Responsible Data Science

Algorithmic Fairness

January 30 & February 1, 2023

Prof. Elisha Cohen
Center for Data Science

NYU Center for Data Science

r/ai

RDS course overview

So what is RDS?

As advertised: ethics, legal compliance, personal responsibility.
But also: **data quality**!

A technical course, with content drawn from:

1. fairness, accountability and transparency
2. data engineering
3. privacy & data protection

We will learn **algorithmic techniques** for data analysis.
We will also learn about recent **laws** / **regulatory frameworks**.

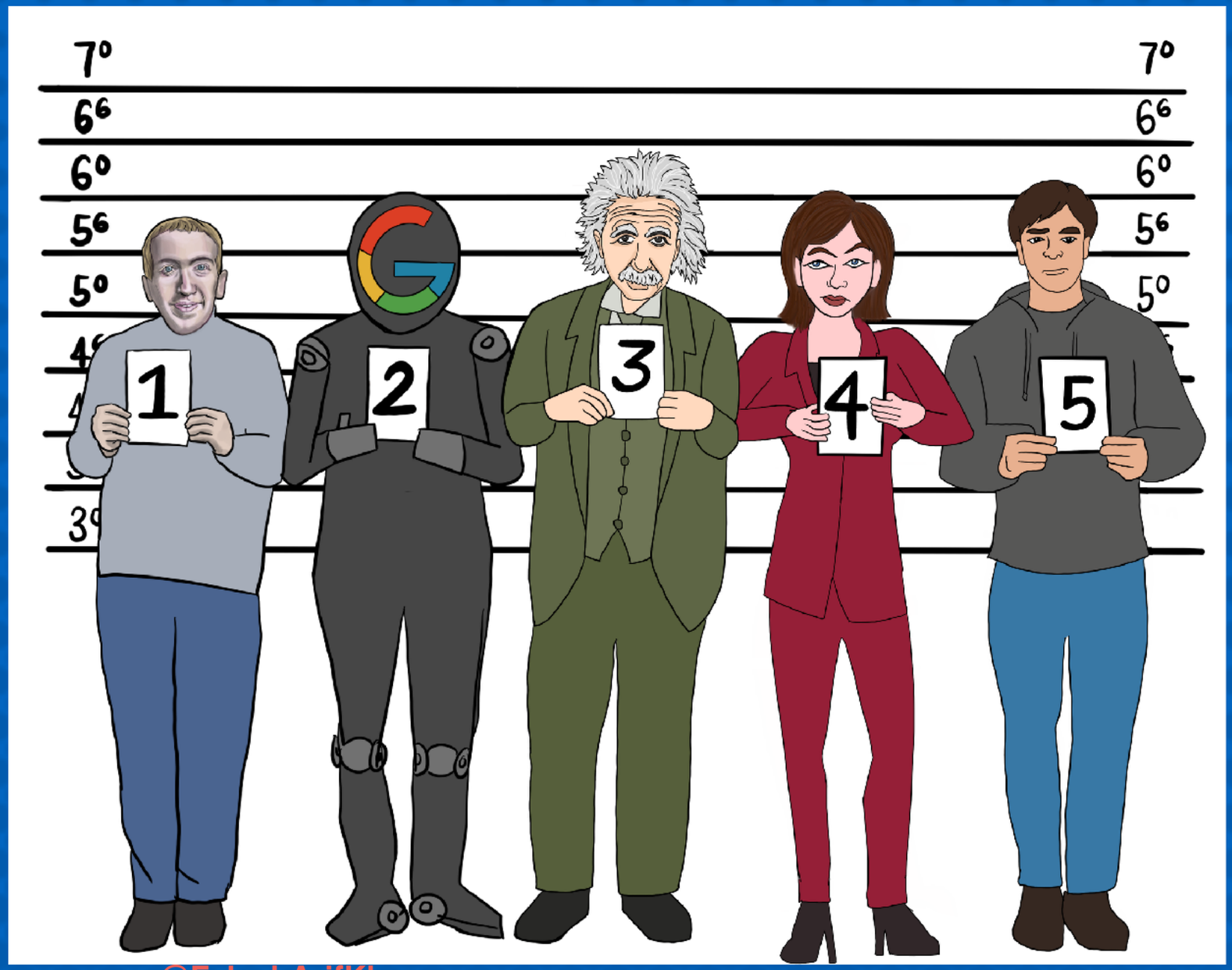
Bottom line: we will learn that many of the problems are **socio-technical**, and so cannot be “solved” with technology alone.

My perspective: a pragmatic engineer, **not** a technology skeptic.

Nuance, please!

r/ai

We all are responsible



@FalaahArifKhan

Reading: Algorithmic bias Week 1

Bias in Computer Systems

BATYA FRIEDMAN
Colby College and The Mina Institute
and
HELEN NISSENBAUM
Princeton University

From an analysis of actual cases, three categories of bias in computer systems have been developed: preexisting, technical, and emergent. Preexisting bias has its roots in social institutions, practices, and attitudes. Technical bias arises from technical constraints or considerations. Emergent bias arises in a context of use. Although others have pointed to bias in particular computer systems and have noted the general problem, we know of no comparable work that examines this phenomenon comprehensively and which offers a framework for understanding and remedying it. We conclude by suggesting that freedom from bias should be counted among the select set of criteria—including reliability, accuracy, and efficiency—according to which the quality of systems in use in society should be judged.

Categories and Subject Descriptors: D.2.0 [**Software**]: Software Engineering; H.1.2 [**Information Systems**]: User/Machine Systems; K.4.0 [**Computers and Society**]: General

General Terms: Design, Human Factors

Additional Key Words and Phrases: Bias, computer ethics, computers and society, design methods, ethics, human values, standards, social computing, social impact, system design, universal design, values

[Friedman & Nissenbaum, Comm ACM (1996)]

Reading: Algorithmic bias Week 2

DOI:10.1145/3376898

A group of industry, academic, and government experts convene in Philadelphia to explore the roots of algorithmic bias.

BY ALEXANDRA CHOULDECHOVA AND AARON ROTH

A Snapshot of the Frontiers of Fairness in Machine Learning

[Chouldechova & Roth, Comm ACM (2020)]

Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg[1], Sendhil Mullainathan[2], and Manish Raghavan[3]

1 Cornell University, Ithaca, USA
kleinber@cs.cornell.edu
2 Harvard University, Cambridge, USA
mullain@fas.harvard.edu
3 Cornell University, Ithaca, USA
manish@cs.cornell.edu

Abstract

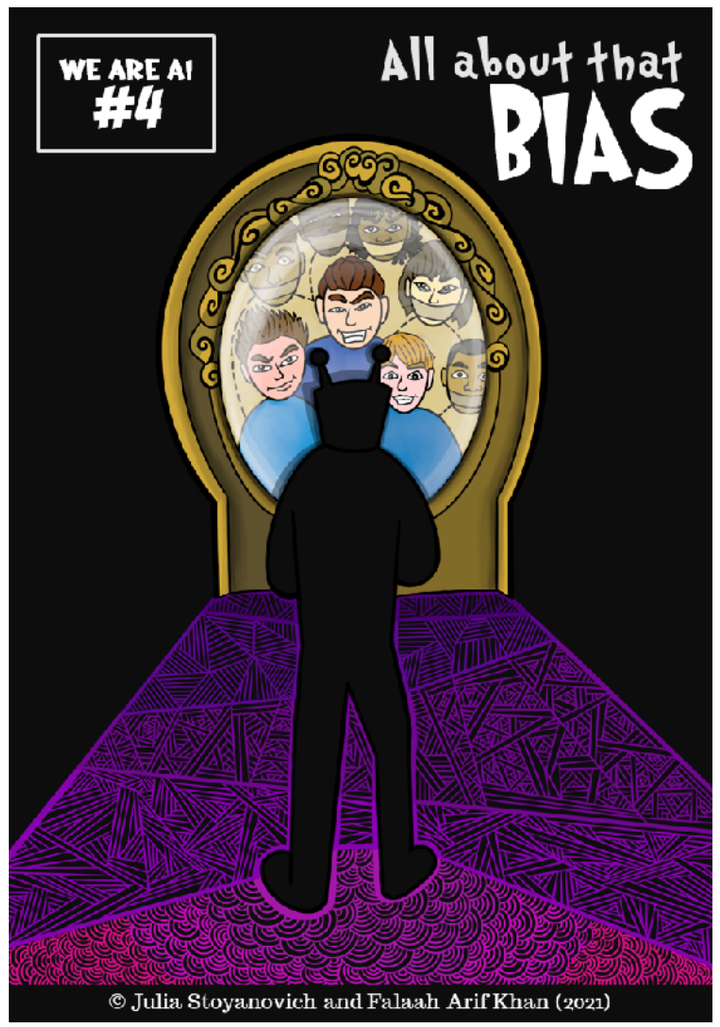
Recent discussion in the public sphere about algorithmic classification has involved tension between competing notions of what it means for a probabilistic classification to be fair to different groups. We formalize three fairness conditions that lie at the heart of these debates, and we prove that except in highly constrained special cases, there is no method that can satisfy these three conditions simultaneously. Moreover, even satisfying all three conditions approximately requires that the data lie in an approximate version of one of the constrained special cases identified by our theorem. These results suggest some of the ways in which key notions of fairness are incompatible with each other, and hence provide a framework for thinking about the trade-offs between them.

1998 ACM Subject Classification H.2.8 Database Applications, J.1 Administrative Data Processing

Keywords and phrases algorithmic fairness, risk tools, calibration

Digital Object Identifier 10.4230/LIPIcs.ITCS.2017.43

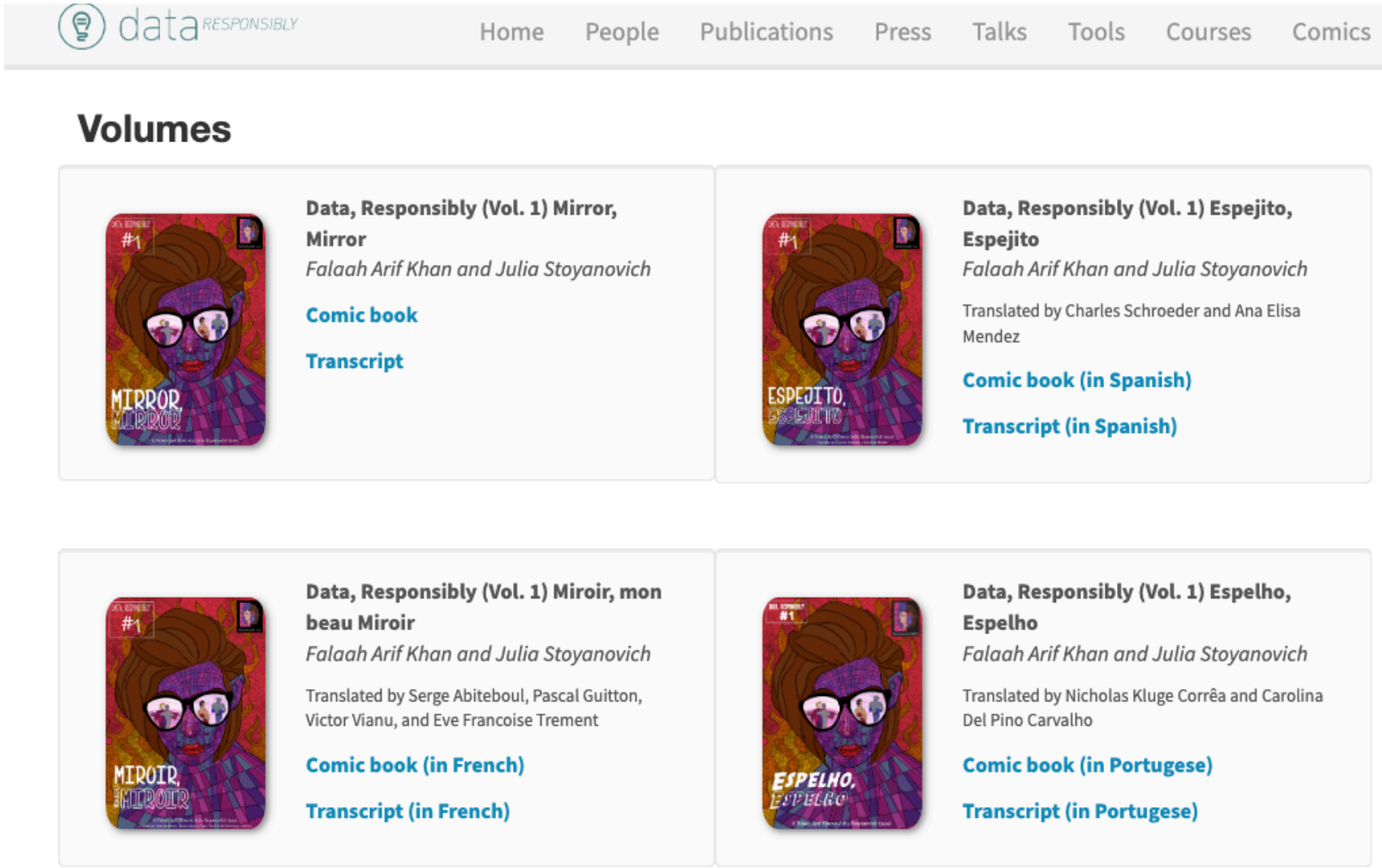
[Kleinberg, Mullainathan & Raghavan, ITCS (2017)]



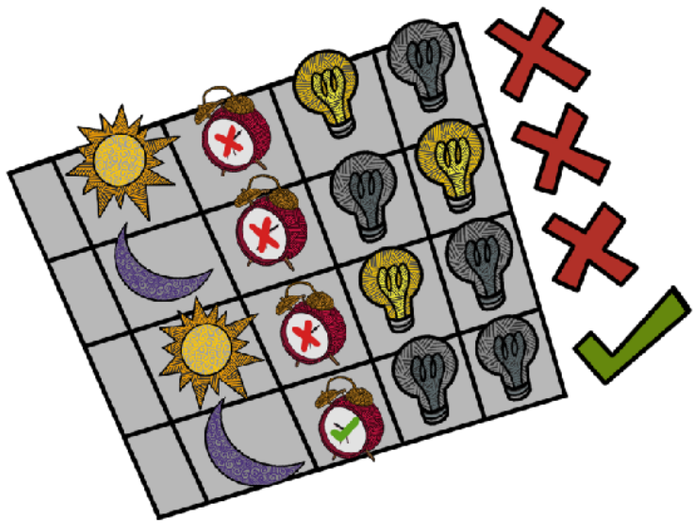
r/ai

Comics

https://dataresponsibly.github.io/comics/



Recall: Individual & cumulative harms

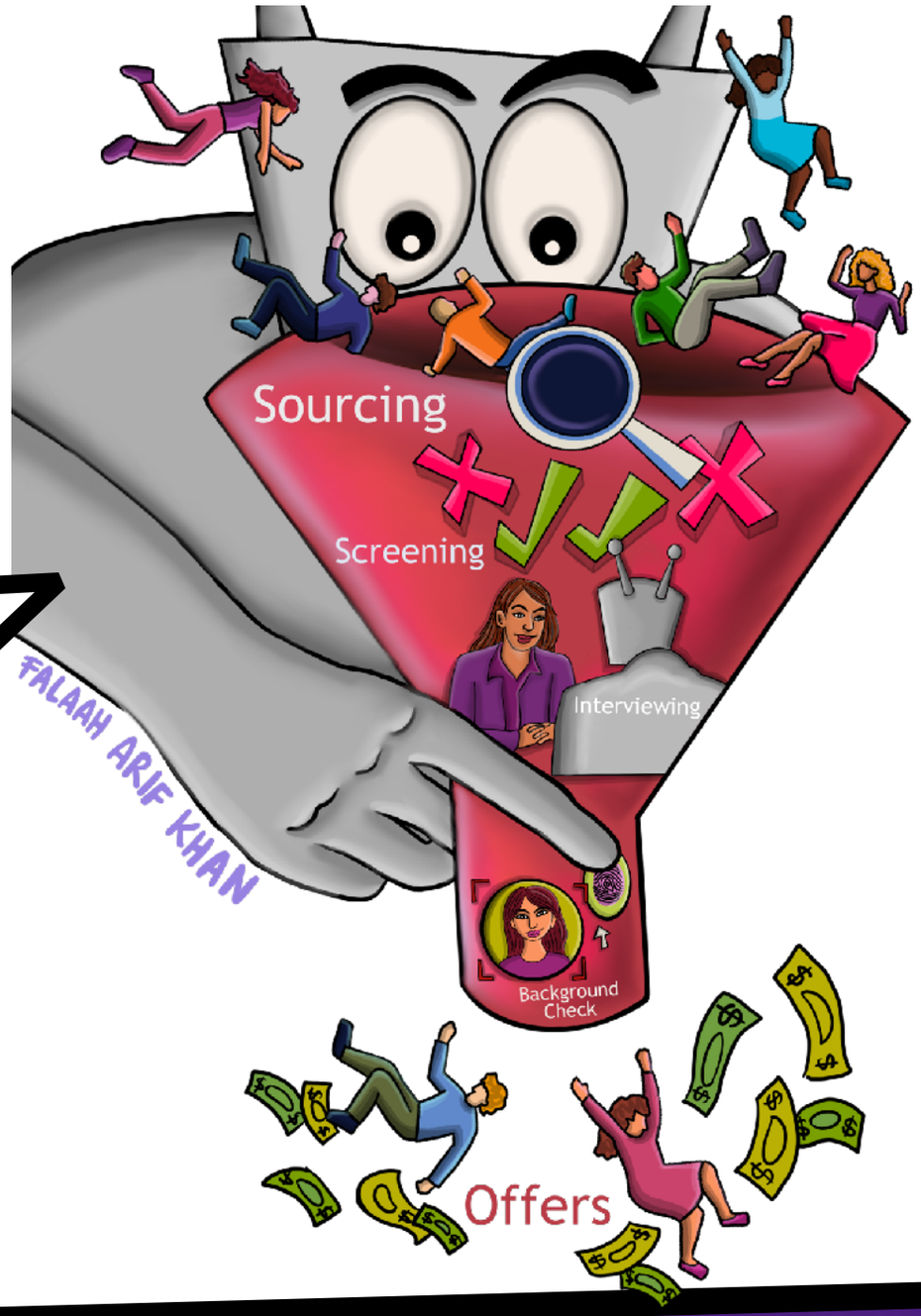


Questions to keep in mind:

what are the **goals** of the AI system?

what are the **benefits** and to **whom**?

what are the **harms** and to **whom**?



r/ai

fairness in classification

r/ai

Vendors and outcomes

Consider a **vendor** assigning positive or negative **outcomes** to individuals.

Positive Outcomes	Negative Outcomes
offered employment	not offered employment
accepted to school	not accepted to school
offered a loan	denied a loan
shown relevant ad for shoes	shown irrelevant ad for shoes

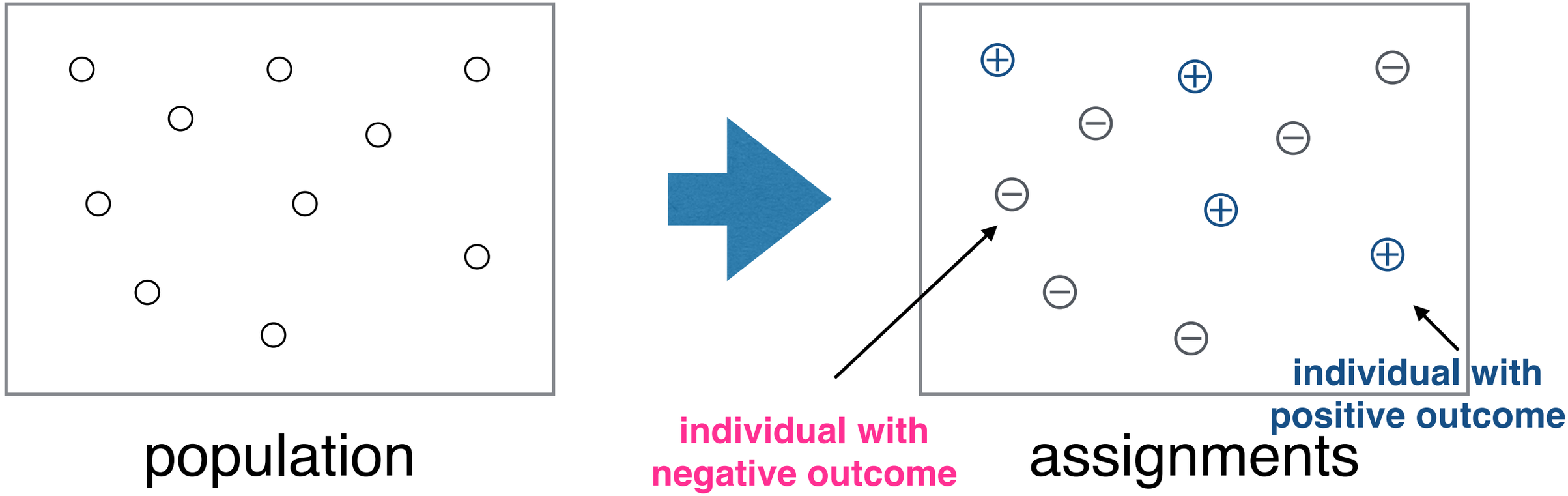
r/ai

Fairness in classification

Fairness in classification is concerned with how outcomes are assigned to a population

positive outcomes

40% of the population

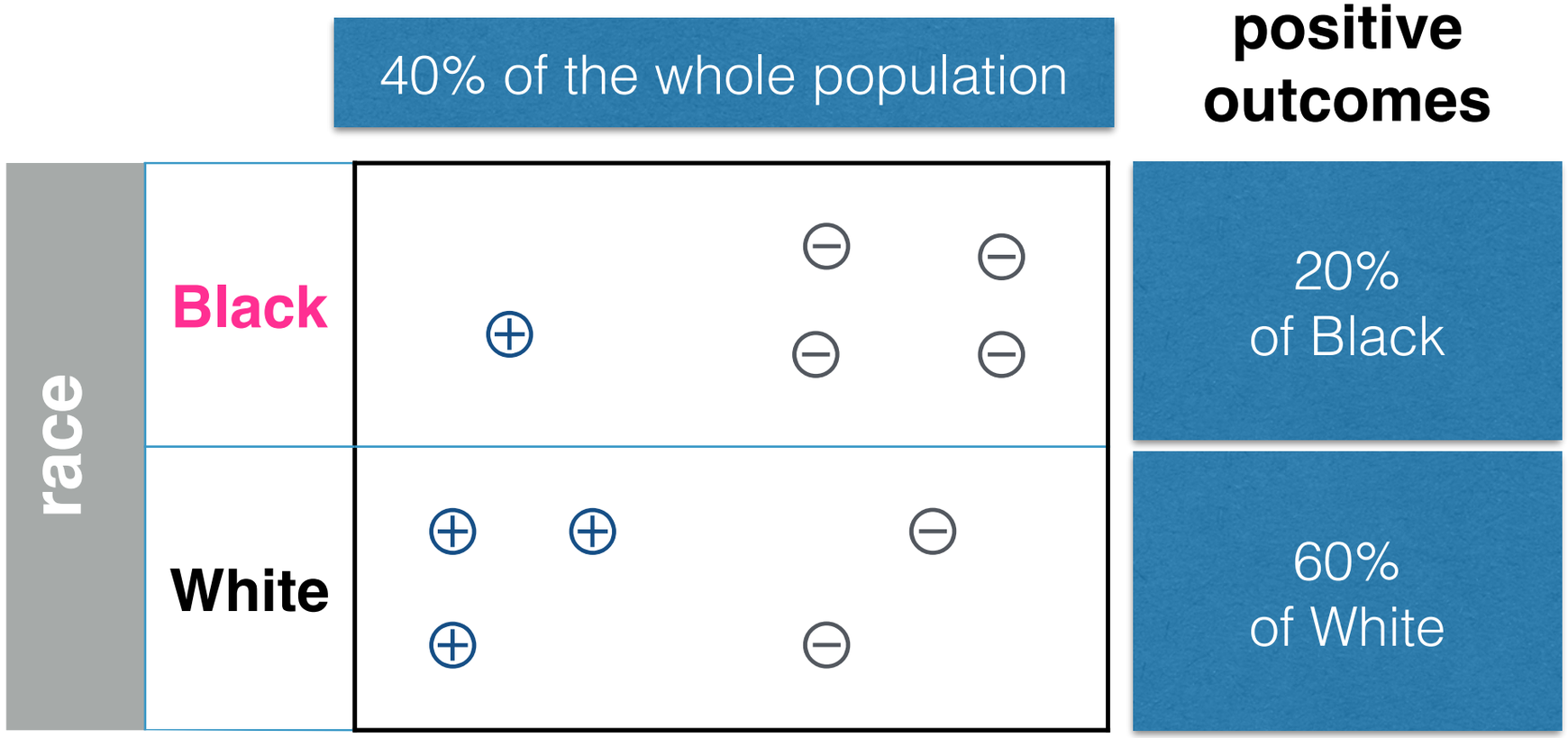


population

assignments

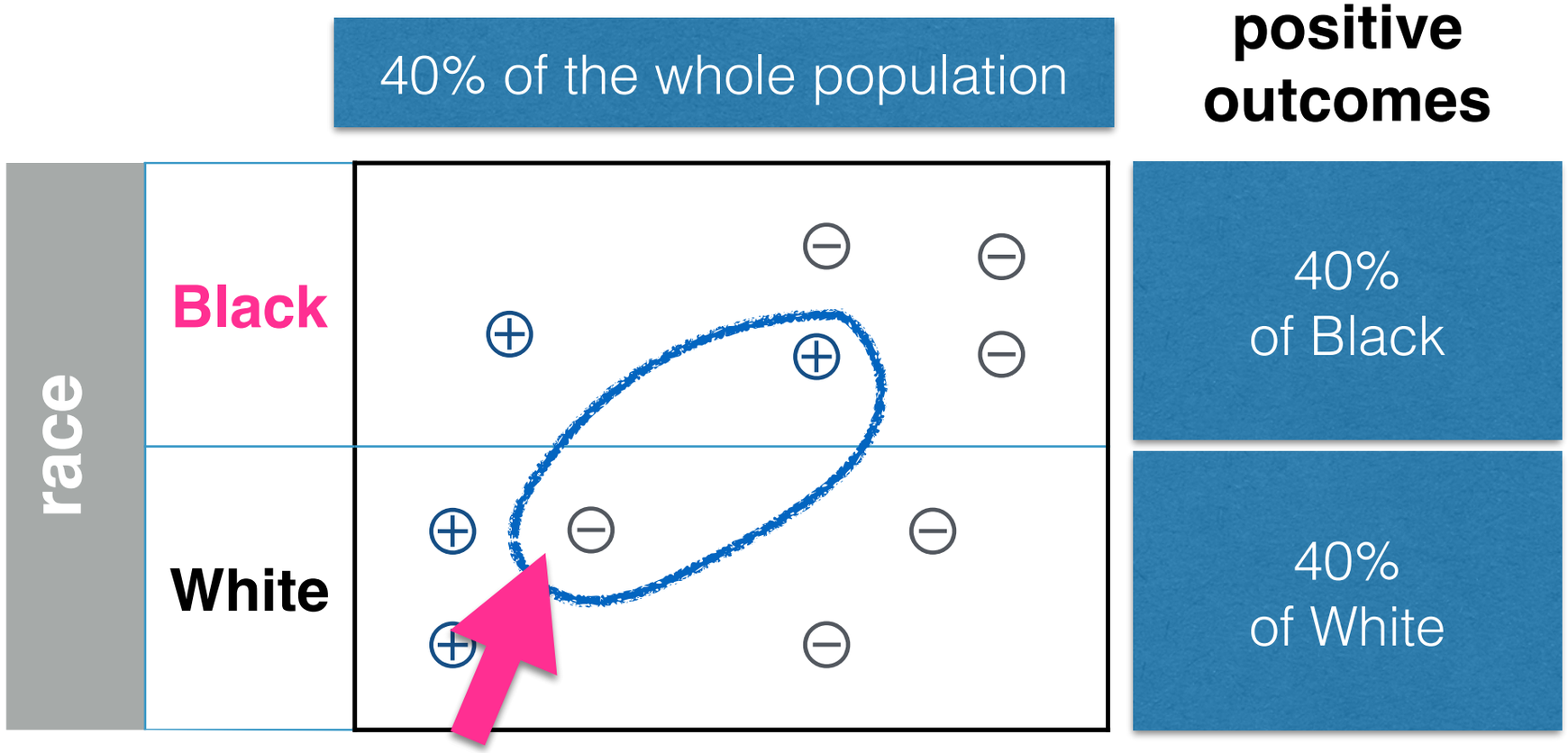
Fairness in classification

Sub-populations may be treated differently



Fairness in classification

Sub-populations may be treated differently



Fairness in classification

Explaining the disparity with proxy variables

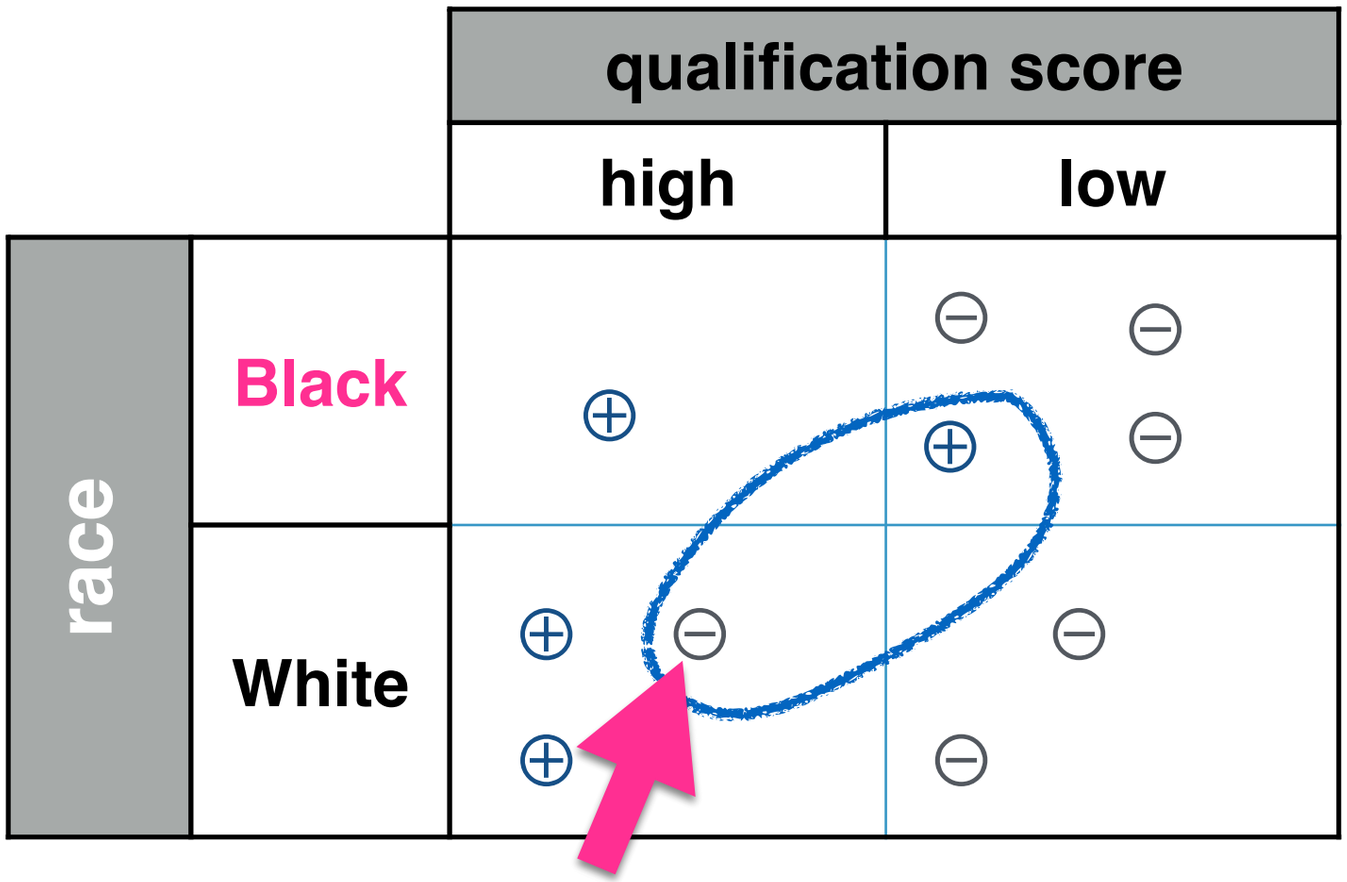
		qualification score: high	qualification score: low	positive outcomes
race	Black	⊕	⊖ ⊖ ⊖	20% of Black
	White	⊕ ⊕ ⊕	⊖ ⊖	60% of White

r/ai

discussion

r/ai

Swapping outcomes



positive outcomes

40%
of Black

40%
of White

Two families of fairness measures

Group fairness (here, **statistical parity)**

demographics of the individuals receiving any outcome - positive or negative - should be the same as demographics of the underlying population

Individual fairness

any two individuals who are similar **with respect to a task** should receive similar outcomes

r/ai

Bias in computer systems

Pre-existing is independent of an algorithm and has origins in society

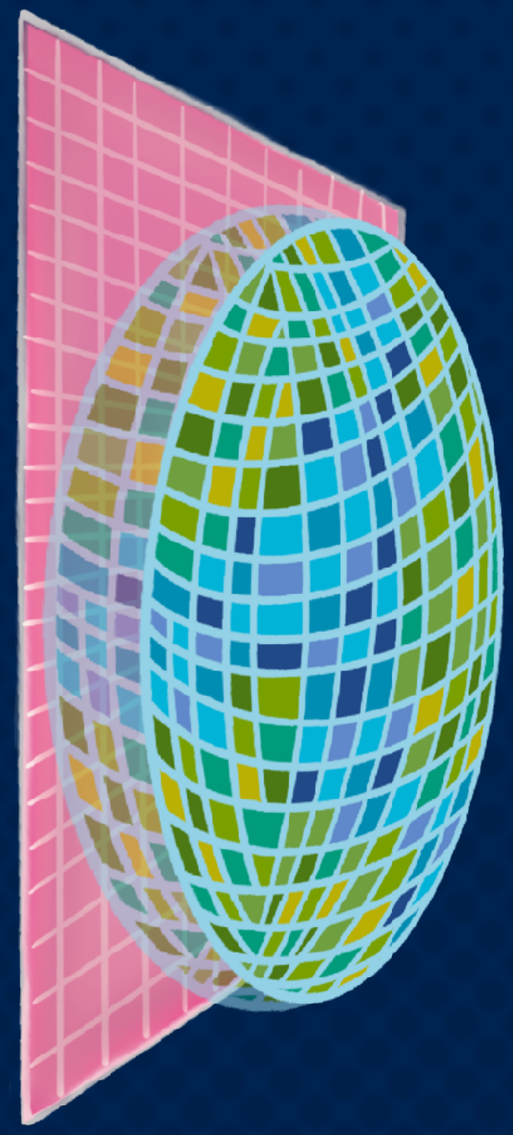
Technical is introduced or exacerbated by the technical properties of an ADS

Emergent arises due to context of use

[Friedman & Nissenbaum (1996)]

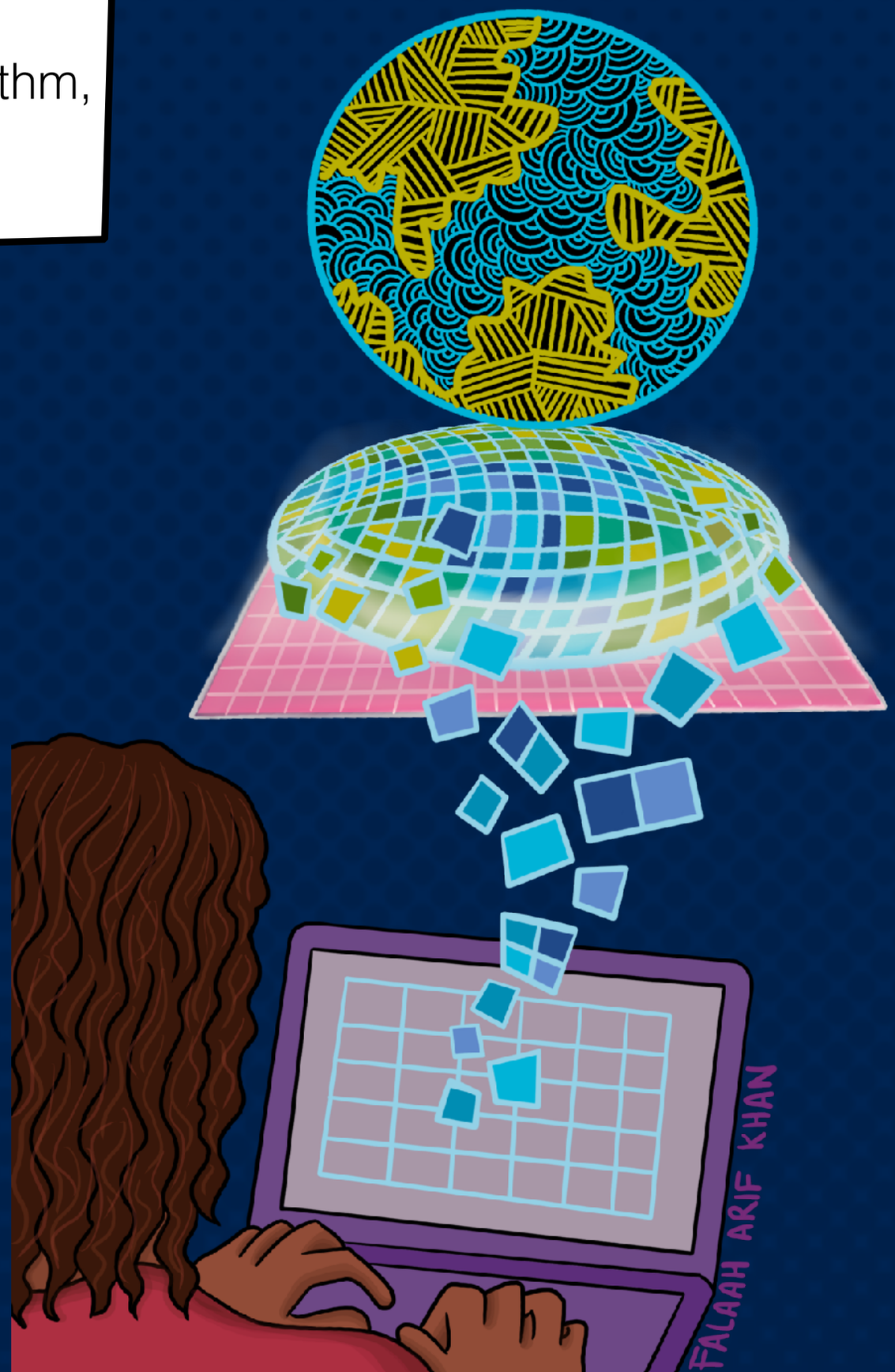
r/ai

Pre-existing bias:
independent of an algorithm, has its origins in society



r/ai

Pre-existing bias:
independent of an algorithm, has its origins in society

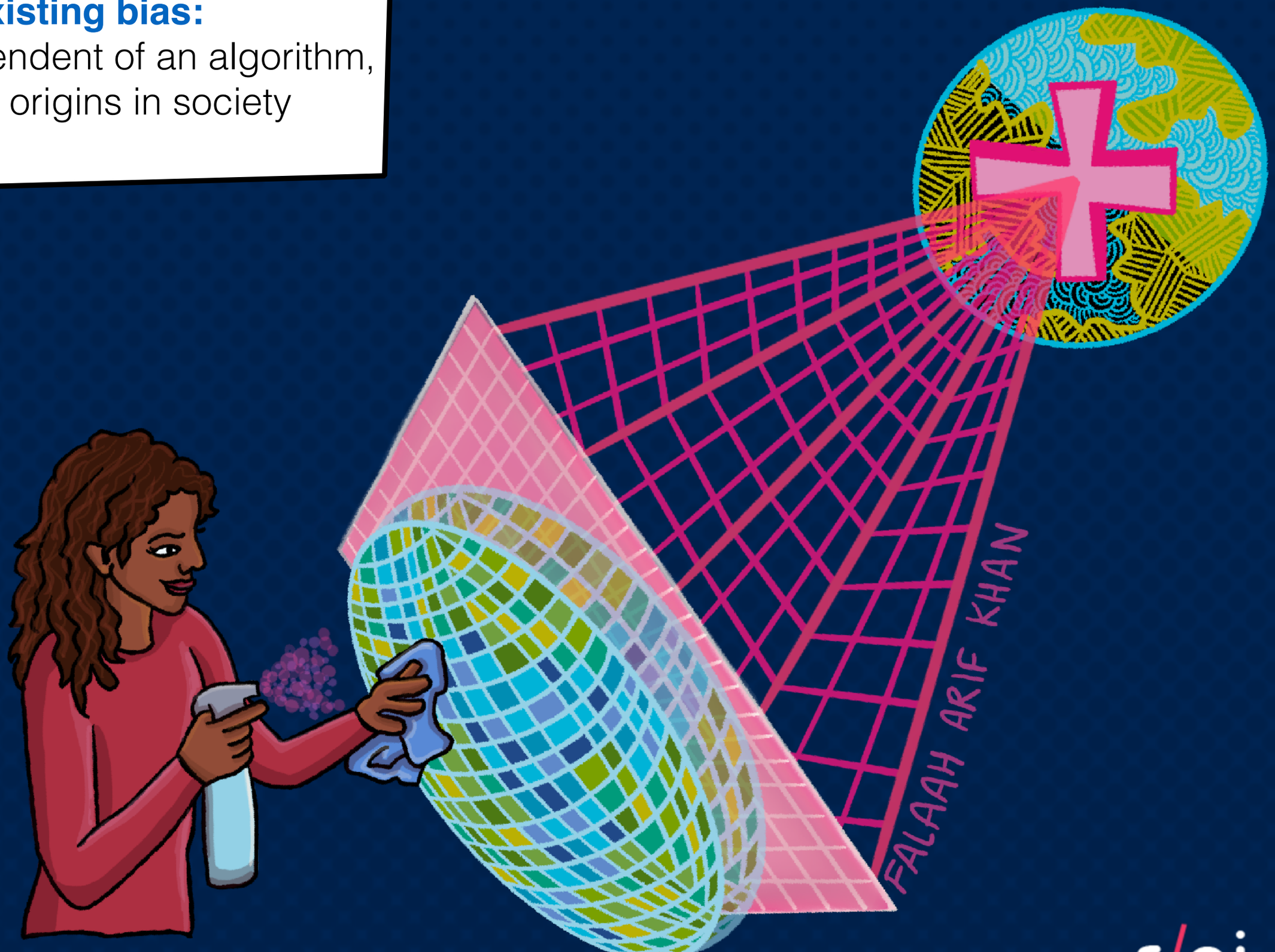


r/ai

Pre-existing bias:
independent of an algorithm, has its origins in society

r/ai

Pre-existing bias:
independent of an algorithm, has its origins in society



r/ai

bias can lead to discrimination

r/ai

The evils of discrimination

Disparate treatment

is the illegal practice of treating an entity, such as a job applicant or an employee, differently based on a **protected characteristic** such as race, gender, age, disability status, religion, sexual orientation, or national origin.

Disparate impact

is the result of systematic disparate treatment, where disproportionate **adverse impact** is observed on members of a **protected class**.

r/ai

Ricci v. DeStefano (2009)

Supreme Court Finds Bias Against White Firefighters

By **ADAM LIPTAK** JUNE 29, 2009

Karen Lee Torre, left, a lawyer who represented the New Haven firefighters in their lawsuit, with her clients Monday at the federal courthouse in New Haven. Christopher Capozziello for The New York Times

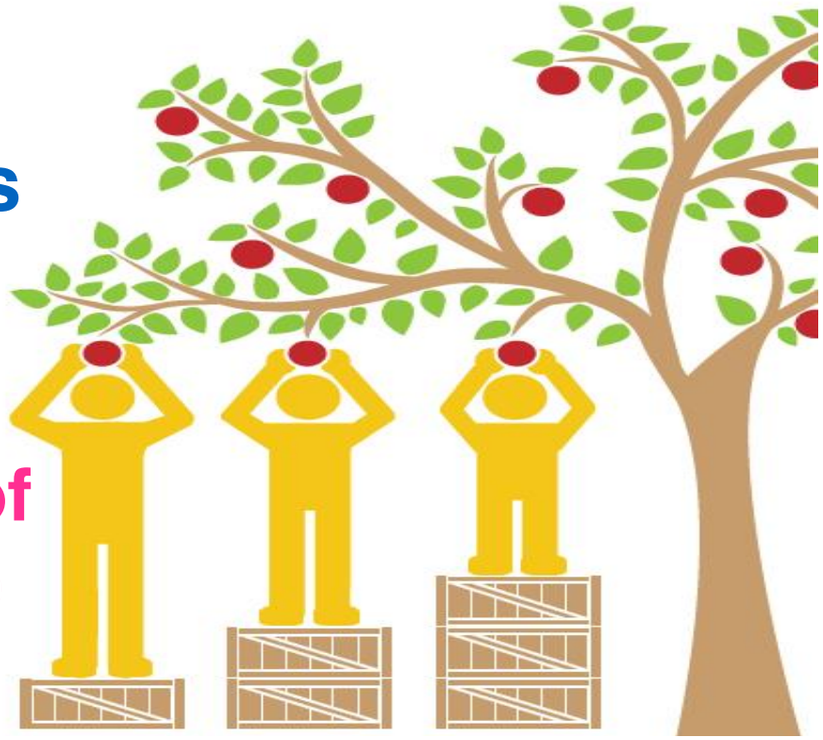
Case opinions	
Majority	Kennedy, joined by Roberts, Scalia, Thomas, Alito
Concurrence	Scalia
Concurrence	Alito, joined by Scalia, Thomas
Dissent	Ginsburg, joined by Stevens, Souter, Breyer
Laws applied	
Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e et seq.	

Fairness and worldviews

More on this in week 4

group fairness

equality of outcome



individual fairness

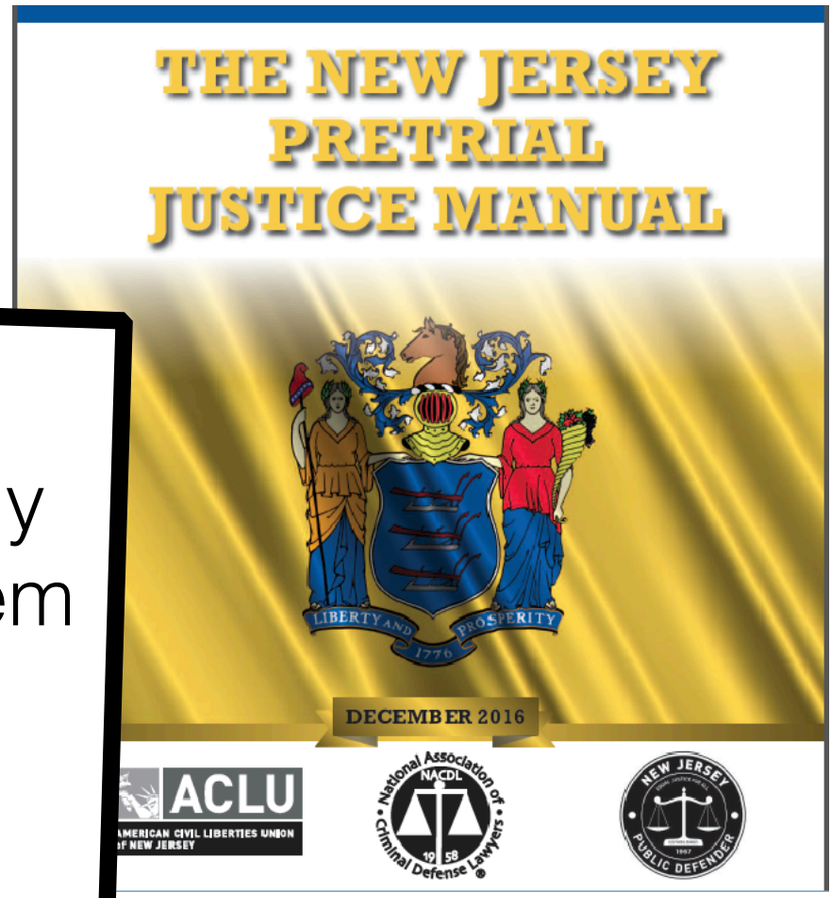
equality of treatment

r/ai

fairness in risk assessment

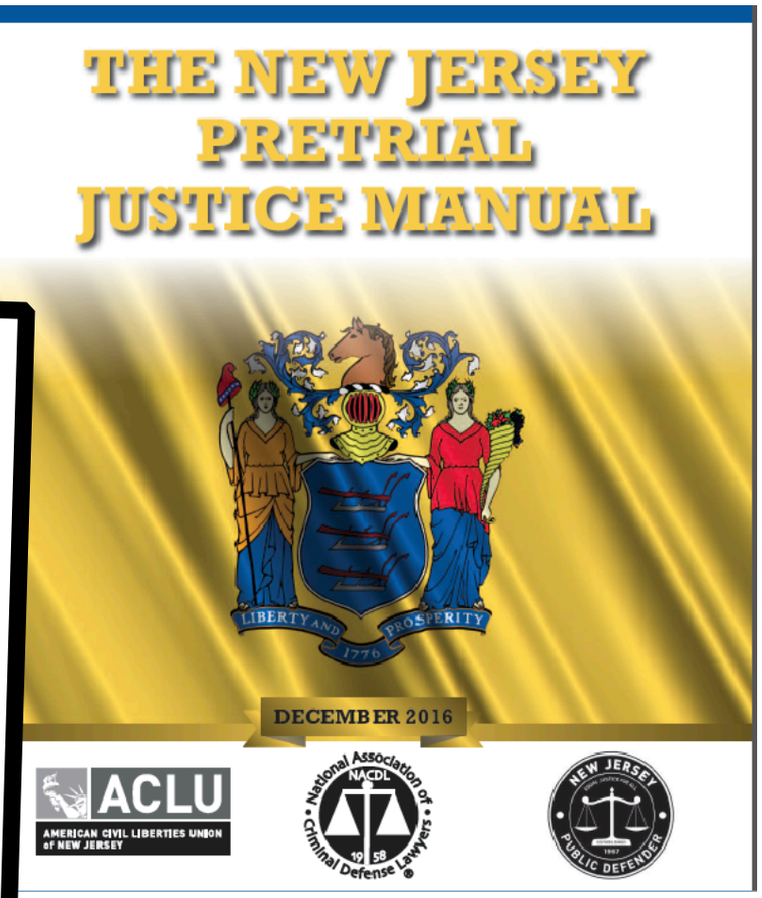
New Jersey bail reform

It is our hope that all defenders, both public and private, use this resource to aggressively and consistently **challenge** the pretrial system that **punishes the accused before conviction**, **forces guilty pleas to obtain release**, and **incarcerates the poor simply because they cannot afford to post a money bond**.



New Jersey bail reform

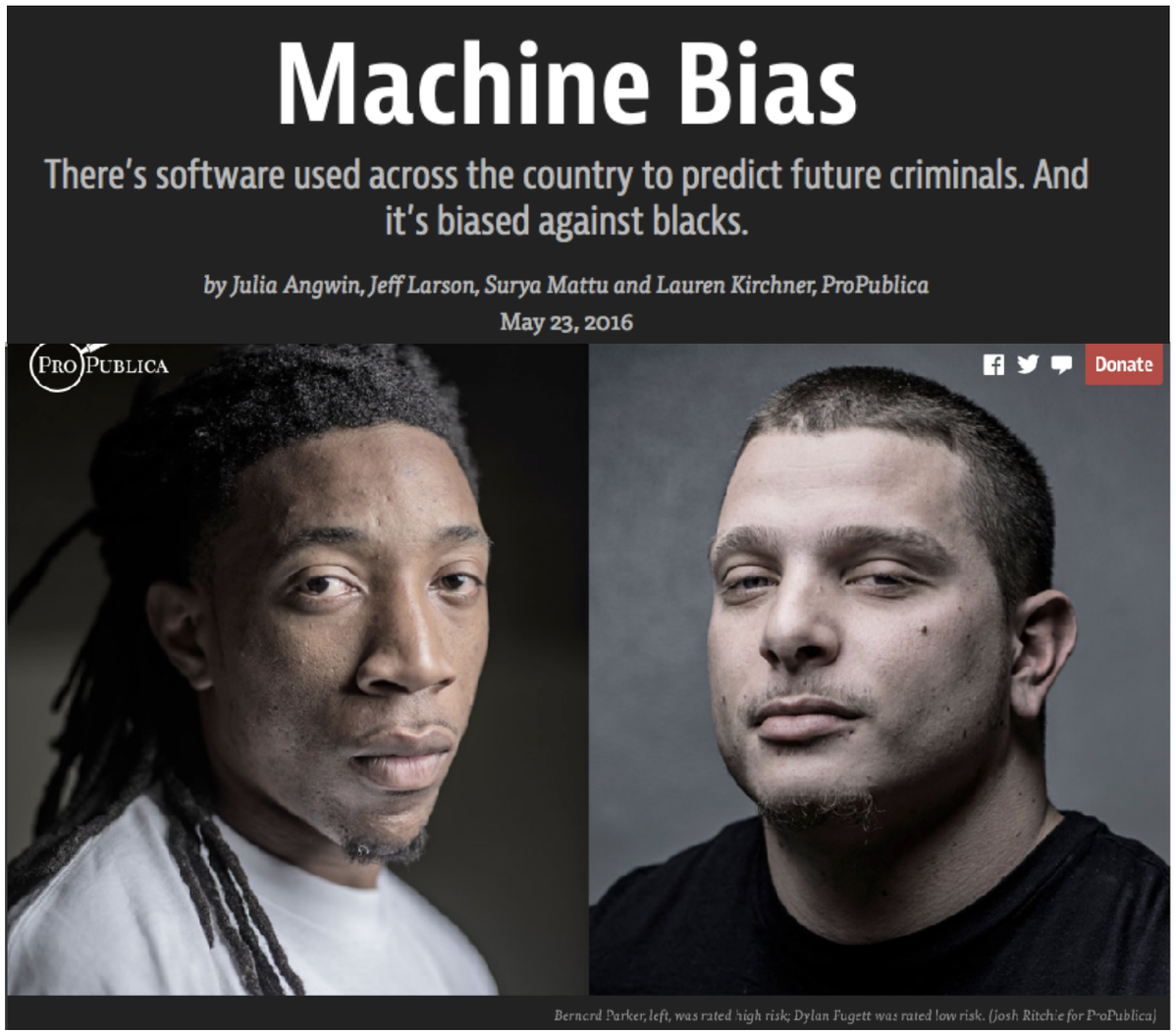
Switching from a system based solely on instinct and experience […] to one in which judges have access to **scientific, objective risk assessment** tools could further the criminal justice system's central goals of increasing public safety, reducing crime, and making the most effective, fair, and efficient use of public resources.



Fairness in risk assessment

- A risk assessment tool gives a probability estimate of a future outcome
- Used in many domains:
  - insurance, criminal sentencing, medical testing, hiring, banking
  - also in less-obvious set-ups, like online advertising
- **Fairness** in risk assessment is concerned with **how different kinds of error are distributed among sub-populations**

ProPublica's COMPAS study



A commercial tool COMPAS predicts some categories of future crime to assist in bail and sentencing decisions.

It uses about 100 factors. Notably, race is not used.

Black people are almost twice as likely as White people to be labeled a higher risk but not actually re-offend.

The tool makes **the opposite mistake among White people**: They are much more likely than black people to be labeled lower risk but go on to commit other crimes.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Distribution of FNR and FPR across groups

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

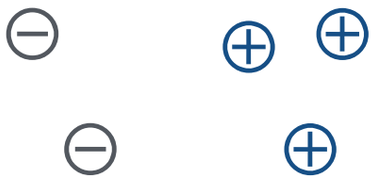
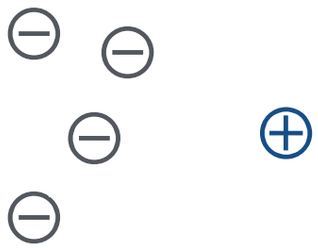
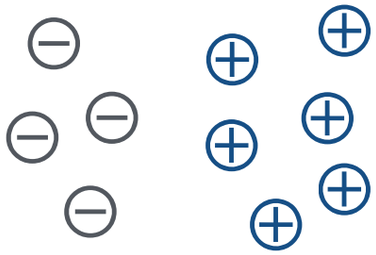
Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

This affects defendant's lives: Those labeled medium- or high-risk are much more likely to be detained while awaiting trial

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Calibration

"+" denotes recidivism

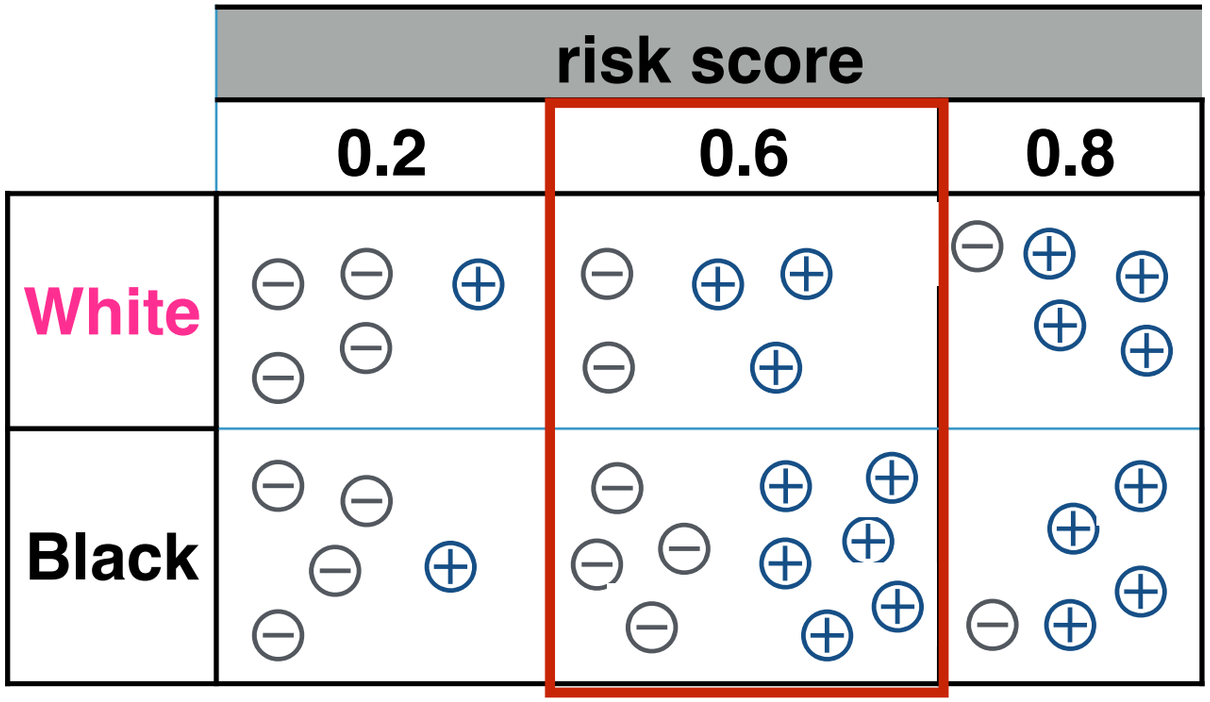
	risk score		
	0.2	**0.6**	**0.8**
White			
Black			

Given the output of a risk tool, likelihood of belonging to the positive class is independent of group membership

- 0.6 risk score means 0.6 for any defendant, no matter which group they belong to

Calibration

"+" denotes recidivism



A score, S, is well calibrated if it reflects the same likelihood of recidivism irrespective of group membership

$$P(Y = 1 \mid S = s, R = b) = P(Y = 1 \mid S = s, R = w)$$

$$P(Y = 1 \mid S = 0.6, R = b) = P(Y = 1 \mid S = 0.6, R = w)$$

$$3/5 = 6/10$$

Calibration

"+" denotes recidivism

	risk score: 0.2	risk score: 0.6	risk score: 0.8
White	⊖ ⊖ ⊕ ⊖ ⊖	⊖ ⊕ ⊕ ⊖ ⊕	⊕ ⊕ ⊕ ⊕ ⊕
Black	⊖ ⊖ ⊖ ⊕ ⊖	⊖ ⊖ ⊕ ⊕ ⊕ ⊖ ⊕ ⊕	⊖ ⊖ ⊕ ⊕ ⊕

Higher FNR

Higher FPR

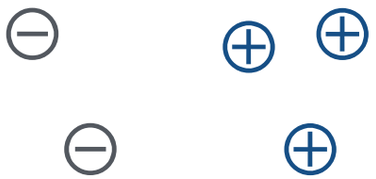
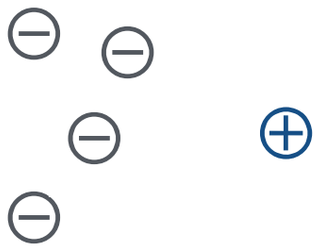
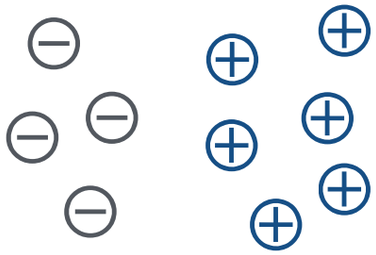
Without calibration!

Note: 80% of those assigned 0.8 risk score recidivate (8/10)

But 100% among White and 60% among Black defendants.

Calibration

"+" denotes recidivism

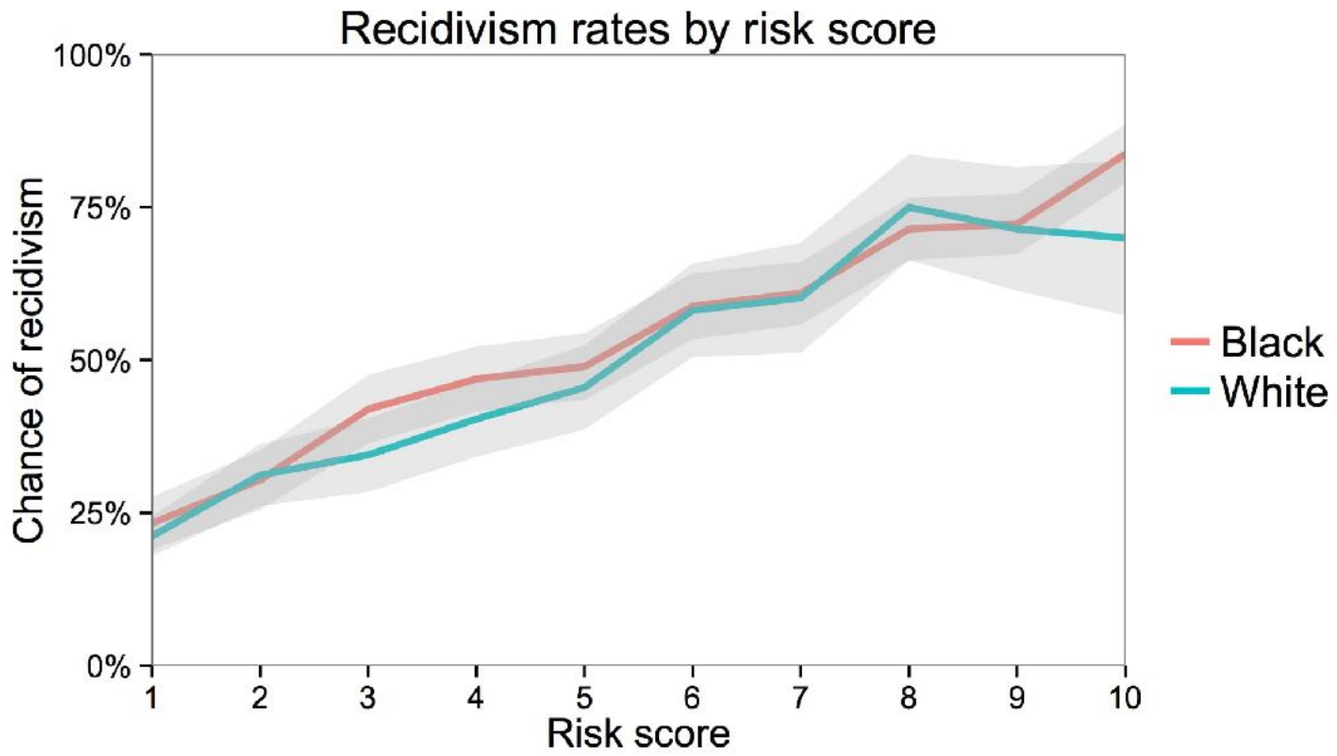
	risk score		
	0.2	**0.6**	**0.8**
White			
Black			

Given the output of a risk tool, likelihood of belonging to the positive class is independent of group membership

why do we want calibration?

COMPAS as a predictive instrument

COMPAS is reasonably well-calibrated:



[plot from Corbett-Davies et al.; *WaPo 2016*]

Distribution of FNR and FPR across groups

FPR

Prediction Fails Differently for Black Defendants

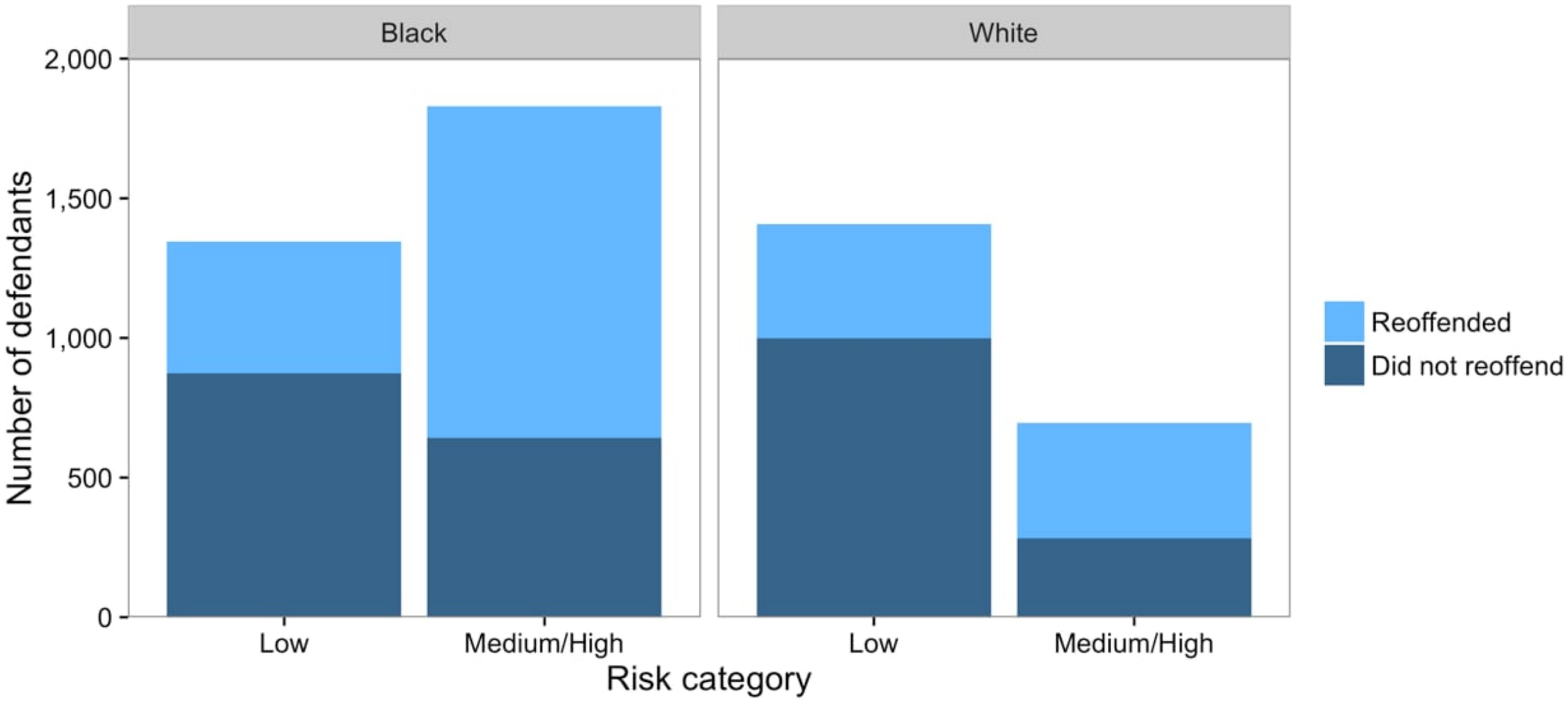
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

This affects defendant's lives: Those labeled medium- or high-risk are much more likely to be detained while awaiting trial

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

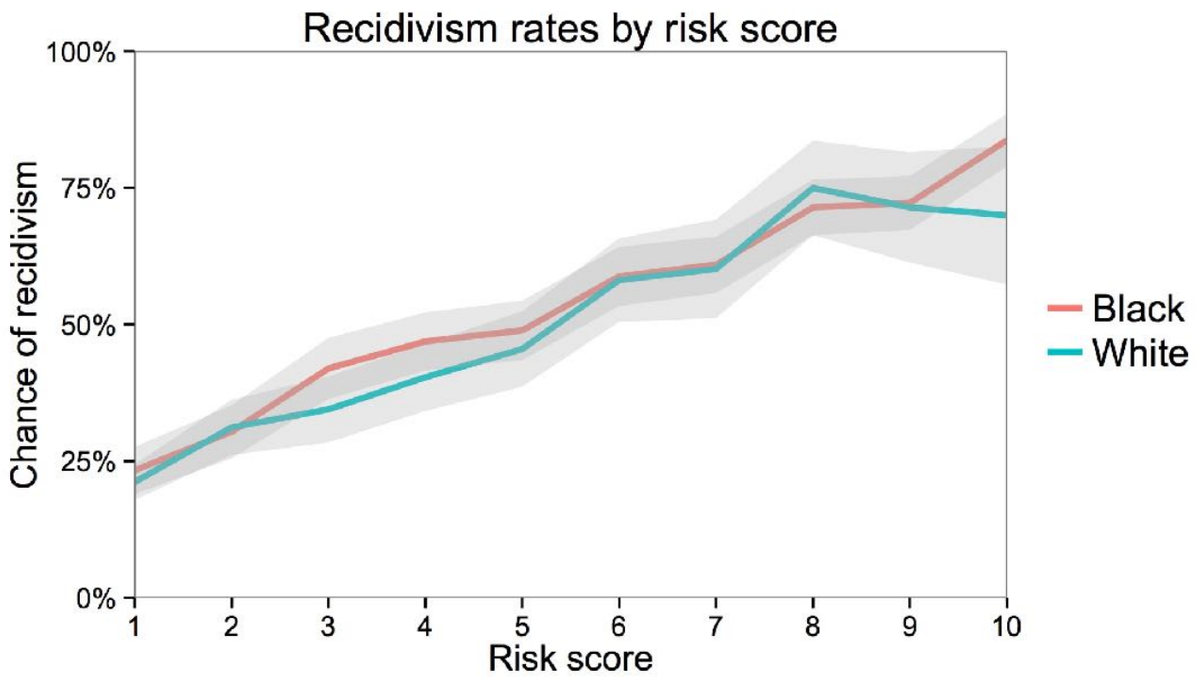
An impossibility result



[plot from Corbett-Davies et al.; *WaPo 2016*]

An impossibility result

COMPAS is reasonably well-calibrated:



But fails differently across groups:

	Risk label	*Did not reoffend*
White	Medium or High	23.5%
Black	Medium or High	44.9%

[plot from Corbett-Davies et al.; *WaPo 2016*]

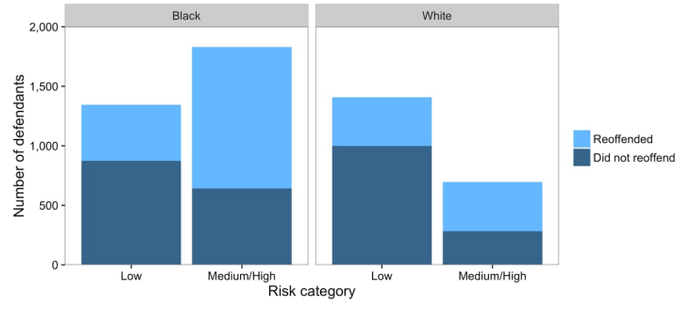
An impossibility result

1. Within each risk category, the proportion of defendants who reoffend is approximately the same regardless of race

2. The overall recidivism rate is higher for Black defendants than White defendants (52% versus 39%)

3. Black defendants are more likely to be classified as medium or high risk (58% versus 33%)

4. Black defendants who don't reoffend are predicted to be riskier than White defendants who don't reoffend



Observations 1 and 2 mathematically guarantee the disparities in 3 and 4

An impossibility result

If a predictive instrument **satisfies predictive parity**, but the **prevalence** of the phenomenon **differs between groups**, then the instrument **cannot achieve** equal false positive rates and equal false negative rates across these groups.

Recidivism rates in the ProPublica dataset are higher for the Black group than for the White group

[A. Chouldechova; arXiv:1610.07524v1 *(2017)*]

A more general statement: Balance

- Balance for the positive class: Positive instances are those who go on to re-offend. The average score of positive instances should be the same across groups.
- Balance for the negative class: Negative instances are those who do not go on to re-offend. The average score of negative instances should be the same across groups.
- Generalization: Both groups should have equal false positive rates and equal false negative rates.
- Different from statistical parity!

the chance of making a mistake does not depend on race

[J. Kleinberg, S. Mullainathan, M. Raghavan; *ITCS 2017*]

Desiderata, re-stated

- For each group, a $\mathbf{v_b}$ fraction in each bin **b** is positive
- Average score of positive class same across groups
- Average score of negative class same across groups

Can we have all these properties?

[J. Kleinberg, S. Mullainathan, M. Raghavan; *ITCS 2017*]

Achievable only in trivial cases

- Perfect information: the tool knows who reoffends (score 1) and who does not (score 0)
- Equal base rates: the fraction of positive-class people is the same for both groups

a negative result, need tradeoffs

proof sketched out in (starts 12 min in)

https://www.youtube.com/watch?v=UUC8tMNxwV8

[J. Kleinberg, S. Mullainathan, M. Raghavan; *ITCS 2017*]

Fairness for whom?

Decision-maker: among those labeled low-risk, how many will recidivate?

Defendant: how likely is it that I will be incorrectly labeled high-risk?

	labeled low-risk	labeled high-risk
did not recidivate	TN	**FP**
recidivated	**FN**	TP

based on a slide by Arvind Narayanan

What's the right answer?

There is no single answer!

Need transparency and public debate

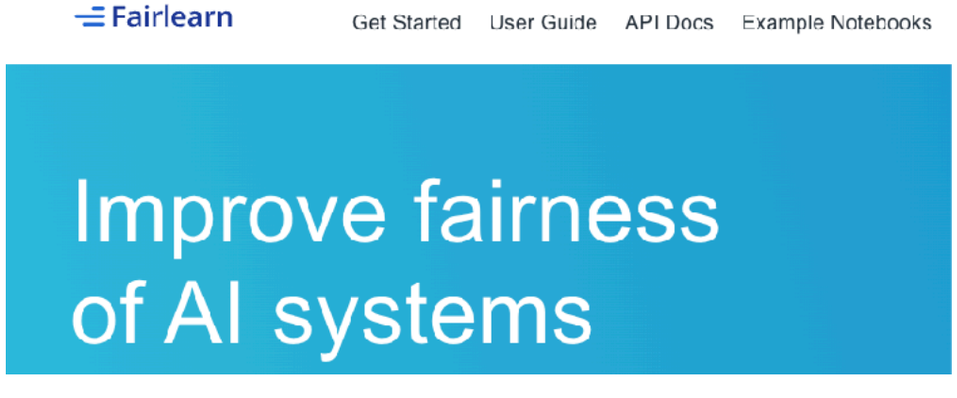
- Consider harms and benefits to different stakeholders
- Being transparent about which fairness criteria we use, how we trade them off
  - Individual vs group fairness
  - Calibration vs FNR/FPR balance

preview:
lab this week

r/ai

Lab 2: Detecting & Mitigating Bias

- Fairlearn toolkit
- Lab 2: Detecting & Mitigating bias in the **pre-processing** stage
- Define protected (sensitive) attribute of interest



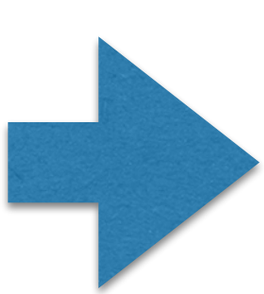
Lab 2: Detecting & Mitigating Bias

- ACS data
- Define protected attribute (gender, race)
- Predicting income (high salary vs low salary) across gender and race

Lab 2: Detecting & Mitigating Bias

One-hot encoding = creating dummy/dichotomous variables

Obs	Gender
1	Man
2	Woman



Obs	Men	Women
1	1	0
2	0	1

Responsible Data Science

Algorithmic Fairness

Thank you!

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai