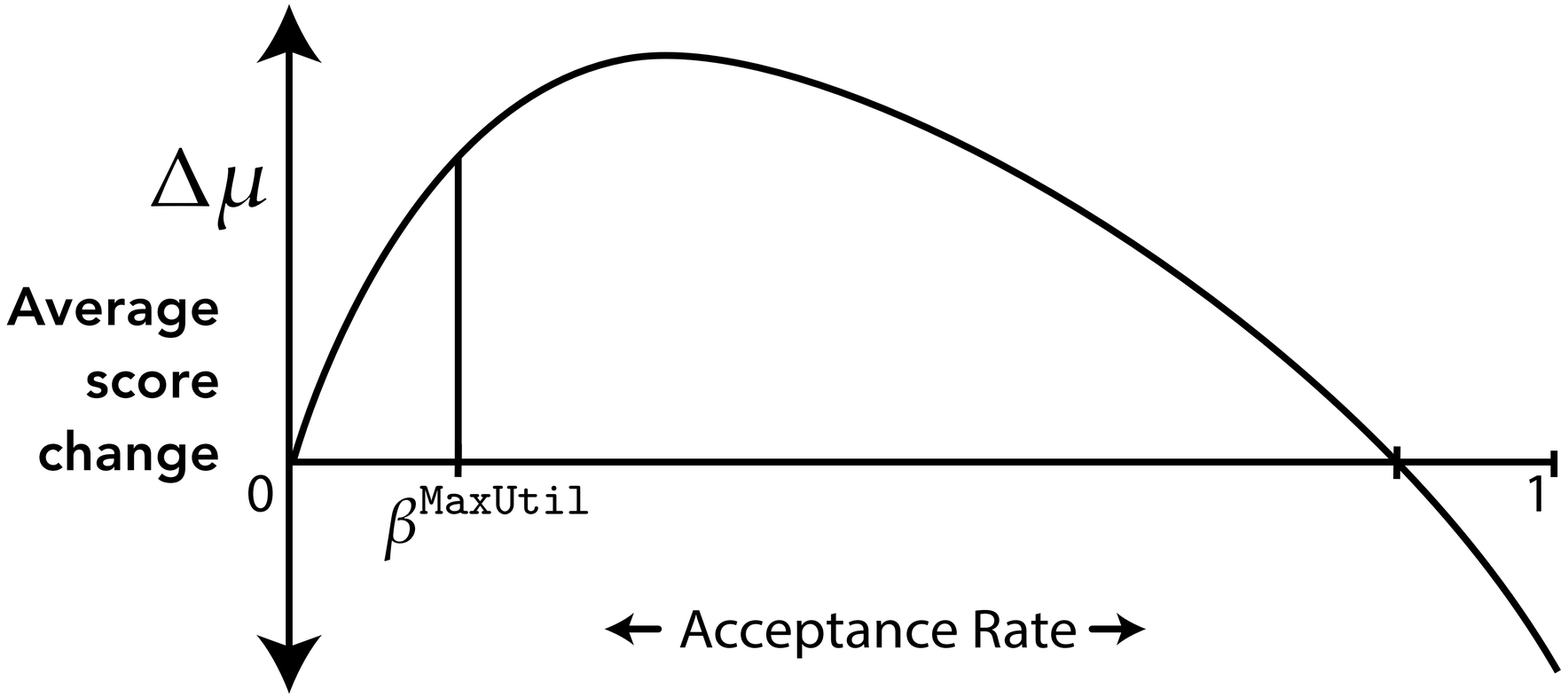
OUTCOME CURVE

Lemma: $\Delta\mu$ is a **concave** function of acceptance rate $\boldsymbol{\beta}$ under mild assumptions.



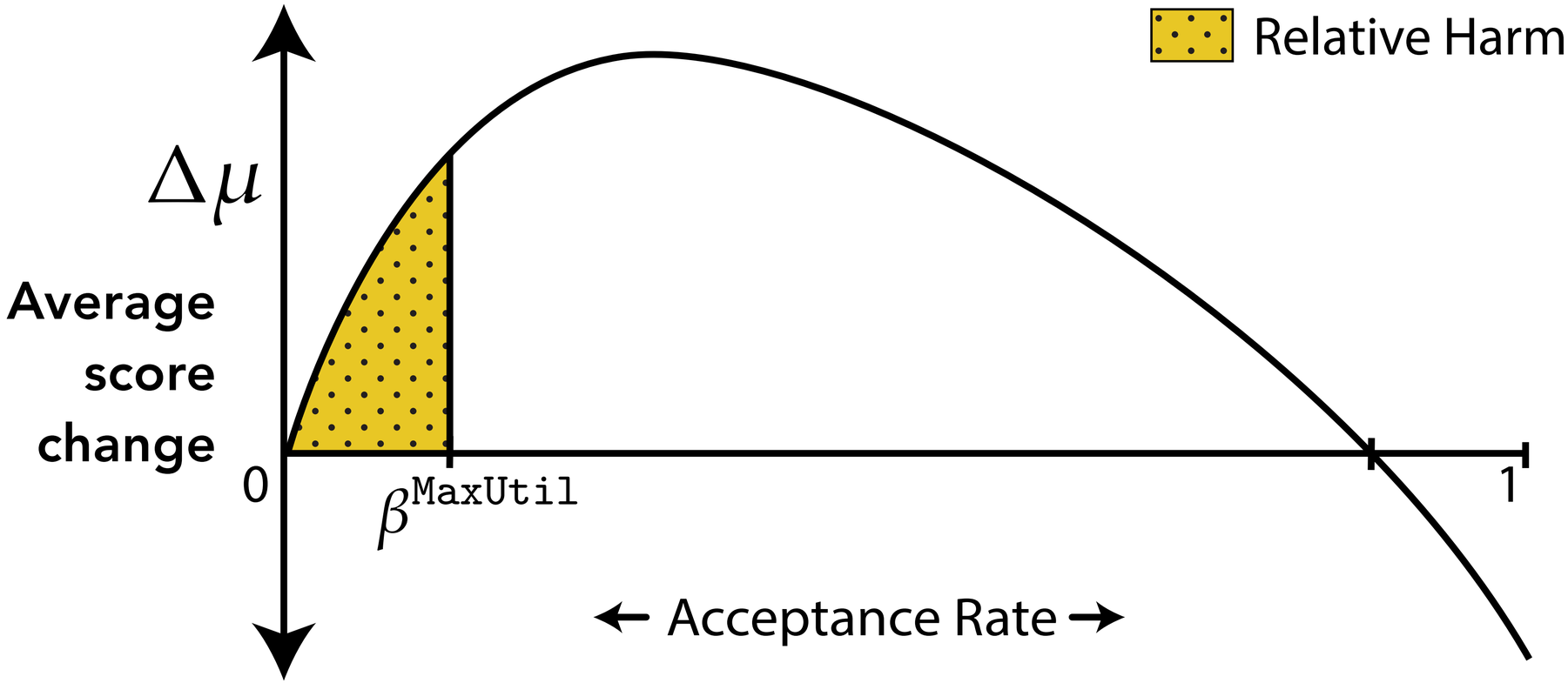
OUTCOME CURVE

Lemma: $\Delta\mu$ is a **concave** function of acceptance rate $\boldsymbol{\beta}$ under mild assumptions.



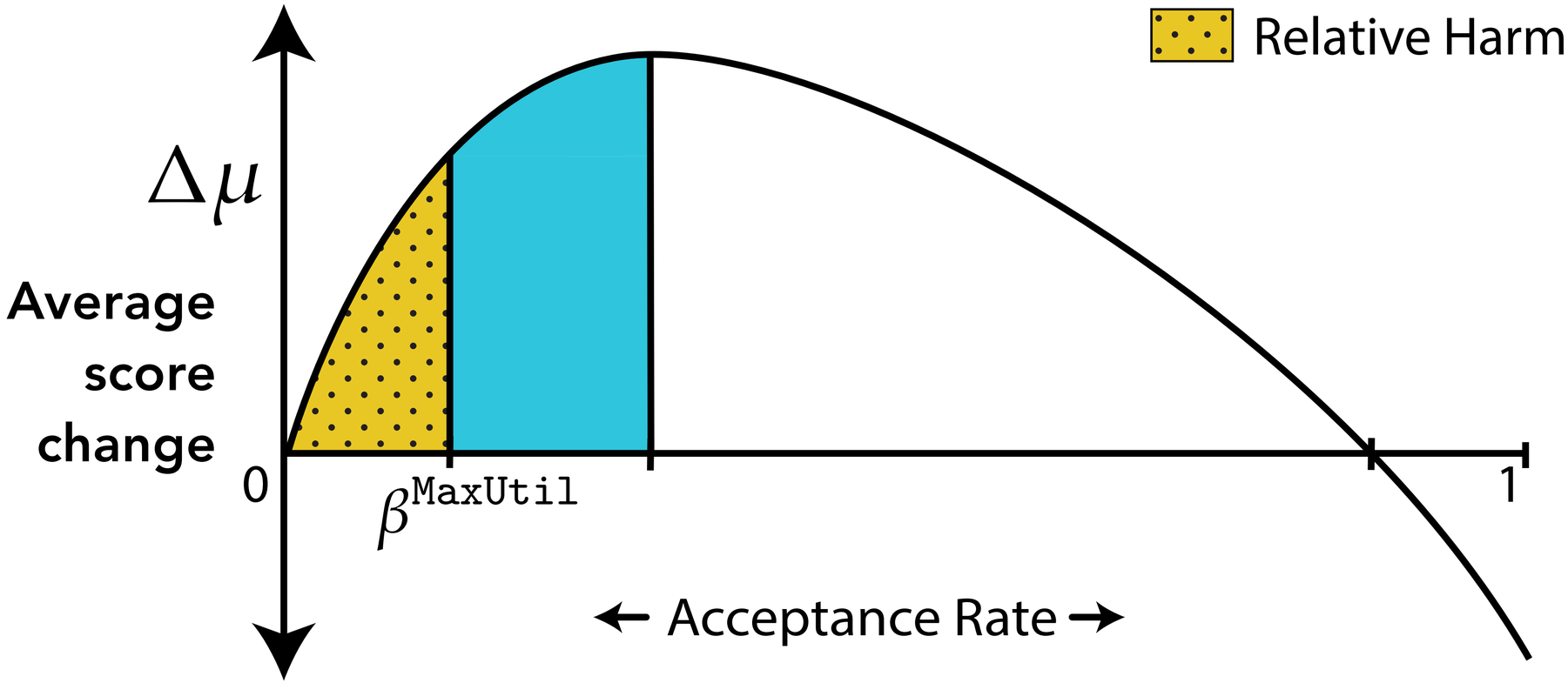
OUTCOME CURVE

Lemma: $\Delta\mu$ is a **concave** function of acceptance rate $\boldsymbol{\beta}$ under mild assumptions.



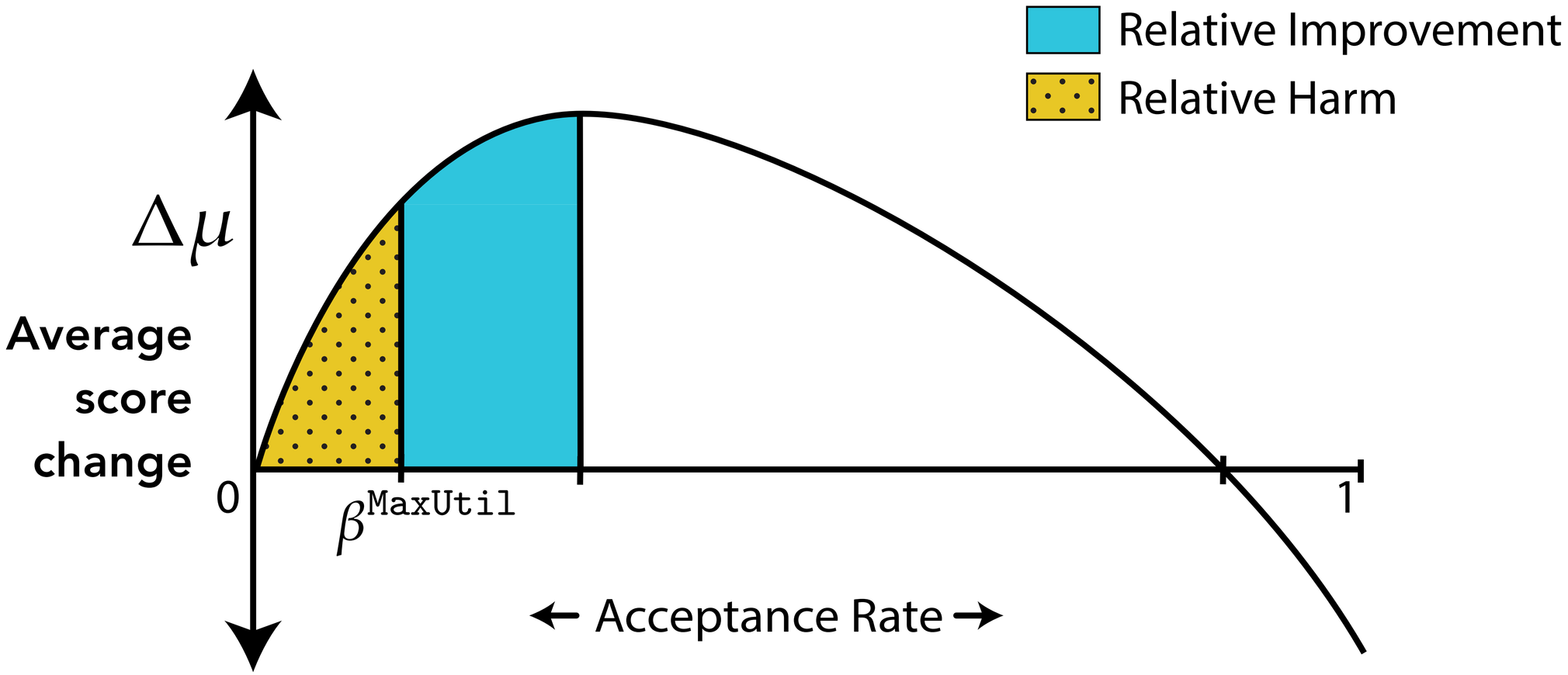
OUTCOME CURVE

Lemma: $\Delta\mu$ is a **concave** function of acceptance rate $\boldsymbol{\beta}$ under mild assumptions.



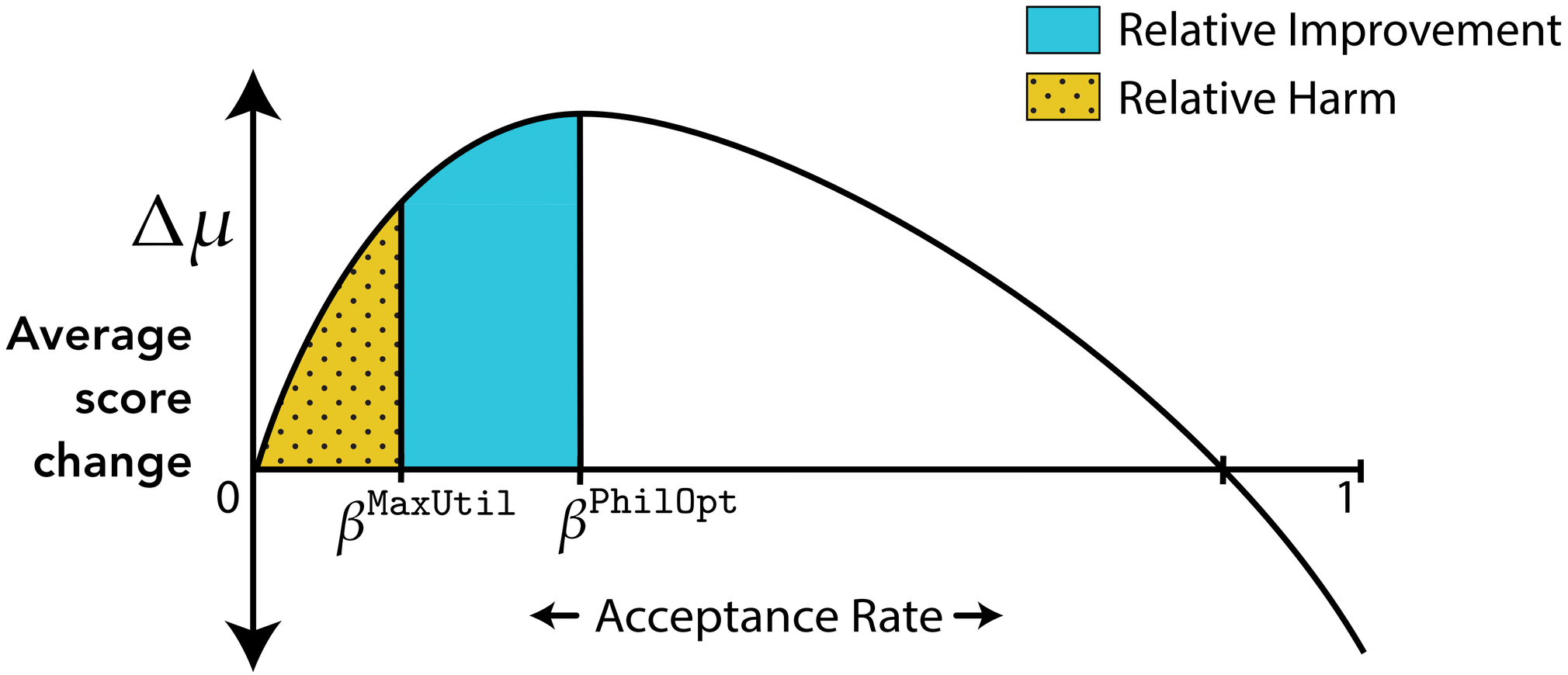
OUTCOME CURVE

Lemma: $\Delta\mu$ is a **concave** function of acceptance rate $\boldsymbol{\beta}$ under mild assumptions.



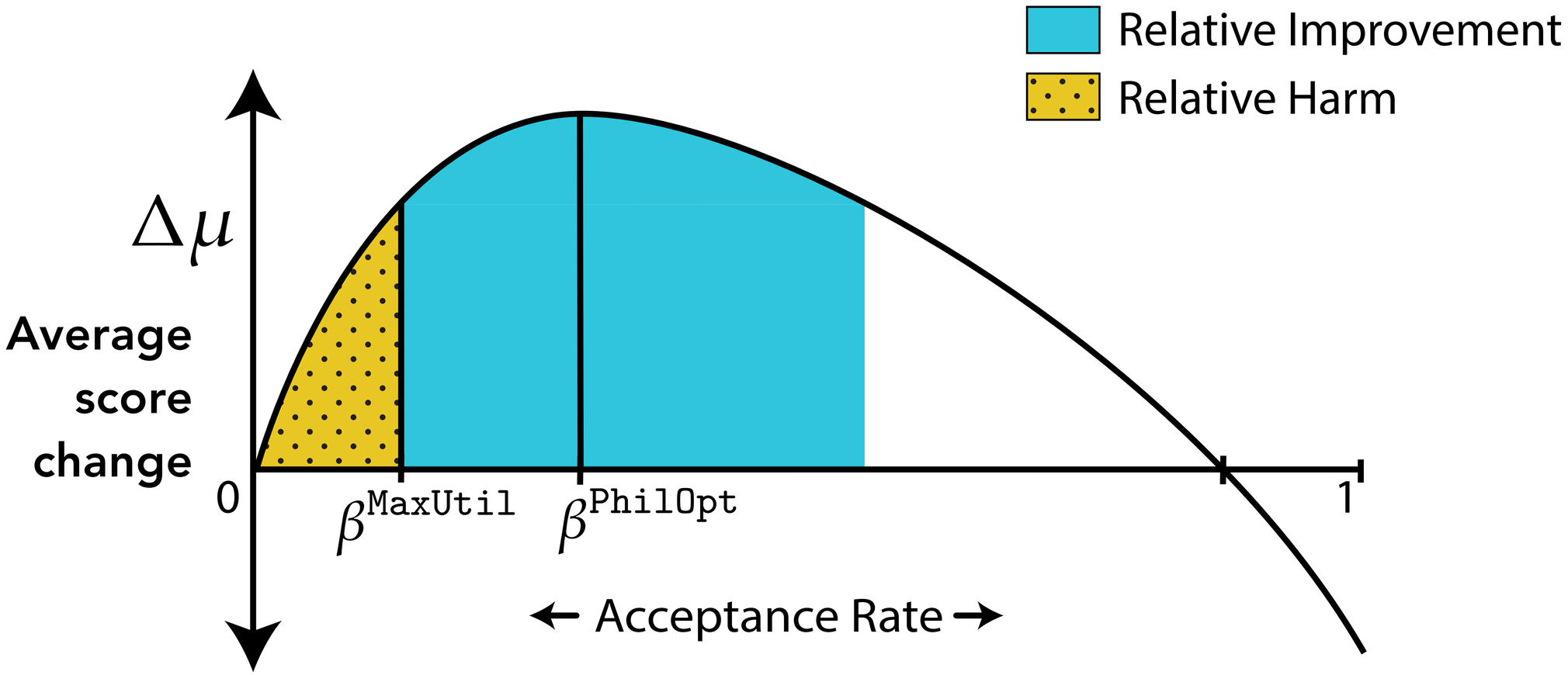
OUTCOME CURVE

Lemma: $\Delta\mu$ is a **concave** function of acceptance rate $\boldsymbol{\beta}$ under mild assumptions.



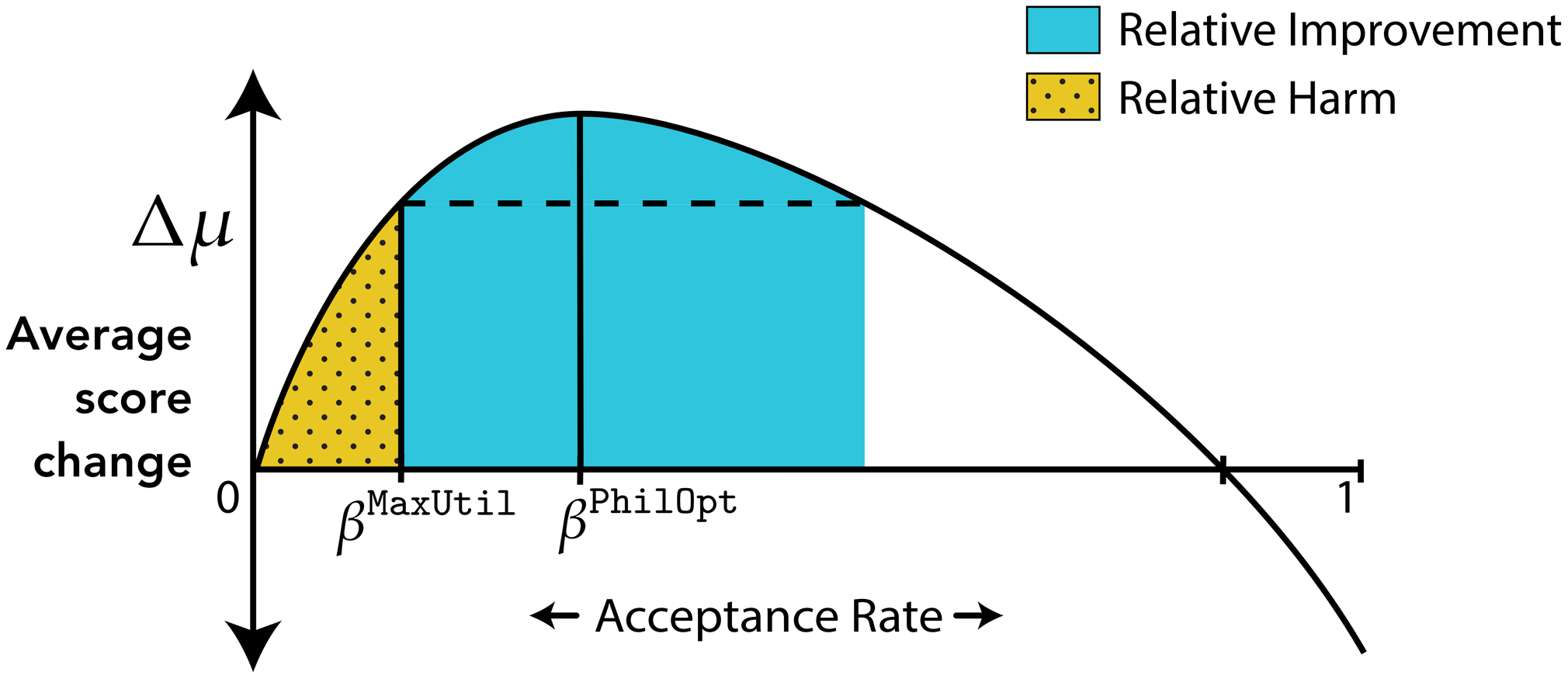
OUTCOME CURVE

Lemma: $\Delta\mu$ is a **concave** function of acceptance rate $\boldsymbol{\beta}$ under mild assumptions.



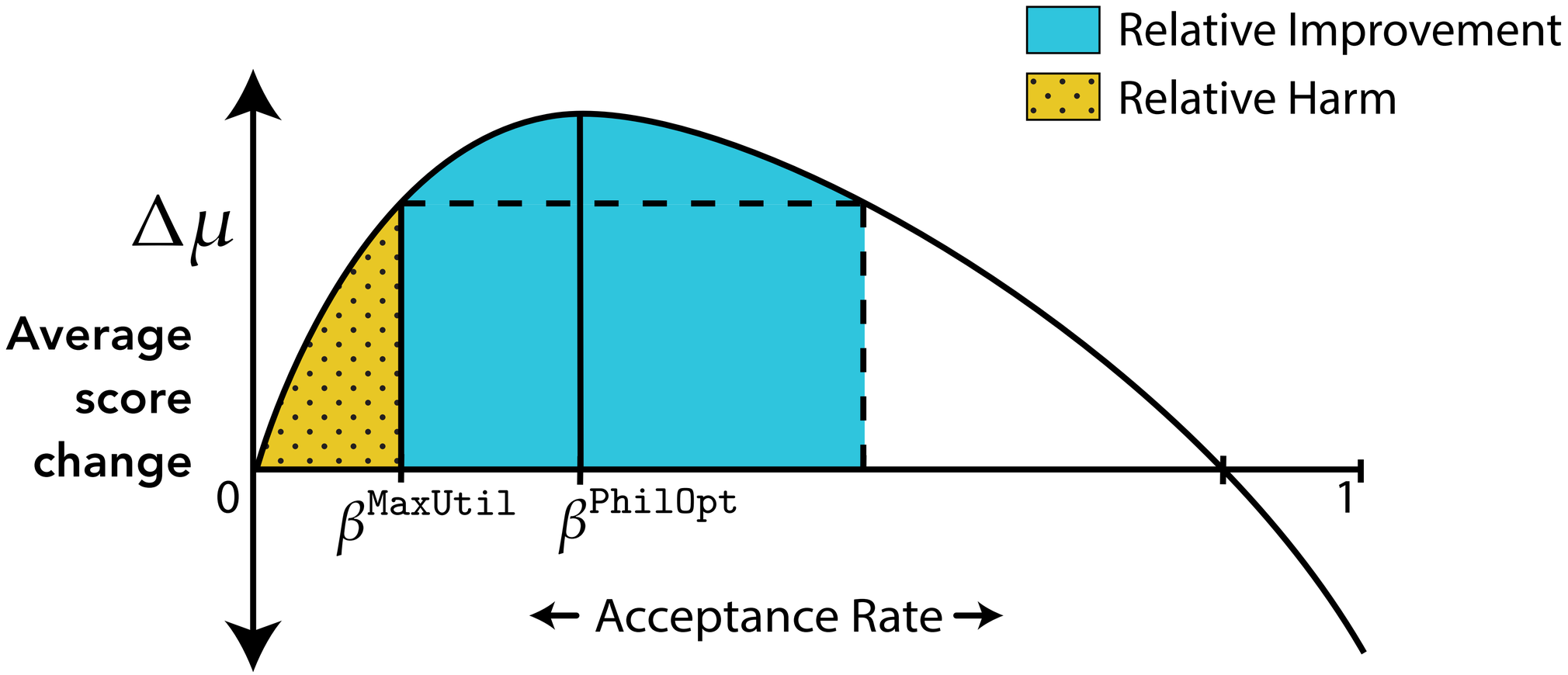
OUTCOME CURVE

Lemma: $\Delta\mu$ is a **concave** function of acceptance rate $\boldsymbol{\beta}$ under mild assumptions.



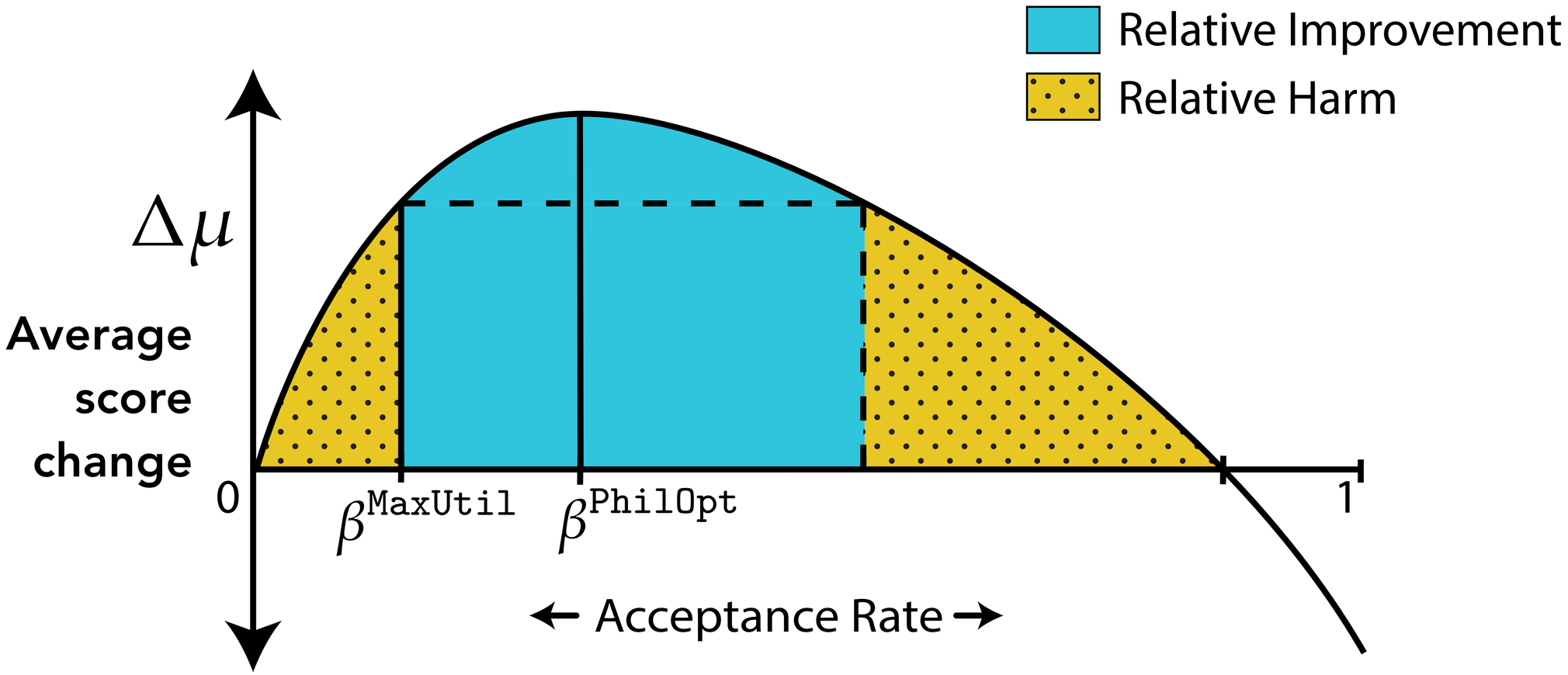
OUTCOME CURVE

Lemma: $\Delta\mu$ is a **concave** function of acceptance rate $\boldsymbol{\beta}$ under mild assumptions.



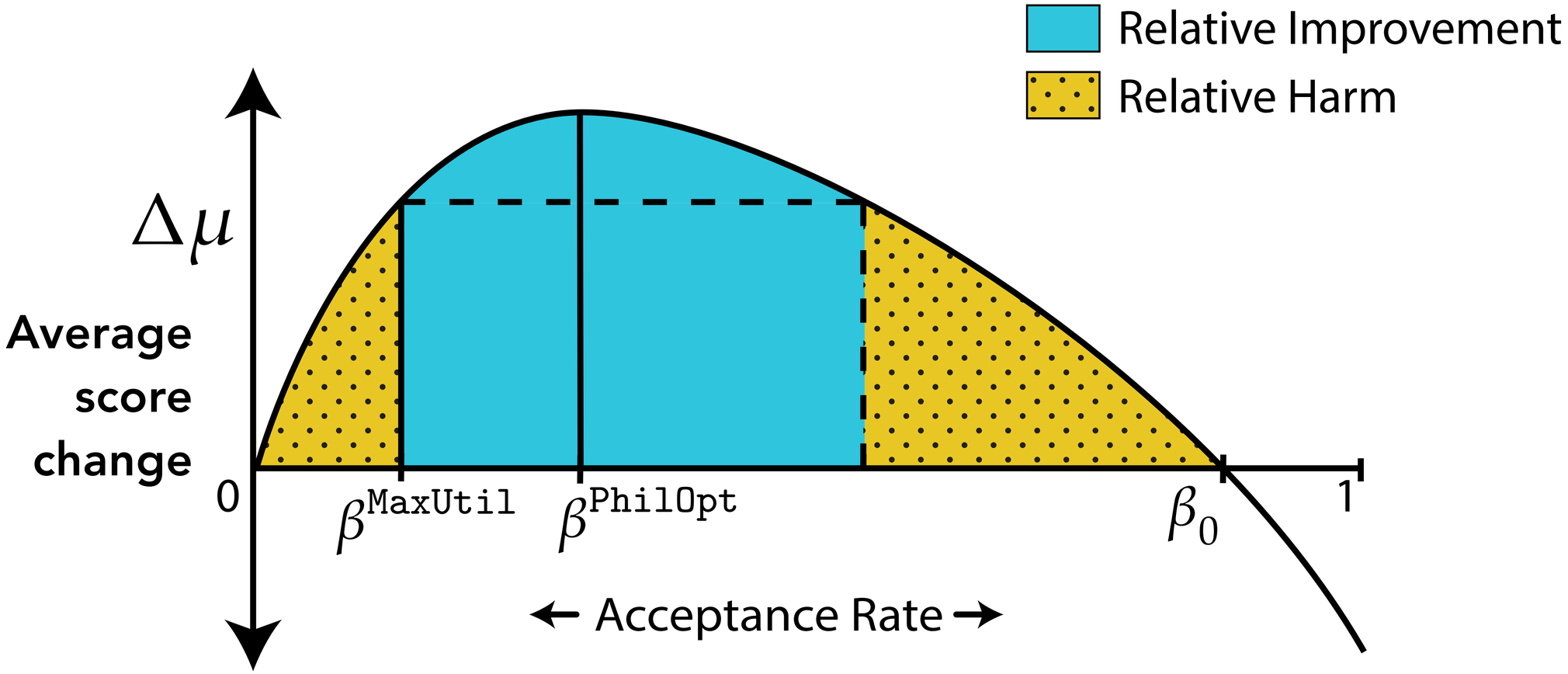
OUTCOME CURVE

Lemma: $\Delta\mu$ is a **concave** function of acceptance rate $\boldsymbol{\beta}$ under mild assumptions.



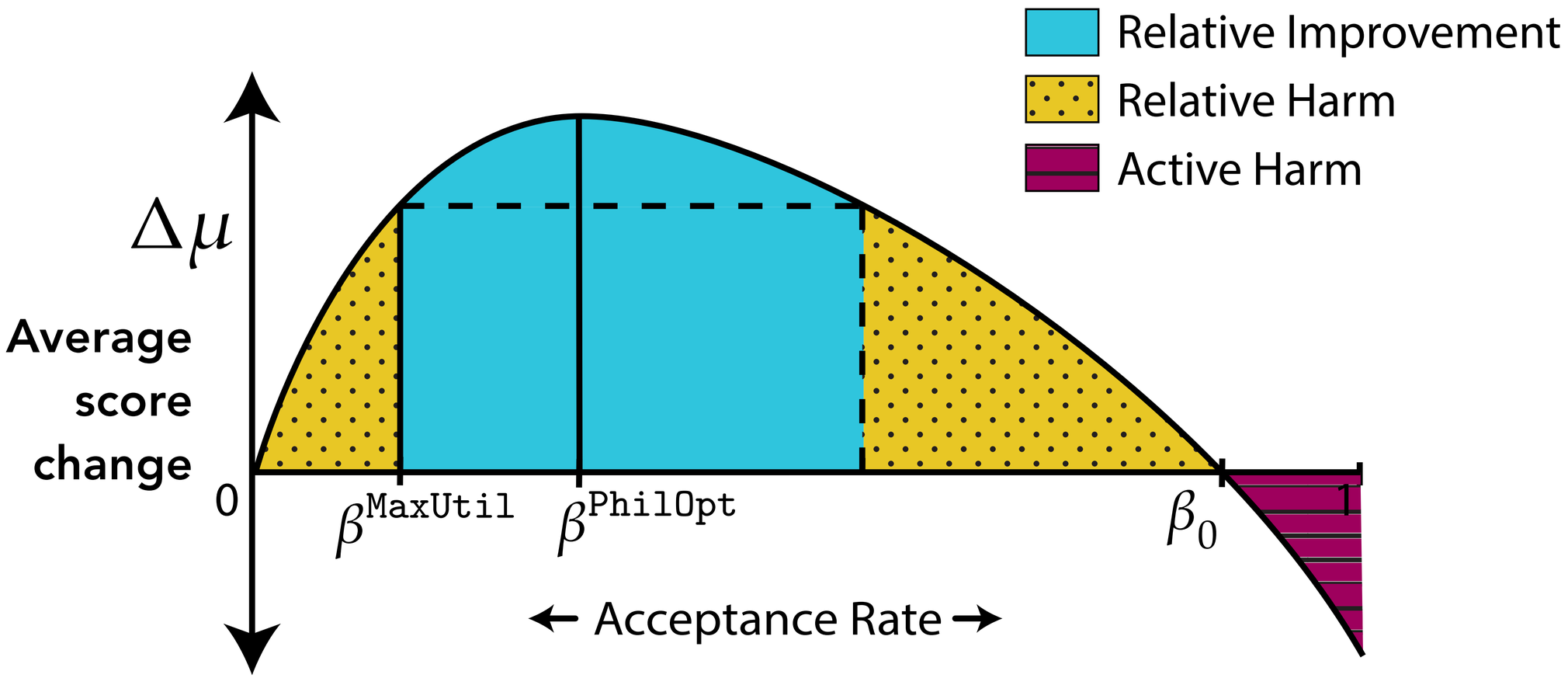
OUTCOME CURVE

Lemma: $\Delta\mu$ is a **concave** function of acceptance rate $\boldsymbol{\beta}$ under mild assumptions.



OUTCOME CURVE

Lemma: $\Delta\mu$ is a **concave** function of acceptance rate $\boldsymbol{\beta}$ under mild assumptions.



FAIRNESS CONSTRAINTS

FAIRNESS CONSTRAINTS

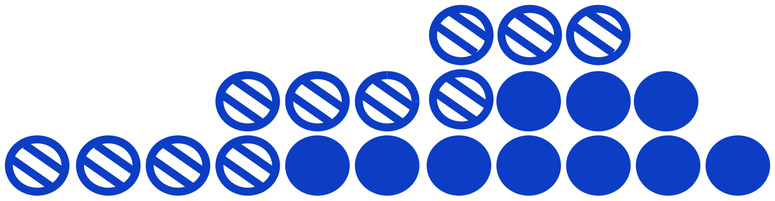
- Alternative to unconstrained utility maximization

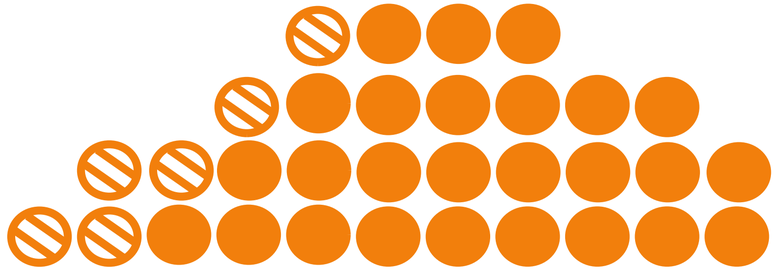
FAIRNESS CONSTRAINTS

- Alternative to unconstrained utility maximization
- ***Demographic Parity***: Equal Acceptance Rate

FAIRNESS CONSTRAINTS

- Alternative to unconstrained utility maximization
- ***Demographic Parity***: Equal Acceptance Rate



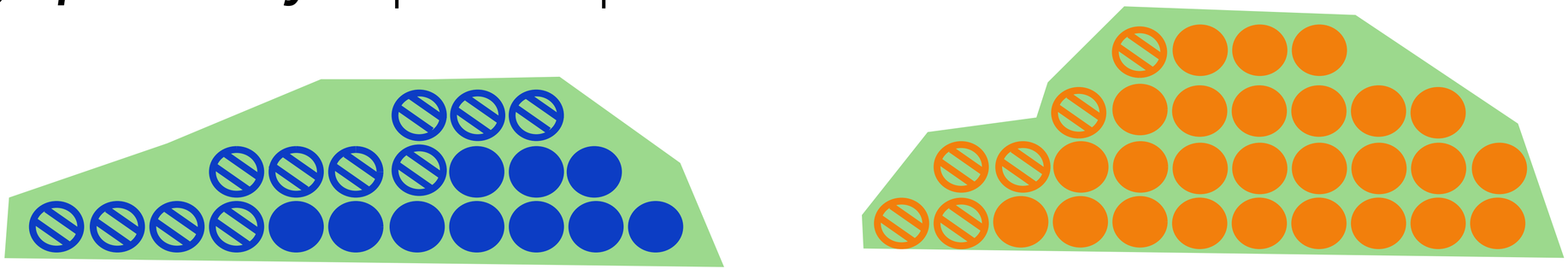


FAIRNESS CONSTRAINTS

- Alternative to unconstrained utility maximization
- ***Demographic Parity***: Equal Acceptance Rate

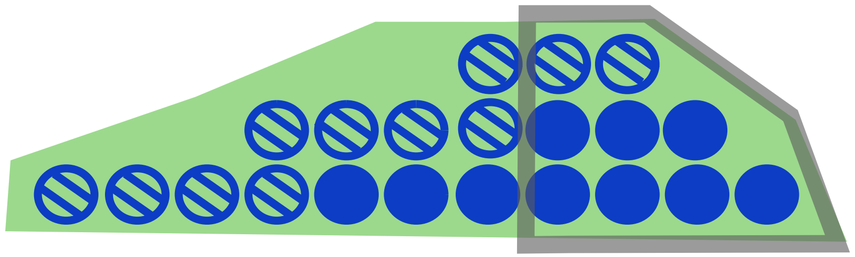
FAIRNESS CONSTRAINTS

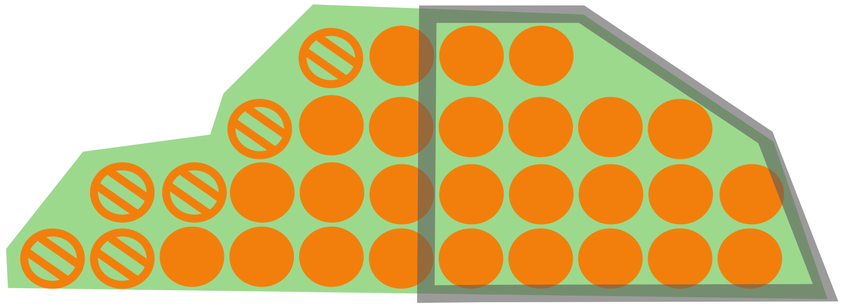
- Alternative to unconstrained utility maximization
- ***Demographic Parity***: Equal Acceptance Rate

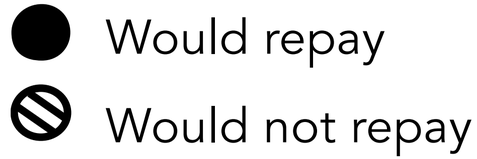


FAIRNESS CONSTRAINTS

- Alternative to unconstrained utility maximization
- ***Demographic Parity***: Equal Acceptance Rate

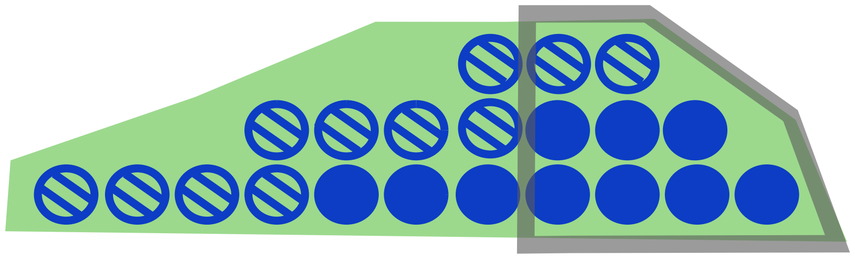


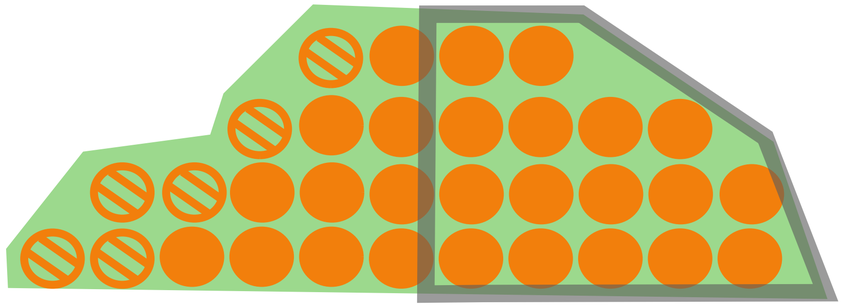




FAIRNESS CONSTRAINTS

- Alternative to unconstrained utility maximization
- ***Demographic Parity***: Equal Acceptance Rate





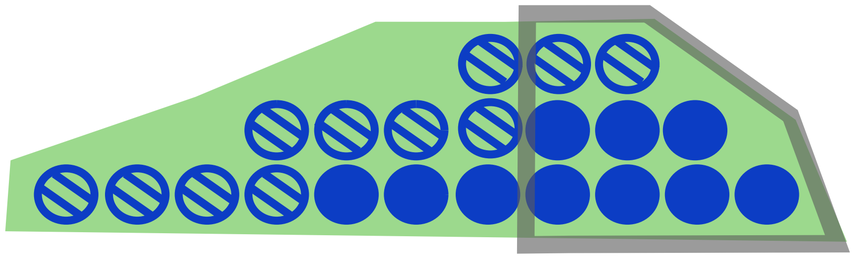
Would repay

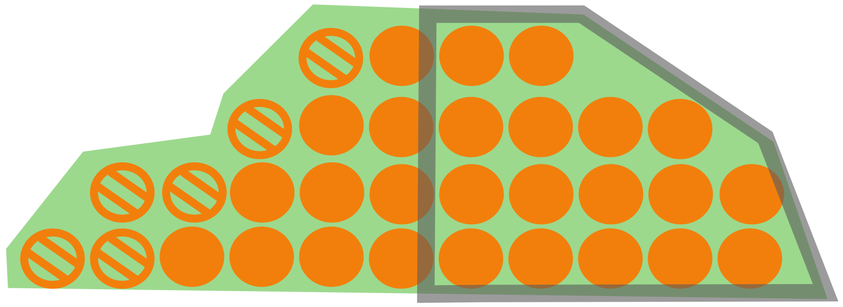
Would not repay

- ***Equal Opportunity***: Equal True Positive Rates
  [Hardt, Price and Srebro 2016]

FAIRNESS CONSTRAINTS

- Alternative to unconstrained utility maximization
- ***Demographic Parity***: Equal Acceptance Rate



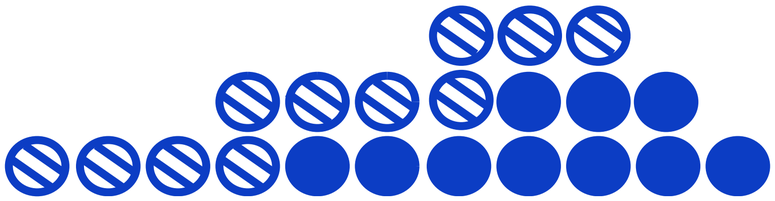


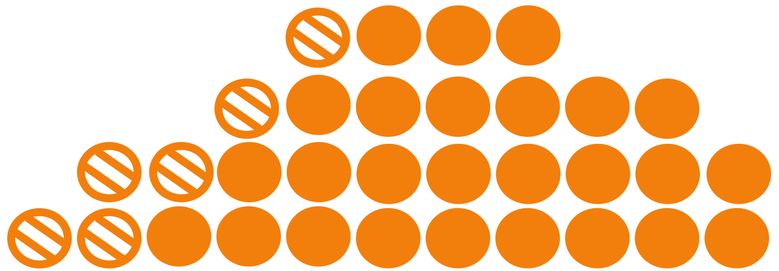
Would repay

Would not repay

- ***Equal Opportunity***: Equal True Positive Rates

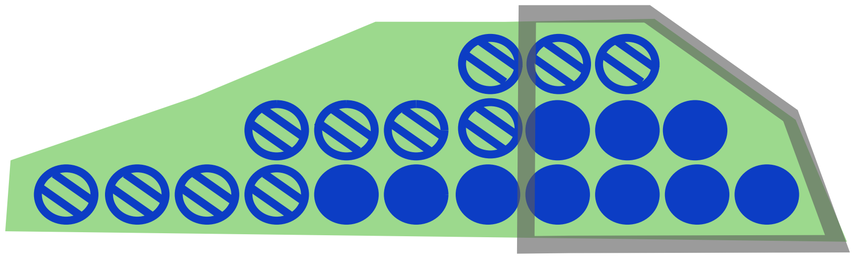
  [Hardt, Price and Srebro 2016]

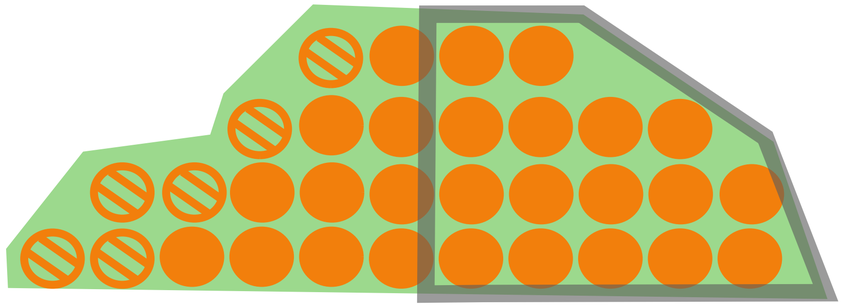


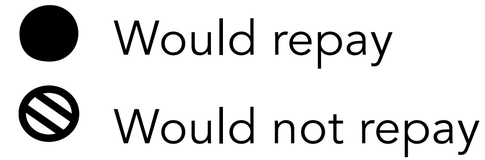


FAIRNESS CONSTRAINTS

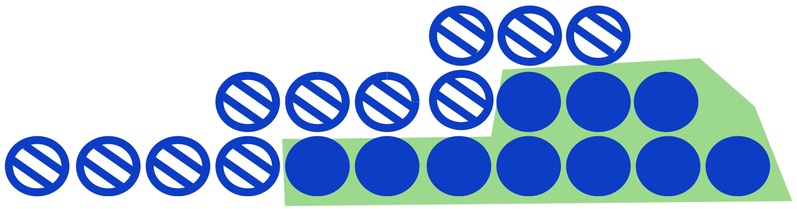
- Alternative to unconstrained utility maximization
- ***Demographic Parity***: Equal Acceptance Rate

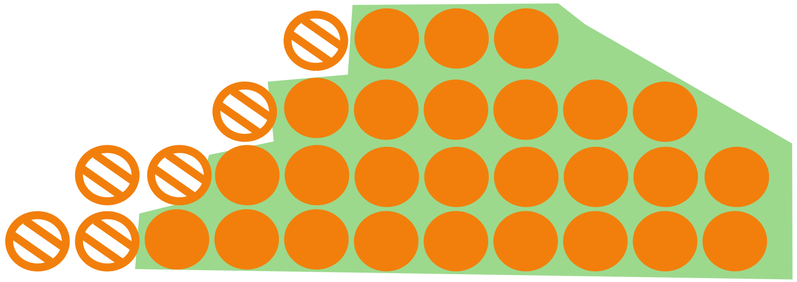






- ***Equal Opportunity***: Equal True Positive Rates
  [Hardt, Price and Srebro 2016]



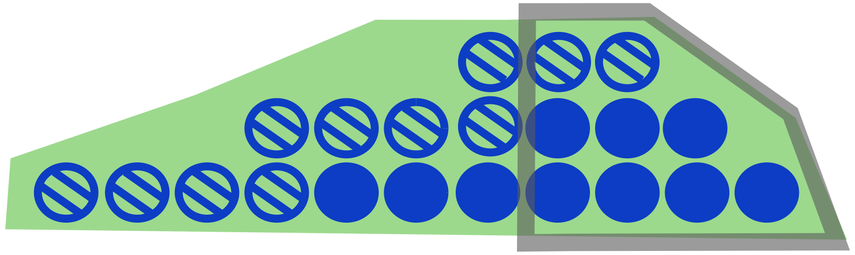


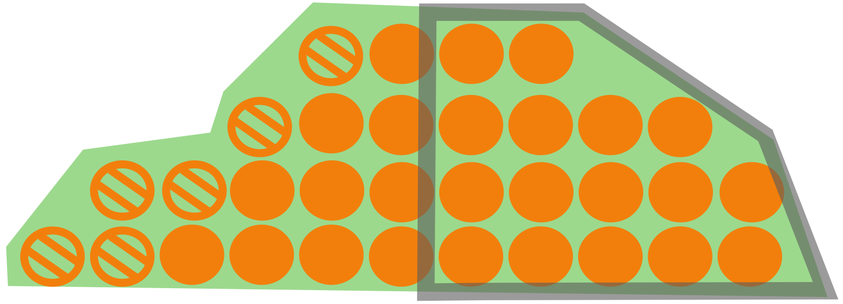
FAIRNESS CONSTRAINTS

- Alternative to unconstrained utility maximization
- ***Demographic Parity***: Equal Acceptance Rate

Would repay

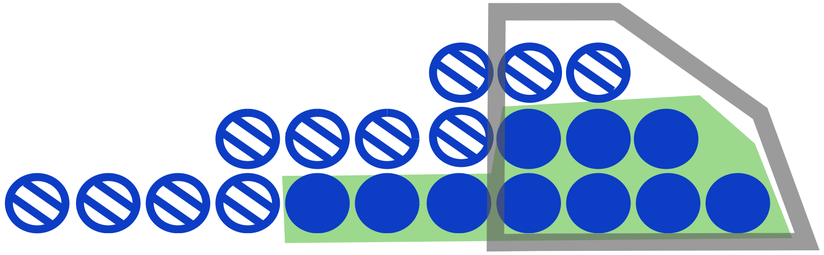
Would not repay

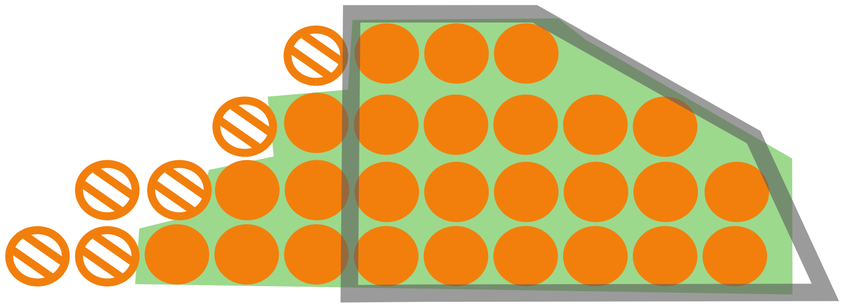




- ***Equal Opportunity***: Equal True Positive Rates

  [Hardt, Price and Srebro 2016]





MAIN THEOREMS

MAIN THEOREMS

Theorem 1 [All outcome regimes are possible]

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

MAIN THEOREMS

Theorem 1 [All outcome regimes are possible]

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

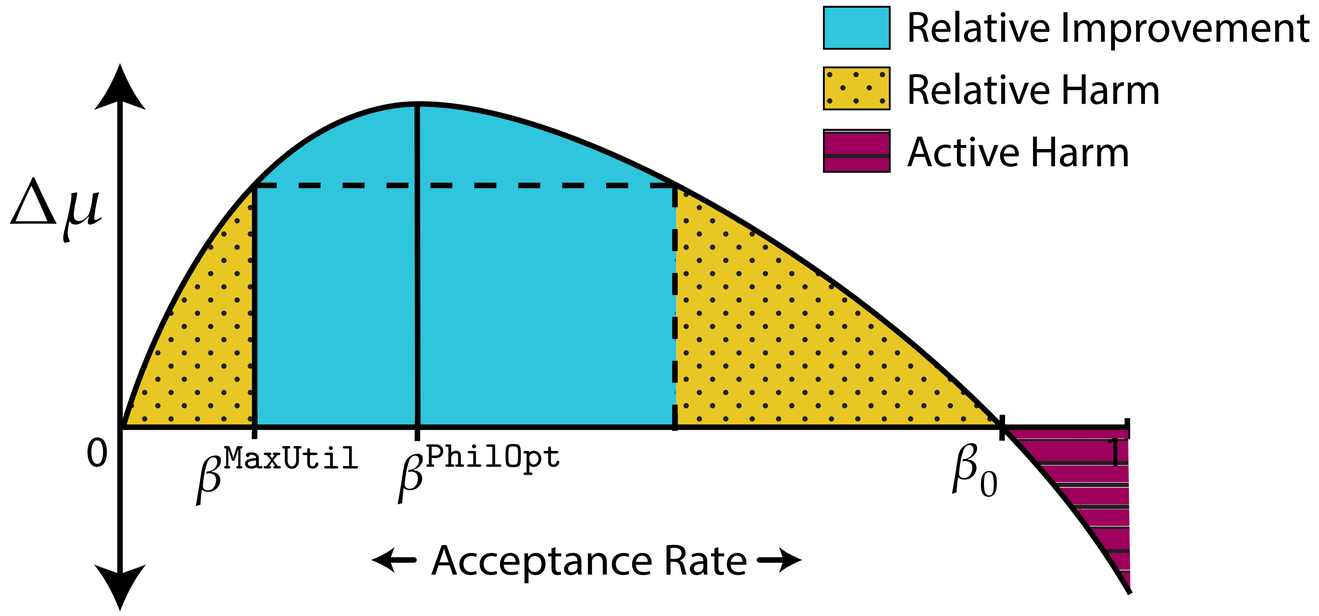
- unconstrained utility maximization never causes active harm.

MAIN THEOREMS

Theorem 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

CHOICE OF FAIRNESS CRITERIA MATTERS

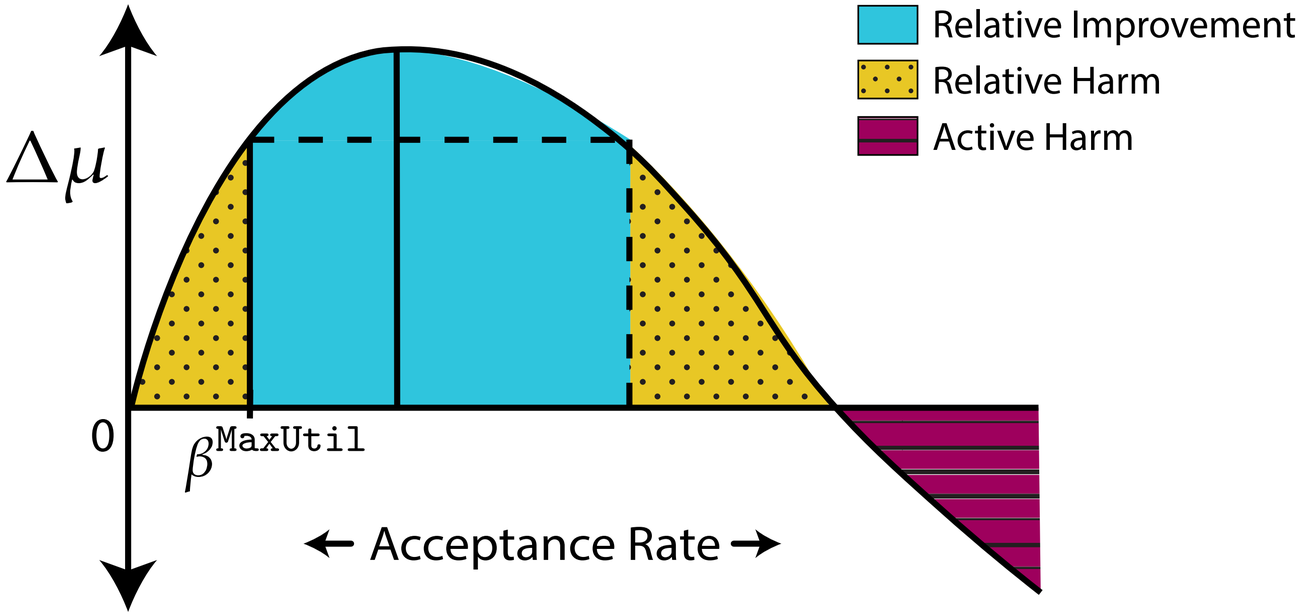
Theorem 2

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

CHOICE OF FAIRNESS CRITERIA MATTERS

Theorem 2

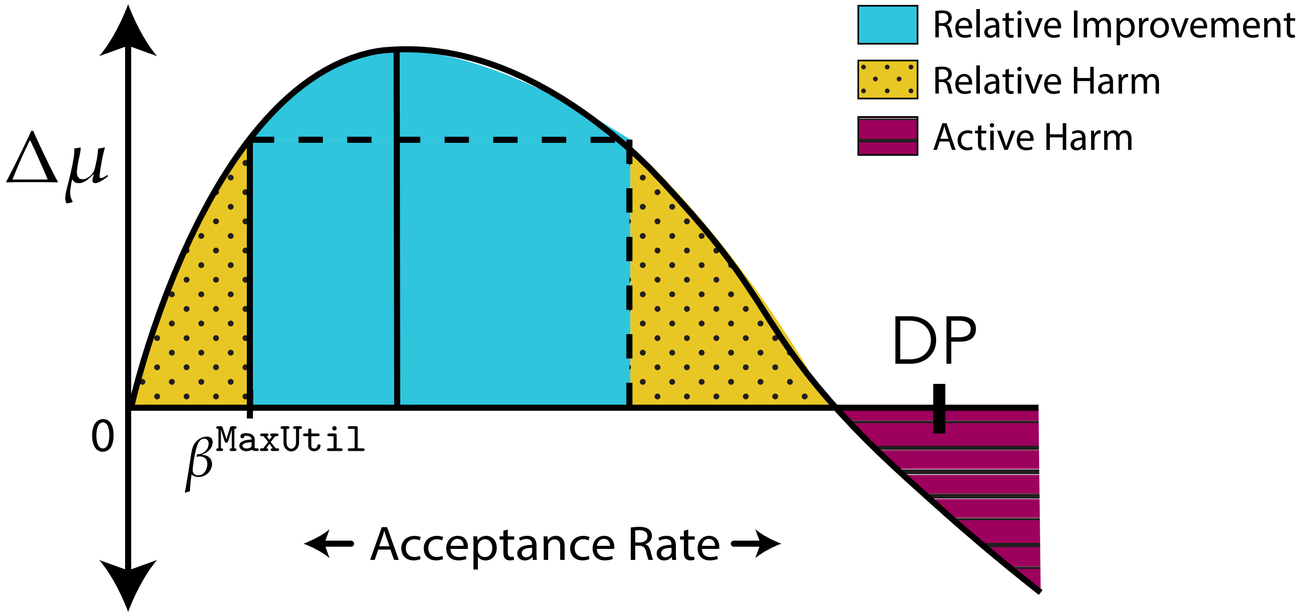
Demographic parity (DP) may cause active or relative harm by **over-acceptance**; equal opportunity (EO) doesn't.



CHOICE OF FAIRNESS CRITERIA MATTERS

Theorem 2

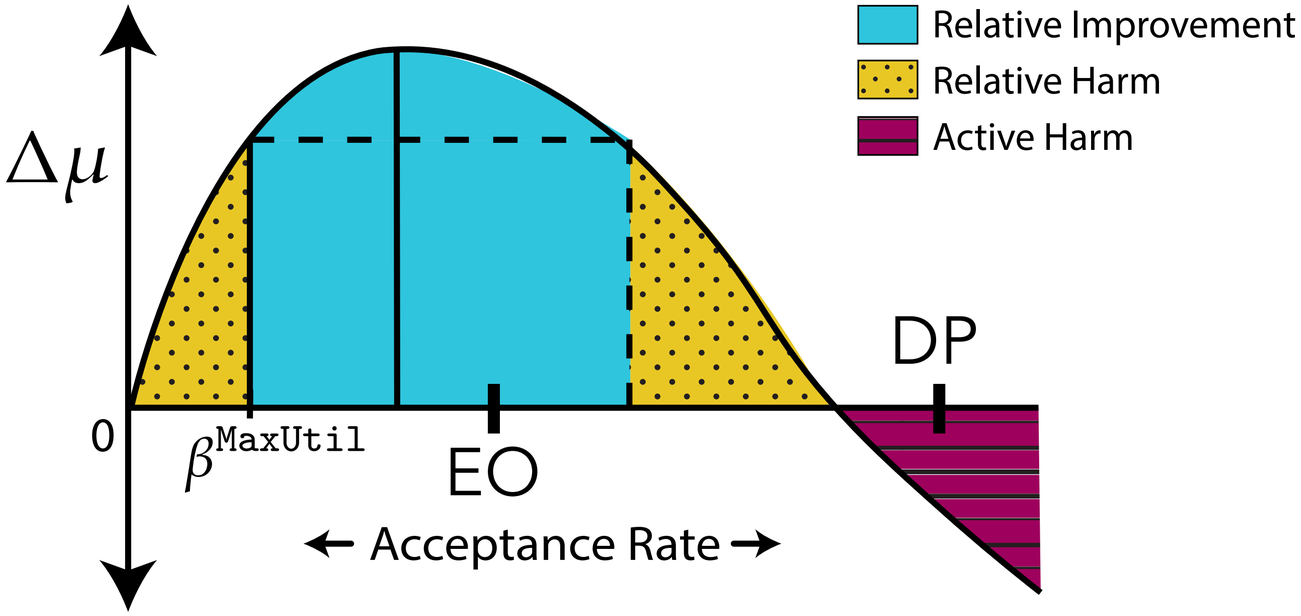
Demographic parity (DP) may cause active or relative harm by **over-acceptance**; equal opportunity (EO) doesn't.



CHOICE OF FAIRNESS CRITERIA MATTERS

Theorem 2

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



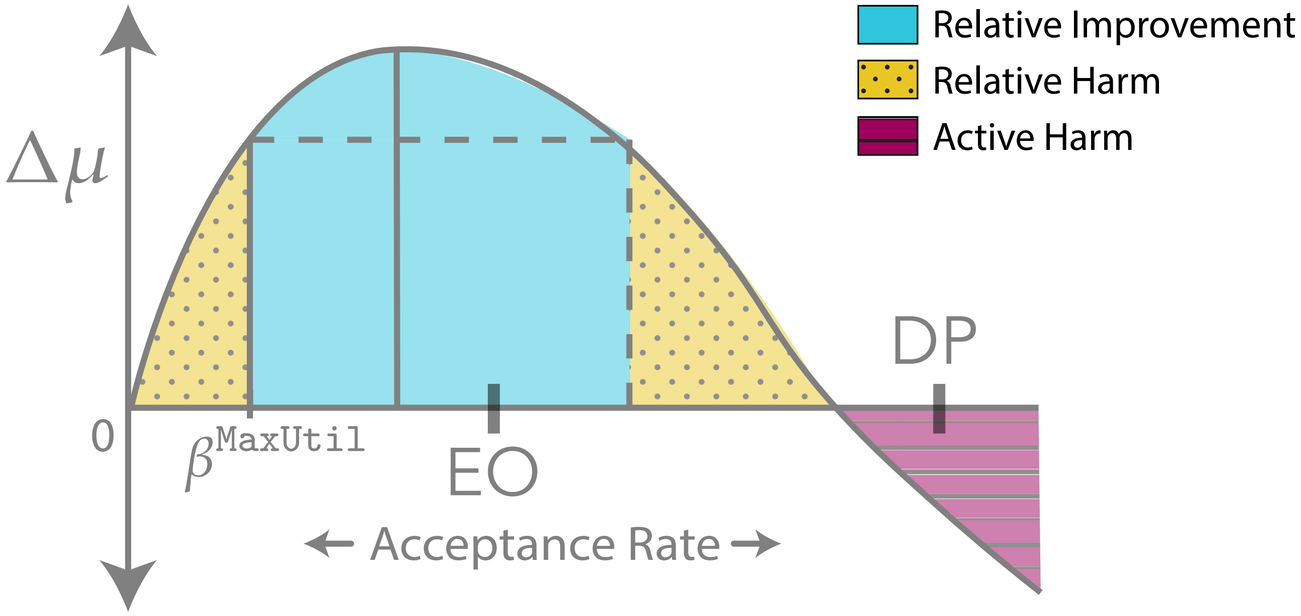
CHOICE OF FAIRNESS CRITERIA MATTERS

Theorem 2

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

Theorem 3

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



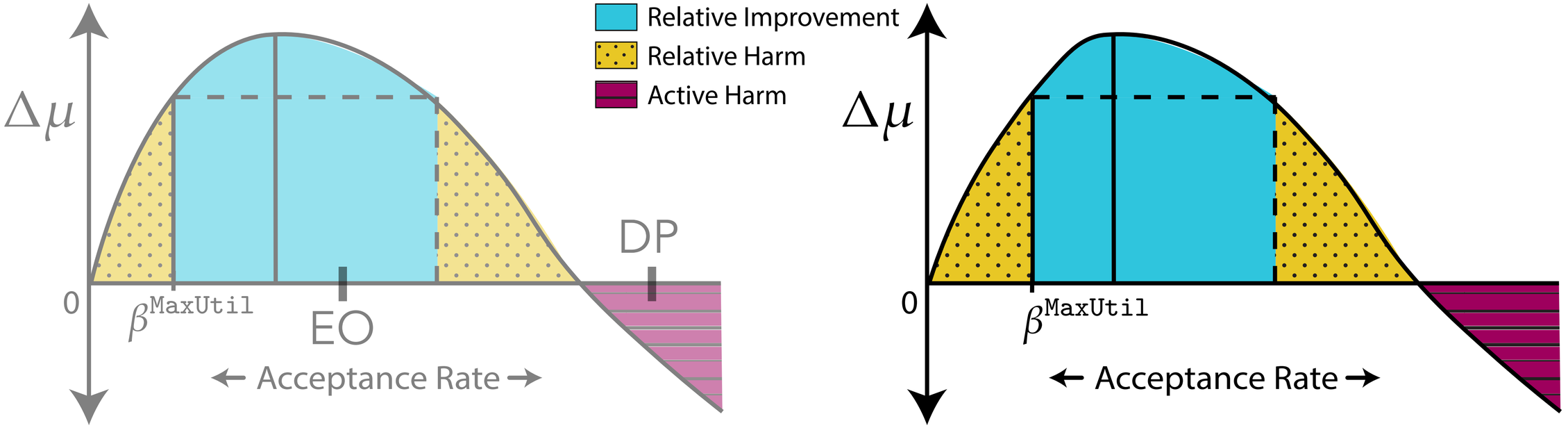
CHOICE OF FAIRNESS CRITERIA MATTERS

Theorem 2

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

Theorem 3

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



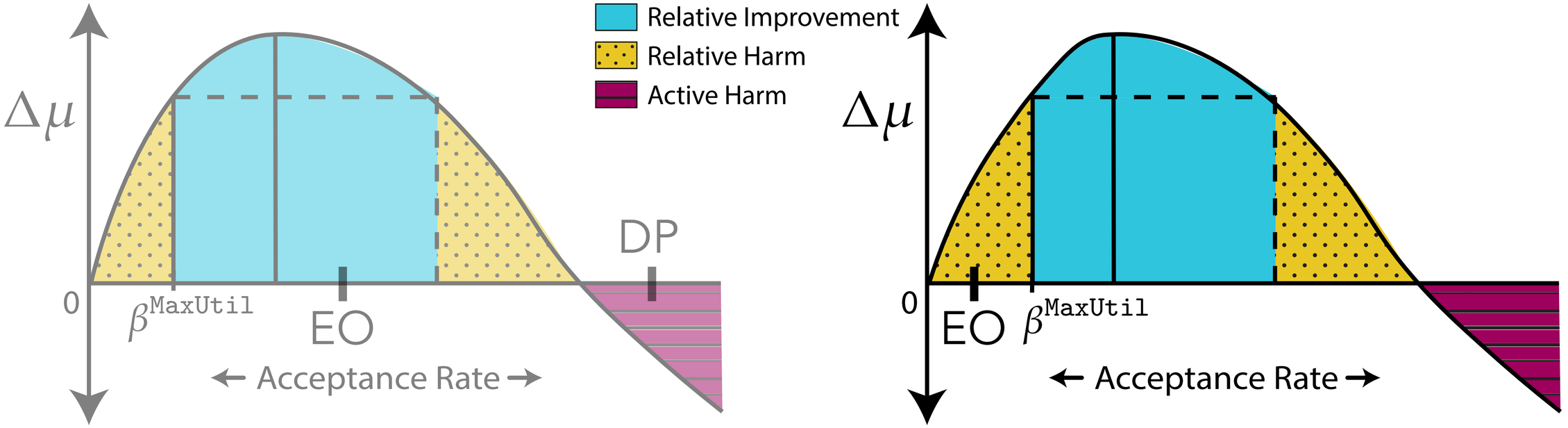
CHOICE OF FAIRNESS CRITERIA MATTERS

Theorem 2

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

Theorem 3

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



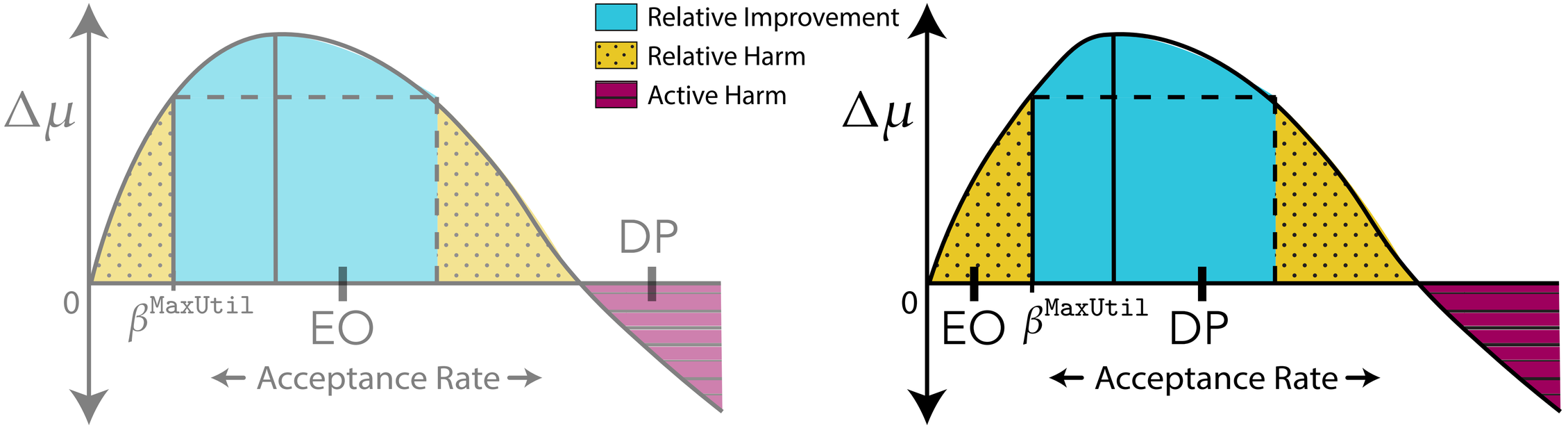
CHOICE OF FAIRNESS CRITERIA MATTERS

Theorem 2

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

Theorem 3

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



EXPERIMENTS ON FICO CREDIT SCORES

EXPERIMENTS ON FICO CREDIT SCORES

- 300,000+ TransUnion TransRisk scores from 2003
- Scores range from 300 to 850 and are meant to predict default risk

EXPERIMENTS ON FICO CREDIT SCORES

- 300,000+ TransUnion TransRisk scores from 2003
- Scores range from 300 to 850 and are meant to predict default risk

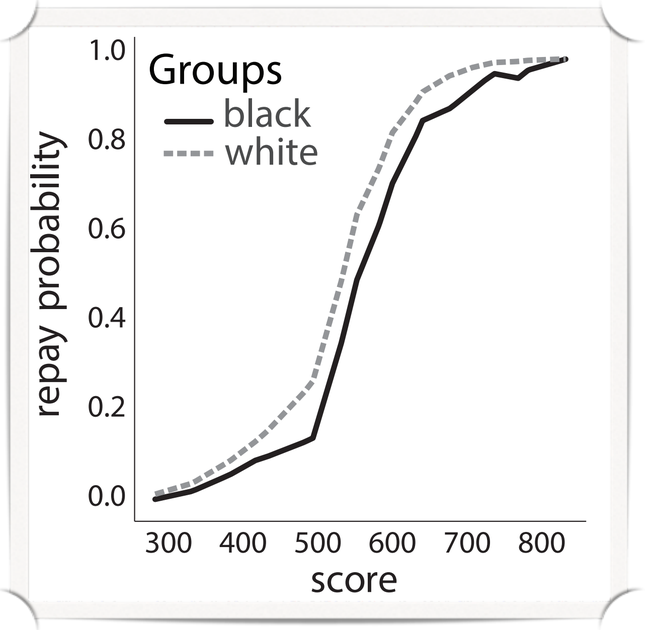
What we did

EXPERIMENTS ON FICO CREDIT SCORES

- 300,000+ TransUnion TransRisk scores from 2003
- Scores range from 300 to 850 and are meant to predict default risk

What we did

- Use empirical data labeled by race (“white” and “black”) to estimate group score distributions, repayment probabilities, and relative sizes

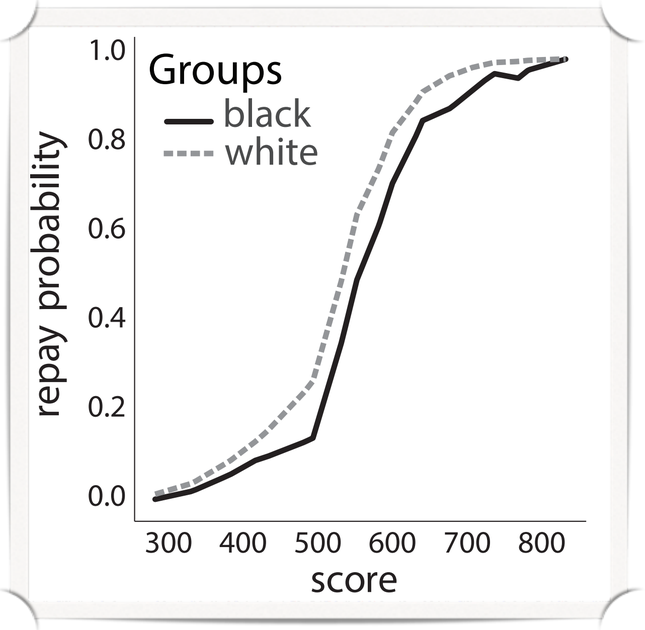


EXPERIMENTS ON FICO CREDIT SCORES

- 300,000+ TransUnion TransRisk scores from 2003
- Scores range from 300 to 850 and are meant to predict default risk

What we did

- Use empirical data labeled by race (“white” and “black”) to estimate group score distributions, repayment probabilities, and relative sizes
- Model the bank’s profit/loss ratio, e.g. +1:-4

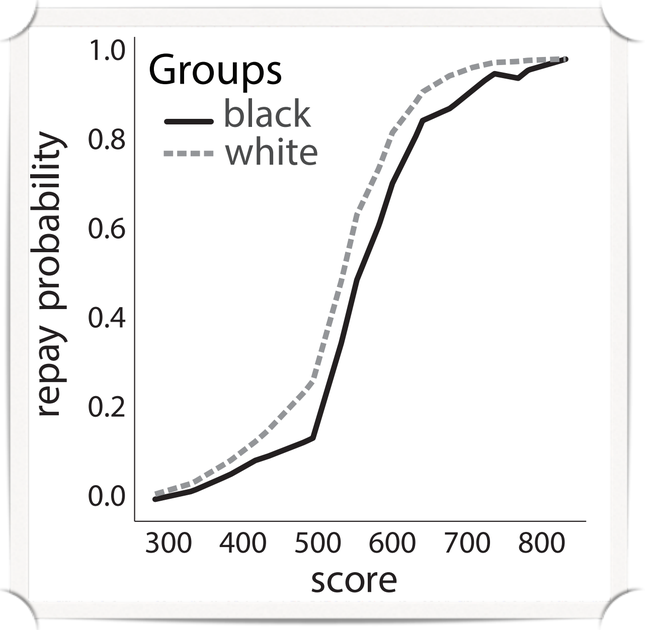


EXPERIMENTS ON FICO CREDIT SCORES

- 300,000+ TransUnion TransRisk scores from 2003
- Scores range from 300 to 850 and are meant to predict default risk

What we did

- Use empirical data labeled by race (“white” and “black”) to estimate group score distributions, repayment probabilities, and relative sizes
- Model the bank’s profit/loss ratio, e.g. +1:-4
- Model the delayed impact of repayment/default on credit score, e.g. +75/-150

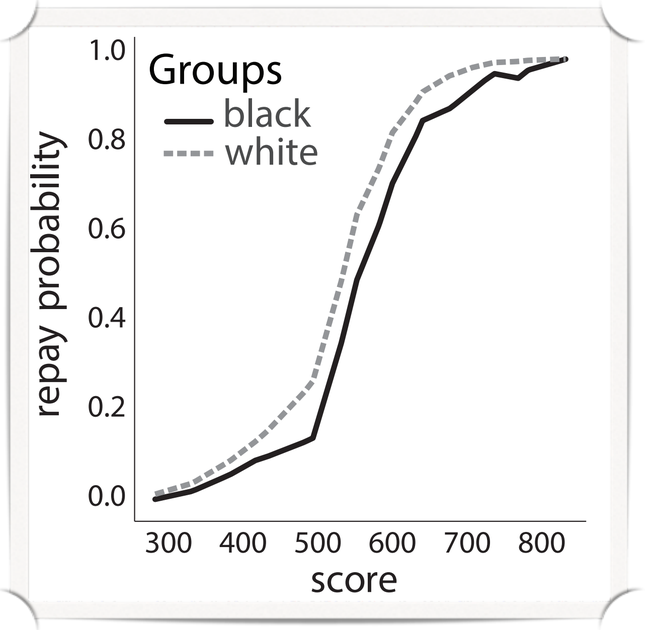


EXPERIMENTS ON FICO CREDIT SCORES

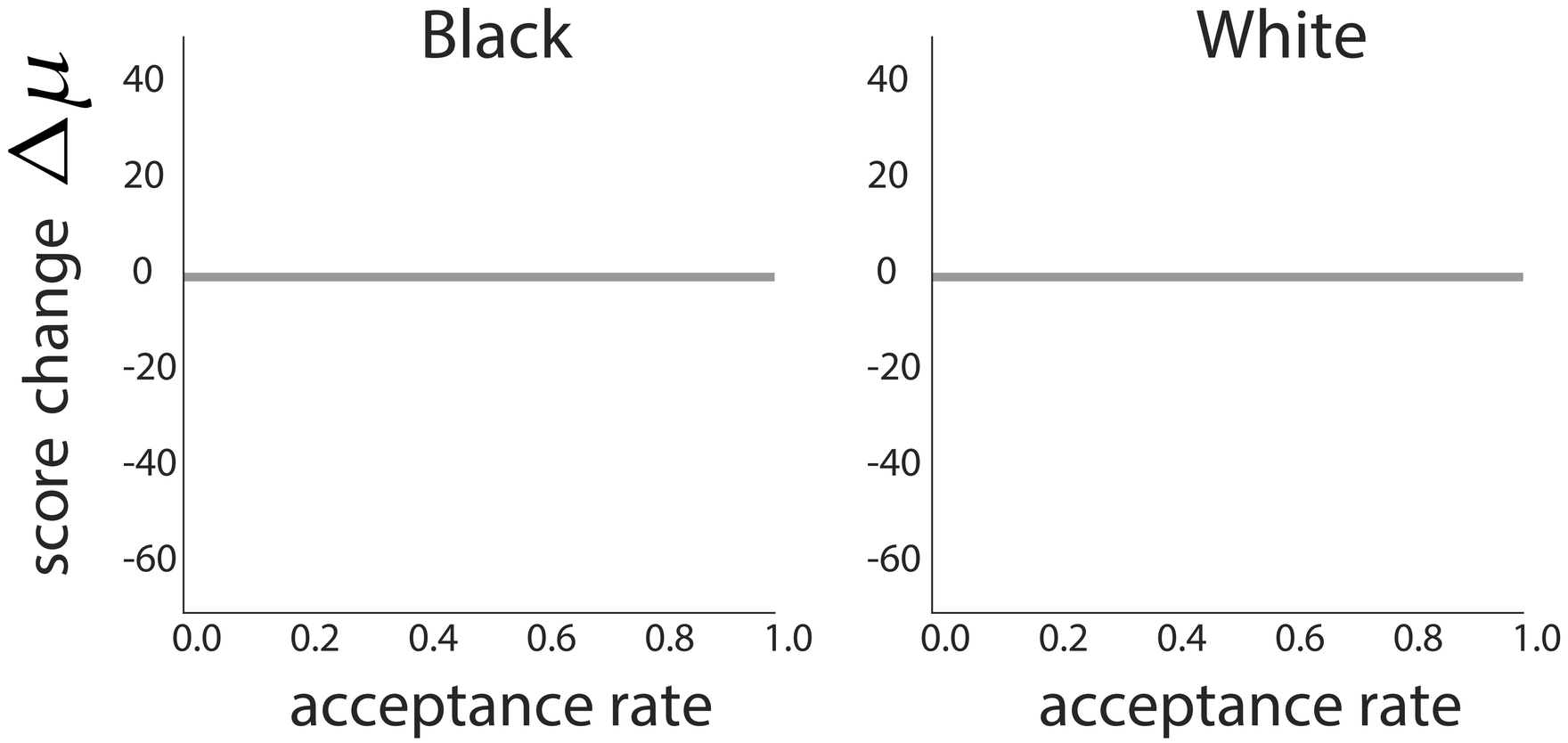
- 300,000+ TransUnion TransRisk scores from 2003
- Scores range from 300 to 850 and are meant to predict default risk

What we did

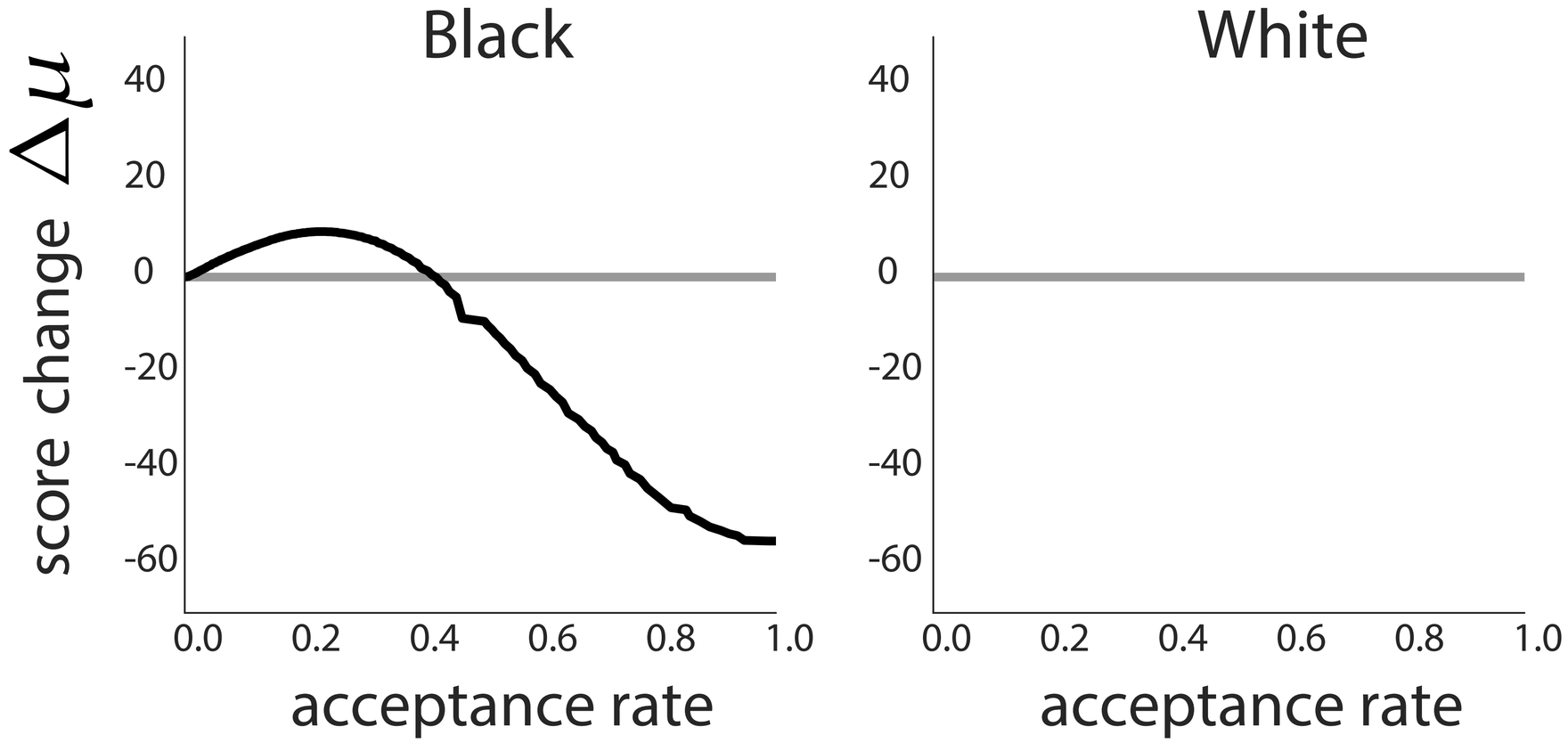
- Use empirical data labeled by race (“white” and “black”) to estimate group score distributions, repayment probabilities, and relative sizes
- Model the bank’s profit/loss ratio, e.g. +1:-4
- Model the delayed impact of repayment/default on credit score, e.g. +75/-150
- Compute “outcome curves” and delayed impact under different fairness criteria



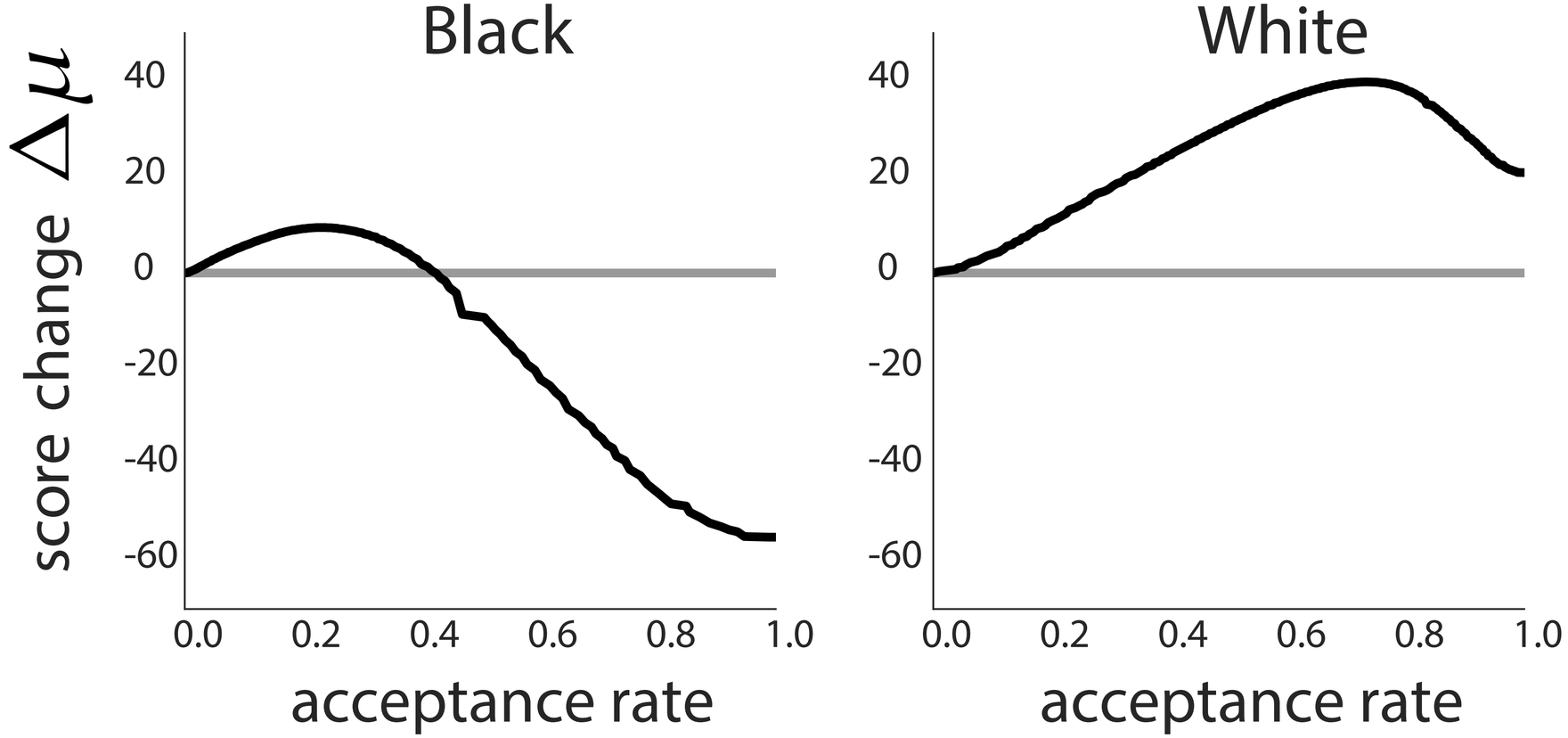
Outcome Curves



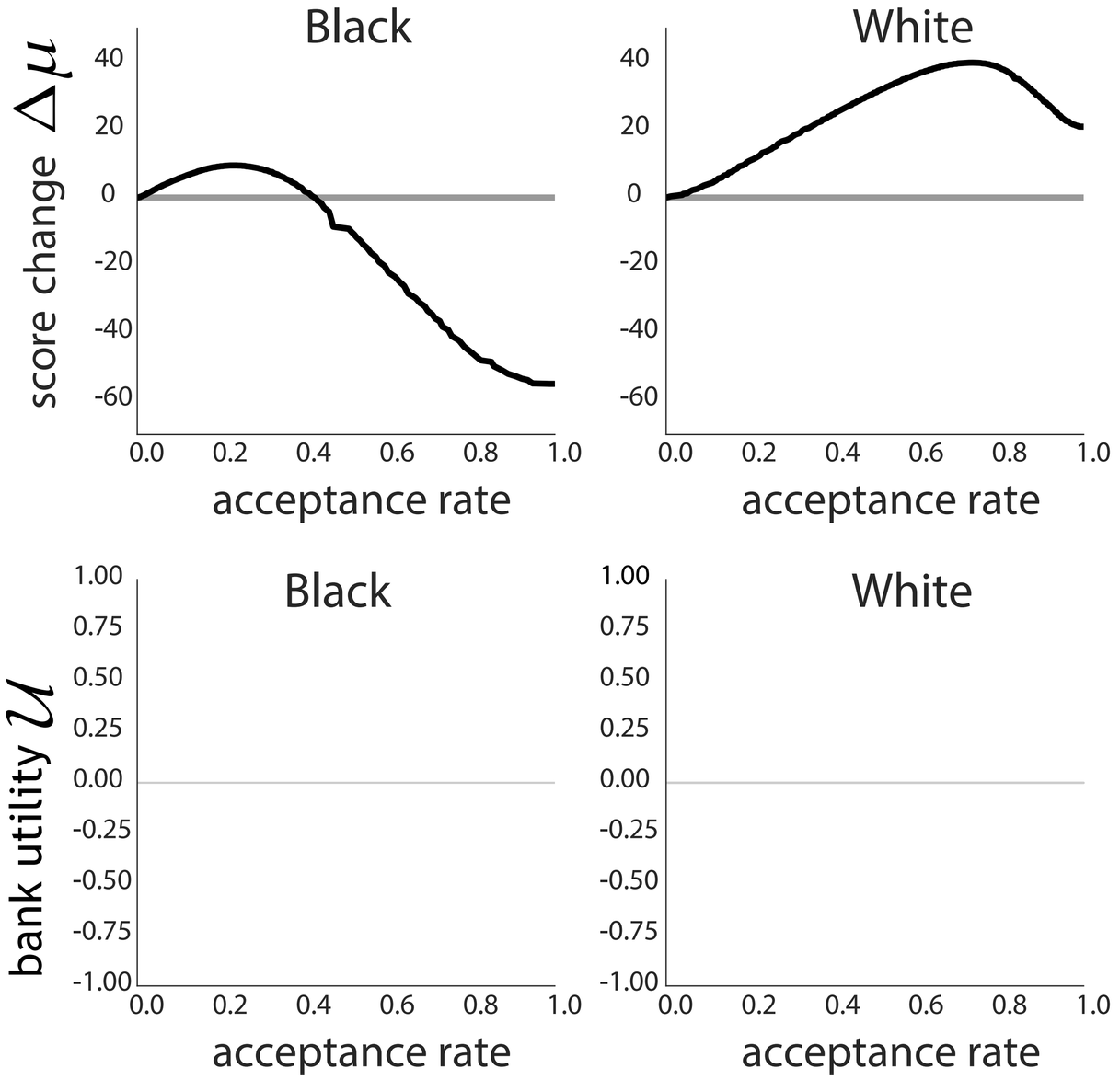
Outcome Curves



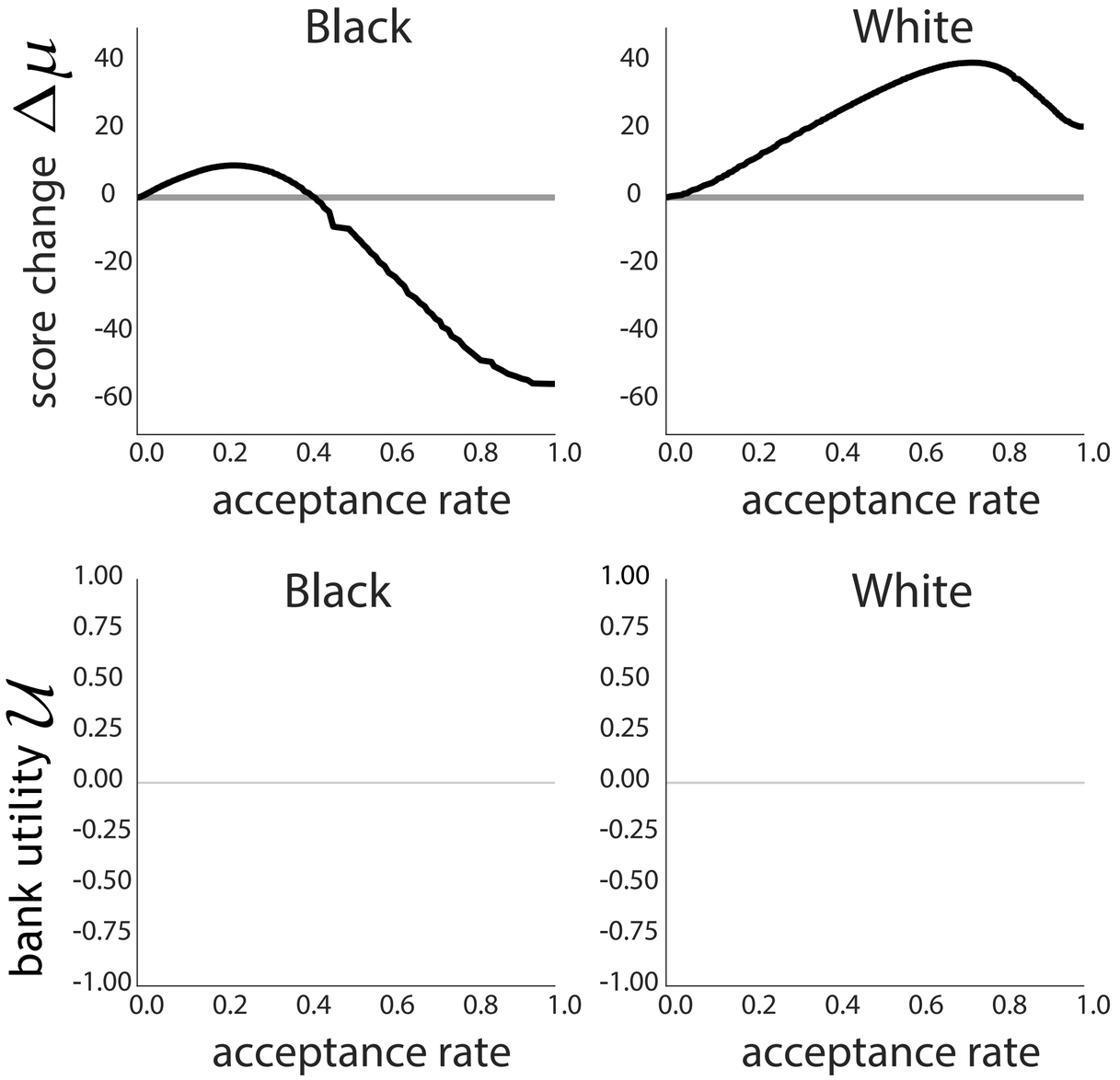
Outcome Curves



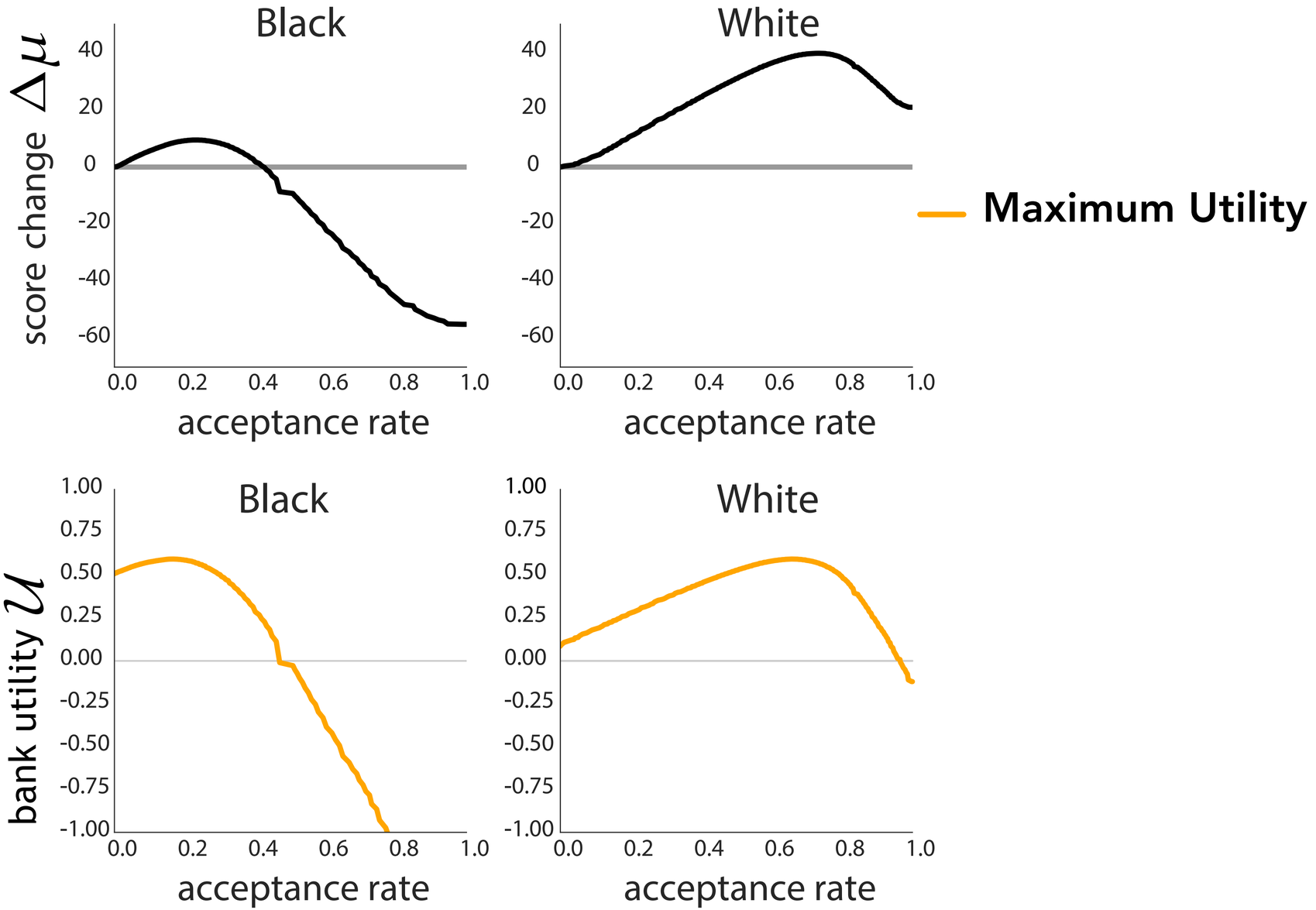
Outcome Curves



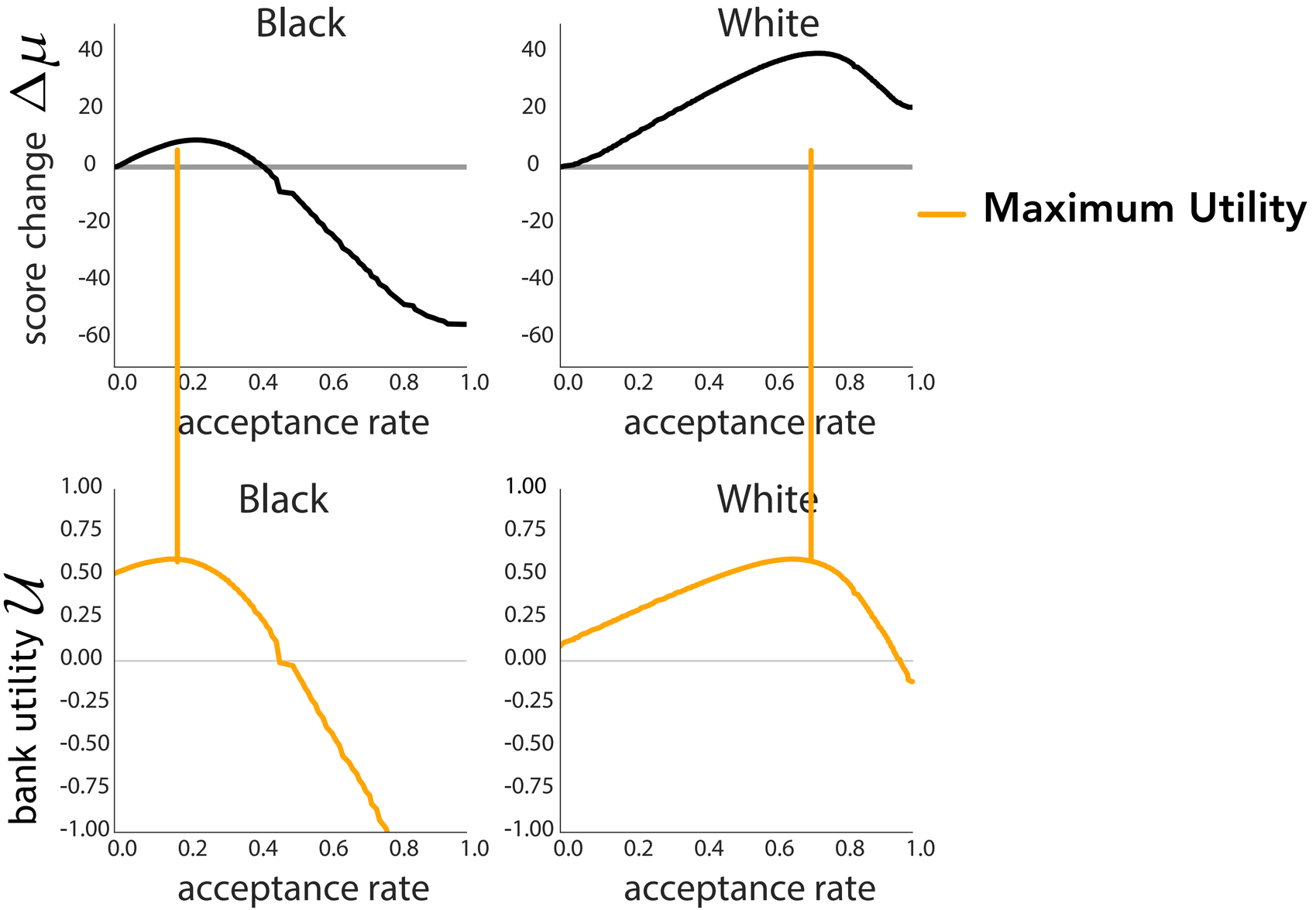
Outcome Curves



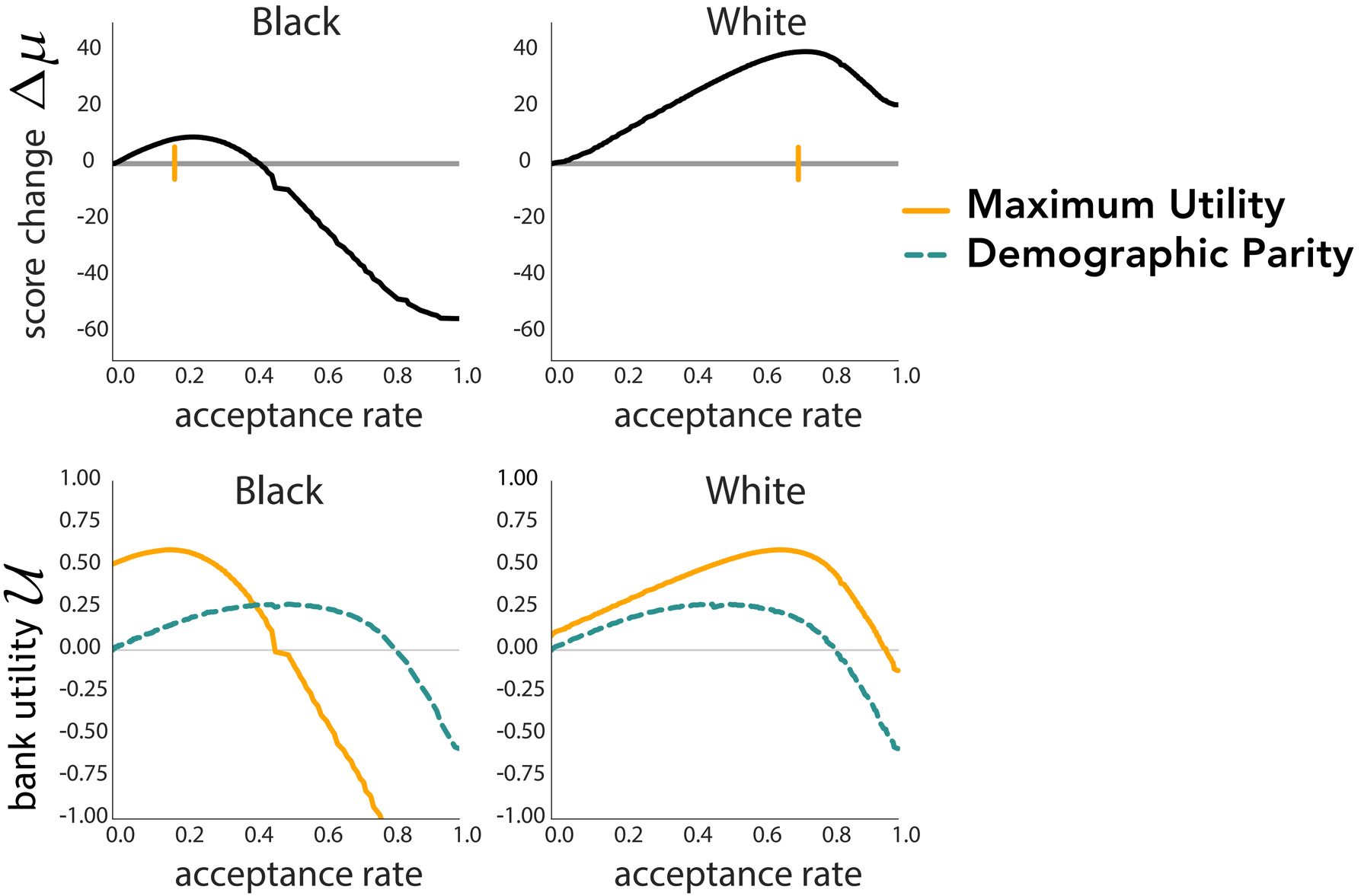
Outcome Curves



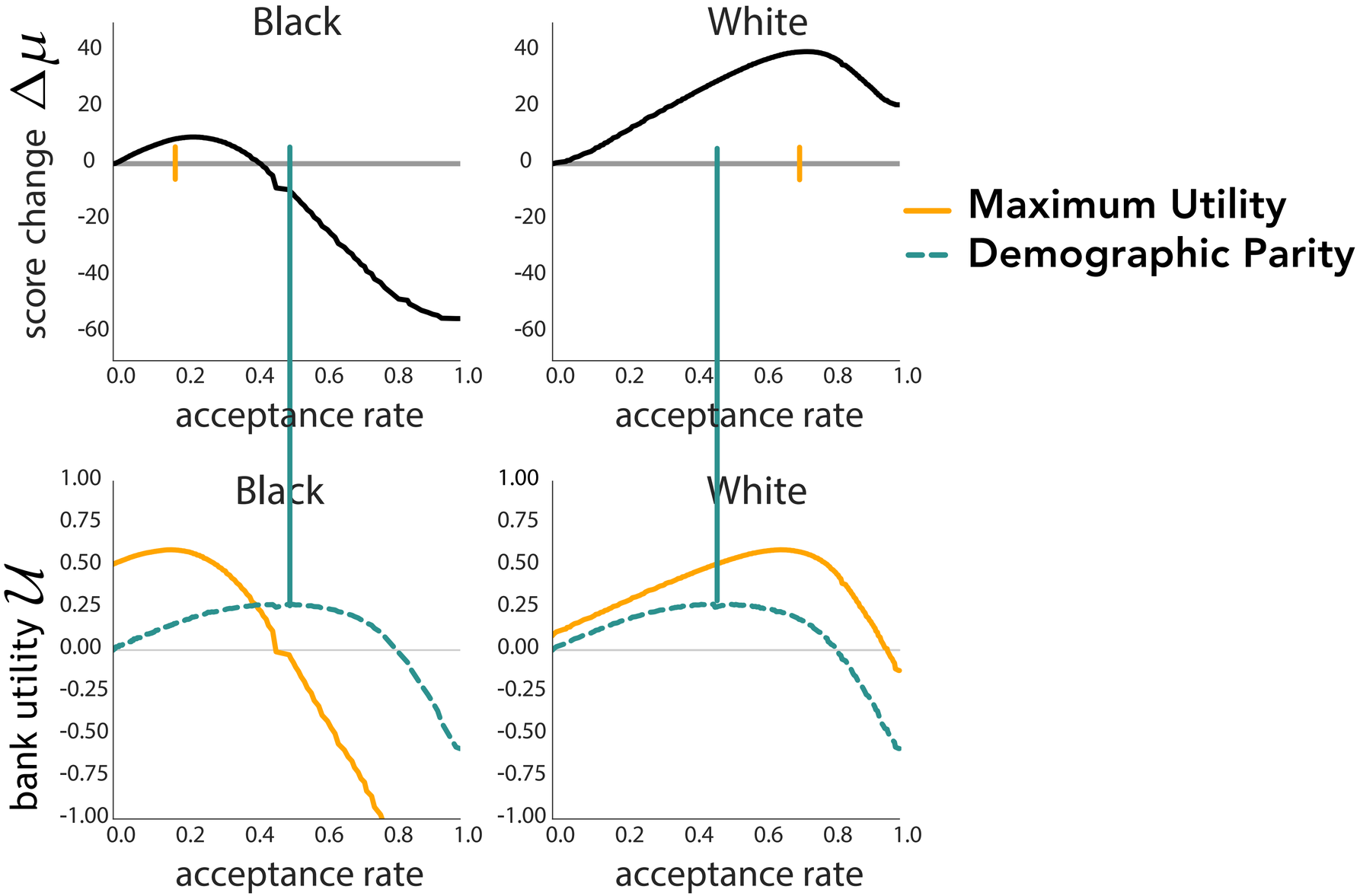
Outcome Curves



Outcome Curves

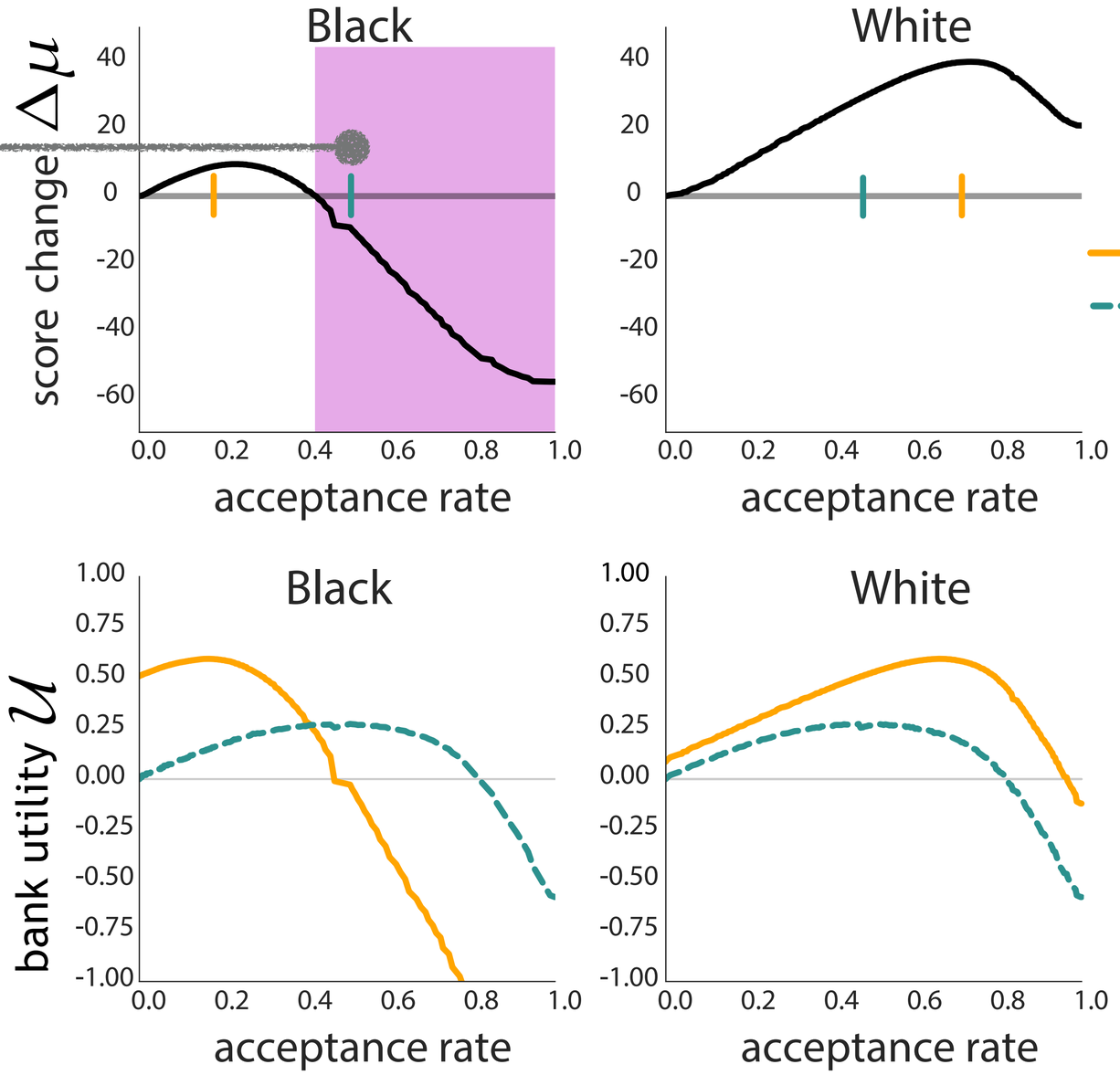


Outcome Curves



Outcome Curves

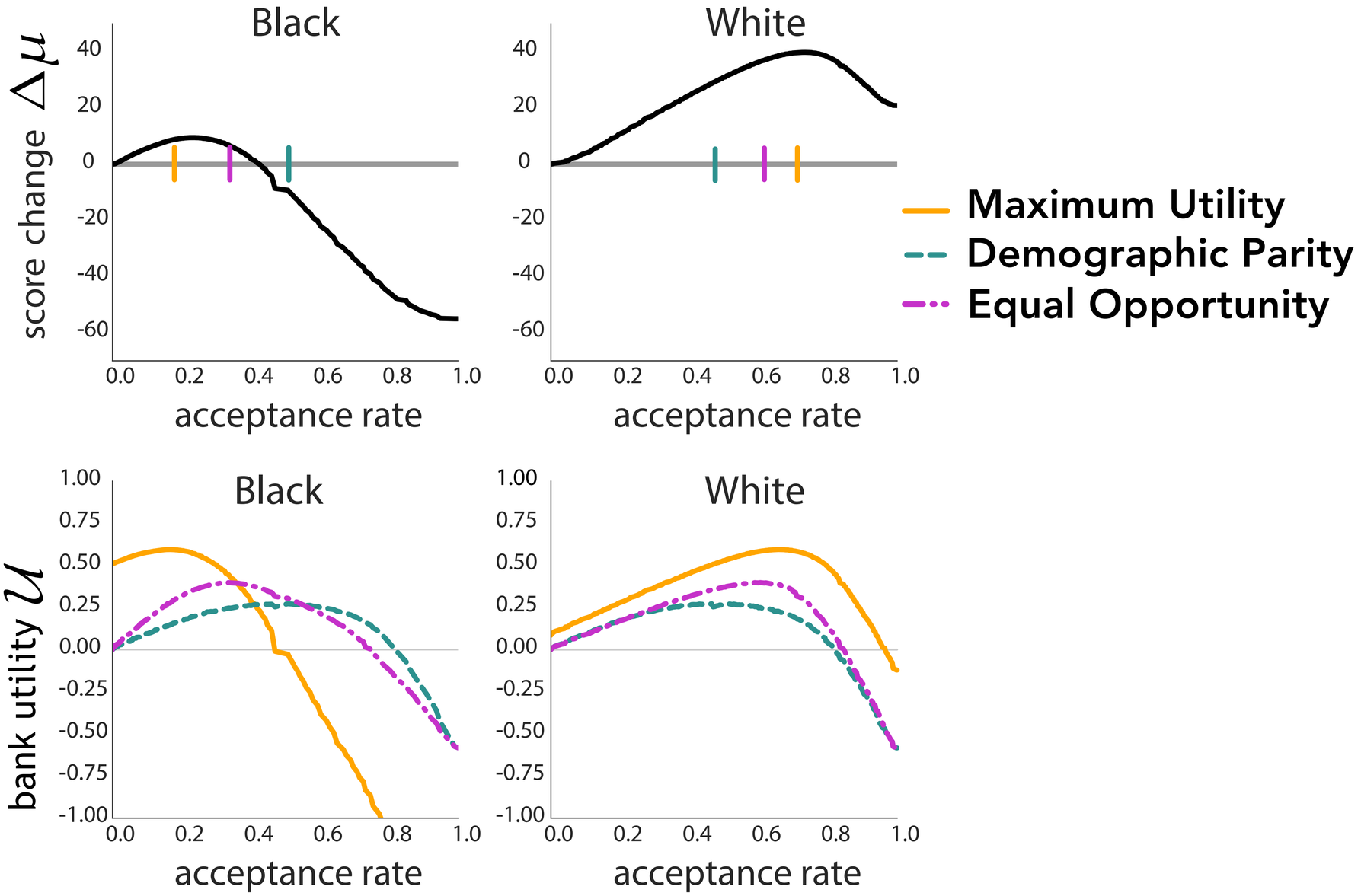
Active harm



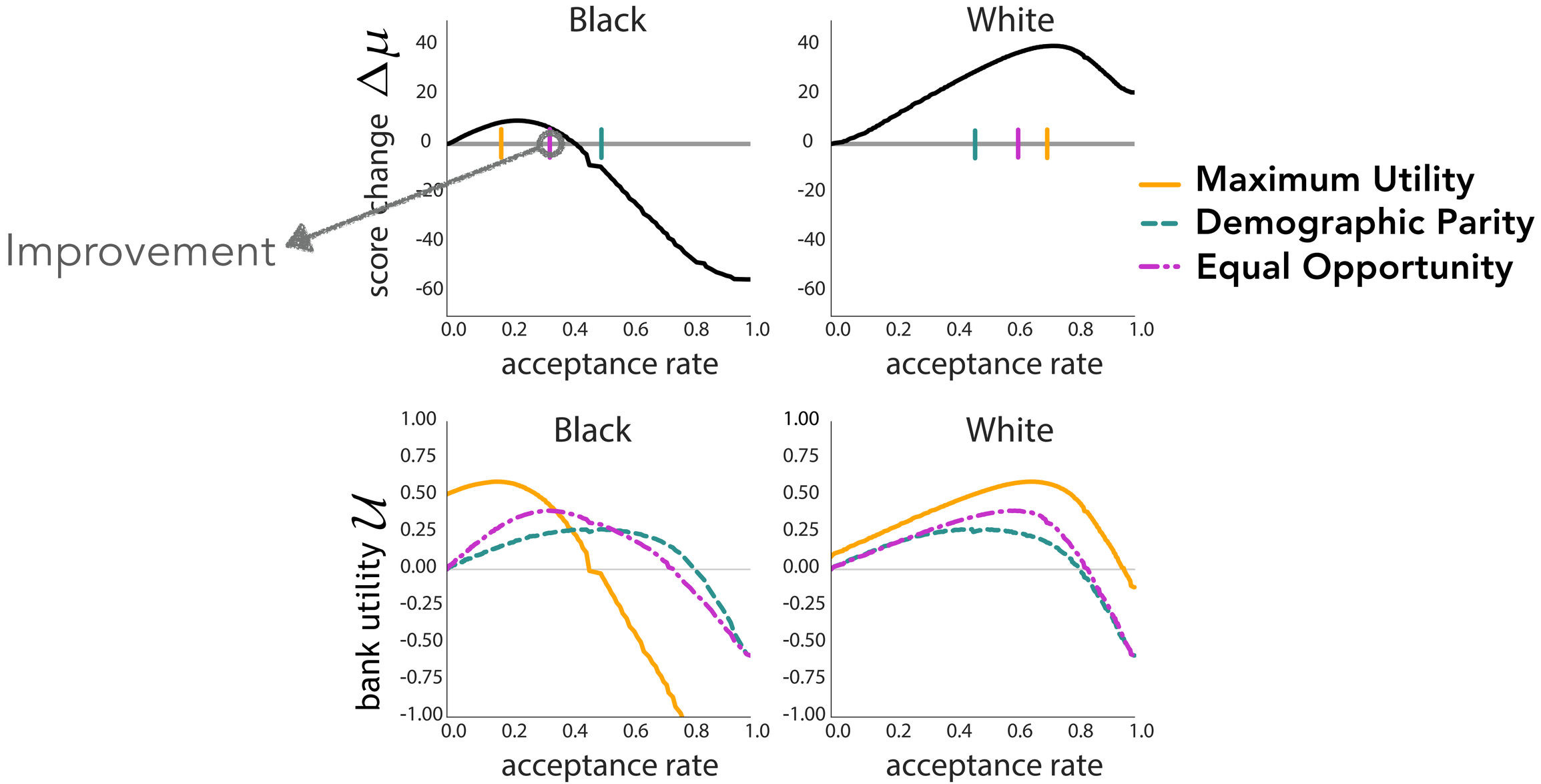
Maximum Utility

Demographic Parity

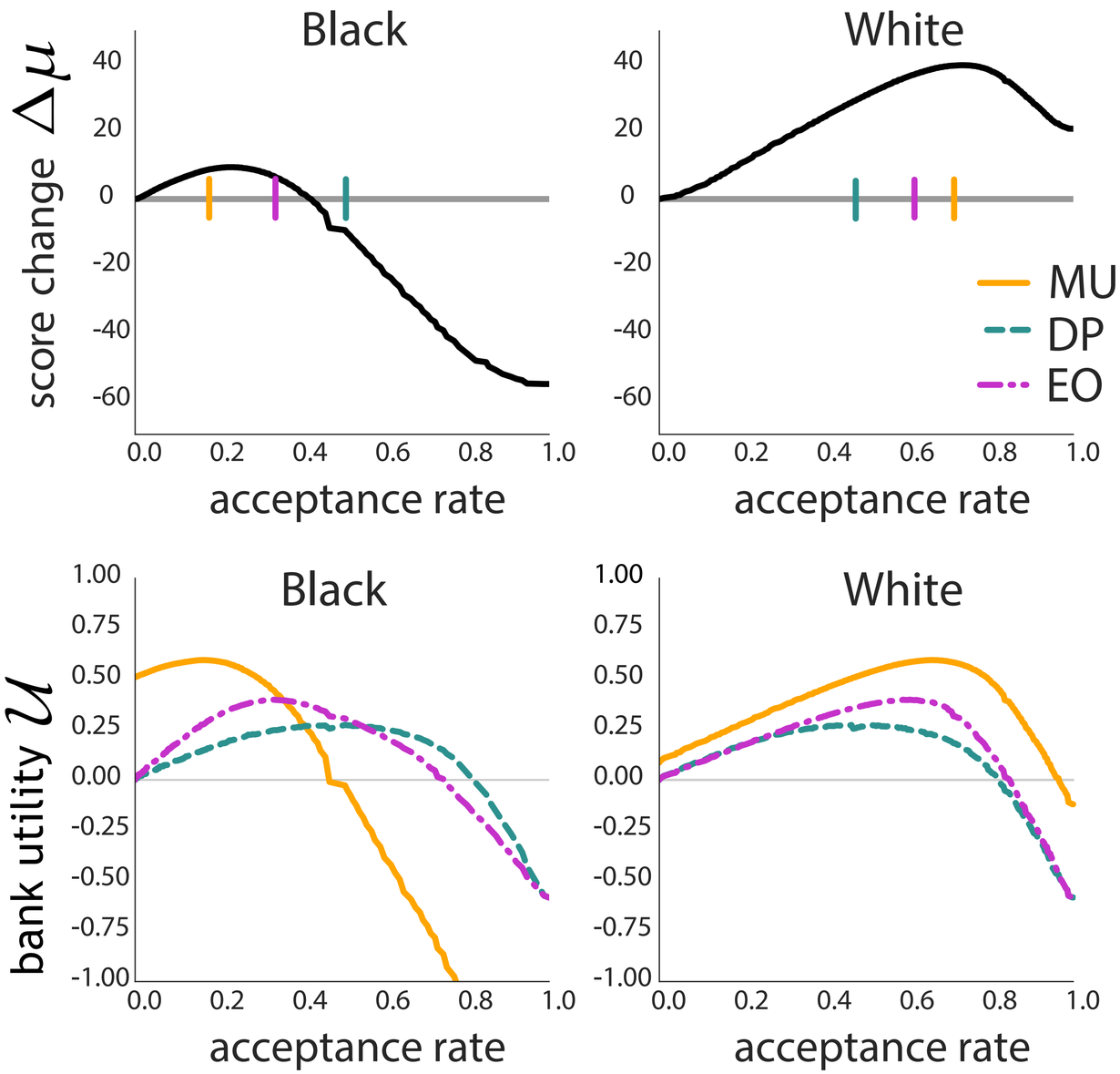
Outcome Curves



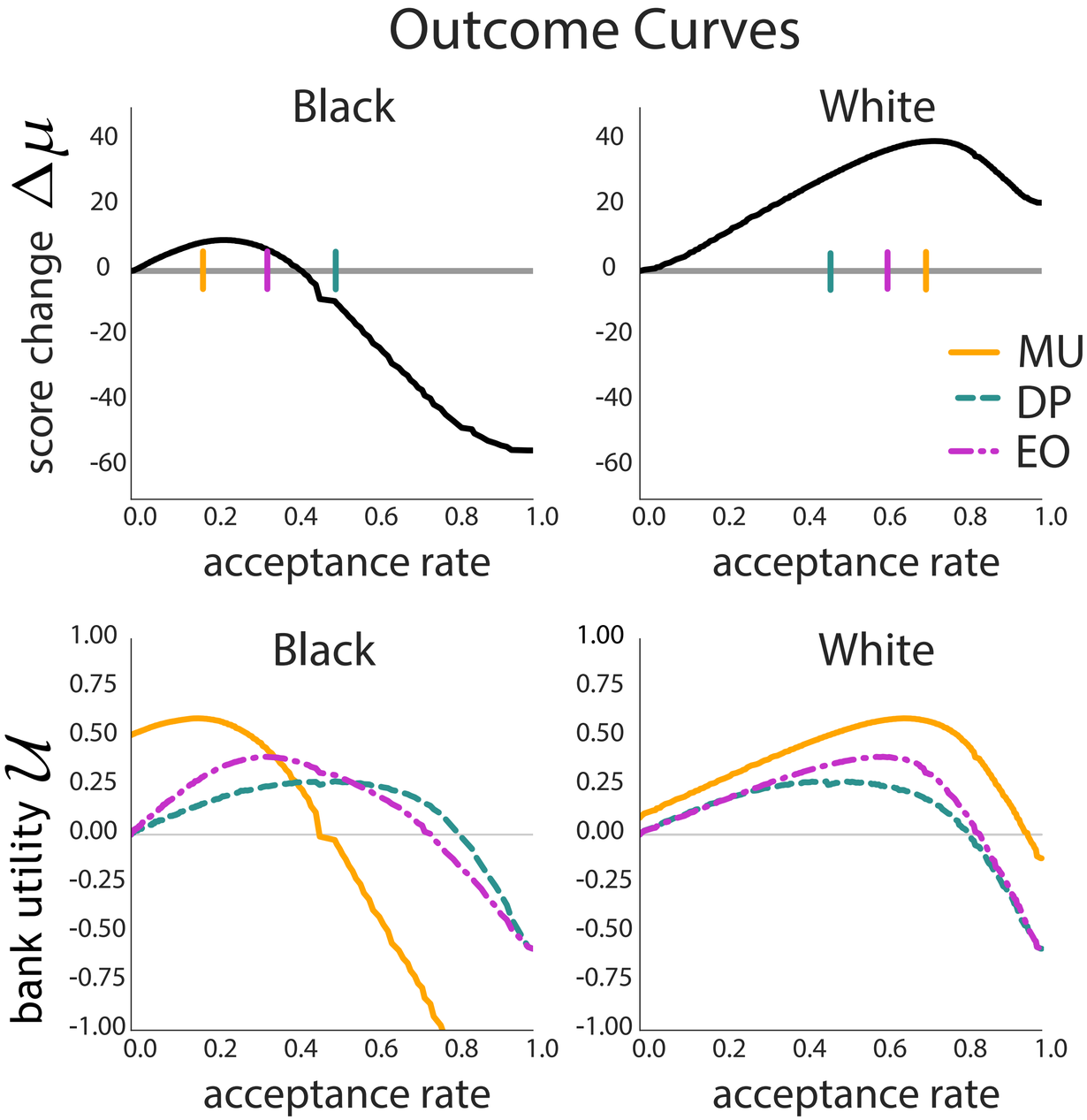
Outcome Curves



Outcome Curves

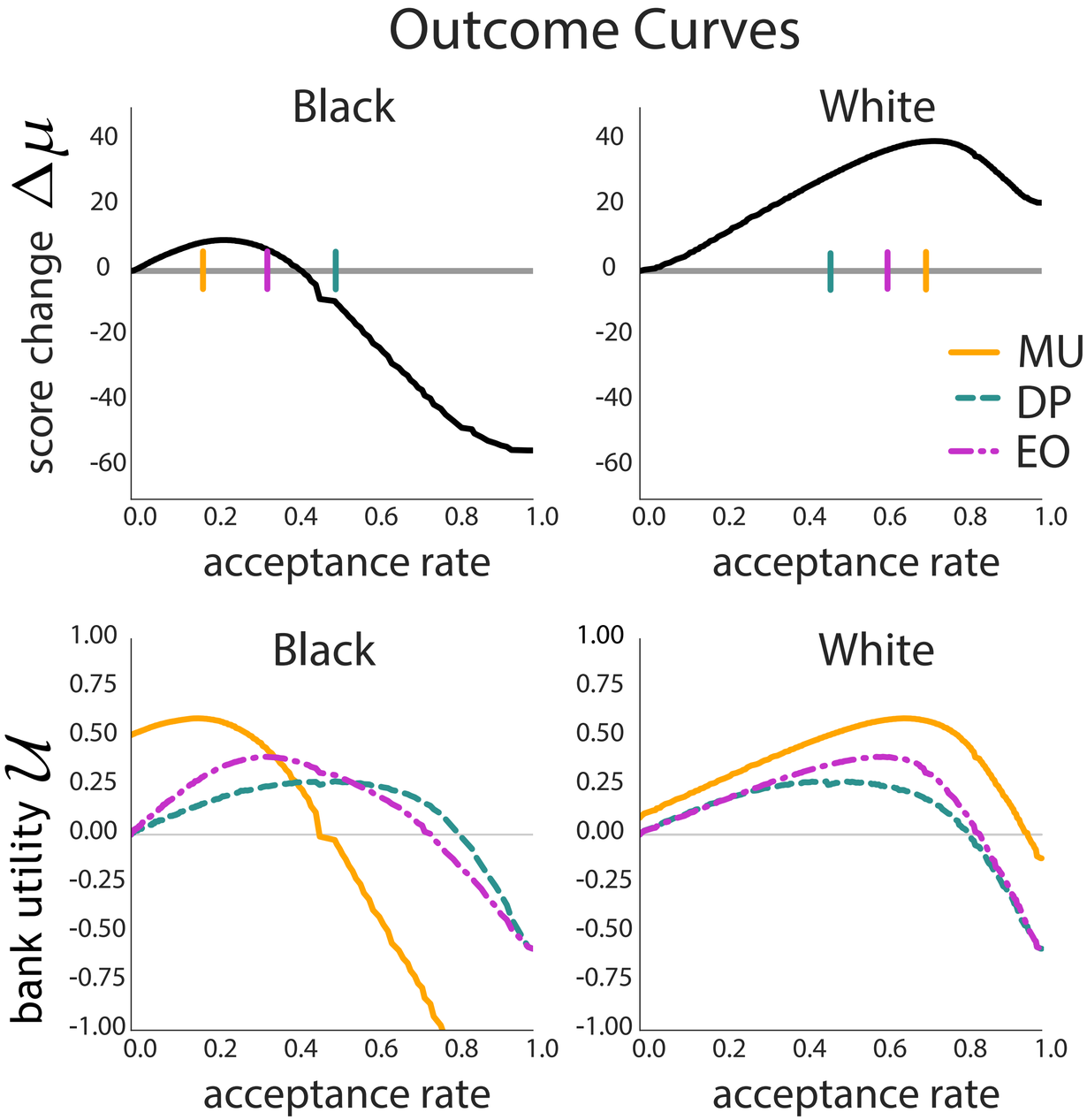


Why the large difference in delayed impact?



Why the large difference in delayed impact?

Maxima of *outcome* and *utility curves under fairness criteria* are **more misaligned** in the minority "black" group



DISCUSSION

DISCUSSION

- **Outcome curves** provide a way to deviate from maximum utility while improving outcomes.

DISCUSSION

- **Outcome curves** provide a way to deviate from maximum utility while improving outcomes.
- Need for **domain-specific** models of delayed impact
  - Context-sensitive nature of fairness in machine learning

DISCUSSION

- **Outcome curves** provide a way to deviate from maximum utility while improving outcomes.
- Need for **domain-specific** models of delayed impact
  - Context-sensitive nature of fairness in machine learning

FUTURE WORK

DISCUSSION

- **Outcome curves** provide a way to deviate from maximum utility while improving outcomes.
- Need for **domain-specific** models of delayed impact
  - Context-sensitive nature of fairness in machine learning

FUTURE WORK

- Moving beyond **binary** decisions

DISCUSSION

- **Outcome curves** provide a way to deviate from maximum utility while improving outcomes.
- Need for **domain-specific** models of delayed impact
  - Context-sensitive nature of fairness in machine learning

FUTURE WORK

- Moving beyond **binary** decisions
- Moving beyond the **mean** score as measure of impact

DISCUSSION

- **Outcome curves** provide a way to deviate from maximum utility while improving outcomes.
- Need for **domain-specific** models of delayed impact
  - Context-sensitive nature of fairness in machine learning

FUTURE WORK

- Moving beyond **binary** decisions
- Moving beyond the **mean** score as measure of impact
- **Dynamics** of the **distributional impact** of machine learning algorithms
  [Ensign et al. 2017; Hu and Chen 2017; Hashimoto et al. 2018]

Thank you!

Poster today | 06:15 -- 09:00 PM | **Hall B #213**

Full version at arxiv.org/abs/**1803.04383**

DELAYED IMPACT OF FAIR MACHINE LEARNING

Lydia T. Liu (UC Berkeley)




Joint work with **Sarah Dean, Esther Rolf, Max Simchowitz, Moritz Hardt**