

# **Responsible Data Science**

## **Algorithmic fairness continued: from values to technical choices & back**

*February 13, 2025*

---

**Professor Umang Bhatt**

Center for Data Science &  
Computer Science and Engineering  
New York University



# Recall: The problem with the trolley problem





back to fairness

r/ai

# Fair resource allocation



executive



sous chefs



line chefs

# Fair resource allocation



# Meet Equality of Opportunity (EO)



♪ Your daddy is rich...  
and your mama's good looking ♪  
...but that won't help you  
in an EO world



# Principles of EO



**Fair contests:** competitions should only judge people based on morally relevant “merit” (i.e., qualifications), not based on morally arbitrary factors (e.g., gender, race, socio-economic status)

**Fair life chances:** level the playing field over a lifetime



# Domains of EO



## **(1) Fairness at a specific decision point**

- distribution of social goods: e.g., employment, loans

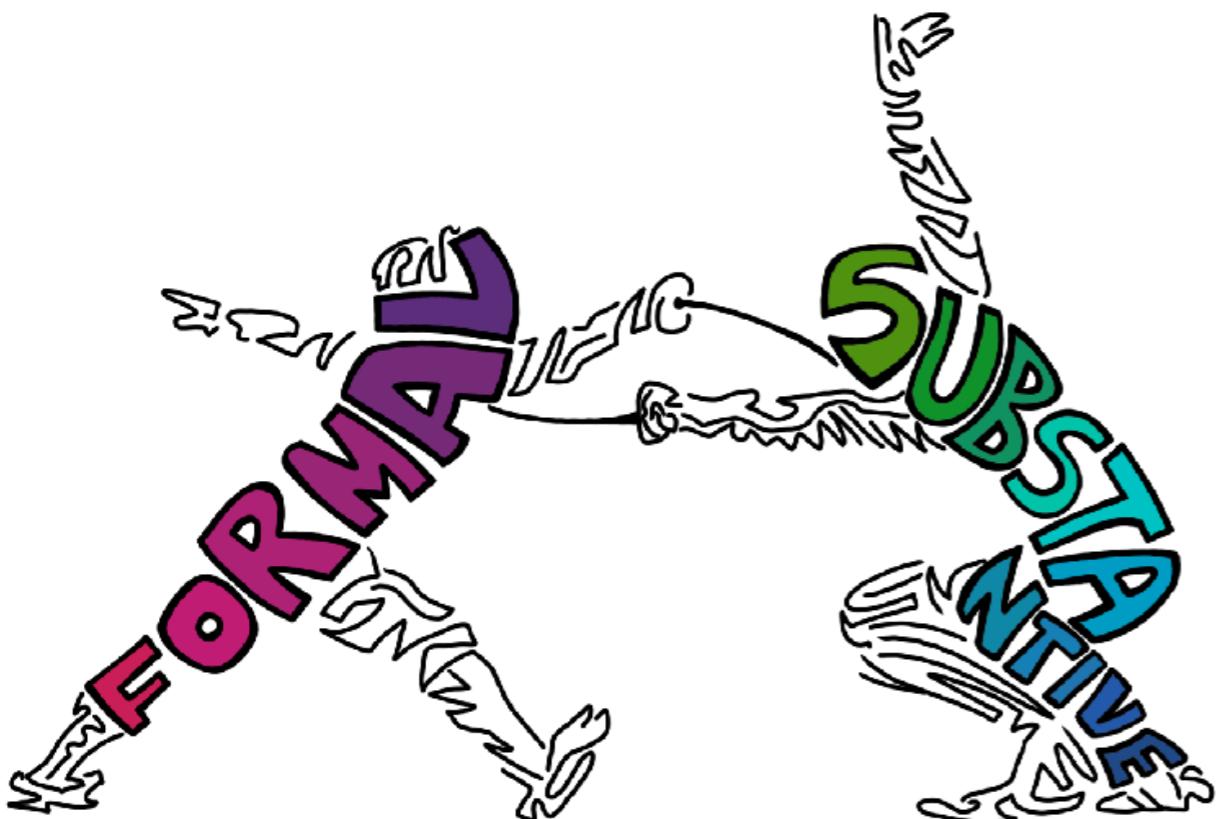
## **(2) Equality in developmental opportunity**

- access to opportunities that shape one's ability to compete for positions at a decision point (1)

## **(3) Equality of opportunity over a lifetime**

- access to comparable opportunity sets over a lifetime

# Summary of EO doctrines



**Formal, formal-plus:** fair contests, at a single decision point

**Substantive:** fair contests, fair life chances, over the course of a lifetime

# Group fairness as EO



## Group fairness

- Protected group membership is irrelevant to correct or positive classification

## Equality of Opportunity / Substantive

- Irrelevant characteristics (such as group membership) don't affect outcomes



# Individual fairness as EO



## Individual fairness

- Similar treatment of similar individuals
- Only irrelevant characteristics separate similar people

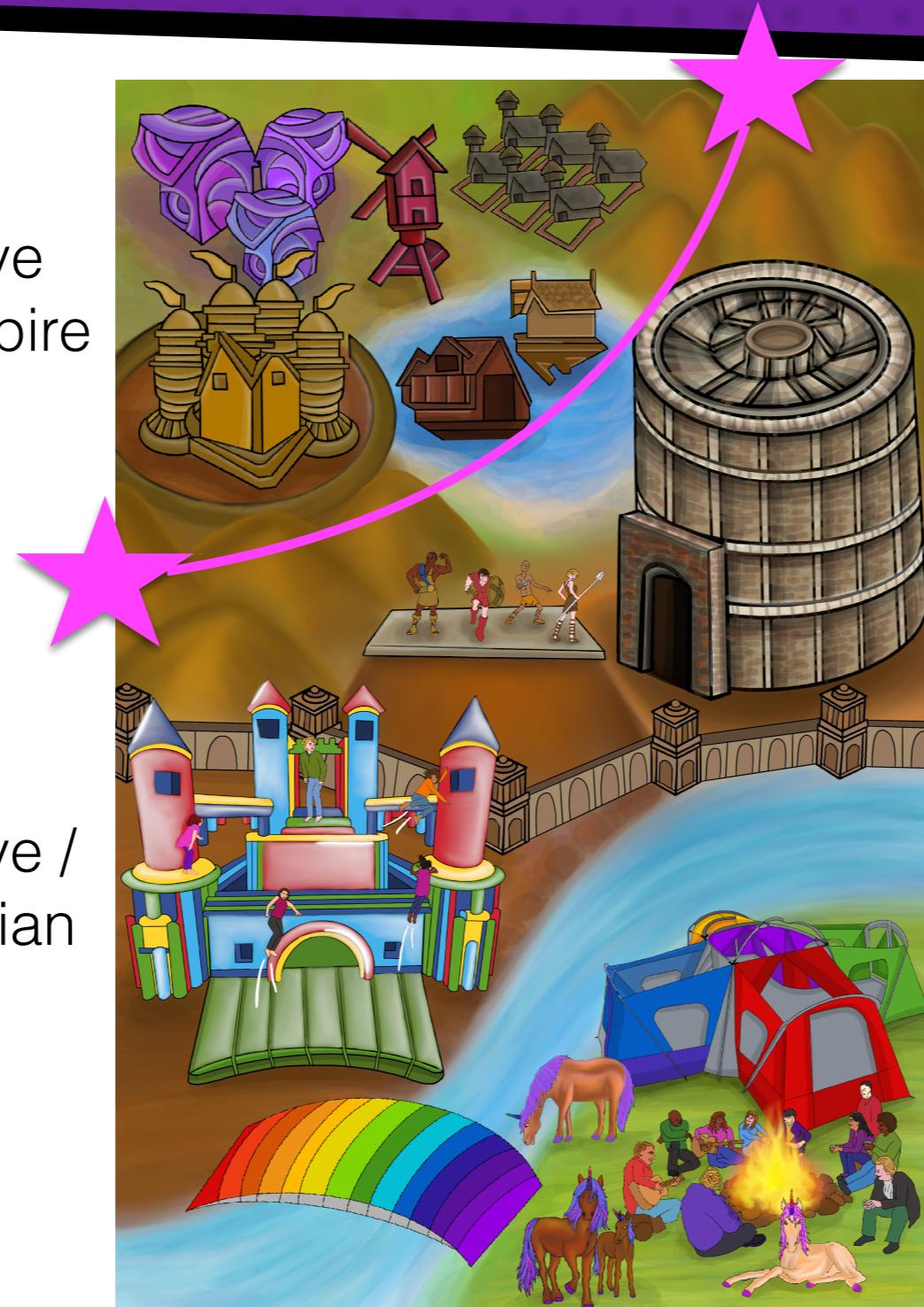
## Equality of Opportunity / Formal

- Irrelevant characteristics don't lead to different treatment of similar people



# The EO Empire

Libertarians now live outside the EO empire



substantive /  
Rawlsian

**Formal-ville**

substantive /  
luck egalitarian

# Formal EO: Careers open to talents



- In any contest, applicants should only be judged by job-relevant qualifications
- “See nothing irrelevant, speak nothing irrelevant, hear nothing irrelevant”
- Codified as “**fairness through blindness**” with its known weaknesses

# Formal EO as calibration



**Calibration:**

$$P(y = 1 | y' = c, s = 0) = P(y = 1 | y' = c, s = 1)$$

If two individuals have the same predicted score  $y'$  (relevant merit) and only differ on group membership  $s$  (morally irrelevant factors), then they are likely to get the same outcome from a well-calibrated test.

# Formal EO as predictive parity



Predictive parity:

$$P(y = 1 | y' > p, s = 0) = P(y = 1 | y' > p, s = 1)$$

If two individuals both have job-relevant qualifications  $y' > p$  (relevant merit) and only differ on group membership  $s$  (morally irrelevant factors), then they are likely to get the same outcome.

# Formal EO: Test validity

- A test that systematically under / over estimates people in a way that tracks group membership violates formal EO
- Measures of accuracy or test validity should be broken out by demographic group



# Formal-plus EO as error rate balance



Error rate balance:

$$P(y' > p \mid y = 0, s = 0) = P(y' > p \mid y = 0, s = 1)$$

$$P(y' \leq p \mid y = 1, s = 0) = P(y' \leq p \mid y = 1, s = 1)$$

A test with balanced error rates at a threshold  $p$  captures formal-plus EO's conception of a fair contest because it ensures that test performance (i.e., false-positive rate and false-negative rate) does not skew with morally irrelevant group membership

**"Equal opportunity"** [Hardt et al. 2016] codifies formal-plus EO

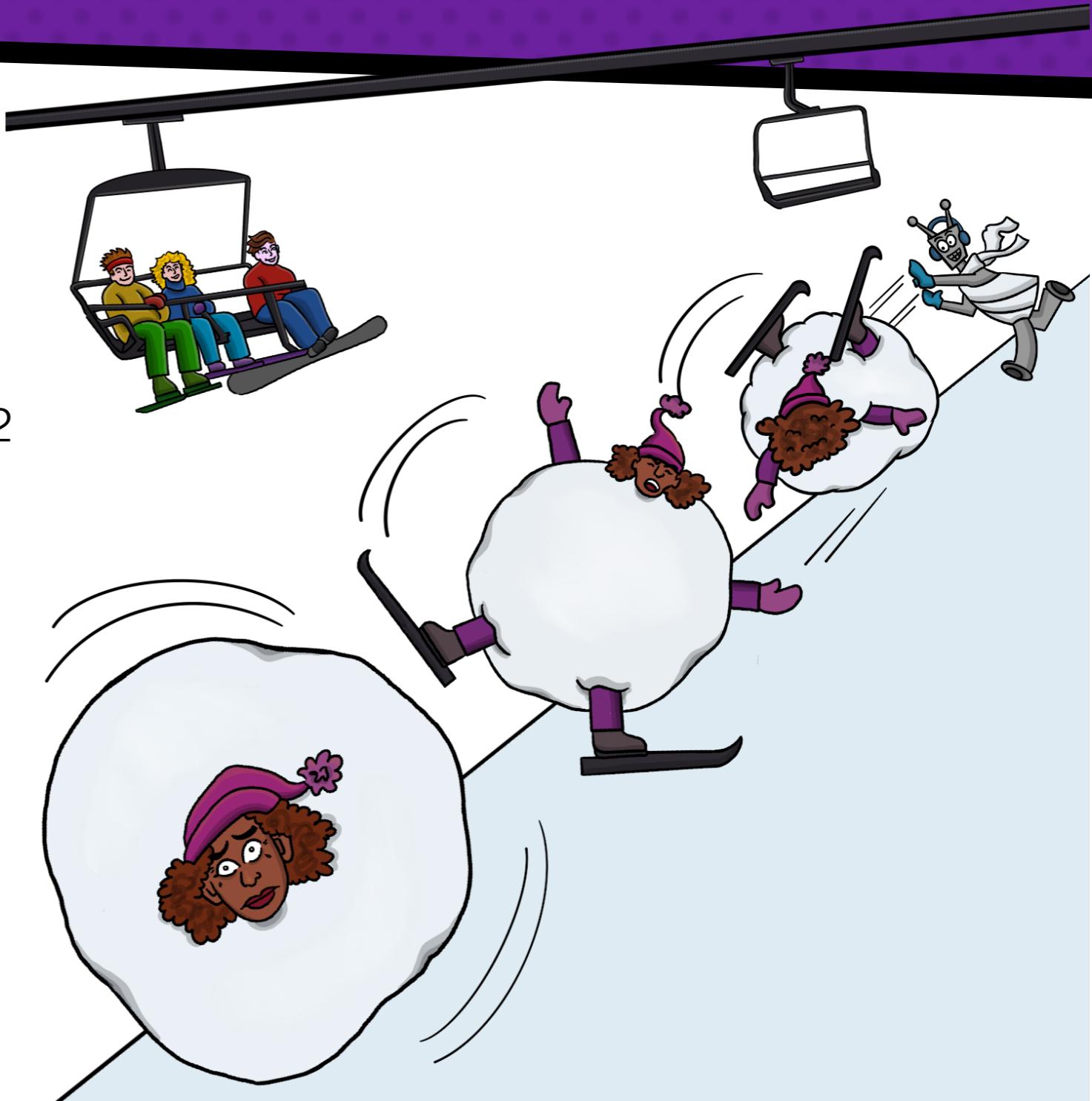
# Limitation of formal EO: the “before” problem

- Formal EO's appeal: relevant skills in, irrelevant characteristics out
- But OK to use irrelevant privileges before competition
- So privileges affect competition outcomes



# Limitation of formal EO: the “after” problem

- Winners at time 1 gain improved characteristics for competing at time 2
- Winners win faster, losers lose faster



# “Before” + “after” → discrimination laundering

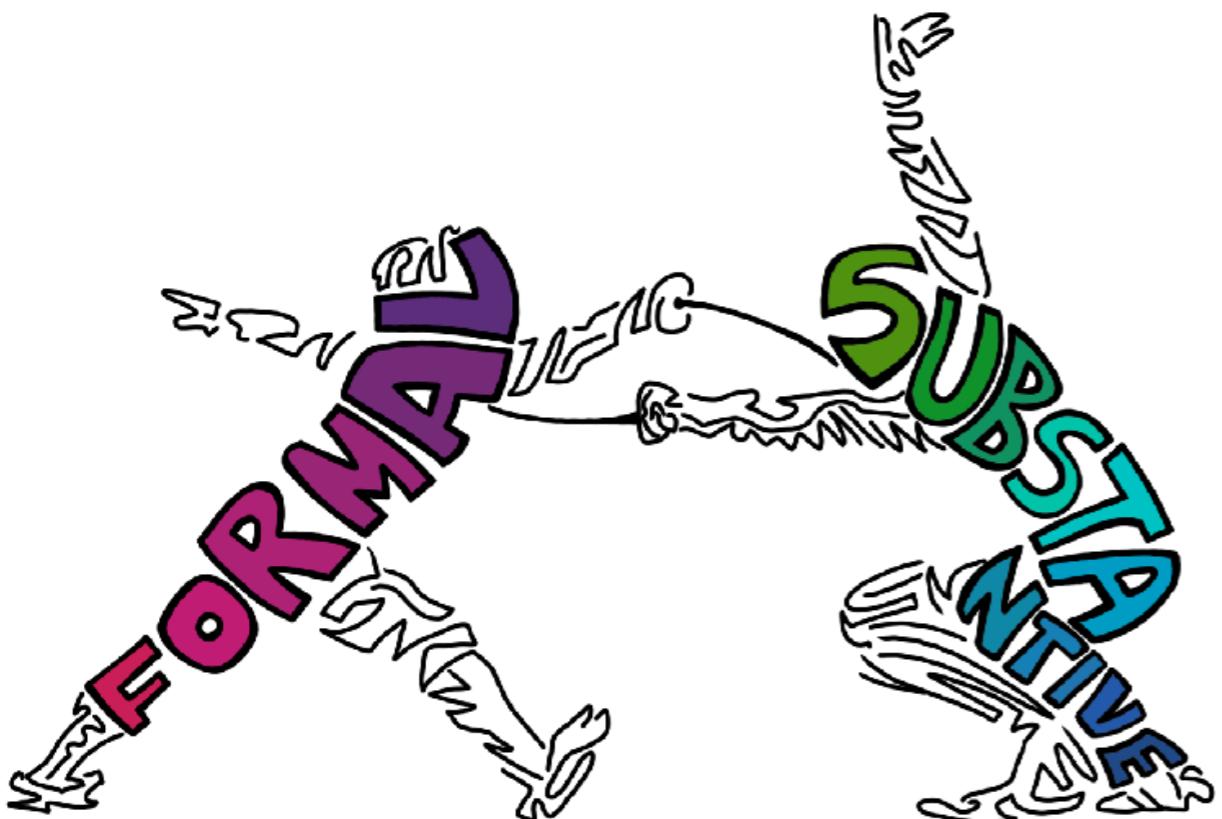
- Real world discrimination against some leads to privileges for others
- According to formal EO, it's OK to convert privileges to qualifications
- Winning on the basis of qualifications leads to more winning on qualifications
- Discrimination recedes from view...

*“Racial discrimination in on-the-job training is illegal; discrimination on the basis of differences in human capital due to differences in on-the-job training is not”*

(Elizabeth Anderson, The Imperative of Integration)



# Summary of EO doctrines



**Formal, formal-plus:** fair contests, at a single decision point

**Substantive:** fair contests, fair life chances, over the course of a lifetime

# The EO Empire

Libertarians now live outside the EO empire

**substantive /  
Rawlsian**



Formal-ville

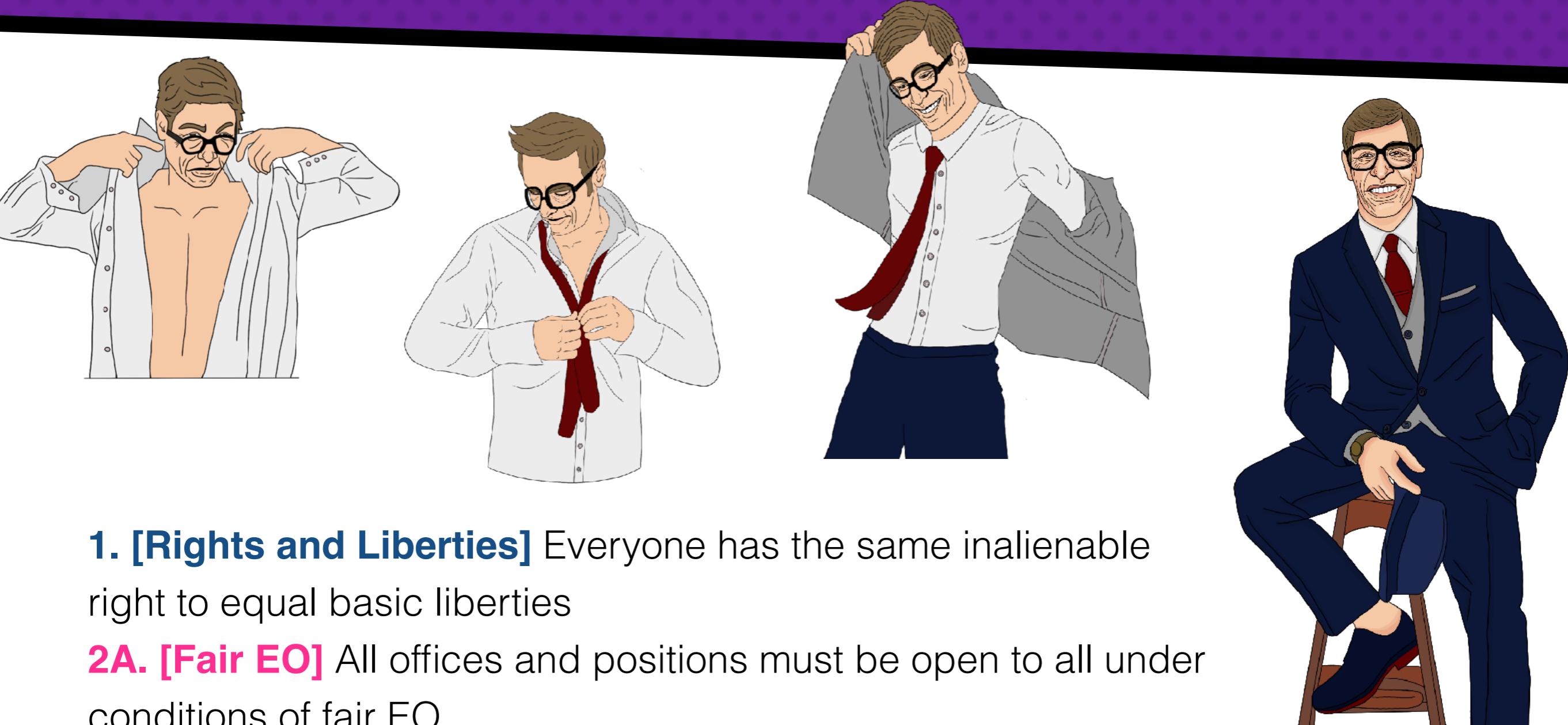
substantive /  
luck egalitarian

# Substantive EO: Rawls



- Equally talented babies must have equal life prospects
- Emphasis is on equality of **developmental opportunities**
- All people - rich or poor - must have the same opportunities to develop their qualifications, so that at the point of competition they are equally likely to succeed

# Rawls' broader view of justice



**1. [Rights and Liberties]** Everyone has the same inalienable right to equal basic liberties

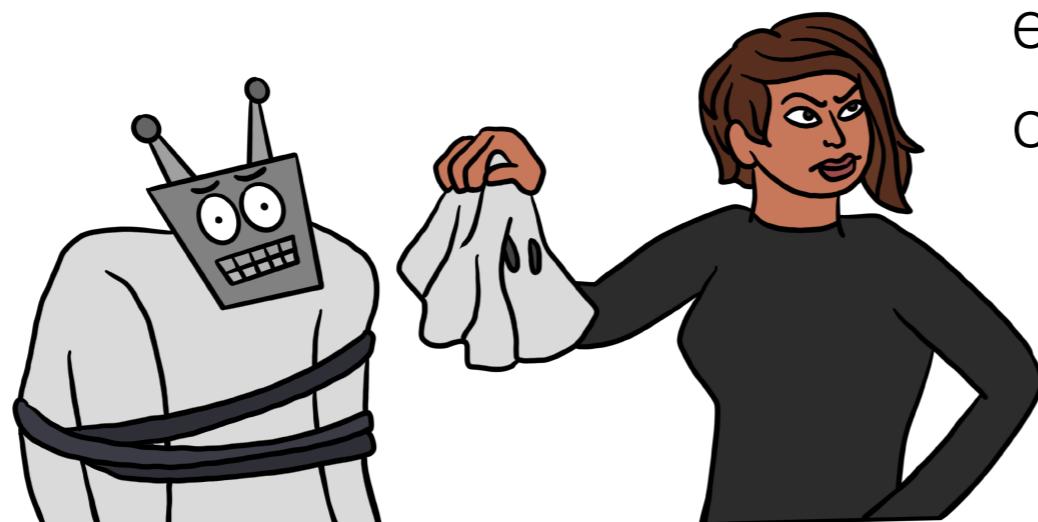
**2A. [Fair EO]** All offices and positions must be open to all under conditions of fair EO

**2B. [Difference Principle]** Social and economic inequalities must be of the greatest benefit to the least advantaged

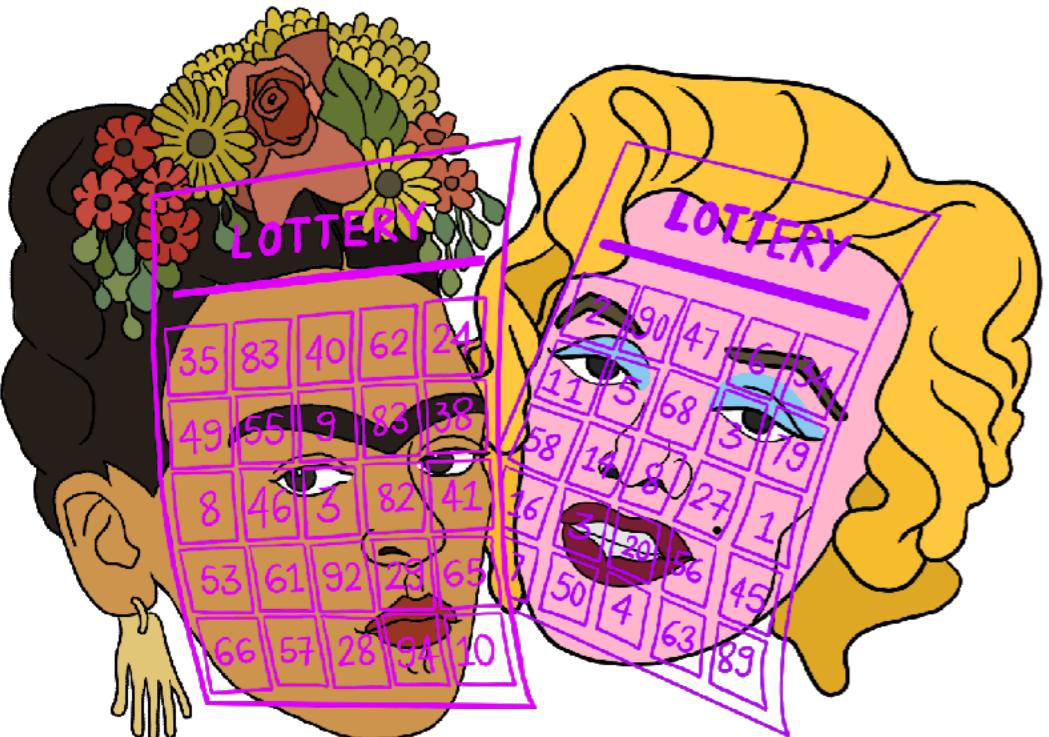
# Misconceptions of Rawls in Fair-ML



- In fair-ML, statistical parity and equality of odds are believed to operationalize Rawlsian fair EO. But this is not so!
- Rawlsian EO is fundamentally about providing developmental opportunities **before** competitions, and about ensuring that opportunity sets are comparable **over a lifetime**



# Substantive EO: Rawls: natural & social lottery



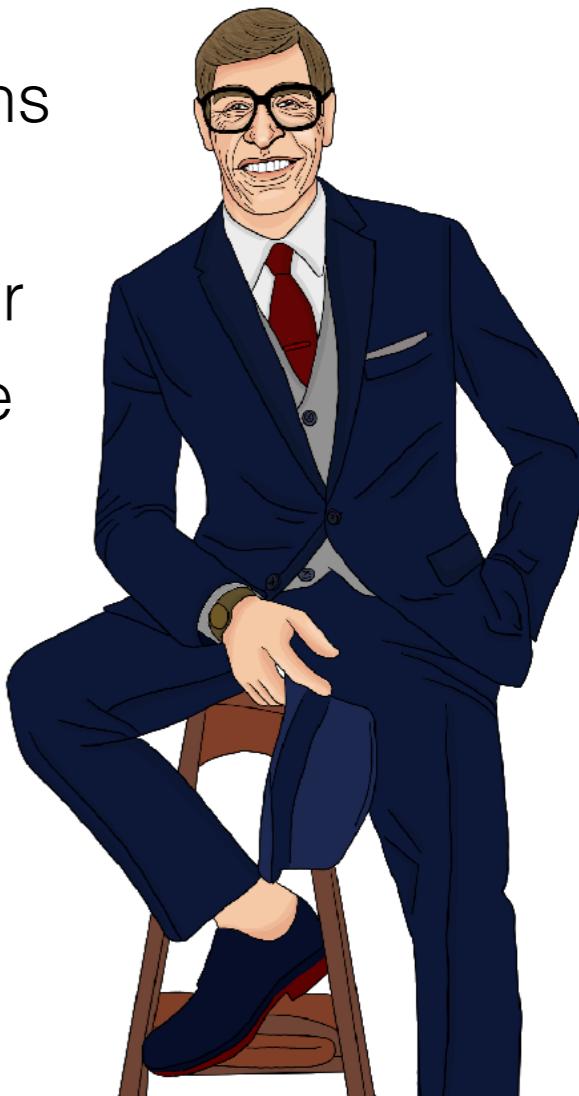
**Difference principle (maximize the minimum):** Since we don't deserve our starting points in life, we must work towards a social system that serves everyone.

**Natural & social lottery:** Talents and fortune are distributed arbitrarily.

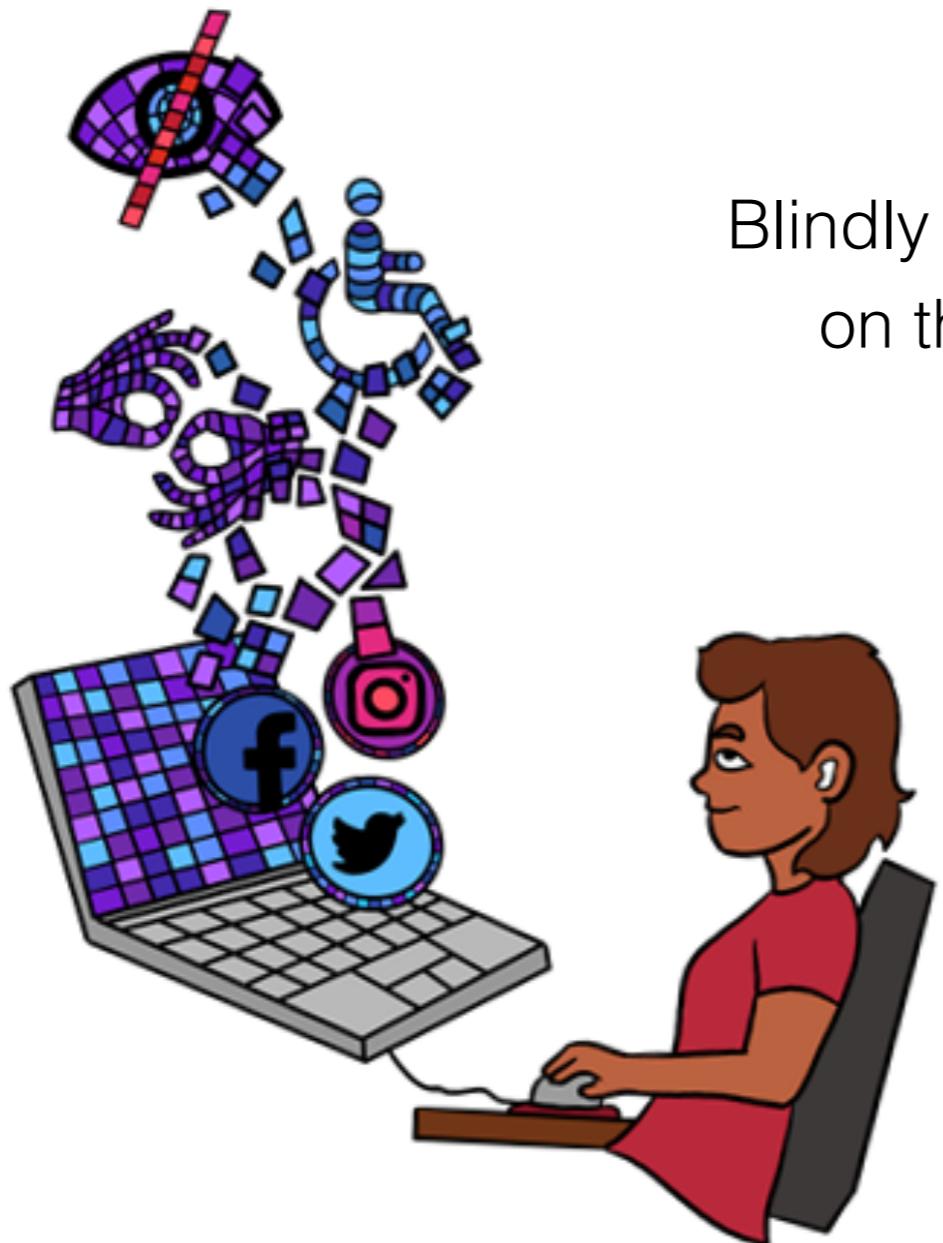


# Rawls' “original position”

**The Veil of Ignorance:** If citizens do not know their race, class, sex, social position (or any other characteristics that might cause them to favor people like themselves), they will advocate for all social positions and their attached privileges to be distributed fairly.



# Broader view of justice

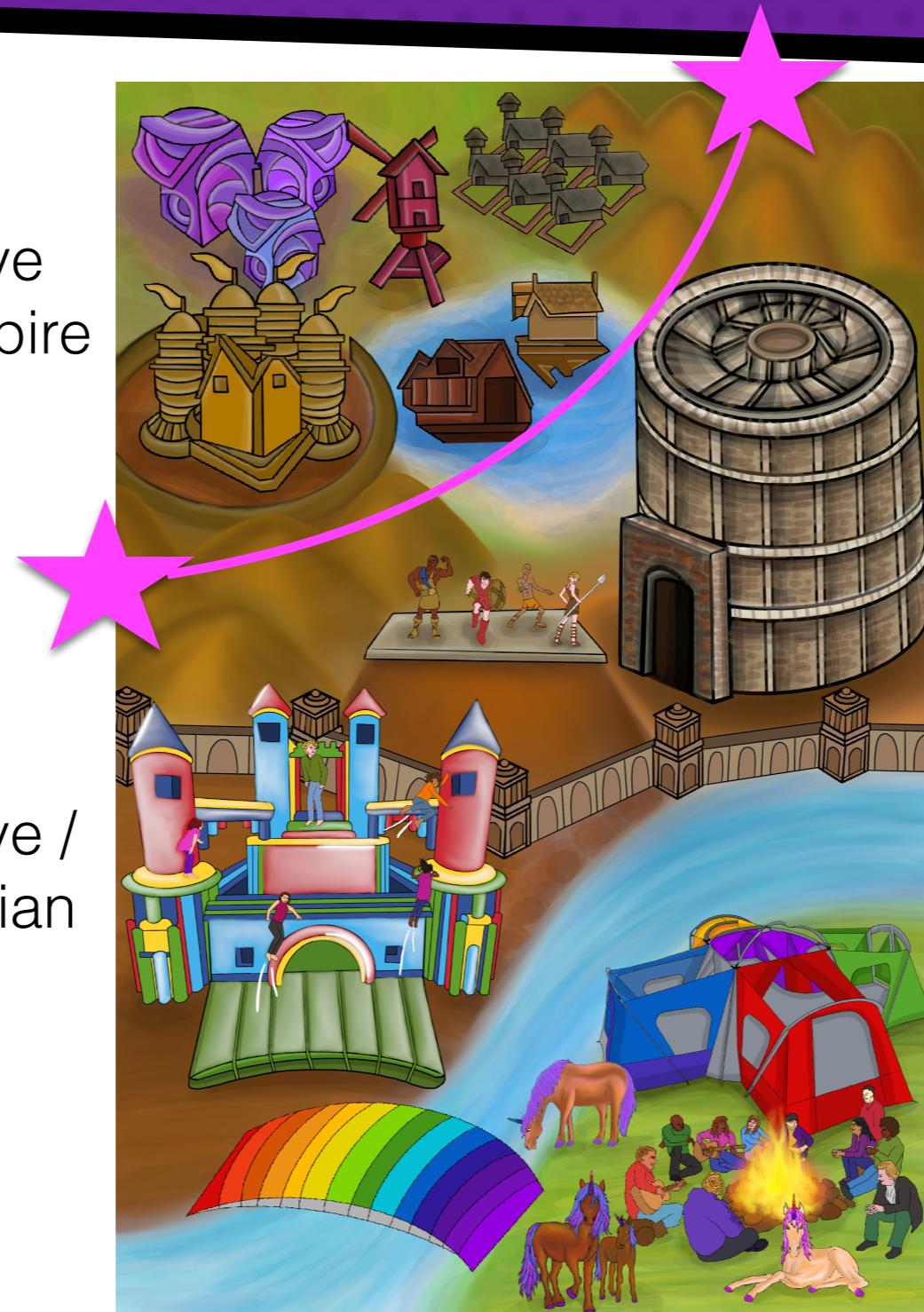


Blindly satisfying EO may infringe  
on the **freedom of speech**.



# The EO Empire

Libertarians now live outside the EO empire



Formal-ville

substantive /  
Rawlsian

substantive /  
luck egalitarian

# Substantive EO: Luck egalitarian



The luck egalitarians gather around the communal fire, forsaking all disparities in talent and effort, in favor of unicorns on rainbows!

# Substantive EO: luck egalitarian

- Outcomes should only be affected by “choice luck” (one’s responsible choices), not by “brute luck”
- But how do we make this separation?



# Substantive EO: luck egalitarian



For which characteristics can we hold an individual accountable?  
**(responsible choice)**

And which matters are completely out of their control?

**(brute luck)**



# Substantive EO: luck egalitarian: Roemer

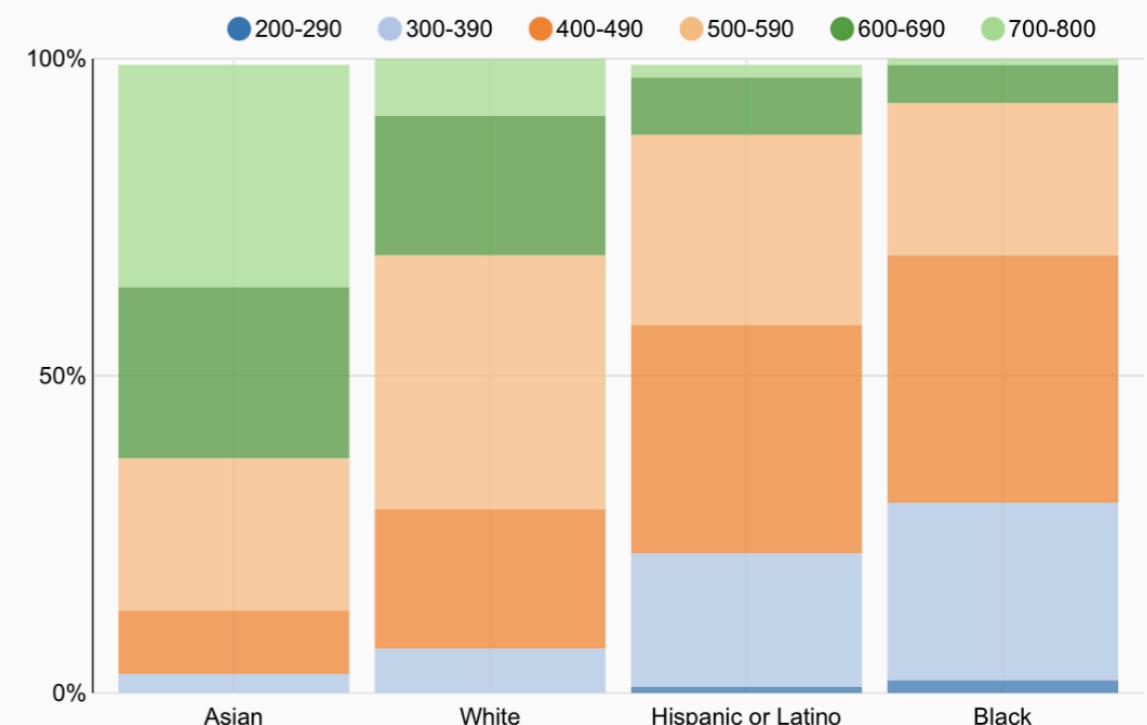
Effort, circumstance, and types

(Roemer, 2002)



## Wide race gaps in SAT math scores

Math score distribution by race or ethnicity



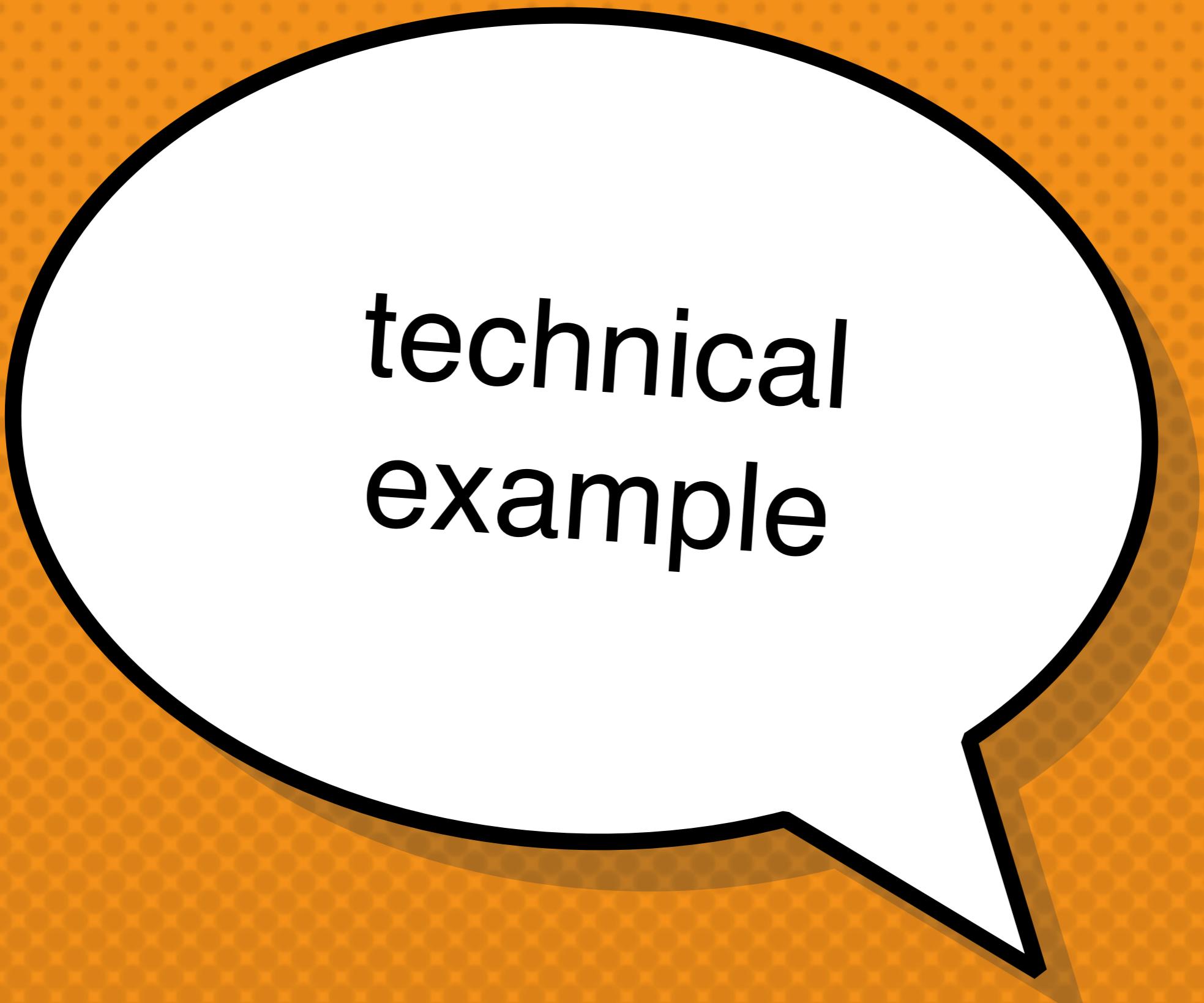
College Board, "SAT Suite of Assessments Annual Report," 2020.

BROOKINGS

# Substantive EO: Luck egalitarian: Roemer

- No split between responsible effort and irrelevant circumstance
- But there is still an apples and oranges problem





technical  
example

# Diverse balanced ranking

## Goals

**diversity**: pick  $k = 4$  candidates, including 2 of each gender, and at least one per race

**utility**: maximize the total score of selected candidates

	Male		Female	
White	A (99)	B (98)	C (96)	D (95)
Black	E (91)	F (91)	G (90)	H (89)
Asian	I (87)	J (87)	K (86)	L (83)

score = 373

score = 372



## Problem

picked the best White and male candidates (A, B) but did not pick the best Black (E, F), Asian (I, J), or female (C, D) candidates

## Beliefs

scores are more informative within a group than across groups - **effort is relative to circumstance**

it is important to **reward effort**

# From beliefs to interventions

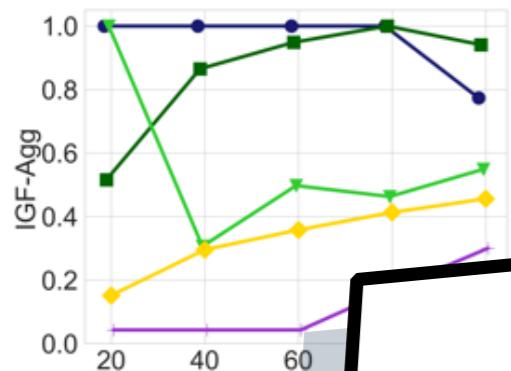
Fairness for female candidates       $83 / 95 = 0.91$

C	D	G	H	K	L
95	95	90	86	83	83

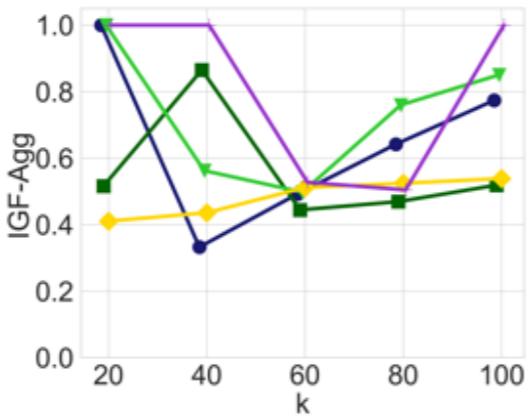
highest-scoring  
skipped

lowest-scoring  
selected

BEFORE: diversity constraints only



AFTER: diversity and fairness constraints



## Beliefs

scores are more informative within a group than across groups -  
**effort is relative to circumstance**

it is important to **reward effort**



another  
example

# Intersectional causal fairness

	gender	race	X	Y
B	m	w	6	12
C	m	<b>b</b>	5	9
D	<b>f</b>	w	6	8
E	m	w	4	7
F	<b>f</b>	<b>b</b>	3	6
K	<b>f</b>	a	5	5
L	m	<b>b</b>	1	3
O	<b>f</b>	w	1	1

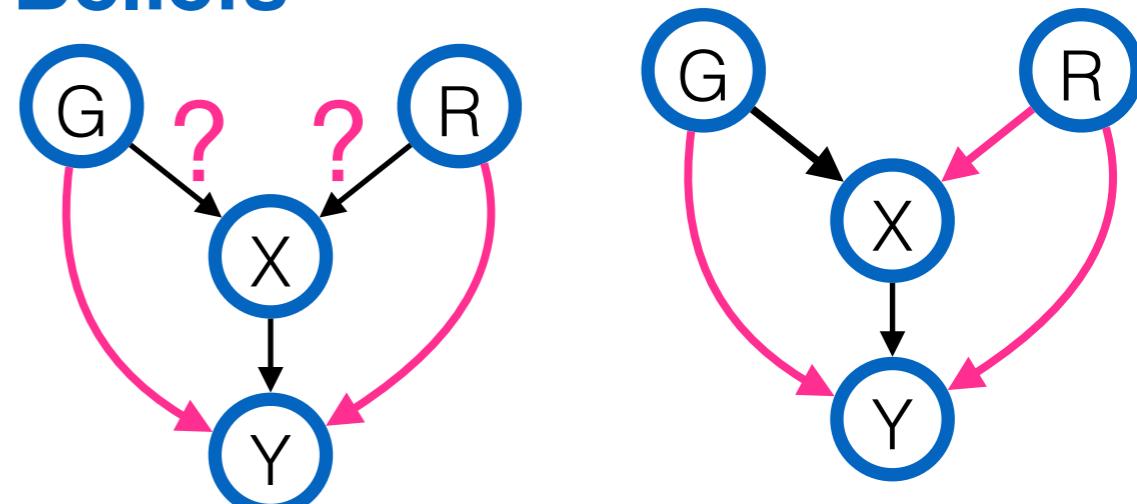
## Problem

weight lifting ability is mapping to qualification score differently depending on gender

## Goal

hire  $k = 4$  best-qualified candidates at a moving company

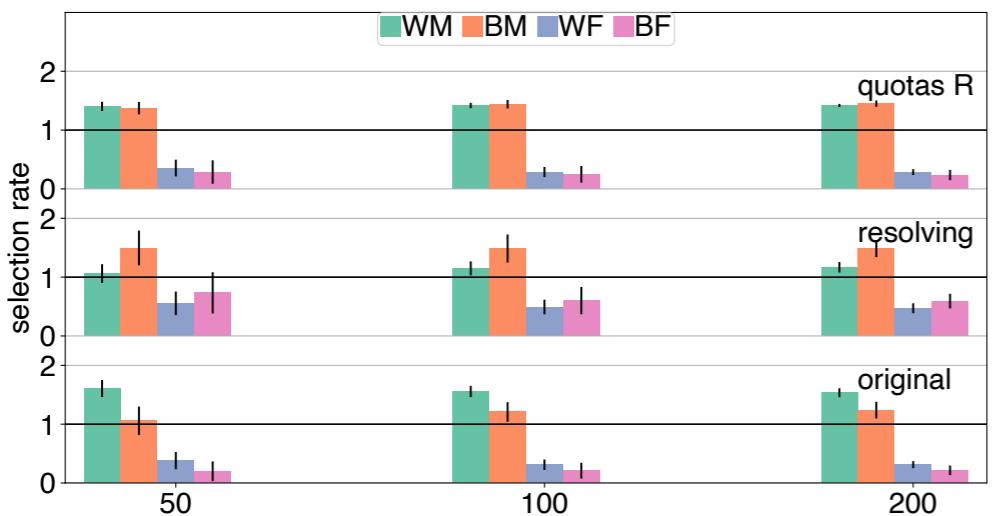
## Beliefs



# From beliefs to interventions

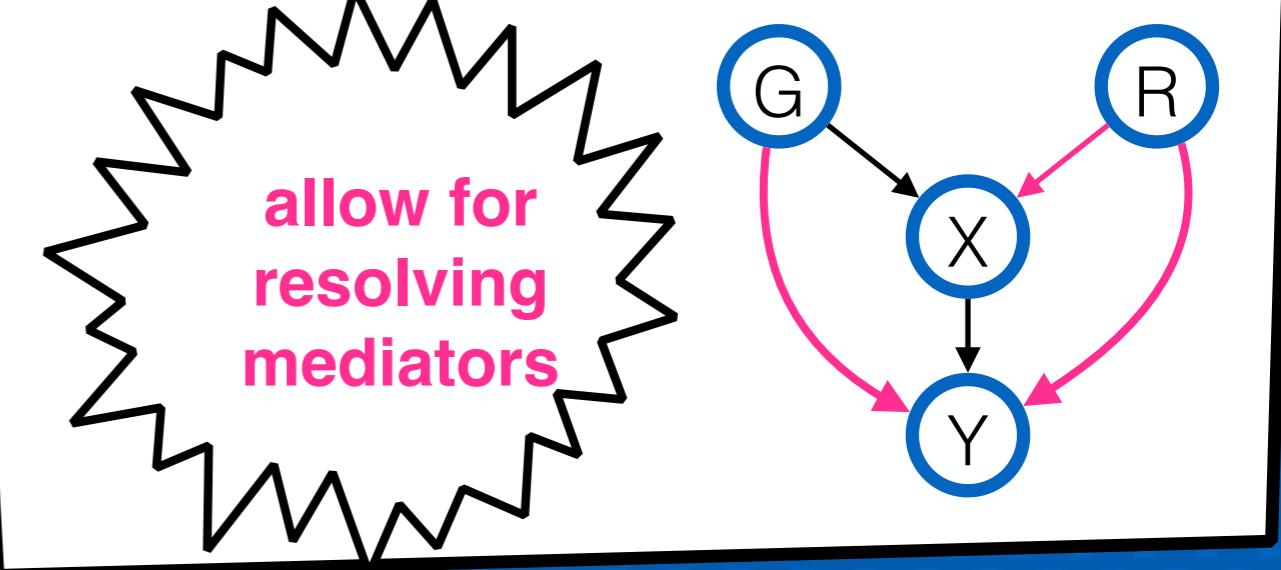
**Idea:** Compute **counterfactual scores**, treating each individual in the sample as though they had belonged to one intersectional group (e.g., Black women). Rank on those scores.

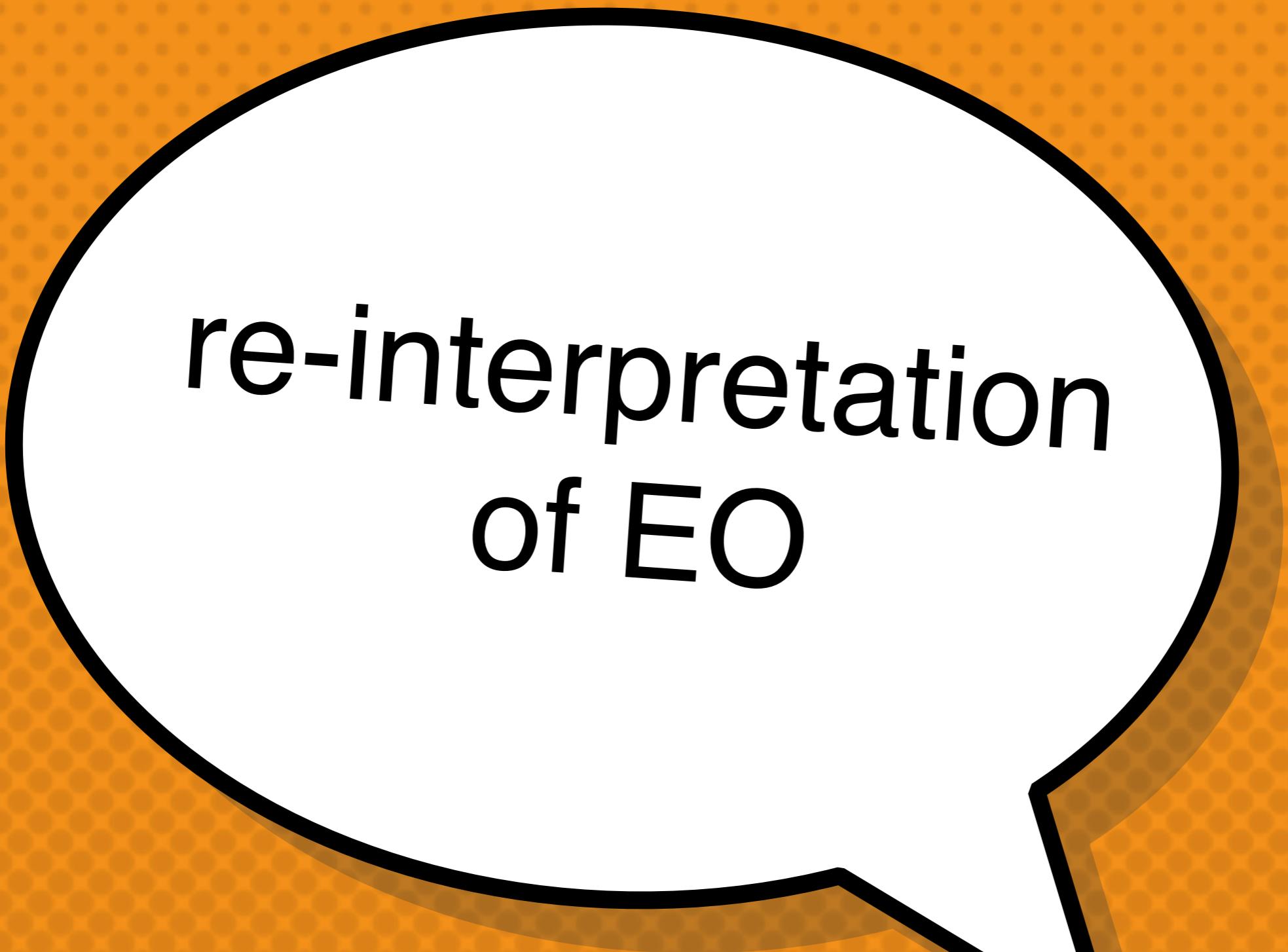
This process produces a **counterfactually fair ranking**.



## Beliefs

**allow for  
resolving  
mediators**





**re-interpretation  
of EO**

# Correcting for the past vs. improving the future

	Backward-facing	Forward-facing
Fair contests	Formal	Formal-plus
Fair life chances	Luck egalitarian	Rawls



# Correcting for the past vs. improving the future

Doctrine	Moral desiderata	Normative approach
Formal	Fair contests should only measure morally relevant qualifications	Accurately measure past performance
Formal-plus	The performance of fair contests should not skew along the lines of morally irrelevant features	Accurately estimate future performance
Substantive: Luck egalitarian	Matters of brute luck should not affect people's outcomes	Distribute outcomes on the basis of effort, after correcting for the past effects of morally arbitrary circumstances
Substantive: Rawls	Equally talented people should have equal prospects of success	Distribute outcomes to equalize future prospects of success of people who have the same native talent, irrespective of arbitrary circumstance



# Responsible Data Science

## Algorithmic Fairness

---

# Thank you!