Responsible Data Science

Algorithmic fairness continued

February 22, 2023

Prof. Elisha Cohen

Center for Data Science
New York University

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai

Overview

- Machine Learning: Train models using labelled data from real world to make predictions/ classifications
- Are decisions made from these models discriminatory or fair?
- Need to formalize definitions of fairness for operationalization

Fairness is an ethical concept

Fairness is not a **technical** or **statistical** concept

No tool or software can fully ‘de-bias’ data or make model ‘fair’

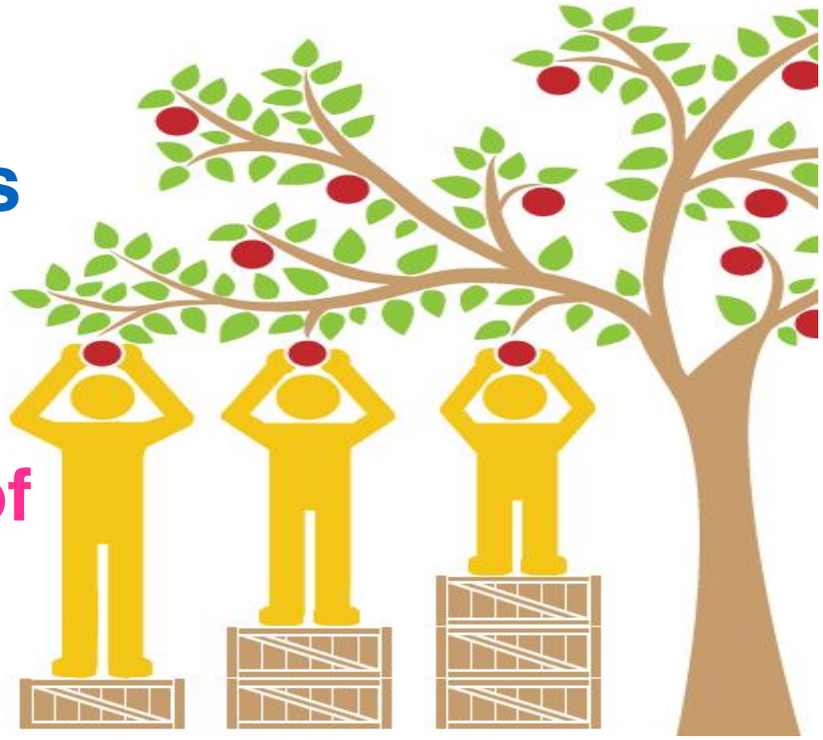
What do we mean by ‘**Fairness**’?

r/ai

Fairness and worldviews

group fairness

equality of outcome



individual fairness

equality of treatment

r/ai

Outcome vs Procedural

Outcome
fairness emphasizes that outcomes meet some requirement

Procedural
fairness emphasizes that the same process be applied to all individuals

Disparate Impact
prohibits unjustified and avoidable disparities in outcomes

Disparate treatment
prohibits procedural unfairness

r/ai

Egalitarianism

Maybe we can get some guidance from political philosophy!

1. Different spheres of justice
    - equal distributions of goods - civil and democratic rights
    - equality of opportunity for competitions for positions and economic goods

Binns "Fairness in Machine Learning: Lessons from Political Philosophy" 2018

Domains of EO

(1) Fairness at a specific decision point

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

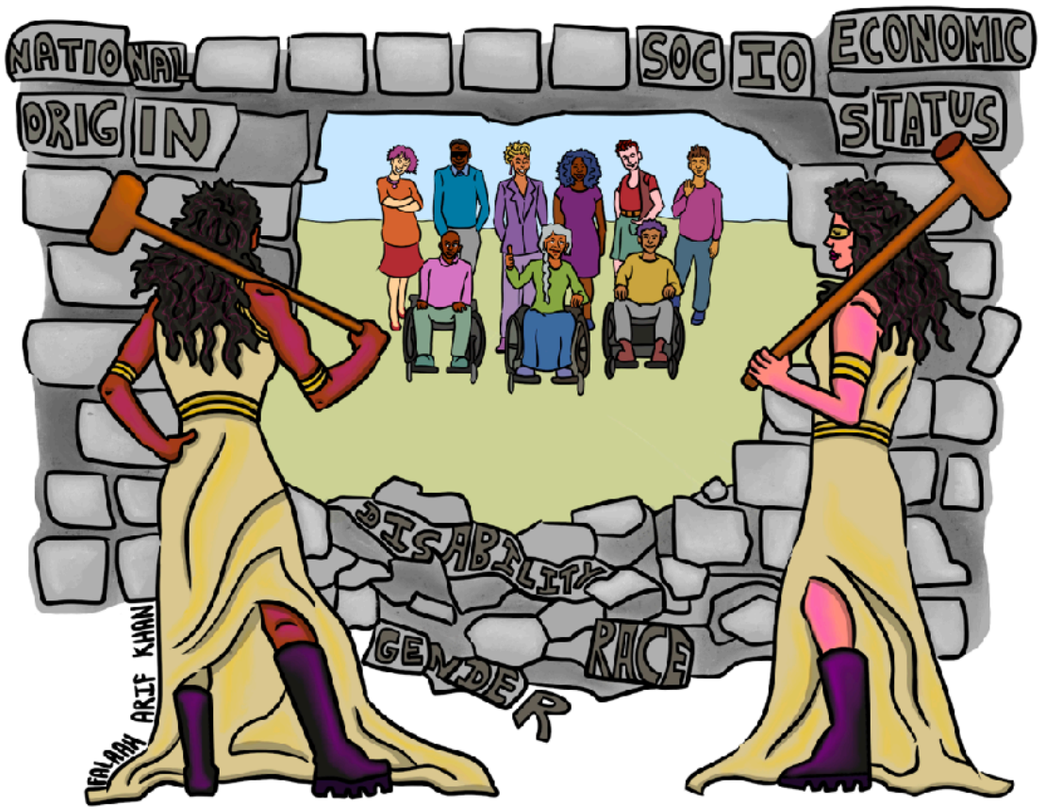
(2) Equality in developmental opportunity

- 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

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

r/ai

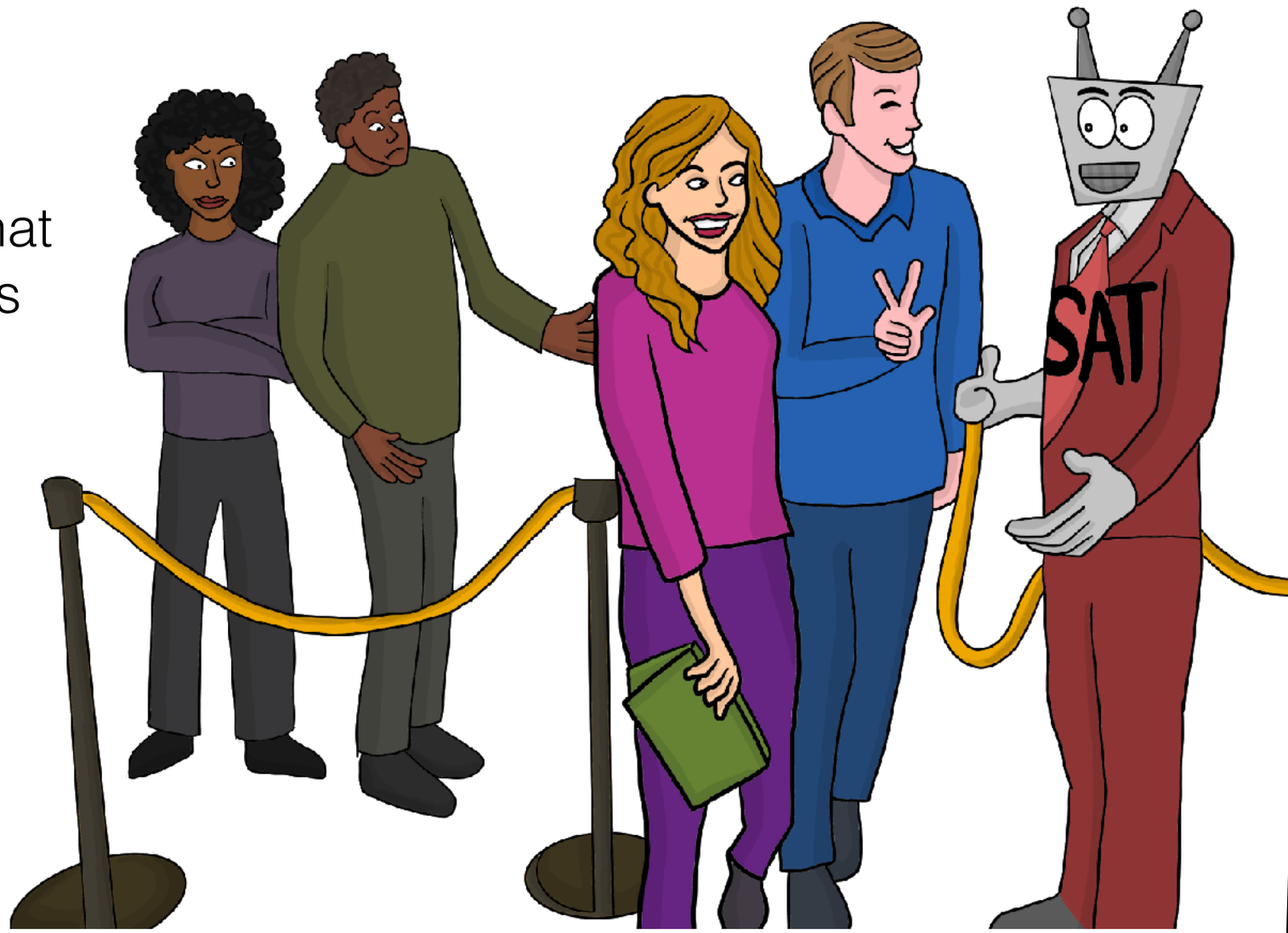
Formal EO: Careers open to talents

- 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: 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” and “after” problem

- Formal EO’s appeal: relevant skills in, irrelevant characteristics out
- But OK to use irrelevant privileges before competition
- So privileges affect competition outcomes
- Winners at time 1 gain improved characteristics for competing at time 2

How do we combine concepts of fair contests with fair life chances?

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

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?

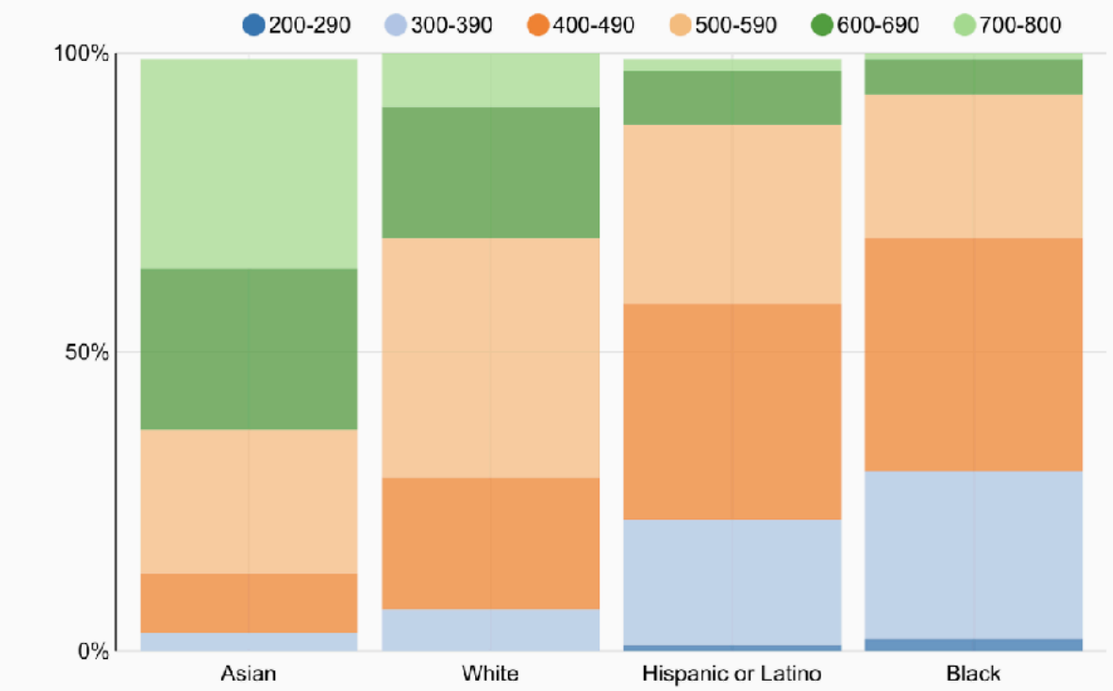
For which characteristics can we hold an individual accountable? (**responsible choice**)

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

Substantive EO: luck egalitarian: Roemer

Effort, circumstance, and types

(Roemer, 2002)



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

BROOKINGS

r/ai

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

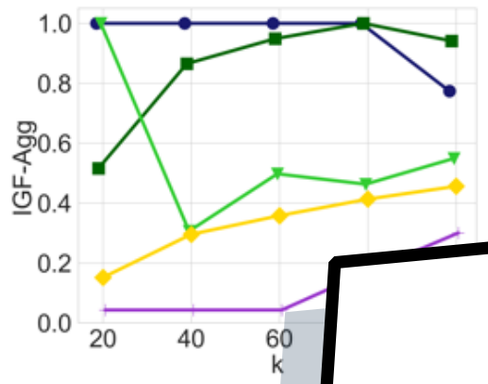
Fairness for female candidates **83 / 95 = 0.91**

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

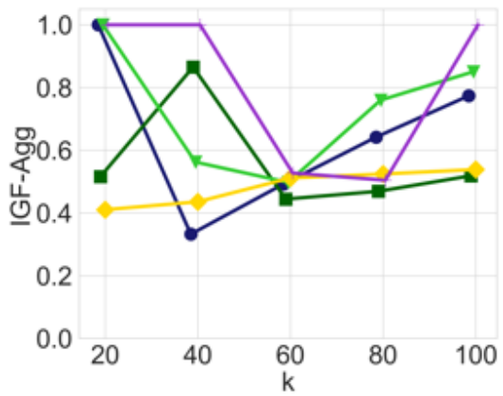
highest-scoring skipped

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)]

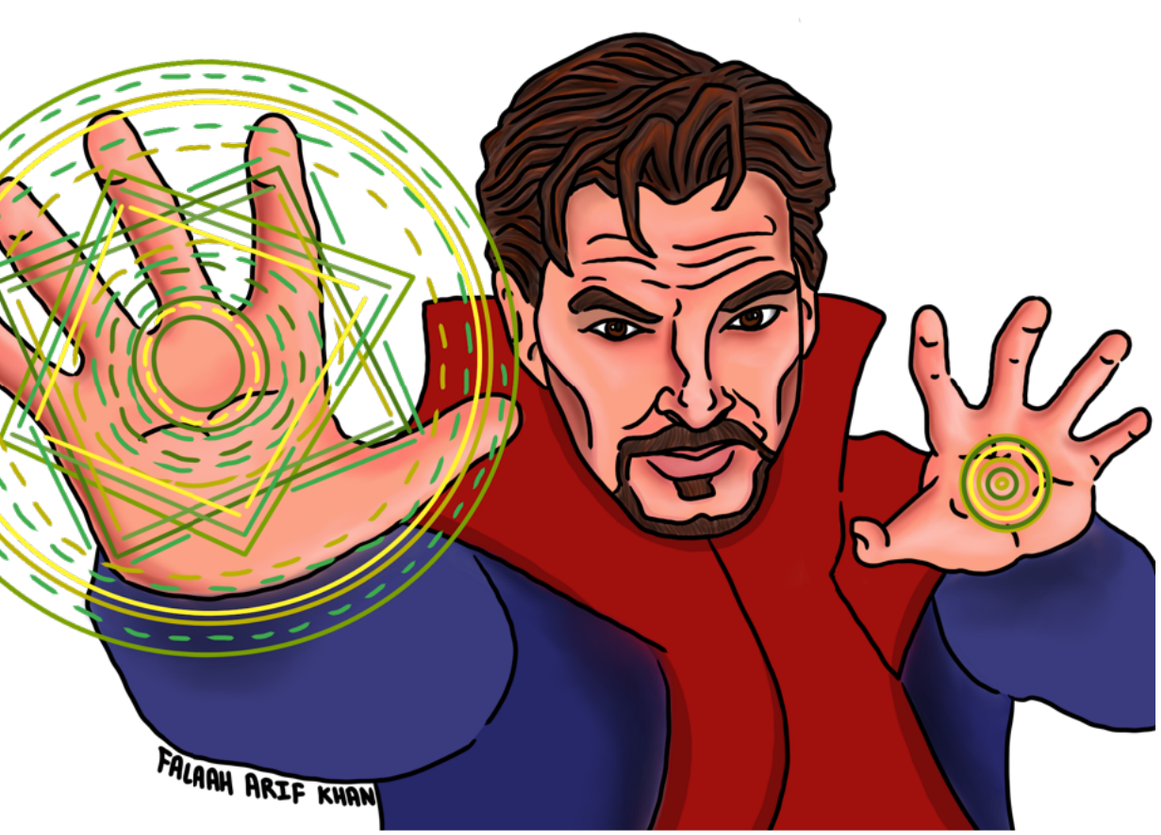
r/ai

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

Fairness module, key ideas

Week 1:

- Goals, benefits, and harms of DS systems
- Stakeholders

Week 2:

- Fairness in classification and risk assessment
- Individual fairness vs group fairness
- Disparate treatment vs disparate impact
- Impossibility result (calibration versus balance of errors)
- Three types of bias in computer systems (pre-existing, technical, emergent)

Week 3:

- Five fairness definitions (FTU, individual fairness, demographic parity, equalized odds, calibration)
- Causal models, causal framework for fairness (causal diagrams, counterfactual fairness)

Week 4:

- Causal framework for fairness continued (causal pathways, counterfactual privilege)
- Philosophical frameworks for fairness
- Fairness as equal opportunity (EOP), formal EOP, substantive EOP

Responsible Data Science

Algorithmic Fairness

Thank you!

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai