Responsible Data Science

Algorithmic fairness continued: from values to technical choices & back

February 13, 2024

Prof. Julia Stoyanovich

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

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai

This week's reading

Towards Substantive Conceptions of Algorithmic Fairness: Normative Guidance from Equal Opportunity Doctrines

Falaah Arif Khan
New York University
New York, NY USA
fa2161@nyu.edu

Eleni Manis
S.T.O.P.
New York, NY USA
eleni@stopspying.org

Julia Stoyanovich
New York University
New York, NY USA
stoyanovich@nyu.edu

ABSTRACT

In this work we use Equal Opportunity (EO) doctrines from political philosophy to make explicit the normative judgements embedded in different conceptions of algorithmic fairness. We contrast formal EO approaches that narrowly focus on *fair contests* at discrete decision points, with substantive EO doctrines that look at people's *fair life chances* more holistically over the course of a lifetime. We use this taxonomy to provide a moral interpretation of the impossibility results as the incompatibility between different conceptions of a *fair contest* — forward-facing versus backward-facing — when people do not have *fair life chances*. We use this result to motivate substantive conceptions of algorithmic fairness and outline two plausible *fair decision procedures* based on the luck egalitarian doctrine of EO, and Rawls's principle of fair equality of opportunity.

ACM Reference Format:
Falaah Arif Khan, Eleni Manis, and Julia Stoyanovich. 2022. Towards Substantive Conceptions of Algorithmic Fairness: Normative Guidance from Equal Opportunity Doctrines. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22), October 6–9, 2022, Arlington, VA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3551624.3555303

1 EQUALITY OF OPPORTUNITY

Equality of Opportunity (EO) is a philosophical doctrine that objects to morally arbitrary and irrelevant factors affecting people's access to desirable positions, and the social goods attached to them (such as opportunity and wealth). In an EO-respecting society, all people, irrespective of their morally arbitrary characteristics, such as socio-economic background, gender, race, or disability status, have comparable access to the opportunities that they desire. Similarly, in fair machine learning (fair-ML), we are usually interested in ensuring that the outputs of algorithmic systems, specially those used in critical social contexts, do not systematically skew along the lines of membership in protected groups based on gender, race, or disability. In so far as protected groups are constructed on the basis of morally arbitrary factors, the moral desiderata of EO doctrines from political philosophy align exactly with the fairness-related concerns in machine learning. In this work, we employ ideas from the rich EO literature from political philosophy [2, 3, 6, 14, 15, 23, 26, 30–33] to clarify the normative foundations of fairness and justice-related interventions, and gauge the efficacy of current algorithmic approaches that attempt to codify these criteria.

1.1 Principles of EO

There are two broad principles of EO, namely, *the principle of fair contests* and *the principle of fair life chances*.

1.1.1 Fair contests. The principle of fair contests, commonly understood as the *nondiscrimination principle*, says that competitions for desirable positions should be open to all and should be adjudicated based on competitors' relevant merits, or qualifications. In any fair contest, the most qualified person wins. Conversely, fair contests do not judge competitors on the basis of irrelevant characteristics, especially excluding morally arbitrary factors such as gender, race, and socio-economic status that are not properly understood as qualifications at all.

The principle of fair contests has been very influential in fair-ML and has guided statistical measures and algorithmic interventions that conceptualize *fairness* as *nondiscrimination*.

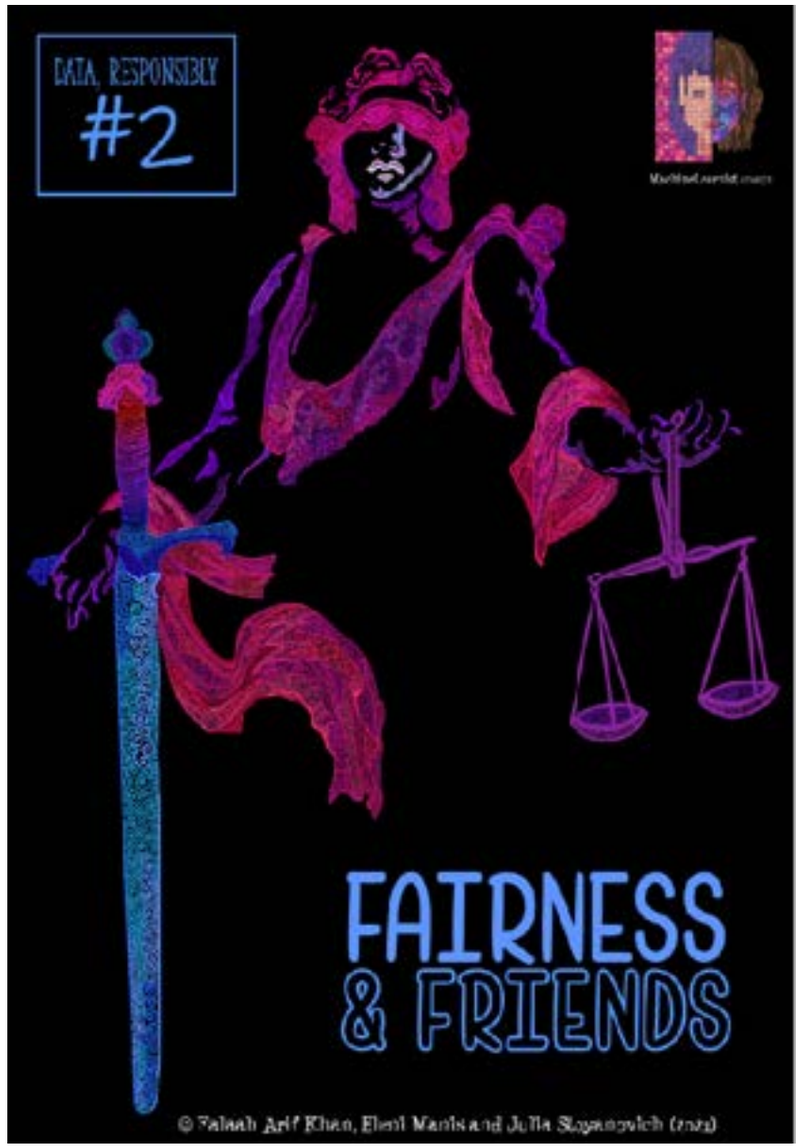
1.1.2 Fair life chances. The principle of fair life chances says that people's chances of success over a lifetime should not depend on morally arbitrary factors. It takes a holistic view of equal opportunity by comparing the *opportunity* sets that people have over the course of a lifetime, and is popularly understood as a principle that *levels the playing field*.

The principle of fair life chances has been almost entirely overlooked in fair-ML, and this omission explains some of the limitations in current approaches, as we will discuss shortly.

1.2 Domains of EO

According to Fishkin [15], there are, broadly, three domains of EO:

1.2.1 Fairness at a specific decision point. The first domain comprises the discrete points at which social goods, such as employment, admissions, and loan decisions are distributed. EO doctrines compel us to think about whether outcomes of decision-making at discrete decision points are influenced by morally arbitrary factors.



r/ai

Recall: The problem with the trolley problem

r/ai

back to fairness

r/ai

Fair resource allocation

executive

sous chefs

line chefs

r/ai

Fair resource allocation

Meet Equality of Opportunity (EO)

Goal: eliminate irrelevant, arbitrary barriers to achievement

♪♪ *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

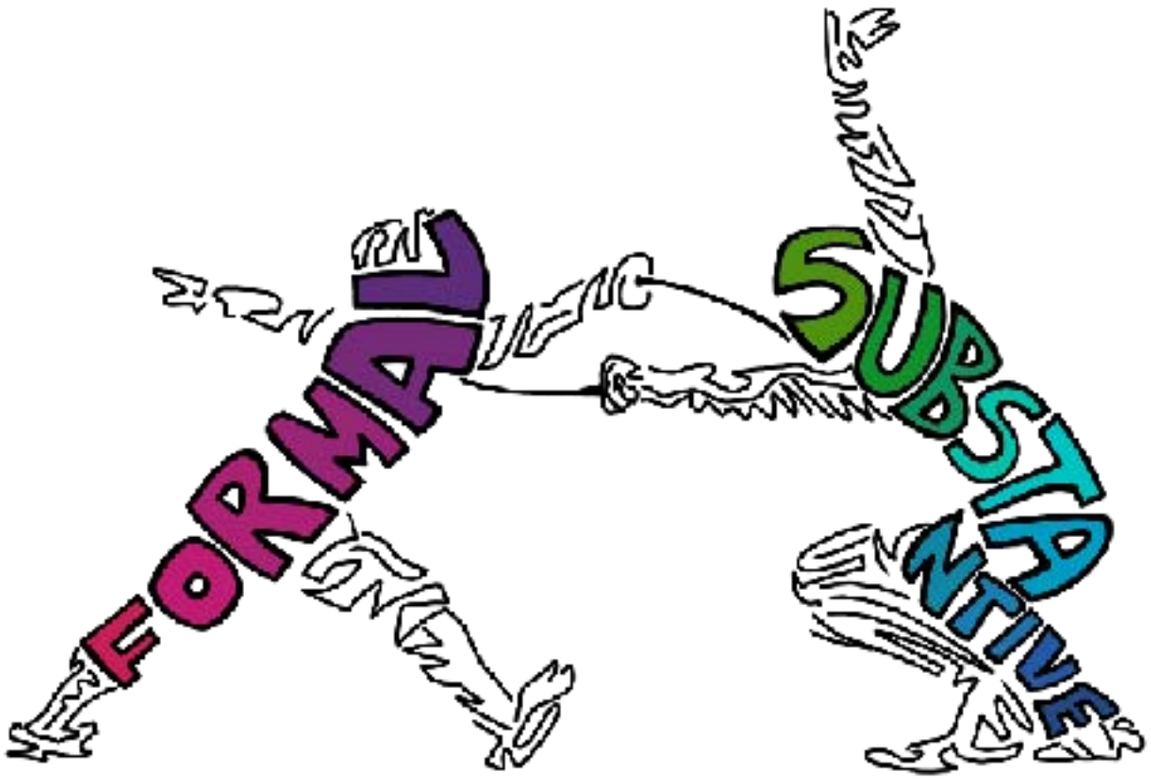
(2) Equality in developmental opportunity

- 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

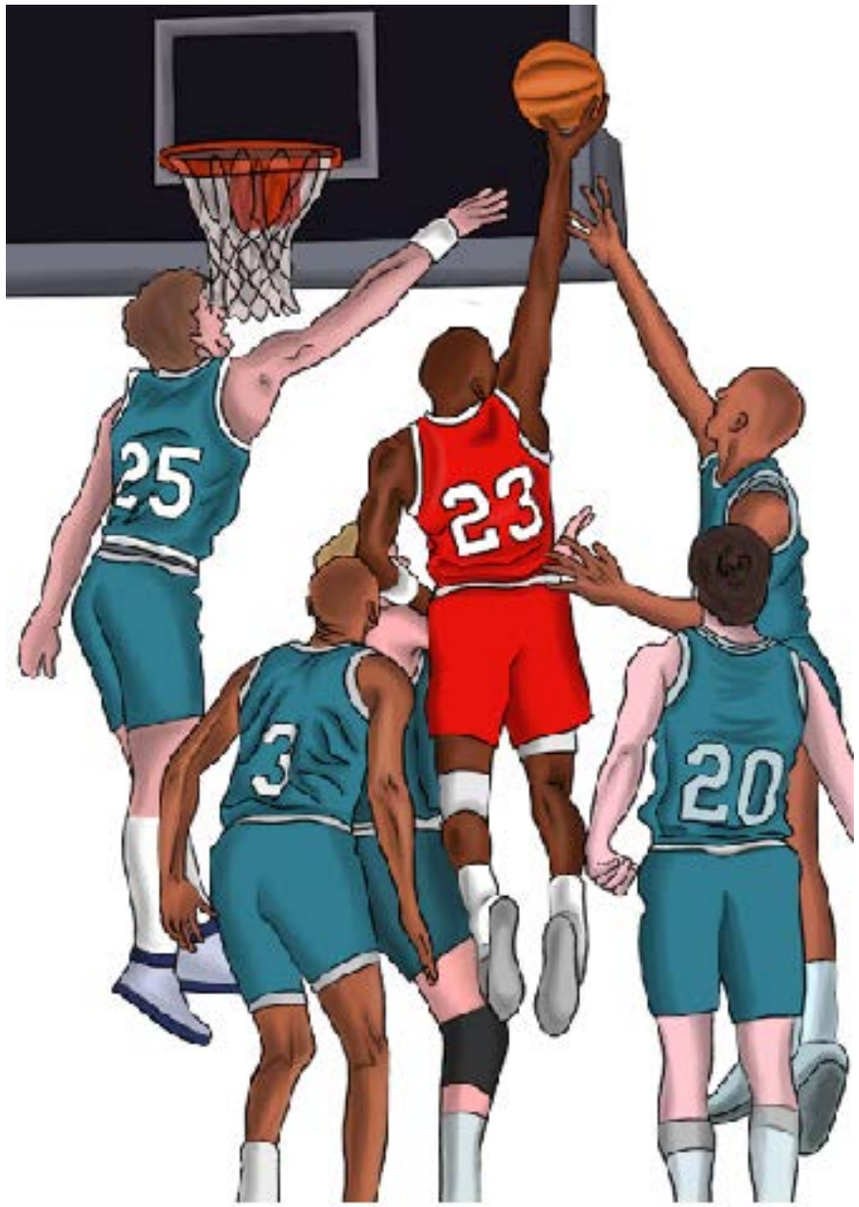
Summary of EO doctrines



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

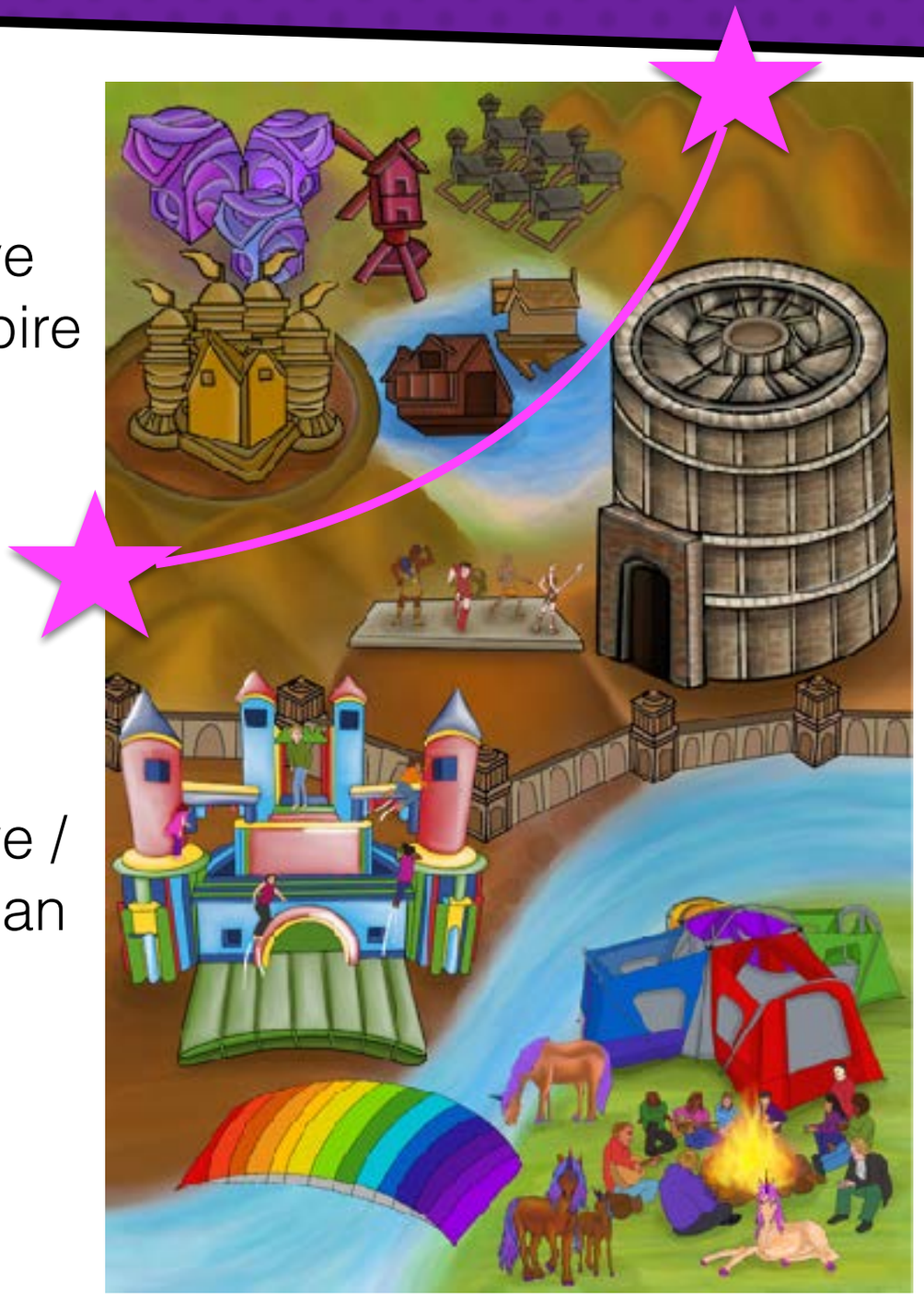
- Similar treatment of similar individuals
- Only irrelevant characteristics separate similar people

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



Formal-ville

substantive / Rawlsian

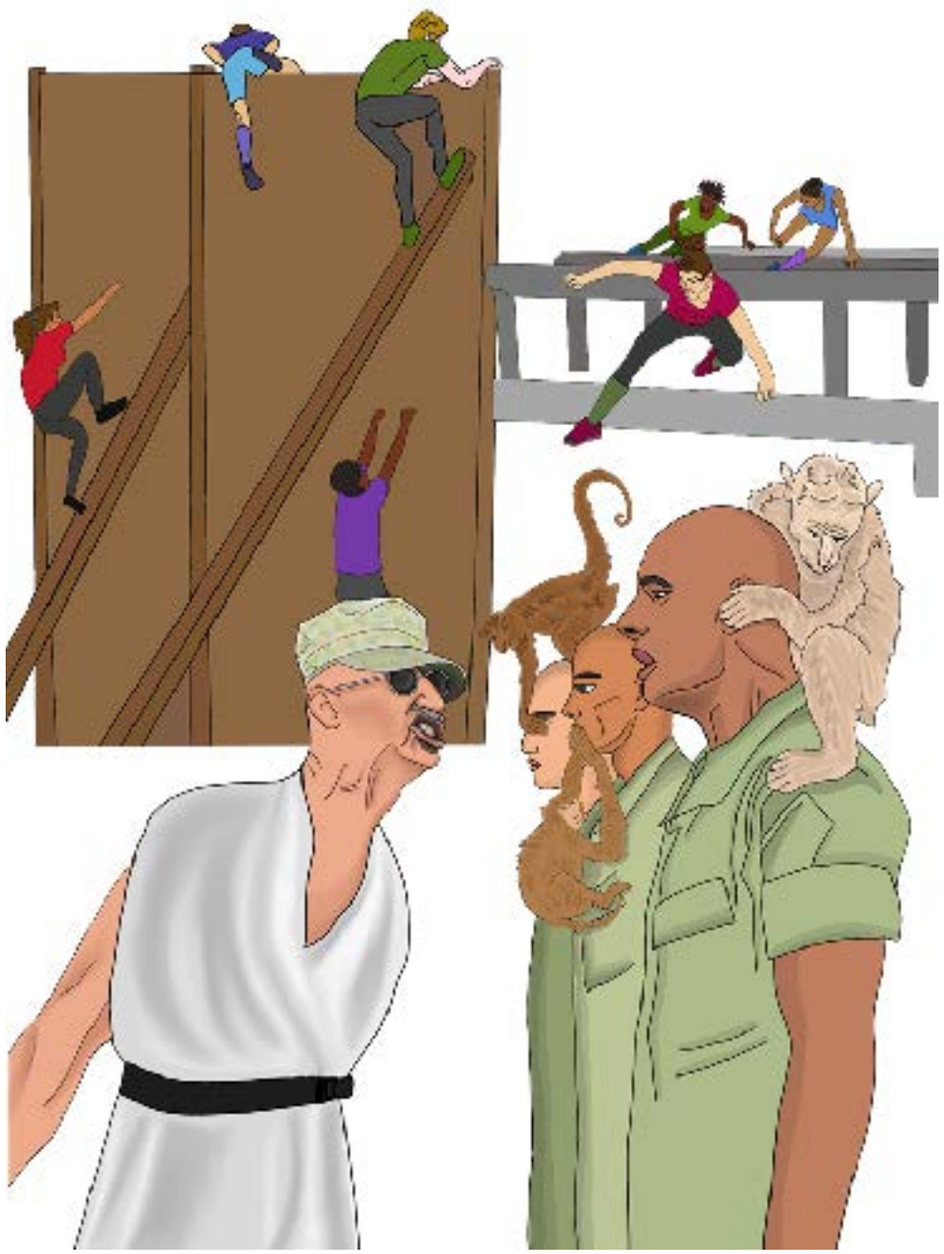
substantive / luck egalitarian

r/ai

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

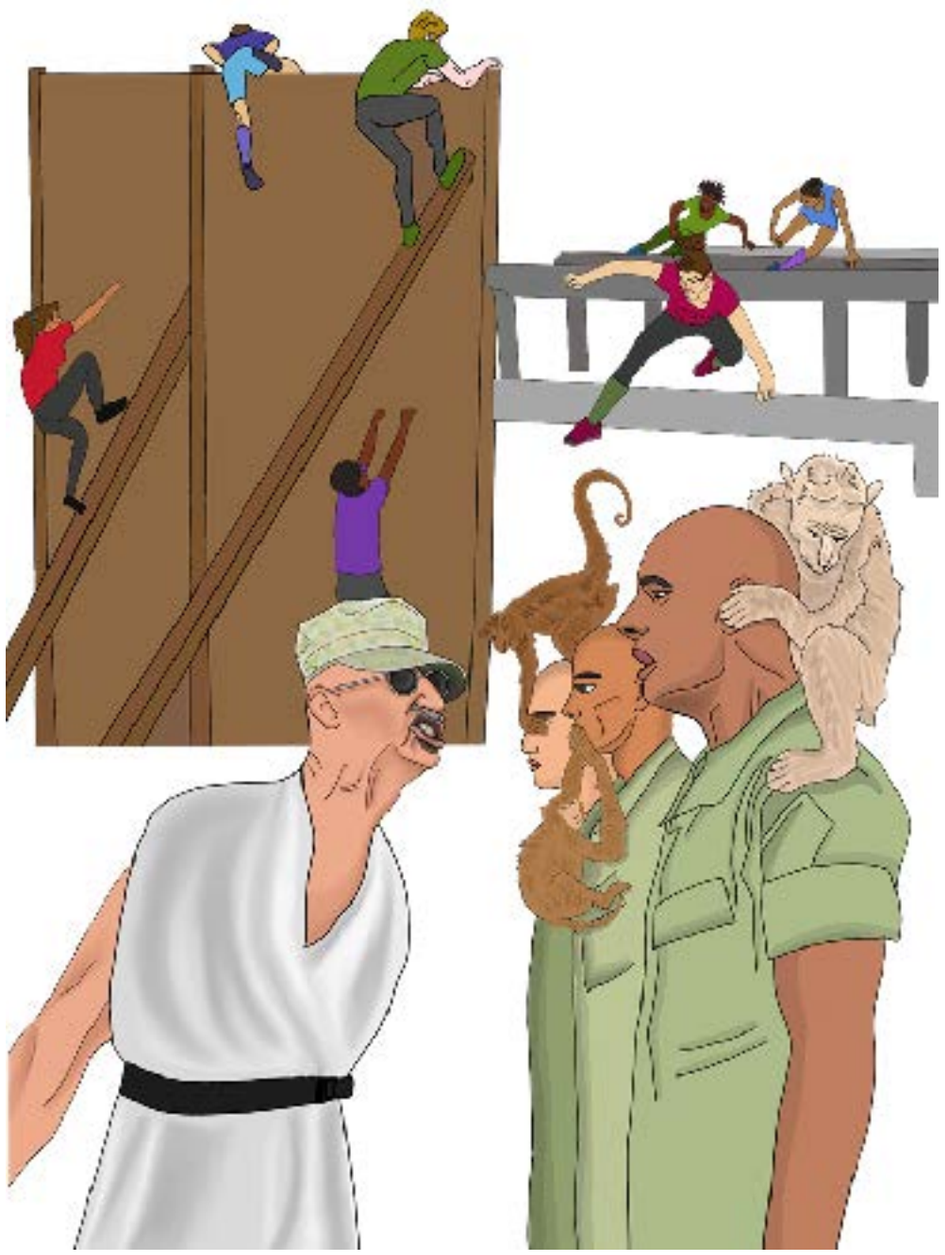


Calibration:

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

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:

$$P\,(y = 1 | y' > \text{p},\, s = 0) = P\,(y = 1 | y' > \text{p},\, s = 1)$$

If two individuals both have job-relevant qualifications $\boldsymbol{y' > p}$ (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

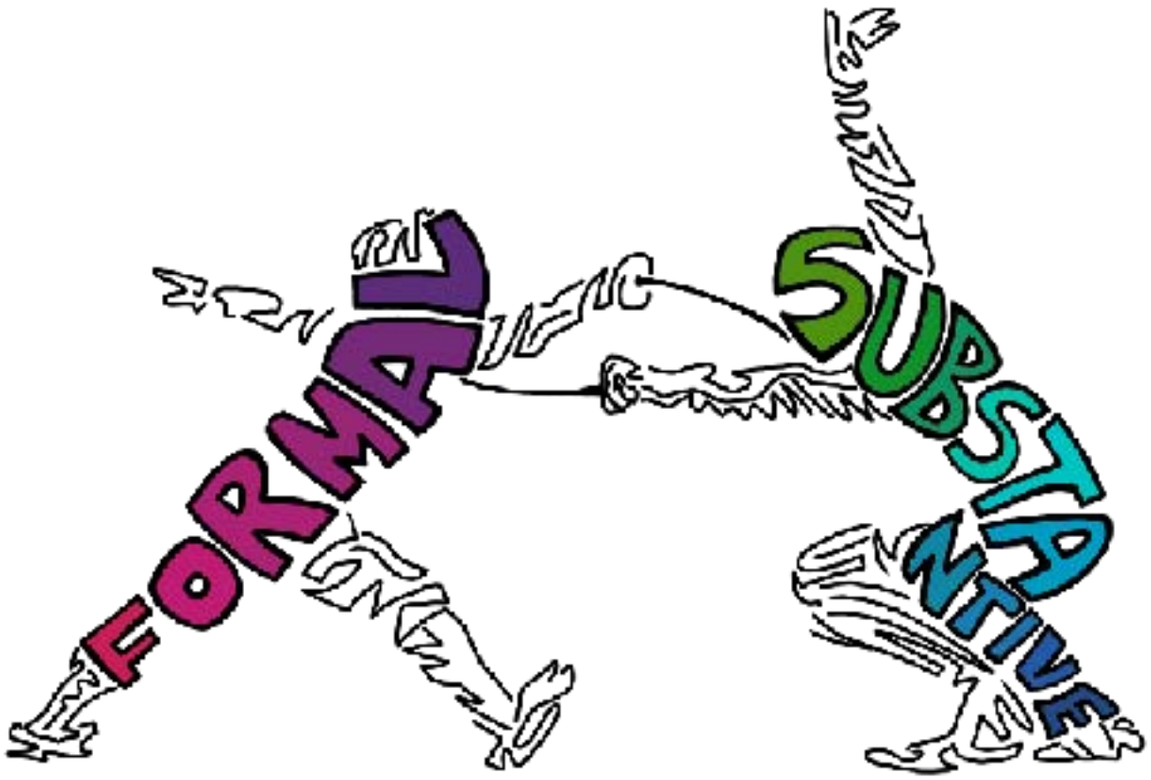
“Before” + “after” → discrimination laundering

- 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

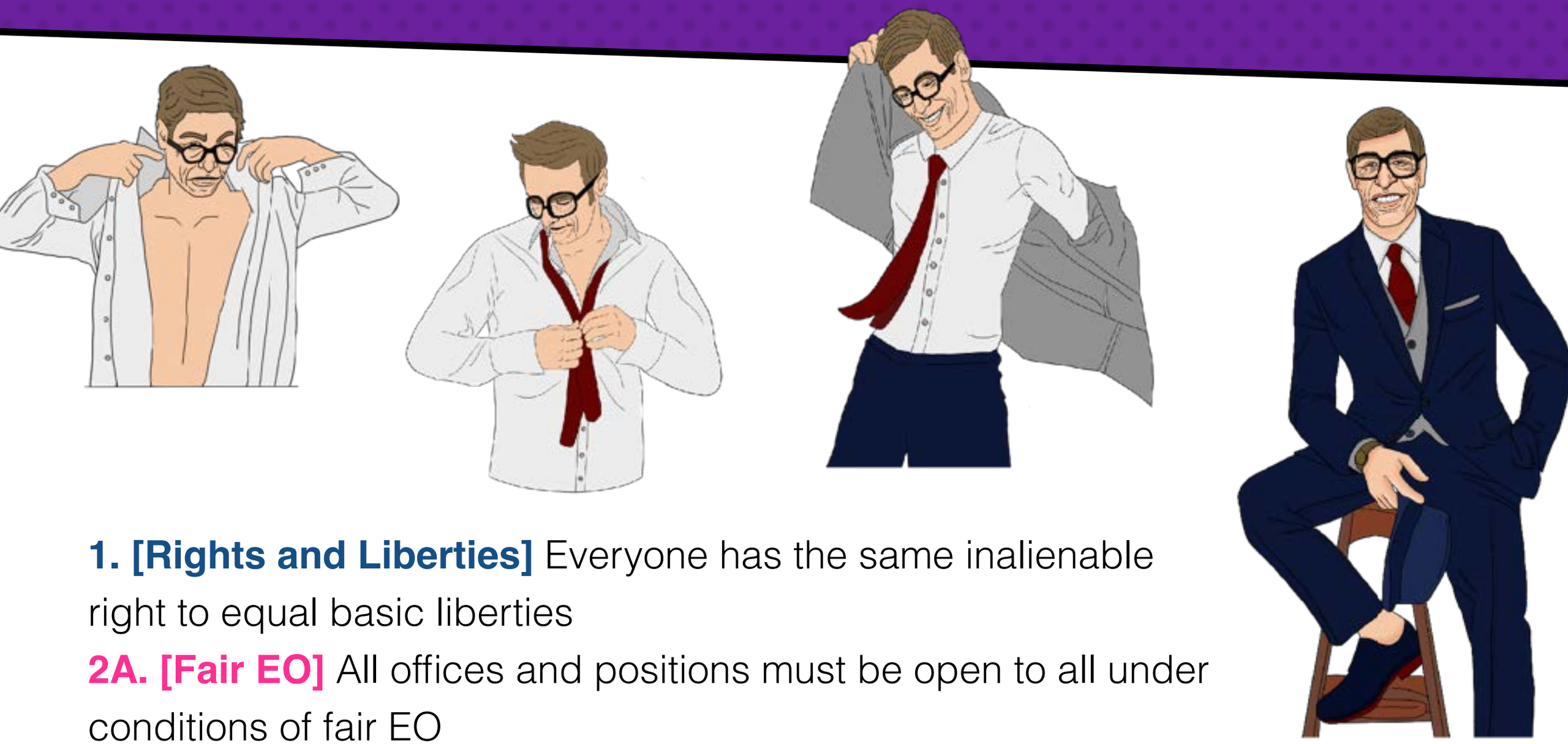
Substantive EO: Rawls

- Equally talented babies myst 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**

r/ai

Substantive EO: Rawls: natural & social lottery

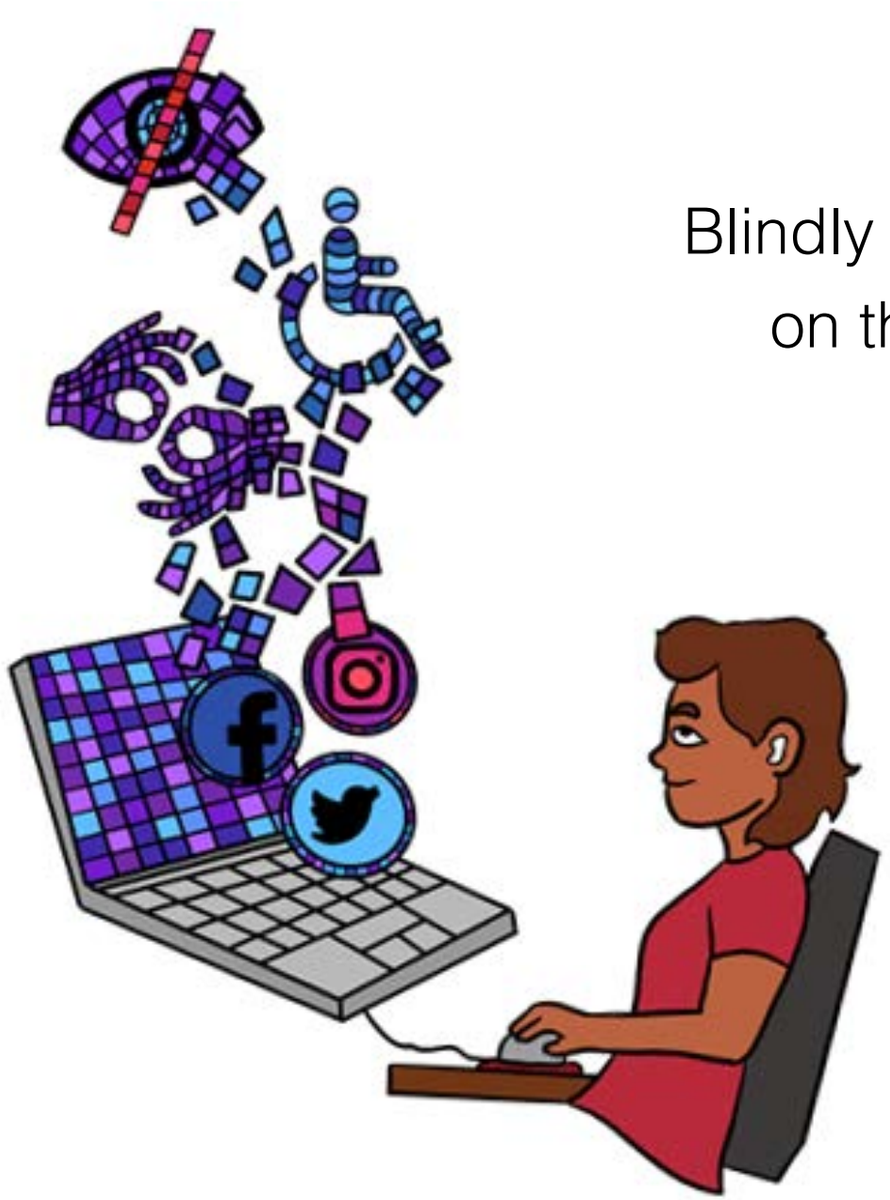
Natural & social lottery: Talents and fortune are distributed arbitrarily.

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

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

r/ai

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

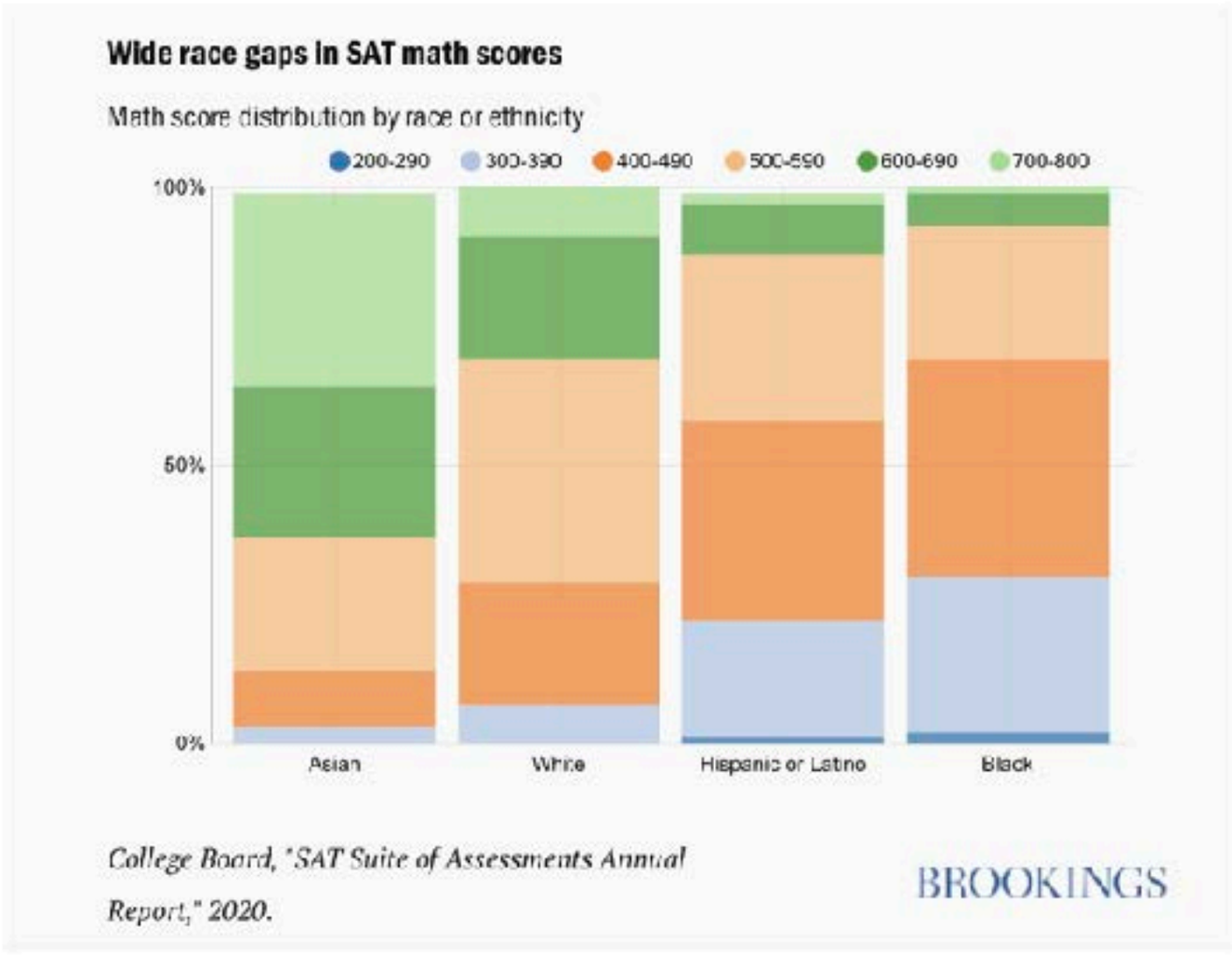
For which characteristics can we hold an individual accountable? (**responsible choice**)

And which matters are completely out of their control? (**brute luck**)

Substantive EO: luck egalitarian: Roemer

Effort, circumstance, and types

(Roemer, 2002)



Substantive EO: Luck egalitarian: Roemer

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

technical example

r/ai

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

score = 373

score = 372

	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)

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**

[Yang, Gkatzelis, Stoyanovich (2019)]

r/ai

From beliefs to interventions

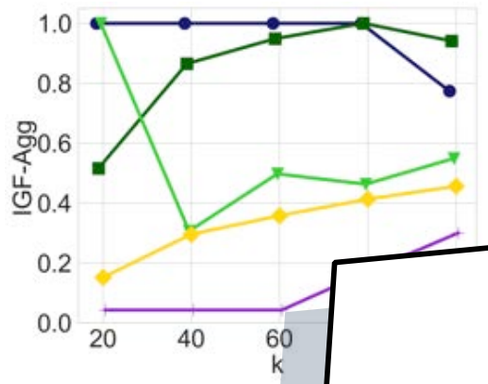
Fairness for female candidates **83 / 95 = 0.91**

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

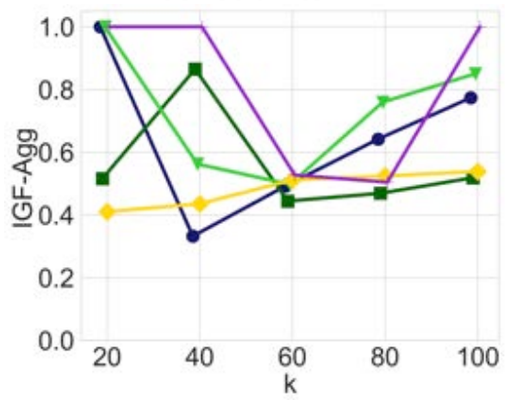
highest-scoring skipped

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**

[Yang, Gkatzelis, Stoyanovich (2019)]

r/ai

another example

Intersectional causal fairness

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

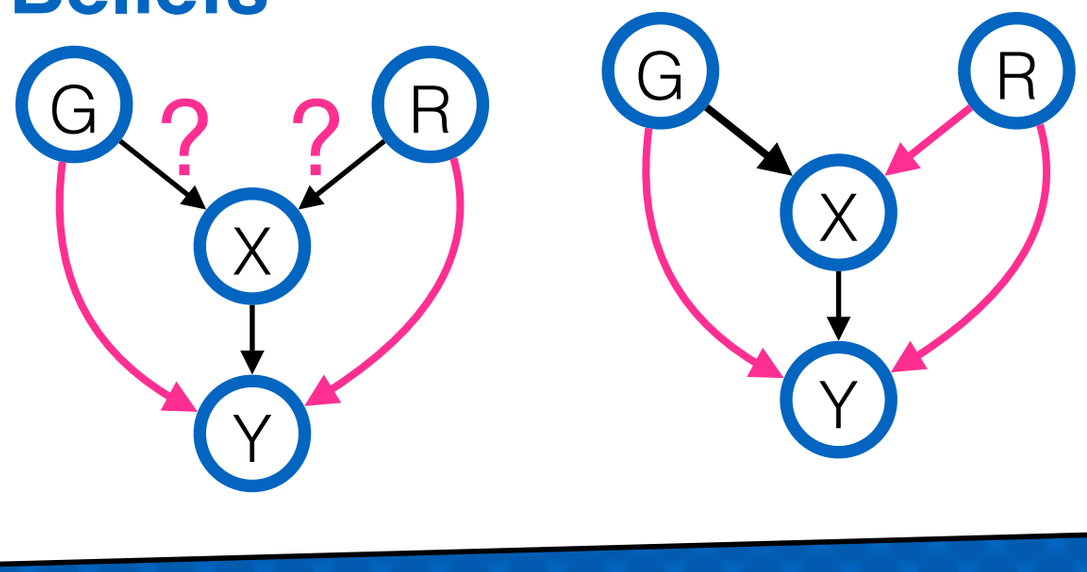
Goal

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

Problem

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

Beliefs



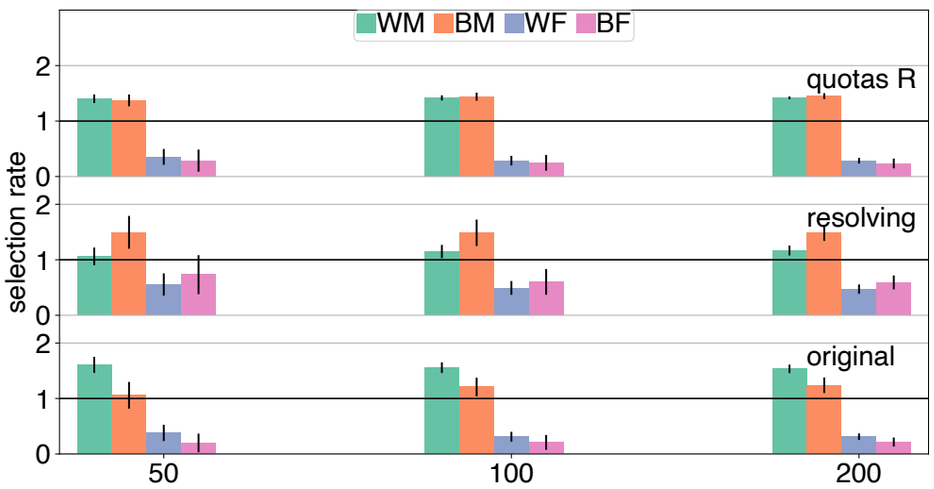
[Yang, Loftus, Stoyanovich (2021)]

r/ai

From beliefs to interventions

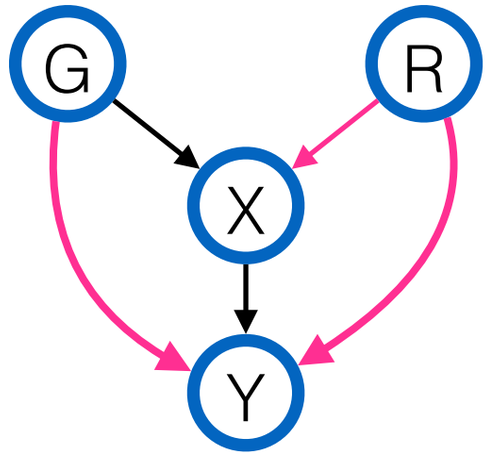
Idea: Compute **counterfactual scores**, treating each individual in the sample as though they had belonged to *one* intersectional group (e.g., Black women). Rank on those scores.

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



Beliefs

allow for resolving mediators



[Yang, Loftus, Stoyanovich (2021)]

r/ai

re-interpretation of EO

r/ai

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

r/ai

Responsible Data Science

Algorithmic Fairness

Thank you!

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai