Causal Conceptions of Fairness and their Consequences

•••

Hamed Nilforoshan*, Johann Gaebler*, Ravi Shroff, Sharad Goel

hamedn@cs.stanford.edu jgaeb@stanford.edu ravi.shroff@nyu.edu sgoel@hks.harvard.edu

(* equal contribution)

[ACIC 2022 / ICML 2022]

Summary

- Unified taxonomy to understand *causal fairness* research field

Summary

- Unified taxonomy to understand *causal fairness* research field
- Prominent causal conceptions of algorithmic fairness, if implemented, can harm the groups they were designed to protect

Stylized Example: College Admissions

Stylized Example: College Admissions

Stylized Example: College Admissions

Test Score			
73			
65			
80			
…			

Stylized Example: College Admissions

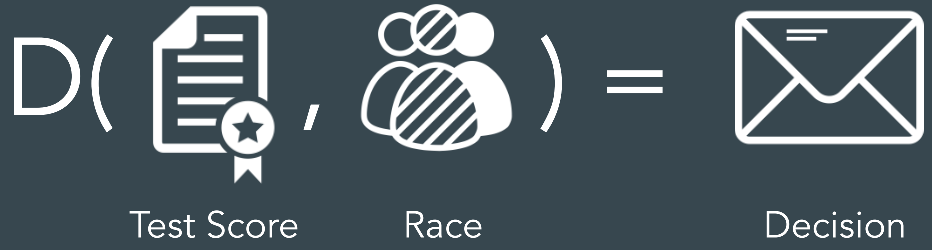
Test Score	Race Group		
73	Minority		
65	Majority		
80	Minority		
…	…		

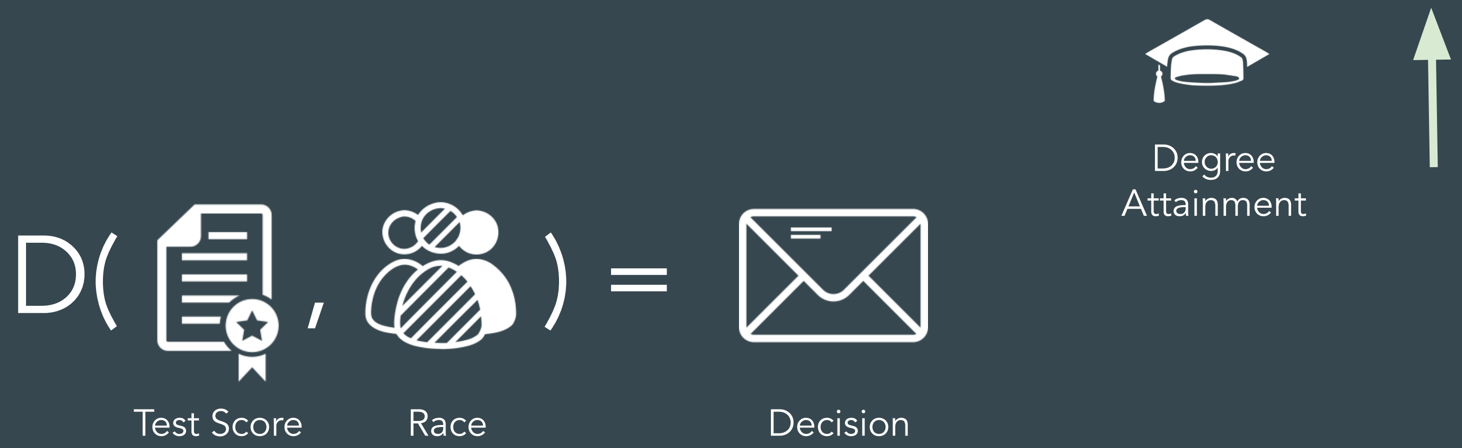
Stylized Example: College Admissions

Test Score	Race Group	Decision	
73	Minority		
65	Majority		
80	Minority		
…	…	…	

Stylized Example: College Admissions

Test Score	Race Group	Decision	Degree Attainment
73	Minority		
65	Majority		
80	Minority		
…	…	…	…


D(,) =
Test Score
Race
Decision


D(,) =
Test Score
Race
Decision
Degree
Attainment

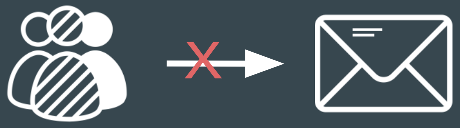

D(,) =
Test Score
Race
Decision
Degree
Attainment
Class Diversity

How to ensure that *D* is fair?

[Part 1: *causal fairness* overview + taxonomy]

Traditional fairness definitions

Anti-classification

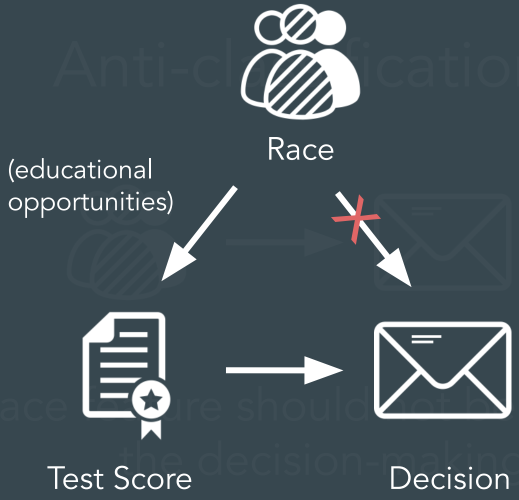


Race Decision

Race feature should not be used in the decision-making

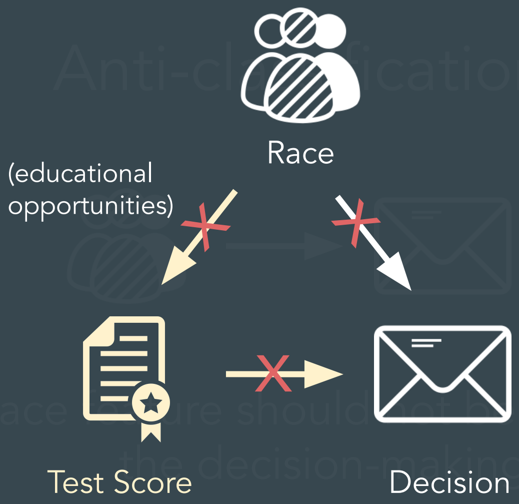
D(=95, =Minority) =
D(=95, =Majority)

Causal Fairness Motivation



Race may still *indirectly* affect decisions

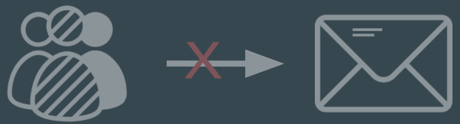
Causal Fairness Taxonomy



Family 1: Limit direct and indirect effects of race on decision

Traditional fairness definitions

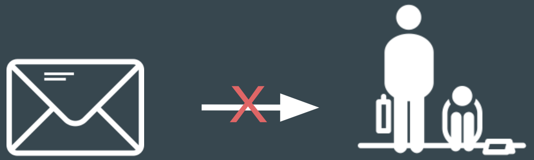
Anti-classification



Race feature should not be used in the decision-making

D(=95, =Minority) =
D(=95, =Majority)

Classification parity

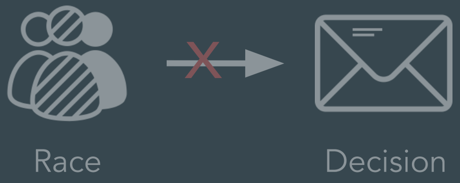


Model performance should be the same across groups

Precision = % of admits who successfully obtain a bachelor's degree

Traditional fairness definitions

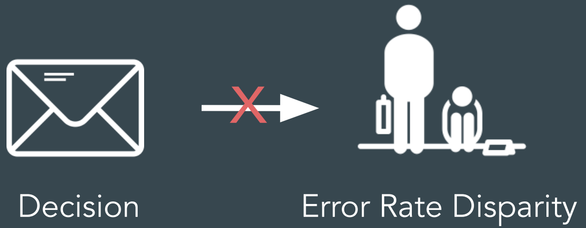
Anti-classification



Race feature should not be used in the decision-making

D(=95, =Minority) =
D(=95, =Majority)

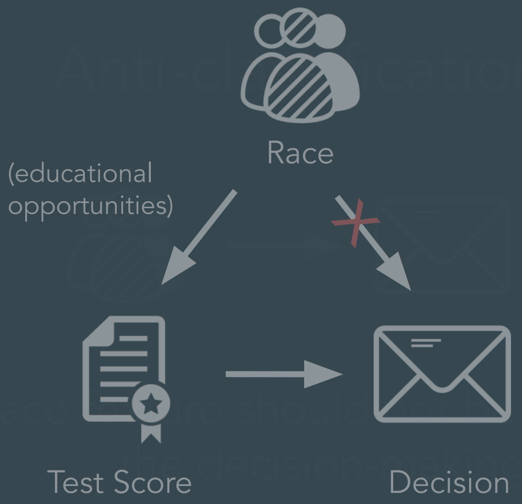
Classification parity



Model performance should be the same across groups

Minority group precision =
Majority group precision

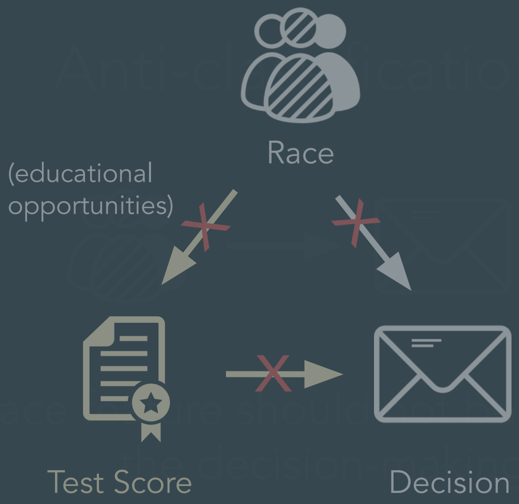
Causal Fairness Motivation



Race may still *indirectly* affect decisions

Decisions may affect graduation, altering error rates

Causal Fairness Taxonomy



Family 1: Limit direct and indirect effects of race on decision

Family 2: Model performance should be counterfactually equal between groups

Causal fairness taxonomy [see paper]

Family 1: Limit direct and indirect effects of race on decision

- Counterfactual fairness
- Path-specific fairness

Family 2: Limit counterfactual disparities between groups

- Counterfactual equalized odds
- Counterfactual predictive parity
- Principal fairness

Causal fairness taxonomy [see paper]

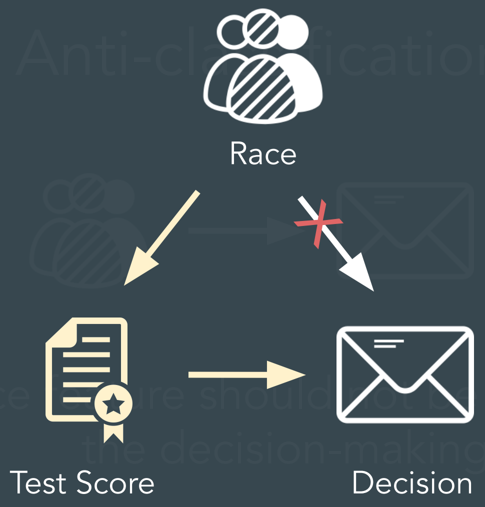
Family 1: Limit direct and indirect effects of race on decision

- Counterfactual fairness
- Path-specific fairness

Family 2: Limit counterfactual disparities between groups

- Counterfactual equalized odds
- Counterfactual predictive parity
- Principal fairness

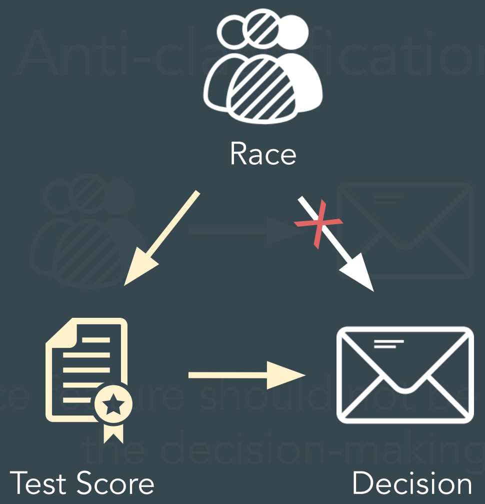
Counterfactual Fairness



Family 1: Limit direct and indirect effects of race on decision

Given a subset of applicants with the exact same feature values, admissions rate should not change *in a counterfactual world in which they belonged to a different race group*

Counterfactual Fairness

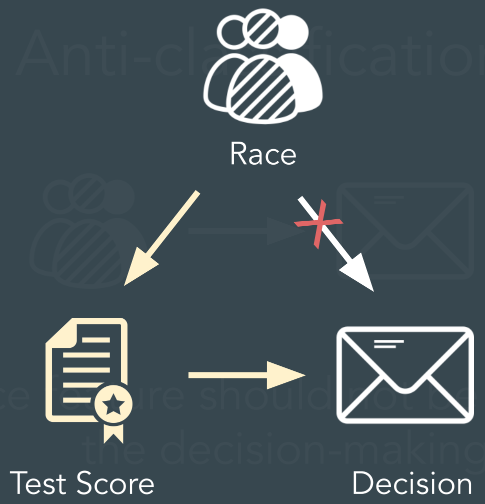


Family 1: Limit direct and indirect effects of race on decision

Given a subset of applicants with the exact same feature values, admissions rate should not change *in a counterfactual world in which they belonged to a different race group*

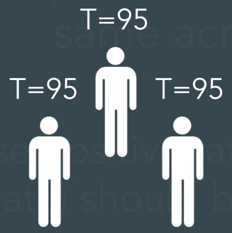
[Important caveat: counterfactuals of race are epistemologically problematic]

Counterfactual Fairness



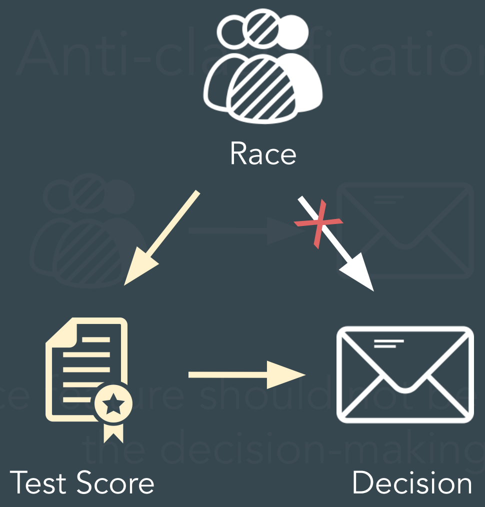
Family 1: Limit direct and indirect effects of race on decision

Given a subset of applicants with the exact same feature values, admissions rate should not change *in a counterfactual world in which they belonged to a different race group*



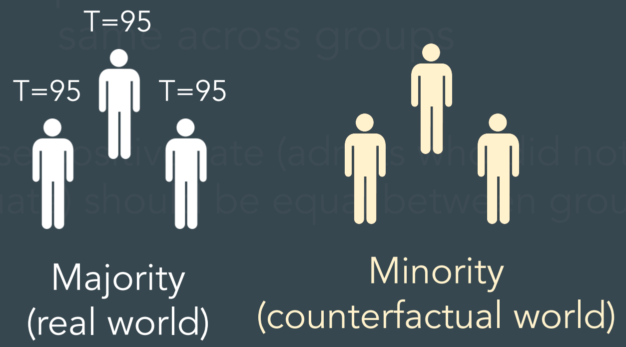
Majority
(real world)

Counterfactual Fairness

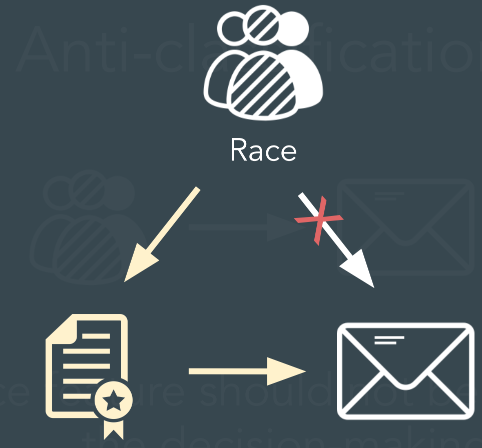


Family 1: Limit direct and indirect effects of race on decision

Given a subset of applicants with the exact same feature values, admissions rate should not change *in a counterfactual world in which they belonged to a different race group*

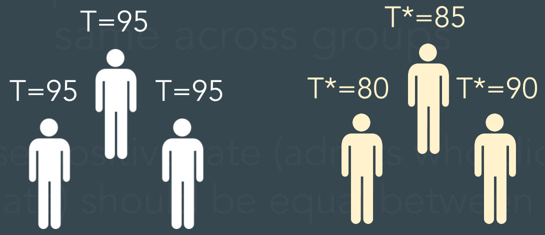


Counterfactual Fairness



Family 1: Limit direct and indirect effects of race on decision

Given a subset of applicants with the exact same feature values, admissions rate should not change *in a counterfactual world in which they belonged to a different race group*

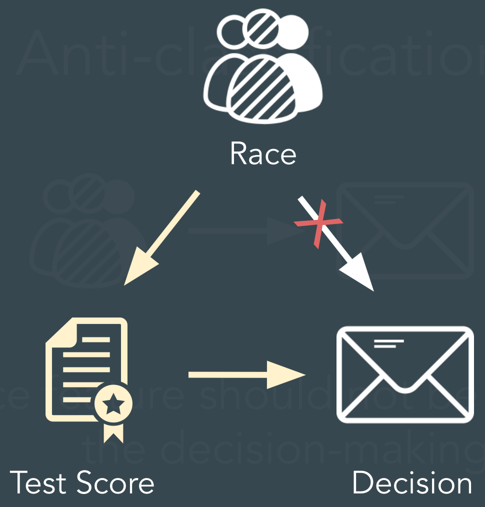


Majority (real world)

Minority (counterfactual world)

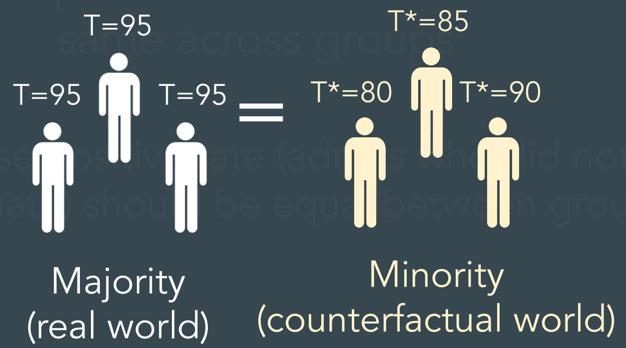
[T* decreases due to reduced access to educational opportunities]

Counterfactual Fairness



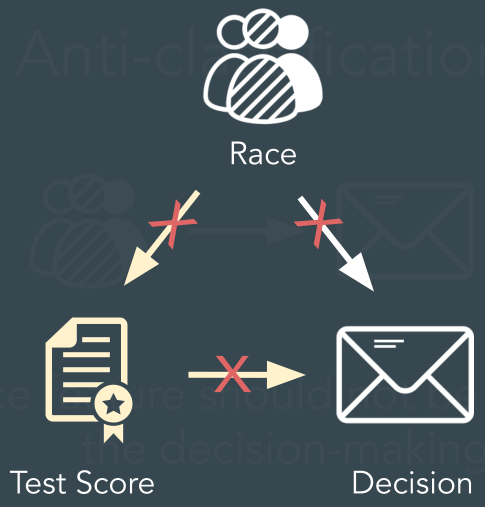
Family 1: Limit direct and indirect effects of race on decision

Given a subset of applicants with the exact same feature values, admissions rate should not change *in a counterfactual world in which they belonged to a different race group*



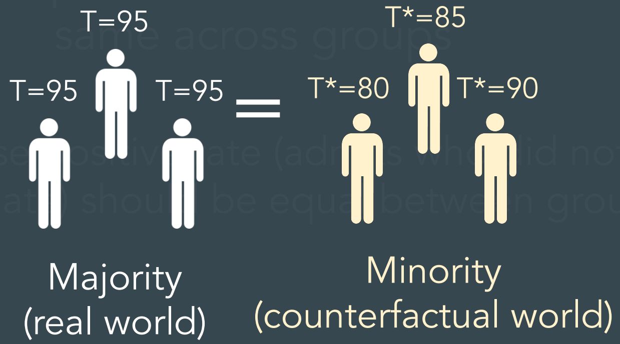
[T* decreases due to reduced access to educational opportunities]

Counterfactual Fairness



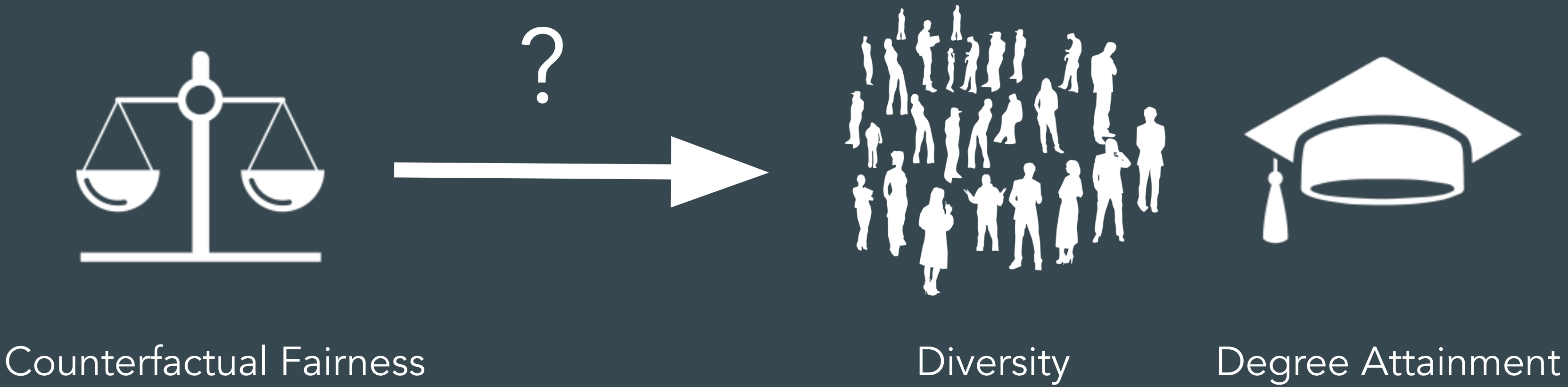
Family 1: Limit direct and indirect effects of race on decision

Given a subset of applicants with the exact same feature values, admissions rate should not change *in a counterfactual world in which they belonged to a different race group*

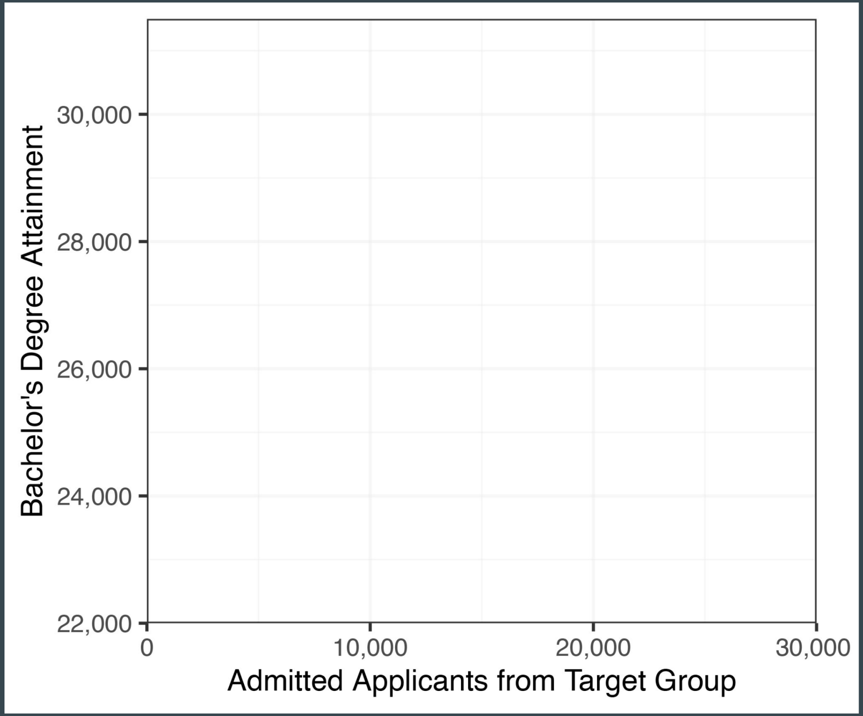


[T* decreases due to reduced access to educational opportunities]

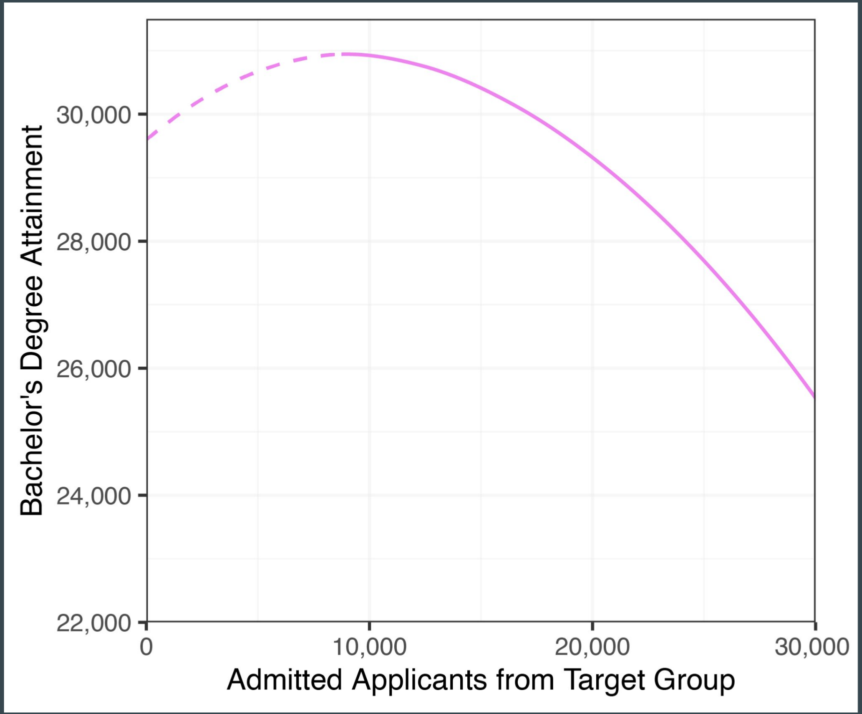
Part 2: What are the downstream consequences of causal fairness?



Illustrative example

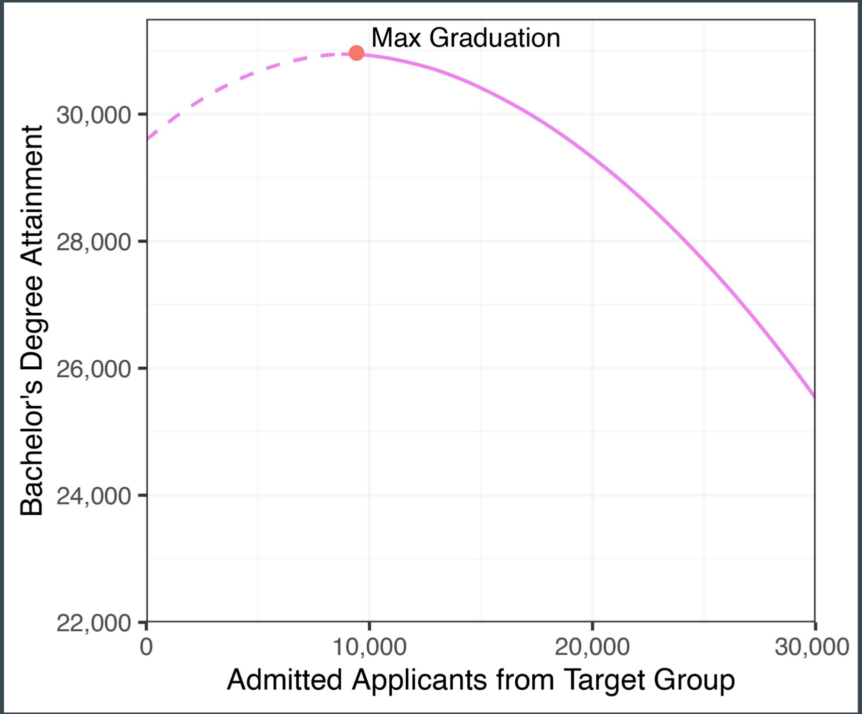


Illustrative example



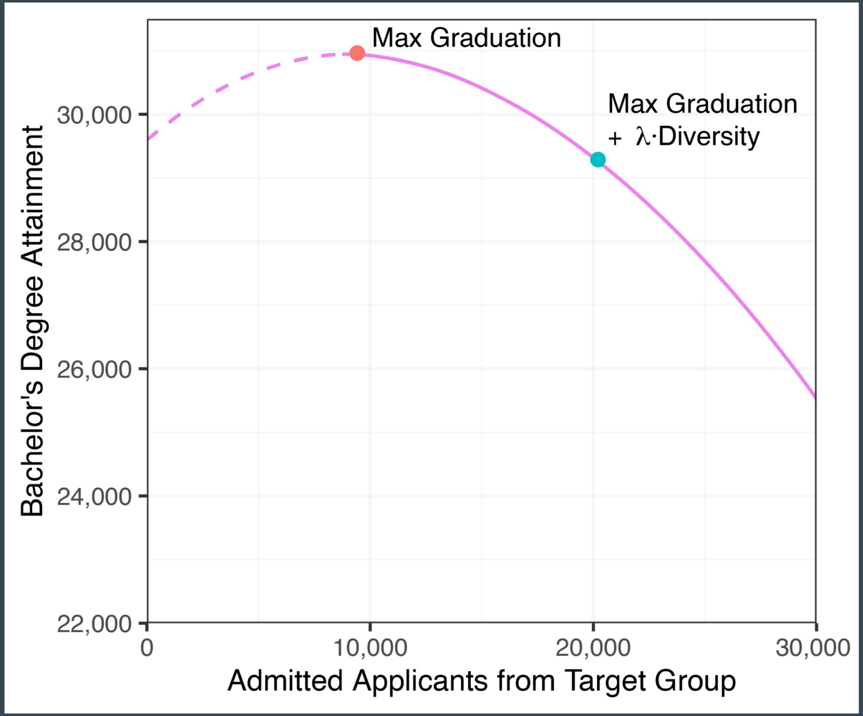
Pareto frontier: different people trade off degree attainment and diversity differently

Illustrative example



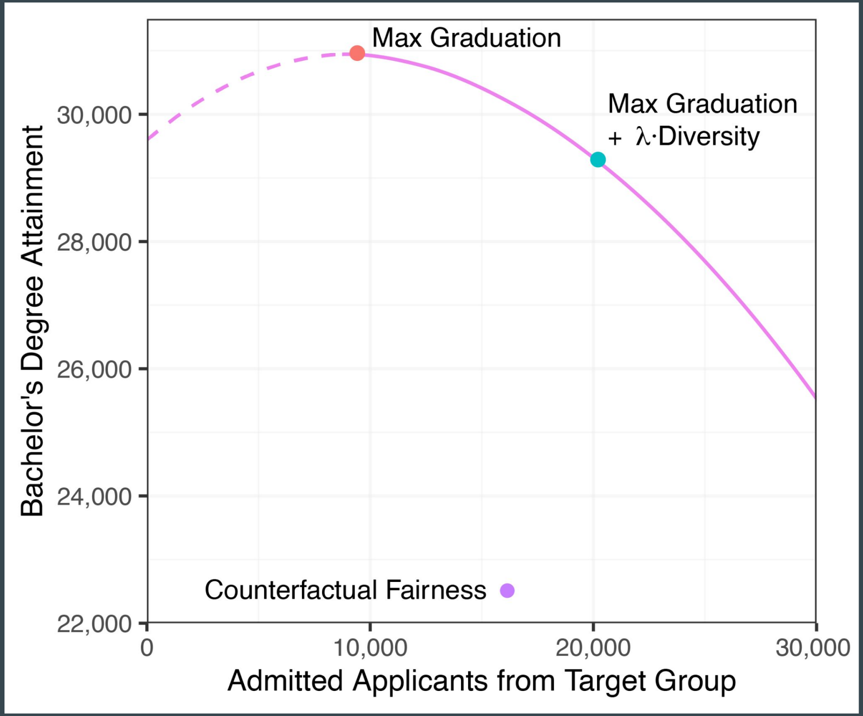
Pareto frontier: different people trade off degree attainment and diversity differently

Illustrative example

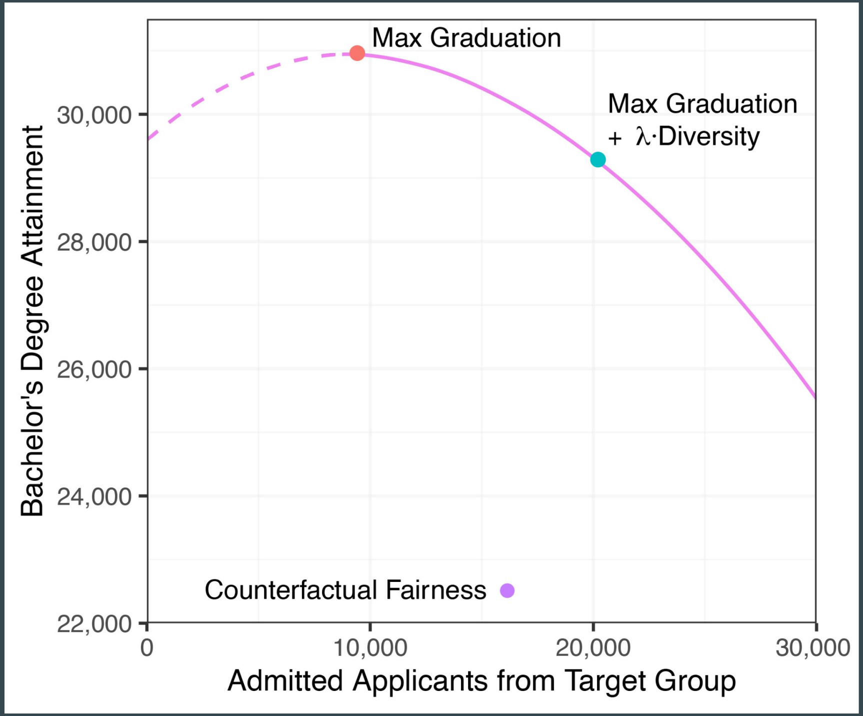


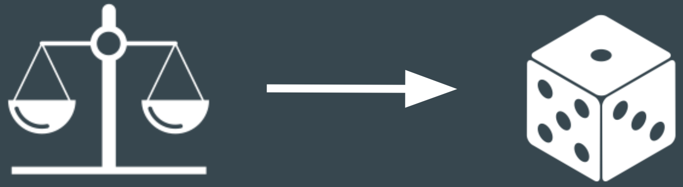
Pareto frontier: different people trade off degree attainment and diversity differently

Illustrative example



Illustrative example

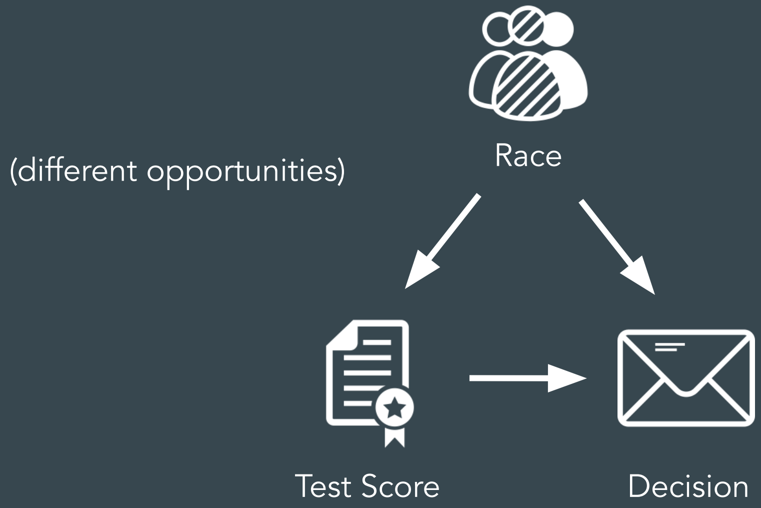




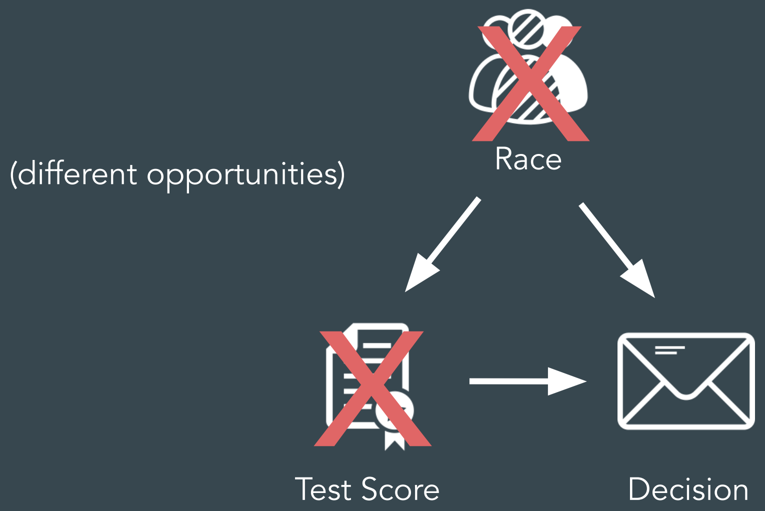
Counterfactual Fairness

Randomized Lottery

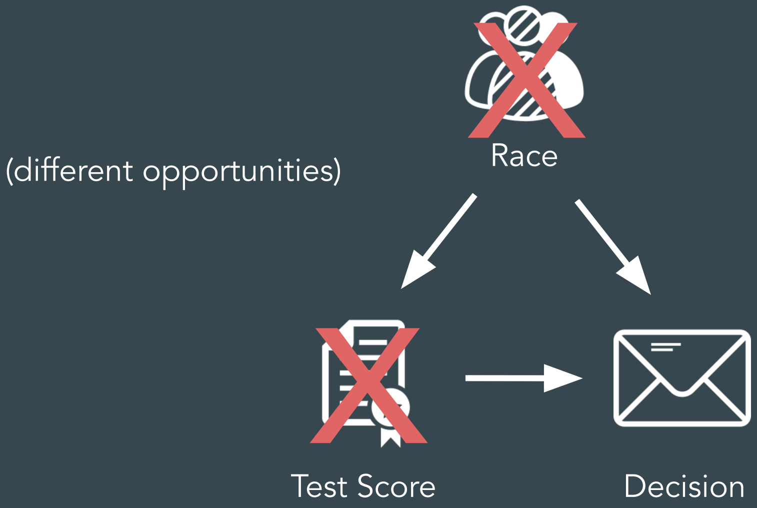
Theoretical result: Under mild assumptions, counterfactual fairness requires decisions to ignore race and all downstream covariates

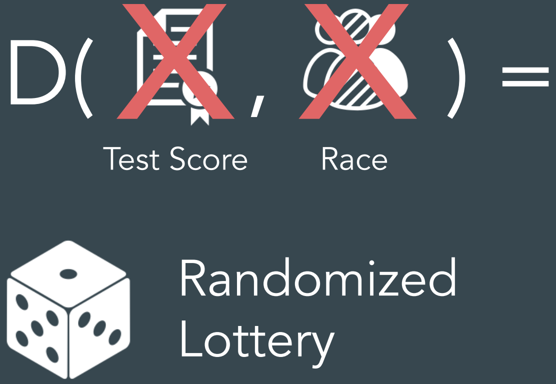


Theoretical result: Under mild assumptions, counterfactual fairness requires decisions to ignore race and all downstream covariates

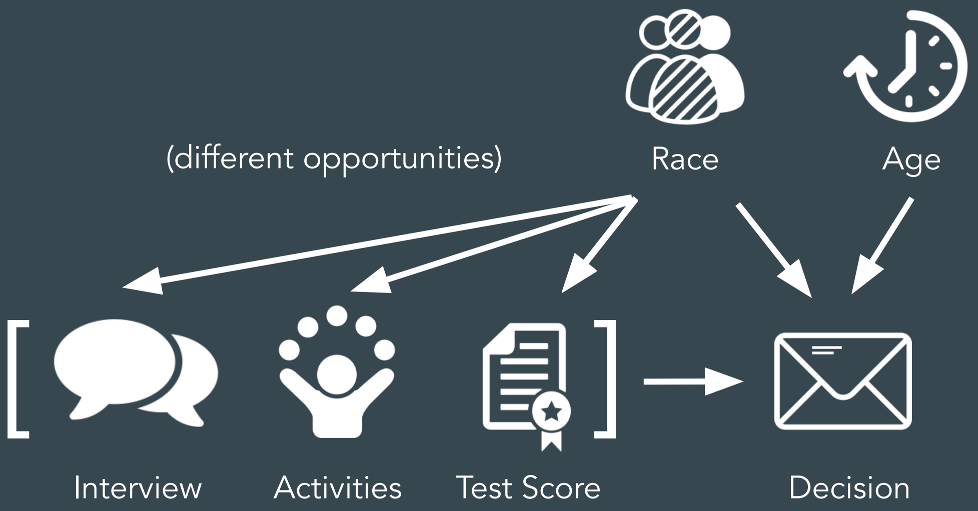


<u>Theoretical result</u>: Under mild assumptions, counterfactual fairness requires decisions to ignore race and all downstream covariates

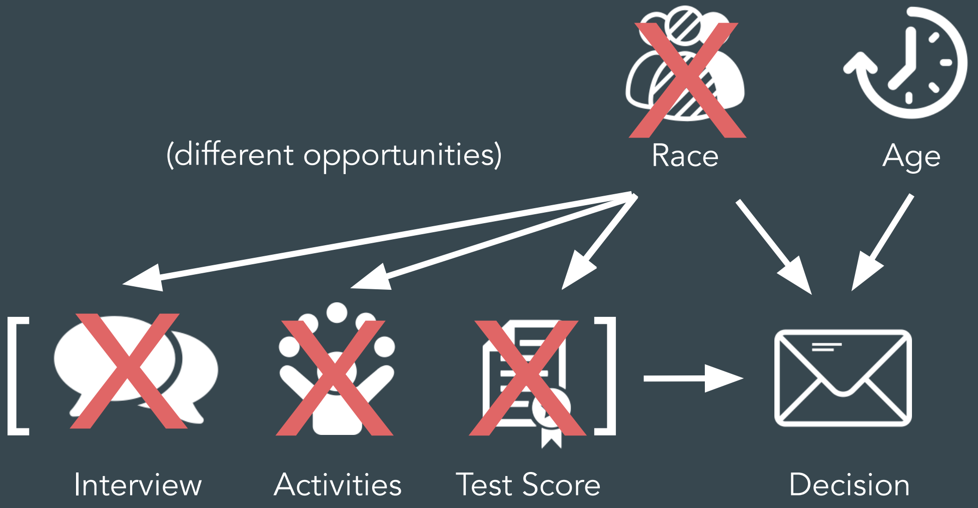




<u>Theoretical result</u>: Under mild assumptions, counterfactual fairness requires decisions to ignore race and all downstream covariates



Theoretical result: Under mild assumptions, counterfactual fairness requires decisions to ignore race and all downstream covariates



Decisions based exclusively on age

Proof sketch

D(T = Low, Race = Majority)

D(T = Low, Race = Minority)

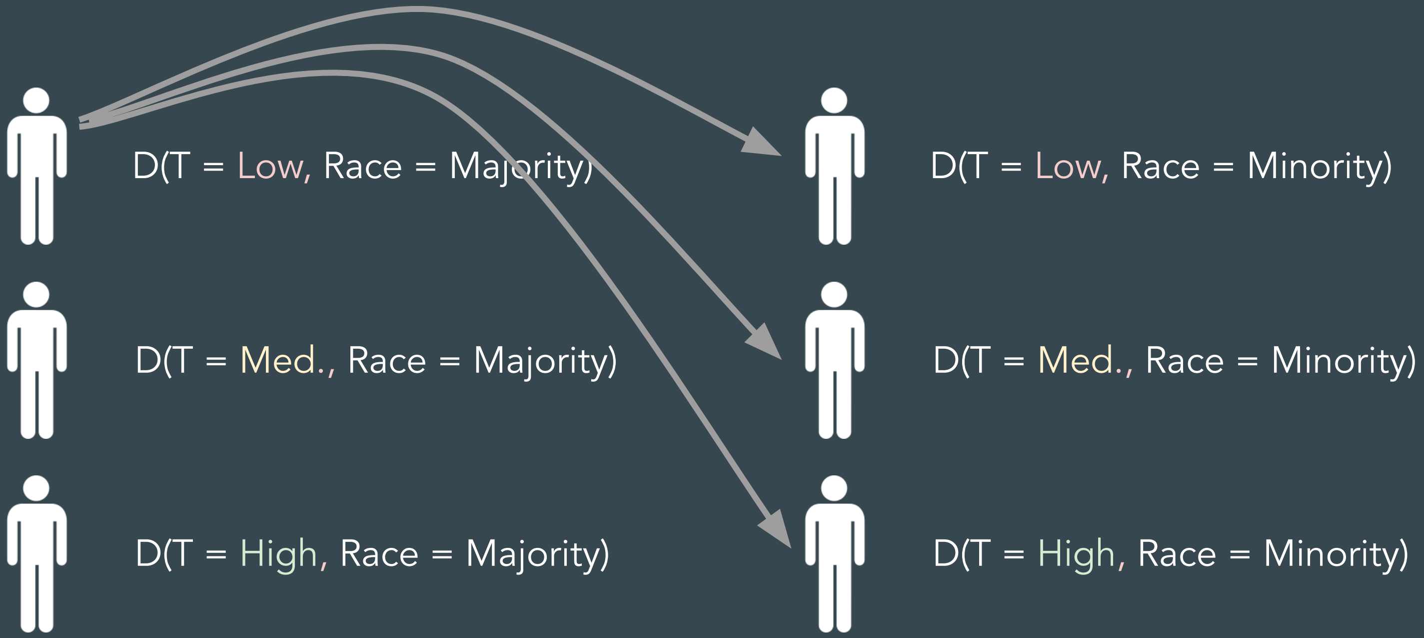
D(T = Med., Race = Majority)

D(T = Med., Race = Minority)

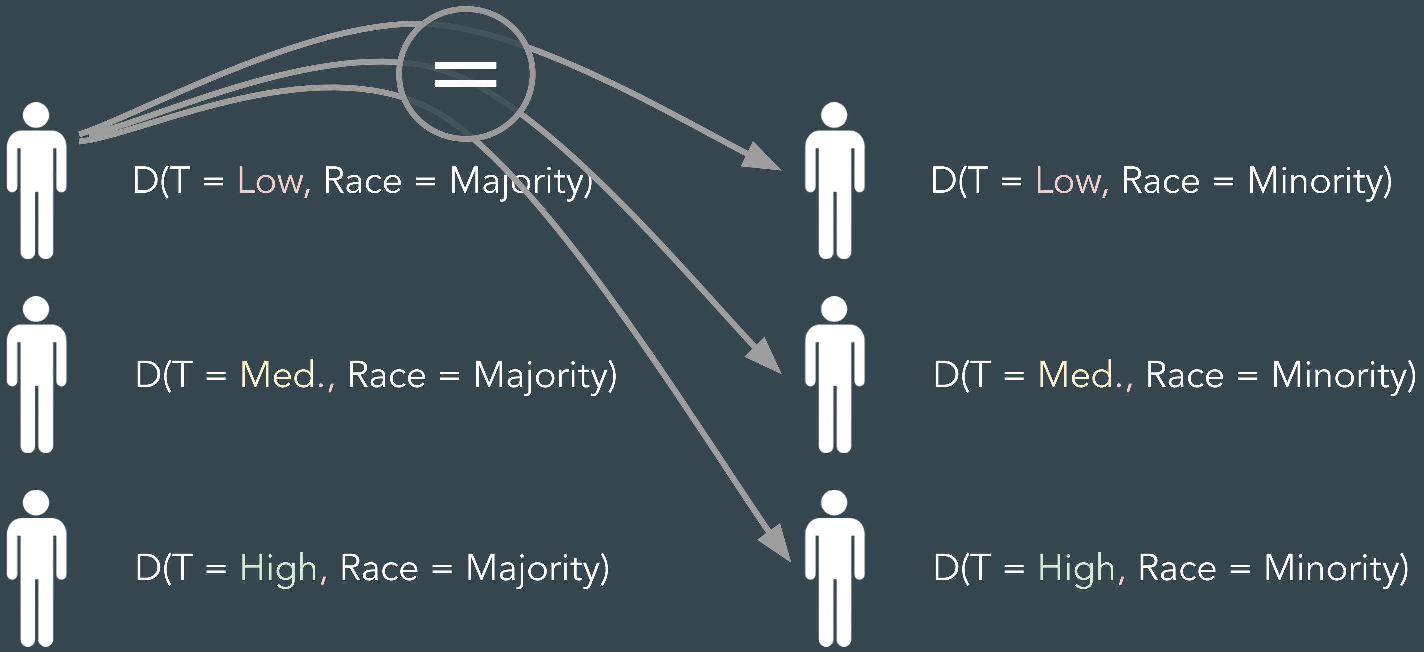
D(T = High, Race = Majority)

D(T = High, Race = Minority)

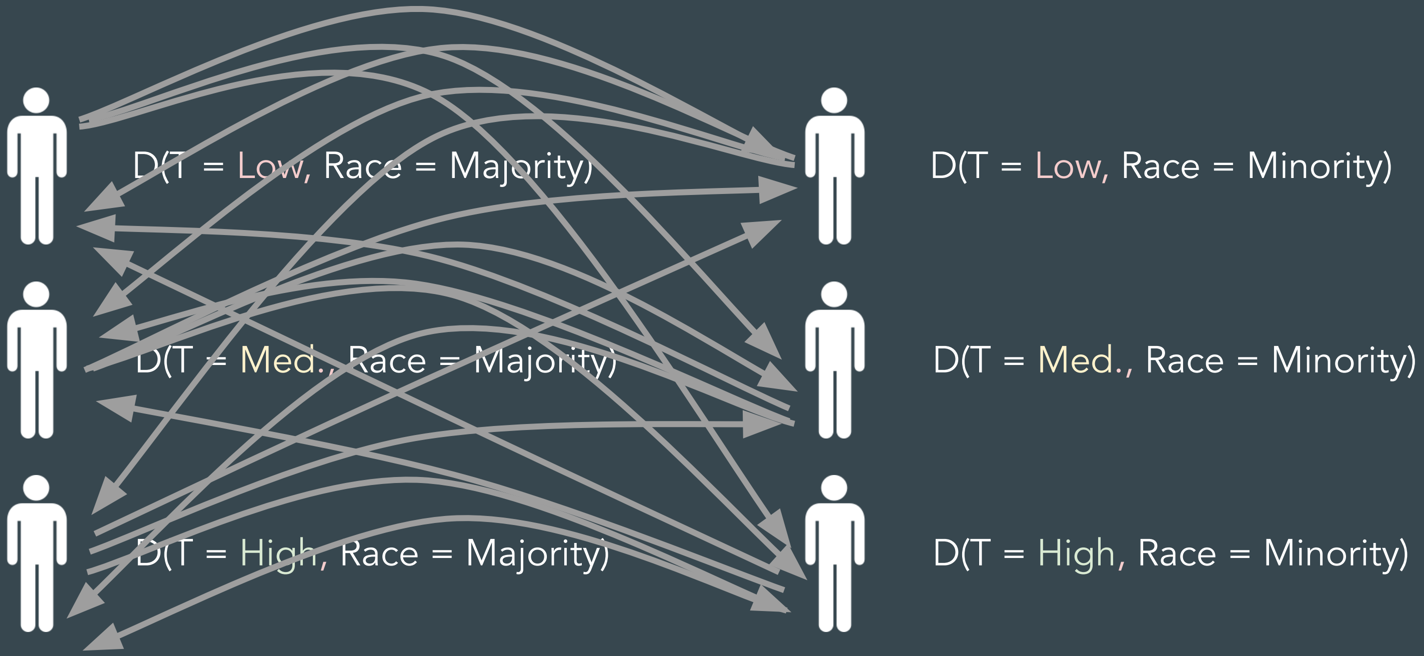
Proof sketch



Proof sketch



Proof sketch



Causal fairness taxonomy [see paper]

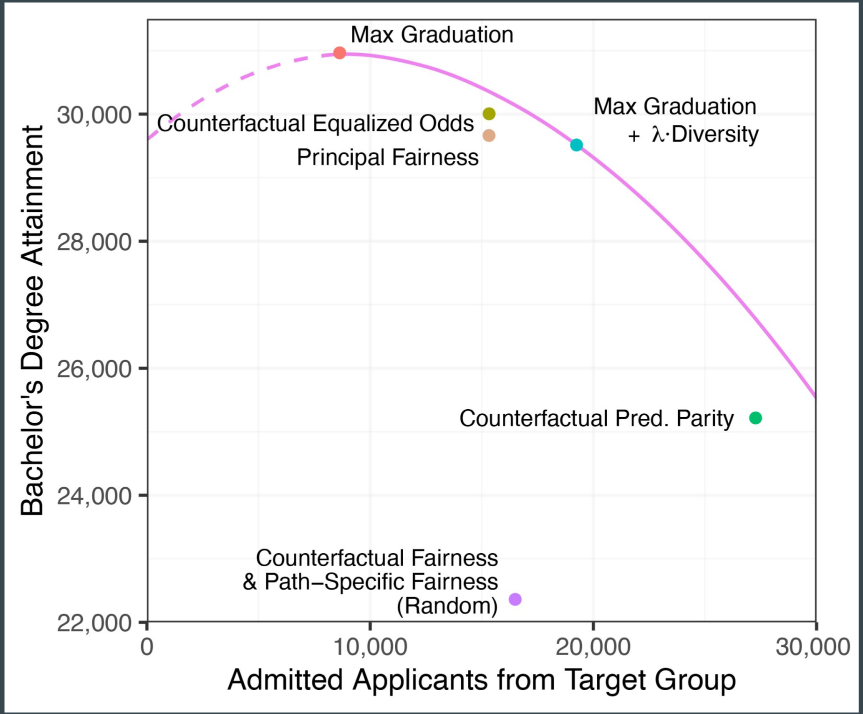
Family 1: Limit direct and indirect effects of race on decision

- Counterfactual fairness
- Path-specific fairness

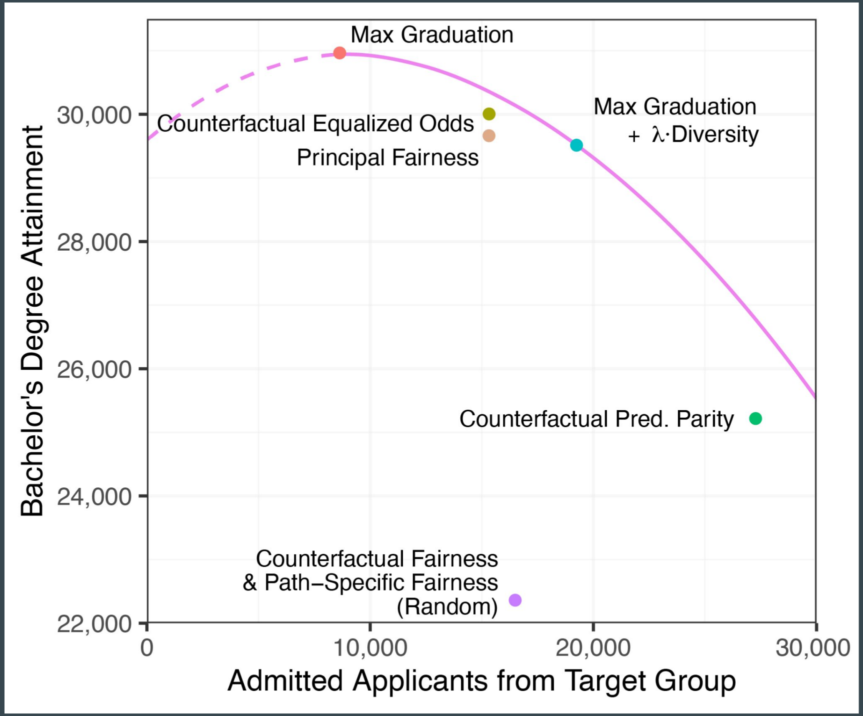
Family 2: Limit counterfactual disparities between groups

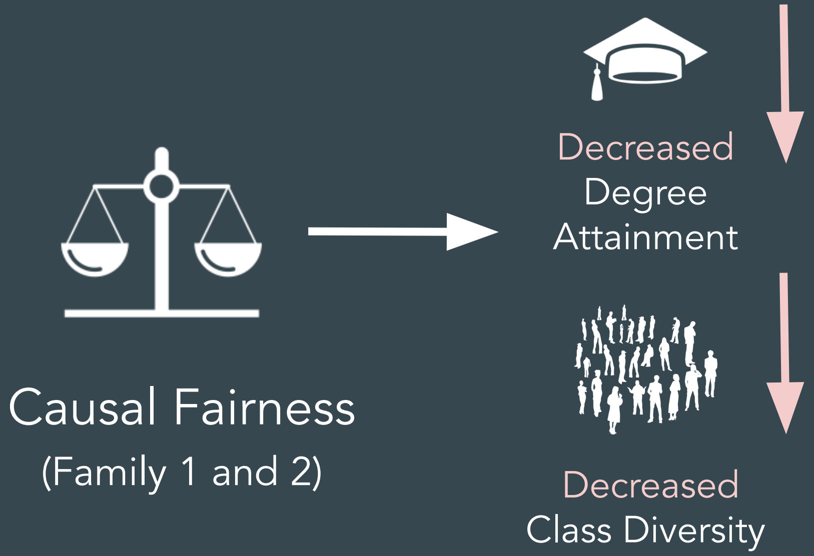
- Counterfactual equalized odds
- Counterfactual predictive parity
- Principal fairness

Key theoretical result #2



Key theoretical result #2





Key theoretical result #2

In *almost every* case (in a measure theoretic sense) it is impossible to satisfy prominent causal fairness definitions and be Pareto optimal

Summary

- Causal fairness definitions lead to Pareto inefficient decisions, perversely harming the groups they were designed to protect

- Directly optimizing for desired outcomes (e.g. degree attainment, diversity) may be preferable

Thank You!

Full Paper

H. Nilforoshan*, J. Gaebler*, R. Shroff, & S. Goel. "Causal Conceptions of Fairness and their Consequences." *International Conference on Machine Learning* (ICML 2022).

Technical Blog Post

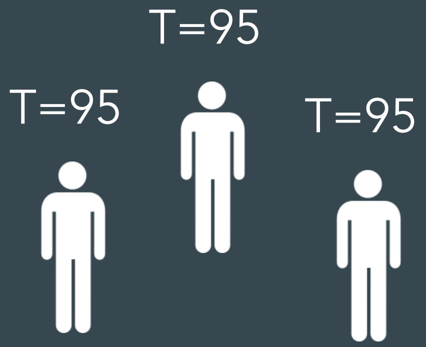
jgaeb.com/2022/07/18/prevalence.html

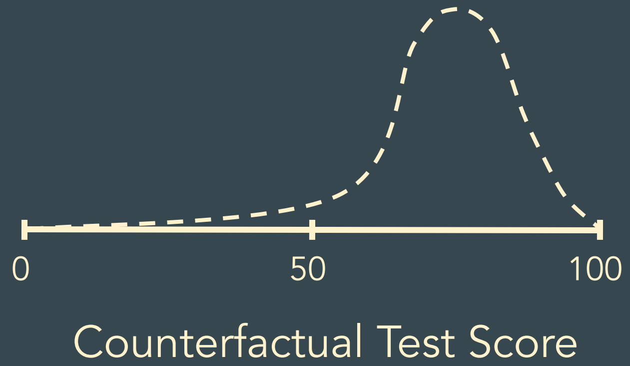
[jgaeb.com jgaeb@stanford.edu]

[hamedn.com hamedn@cs.stanford.edu]

Assumptions

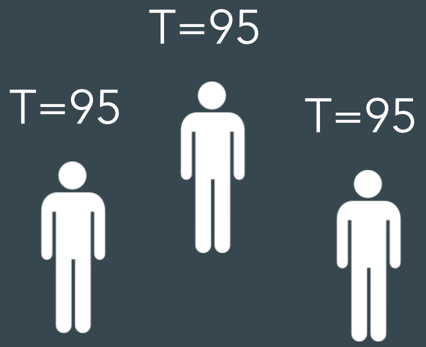
There is variance in the counterfactual distribution of covariates

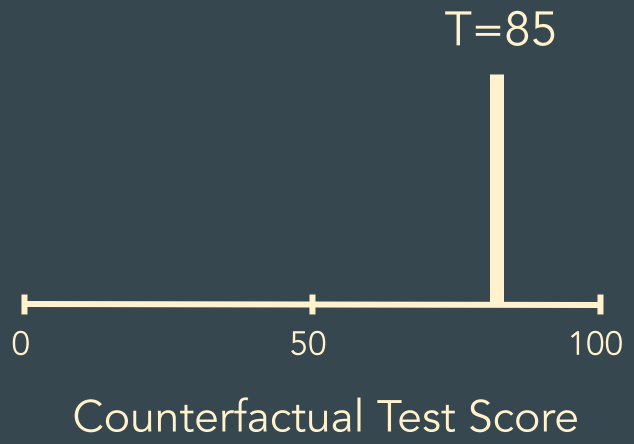




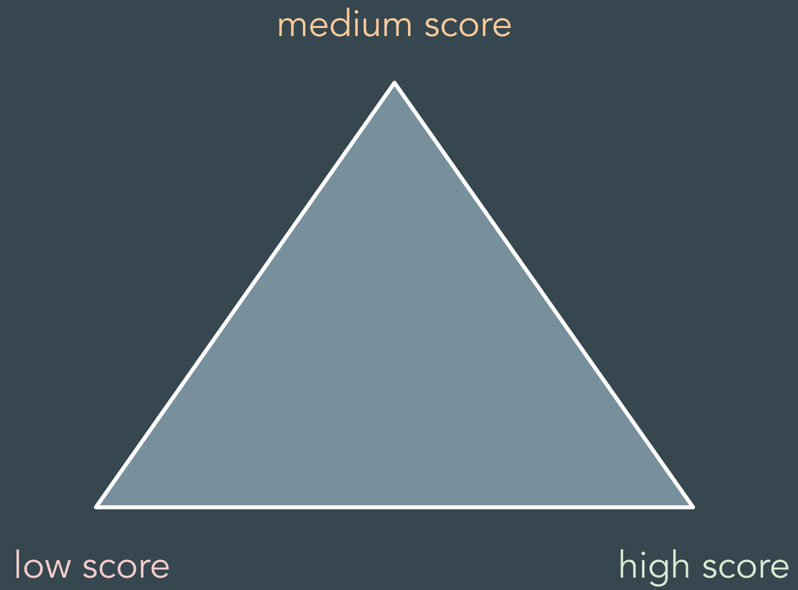
Assumptions

There is variance in the counterfactual distribution of covariates

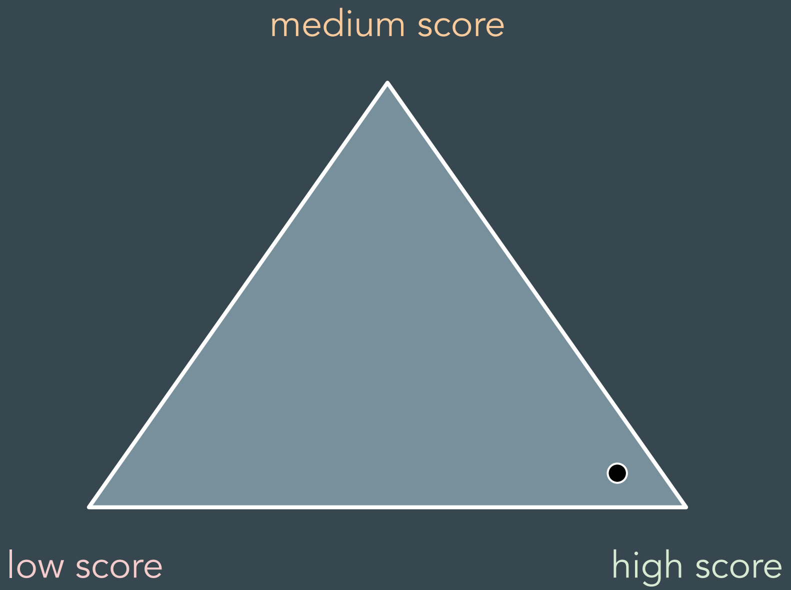




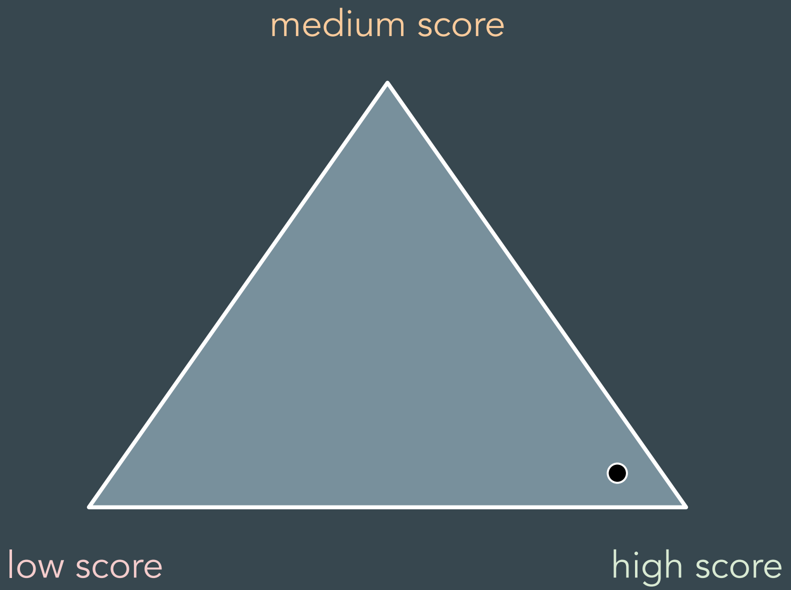
What do we mean by "almost every"?



What do we mean by "almost every"?



What do we mean by “almost every”?

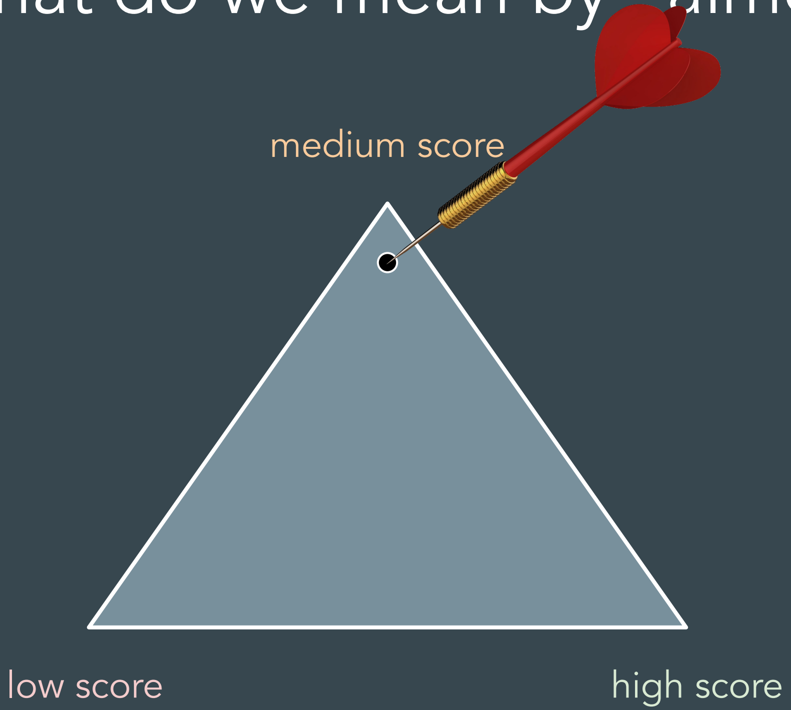


P( = low) = 0.05

P( = medium) = 0.05

P( = high) = 0.90

What do we mean by “almost every”?

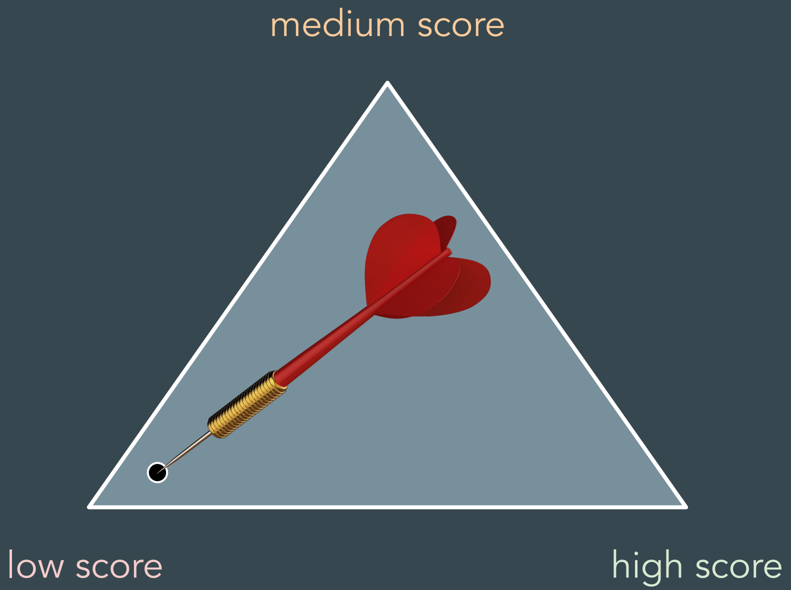


P( = low) = 0.05

P( = medium) = 0.90

P( = high) = 0.05

What do we mean by "almost every"?



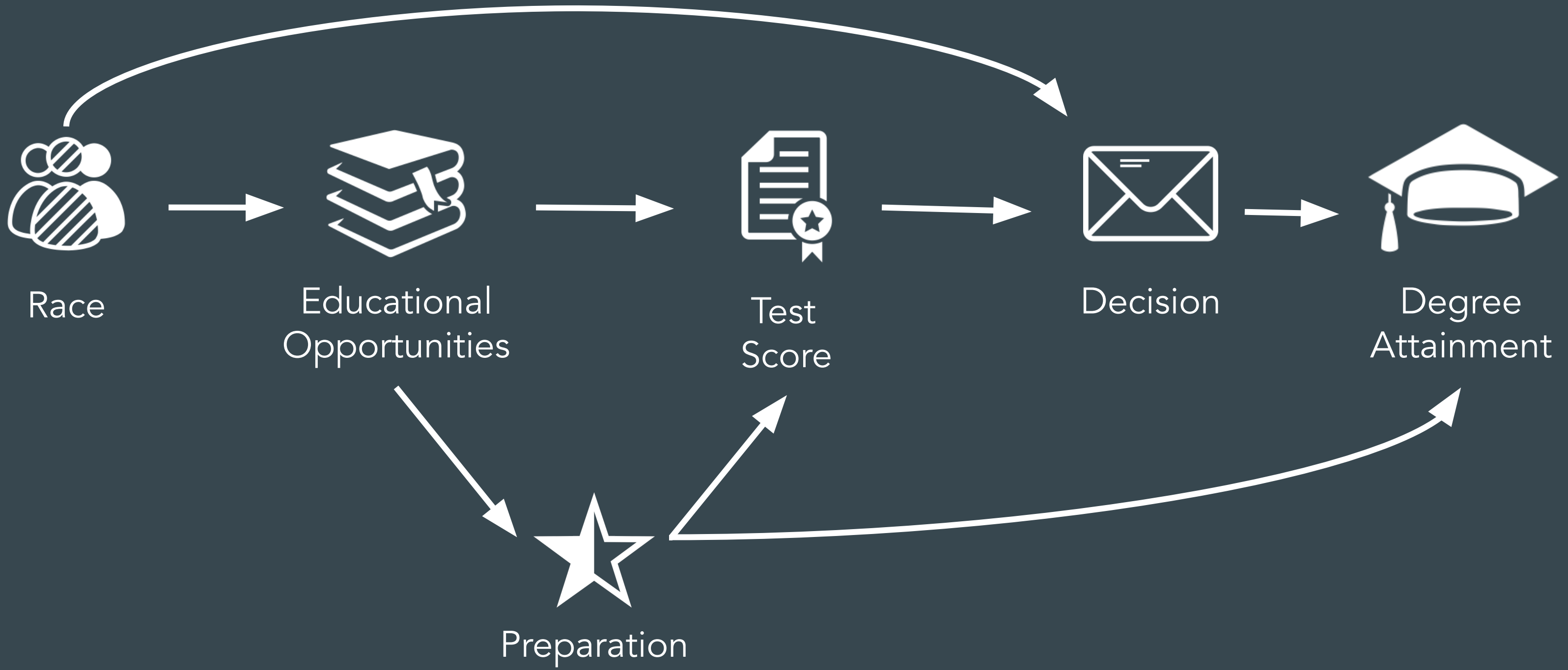
P(= low) = 0.90

P(= medium) = 0.05

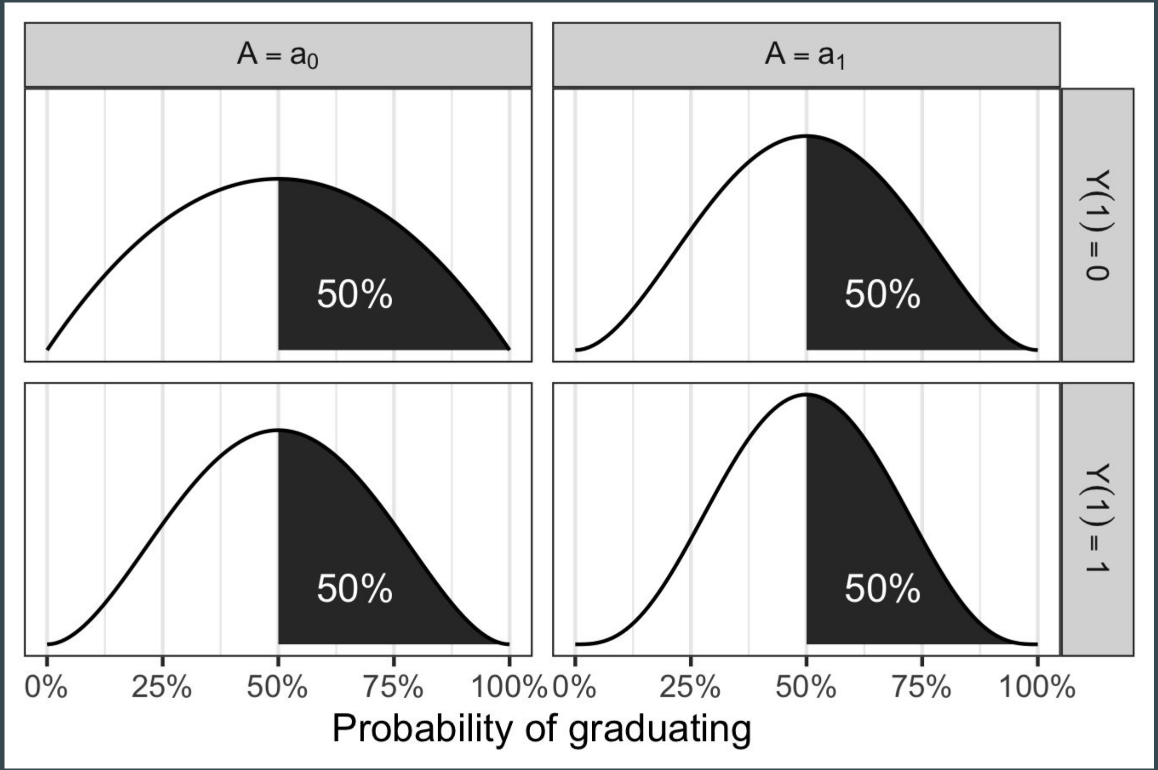
P(= high) = 0.05

P(Pareto Inefficient | Randomly Chosen Distribution) = 1.0

Simulation variables

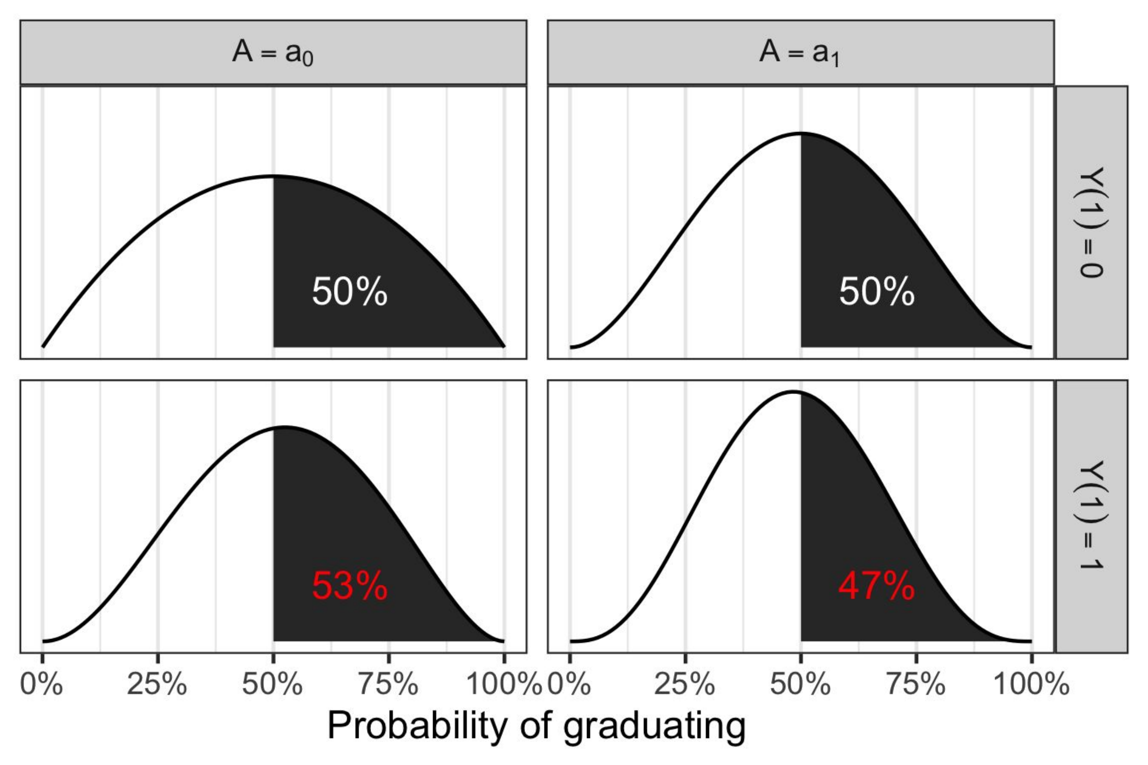


Key idea



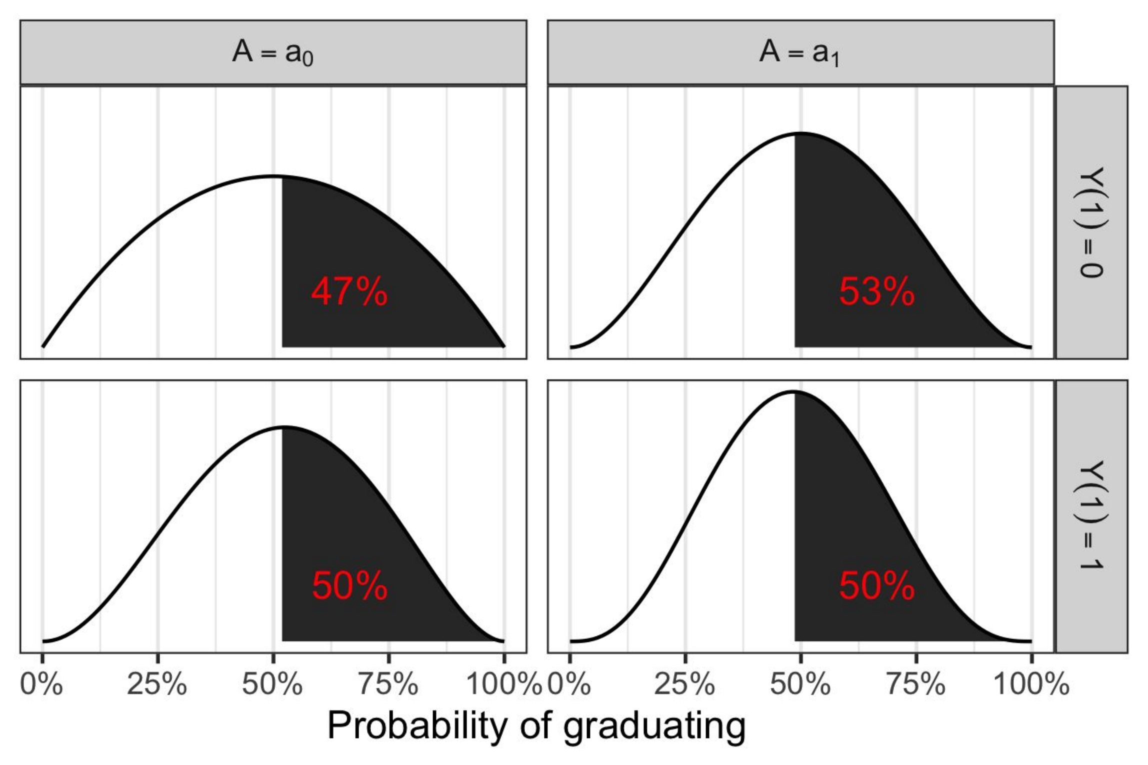
A = a0
A = a1
Y(1) = 0
Y(1) = 1
50%
50%
50%
50%
0%
25%
50%
75%
100%
0%
25%
50%
75%
100%
Probability of graduating

Key idea



A = a0
A = a1
Y(1) = 0
Y(1) = 1
50%
50%
53%
47%
0%
25%
50%
75%
100%
0%
25%
50%
75%
100%
Probability of graduating

Key idea



A = a0
A = a1
47%
53%
50%
50%
Y(1) = 0
Y(1) = 1
0%
25%
50%
75%
100%
Probability of graduating