Responsible Data Science

Algorithmic Fairness

~~January 30 & February 6, 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

RDS course overview

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

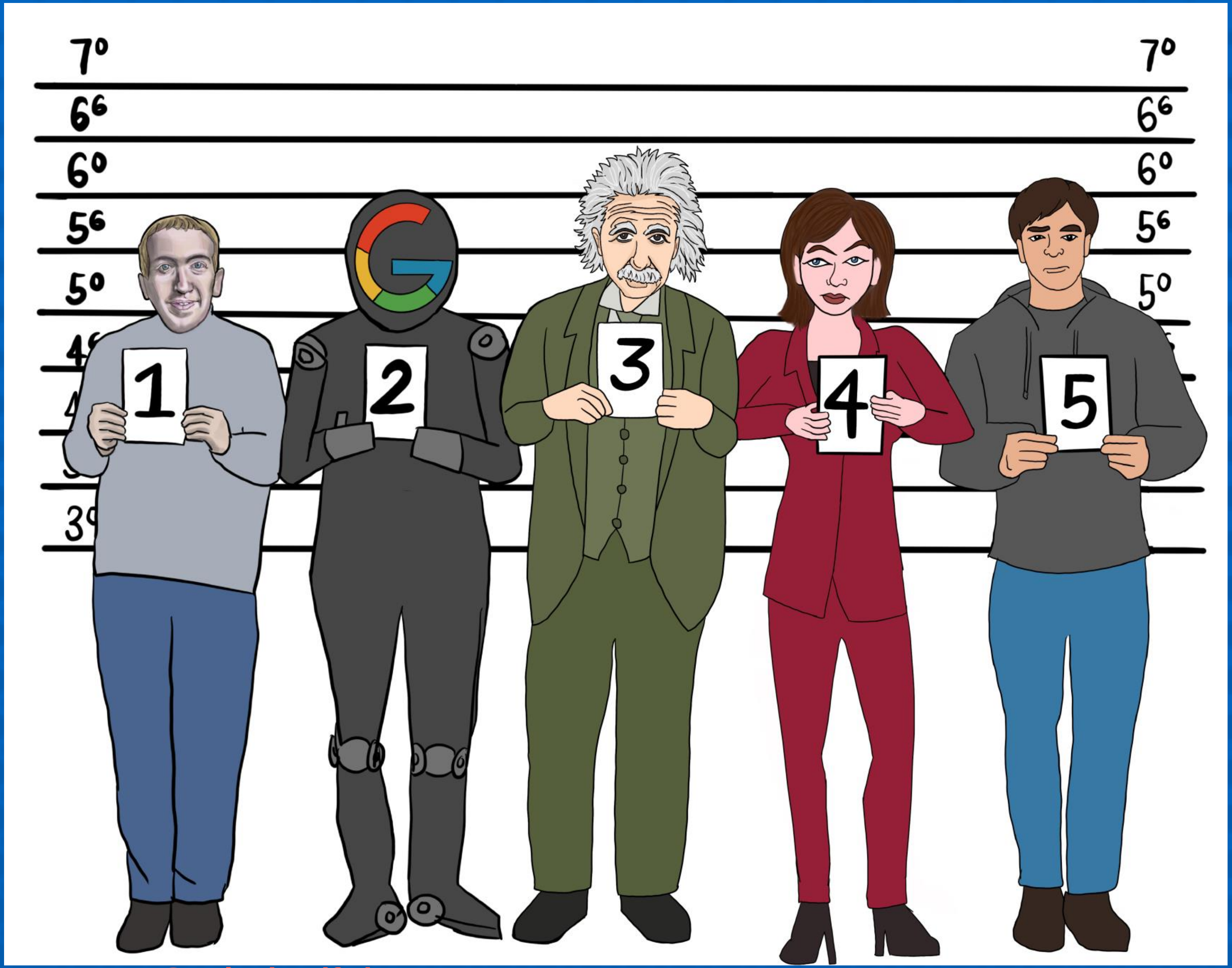
We will learn **algorithmic techniques** for data analysis.
We will also learn about recent **laws** / **regulatory frameworks**.

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.

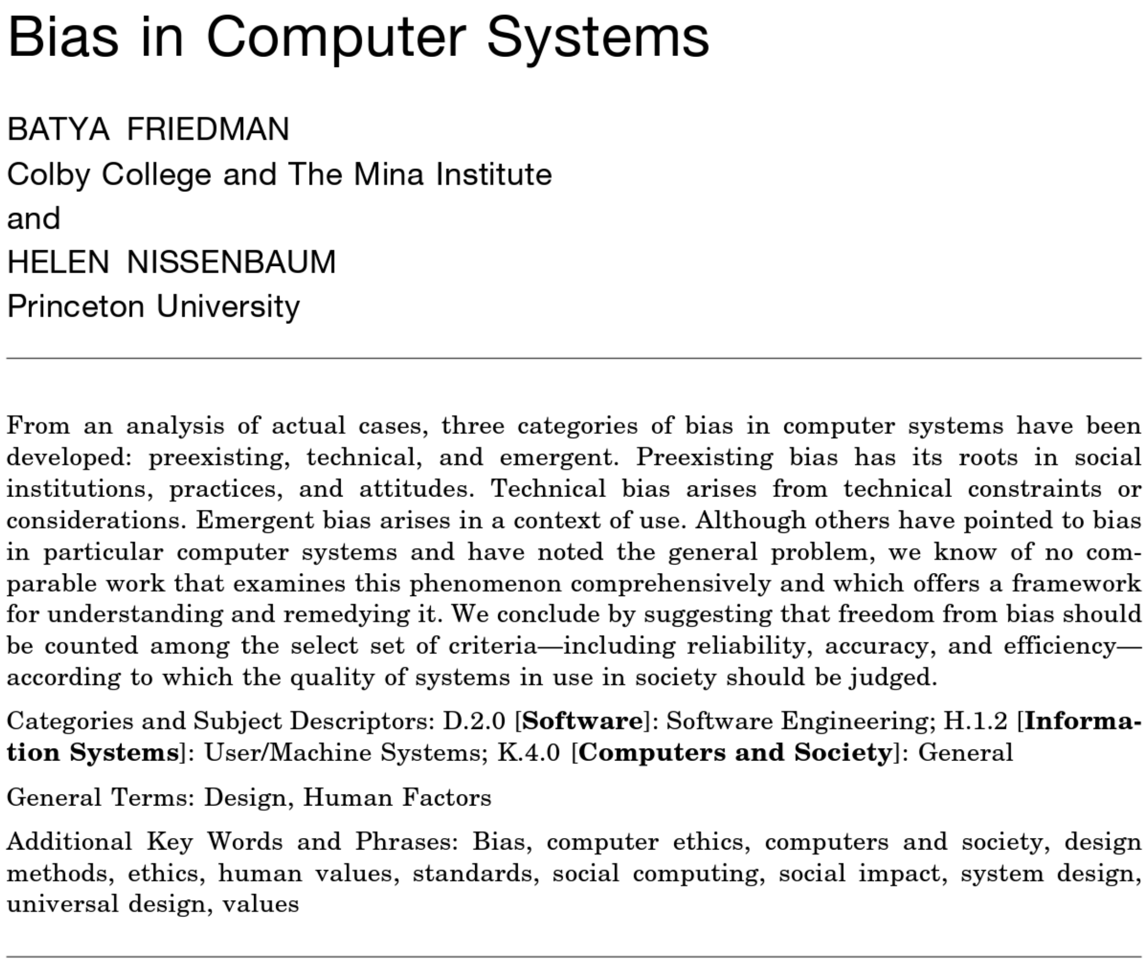
Nuance, please!

We all are responsible



@FalaahArifKhan

Reading: Algorithmic bias



Bias in Computer Systems

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

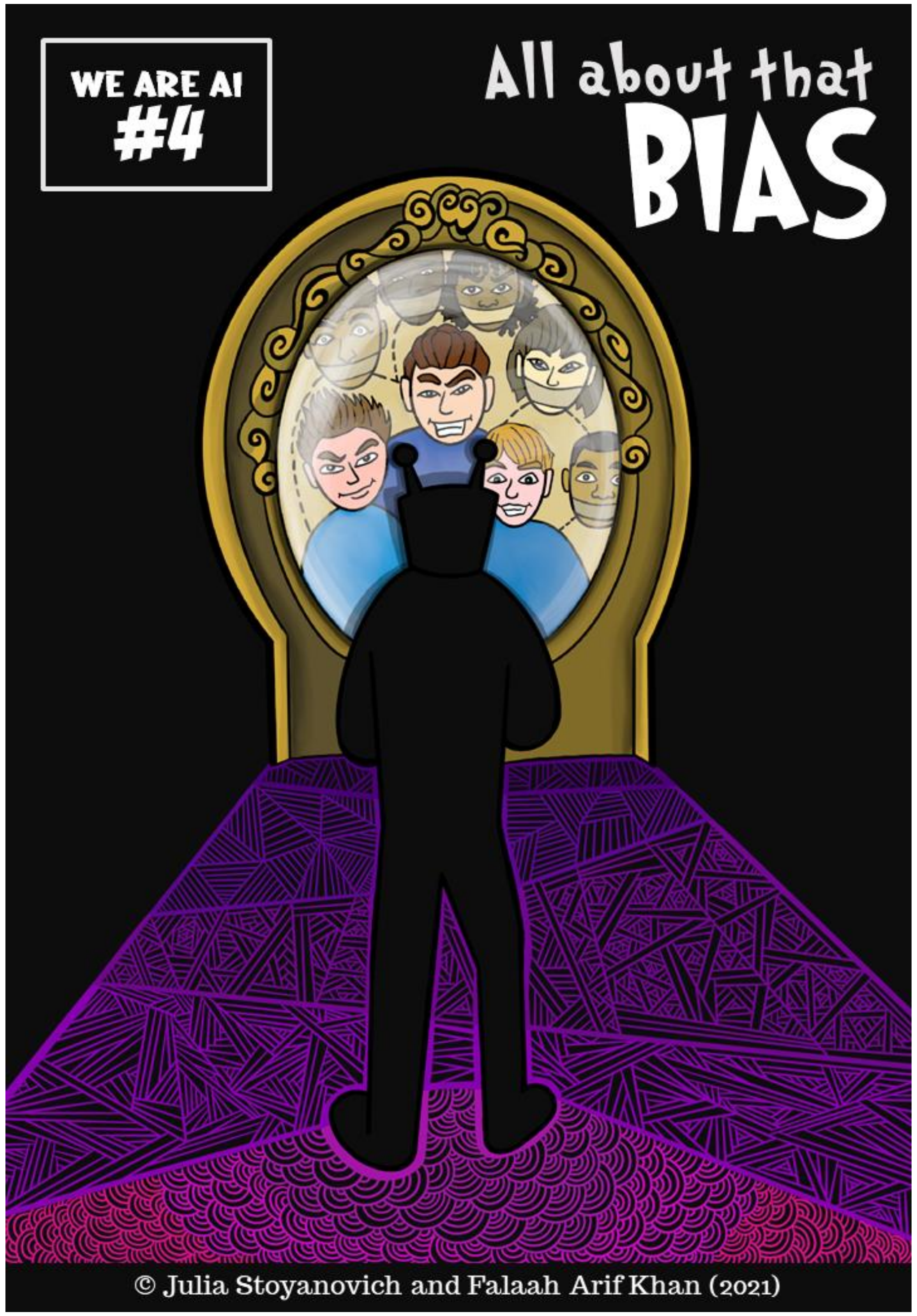
From an analysis of actual cases, three categories of bias in computer systems have been developed: preexisting, technical, and emergent. Preexisting bias has its roots in social institutions, practices, and attitudes. Technical bias arises from technical constraints or considerations. Emergent bias 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)]



© Julia Stoyanovich and Falaah Arif Khan (2021)

Reading: Algorithmic fairness

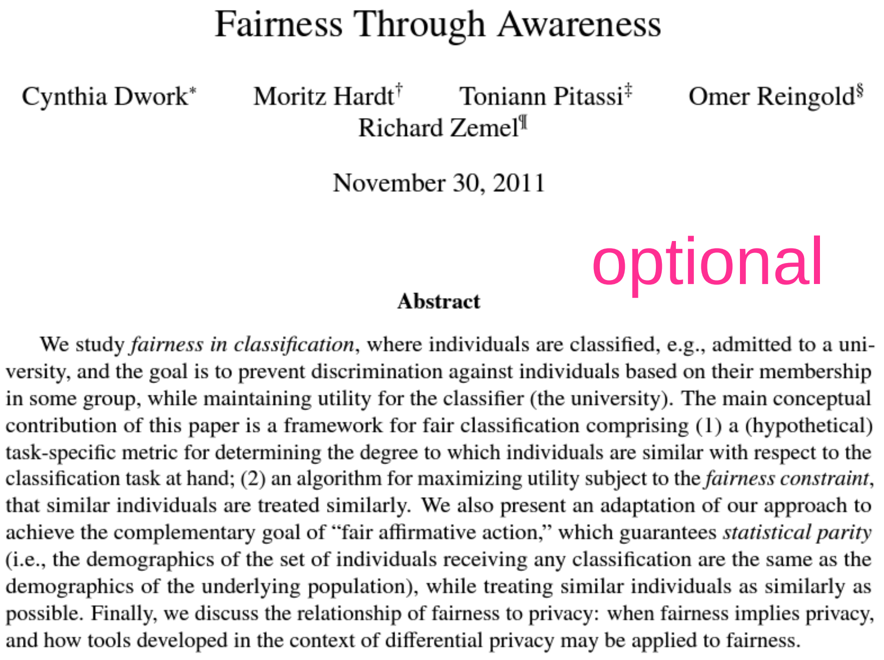
DOI:10.1145/3376898

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)]



Fairness Through Awareness

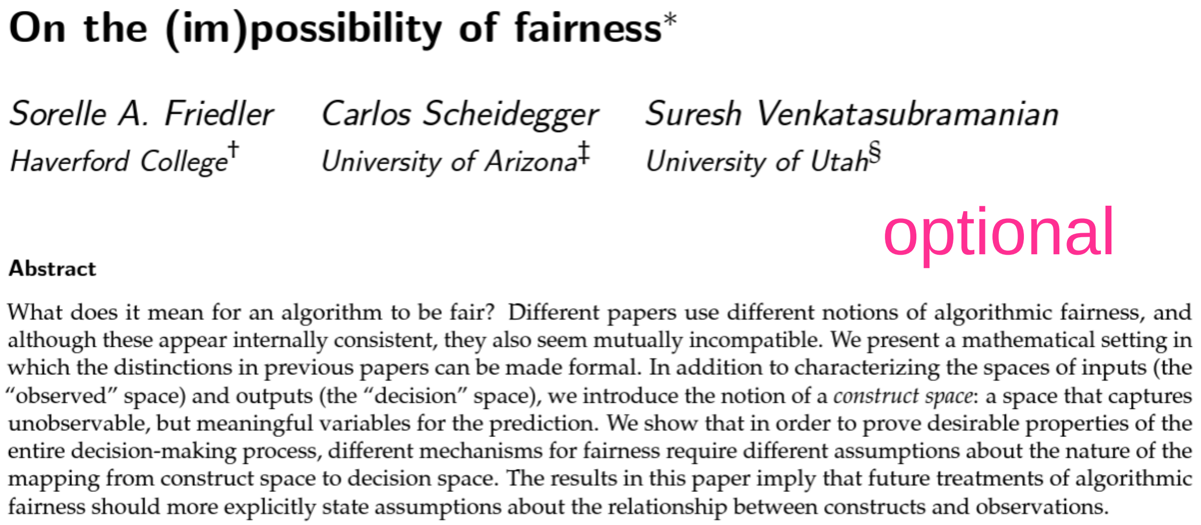
Cynthia Dwork* Moritz Hardt† Toniann Pitassi‡ Omer Reingold§ Richard Zemel¶

November 30, 2011

optional

Abstract

We study *fairness in classification*, where individuals are classified, e.g., admitted to a university, and the goal is to prevent discrimination against individuals based on their membership in some group, while maintaining utility for the classifier (the university). The main conceptual contribution of this paper is a framework for fair classification comprising (1) a (hypothetical) task-specific metric for determining the degree to which individuals are similar with respect to the classification task at hand; (2) an algorithm for maximizing utility subject to the *fairness constraint*, that similar individuals are treated similarly. We also present an adaptation of our approach to achieve the complementary goal of "fair affirmative action," which guarantees *statistical parity* (i.e., the demographics of the set of individuals receiving any classification are the same as the demographics of the underlying population), while treating similar individuals as similarly as possible. Finally, we discuss the relationship of fairness to privacy: when fairness implies privacy, and how tools developed in the context of differential privacy may be applied to fairness.



On the (im)possibility of fairness*

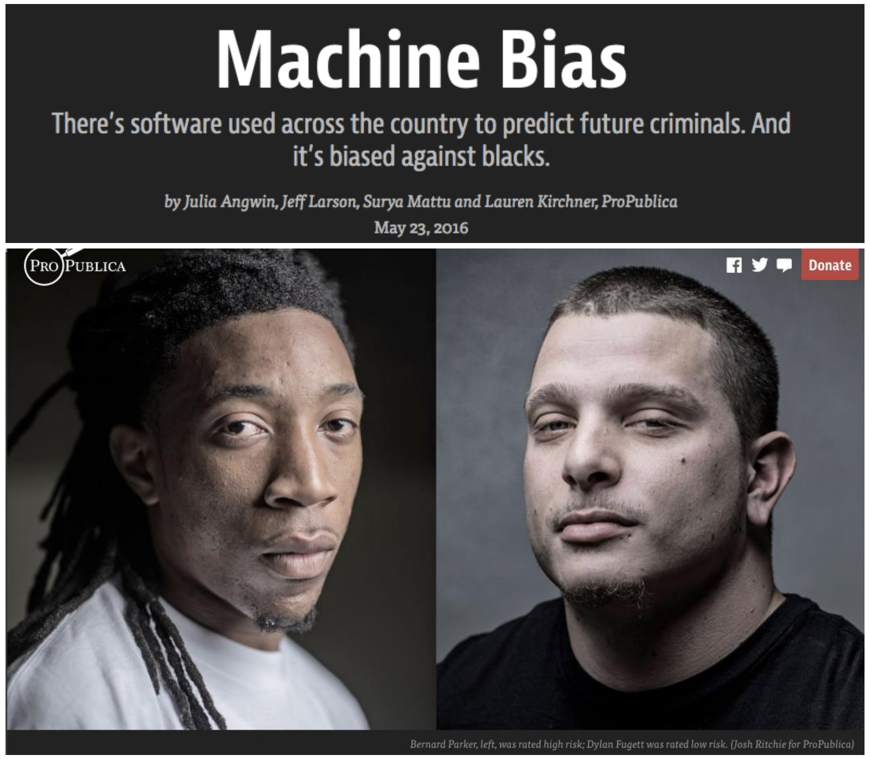
Sorelle A. Friedler — *Haverford College*†
Carlos Scheidegger — *University of Arizona*‡
Suresh Venkatasubramanian — *University of Utah*§

optional

Abstract

What does it mean for an algorithm to be fair? Different papers use different notions of algorithmic fairness, and although these appear internally consistent, they also seem mutually incompatible. We present a mathematical setting in which the distinctions in previous papers can be made formal. In addition to characterizing the spaces of inputs (the "observed" space) and outputs (the "decision" space), we introduce the notion of a *construct space*: a space that captures unobservable, but meaningful variables for the prediction. We show that in order to prove desirable properties of the entire decision-making process, different mechanisms for fairness require different assumptions about the nature of the mapping from construct space to decision space. The results in this paper imply that future treatments of algorithmic fairness should more explicitly state assumptions about the relationship between constructs and observations.

Reading: Fairness in risk assessment



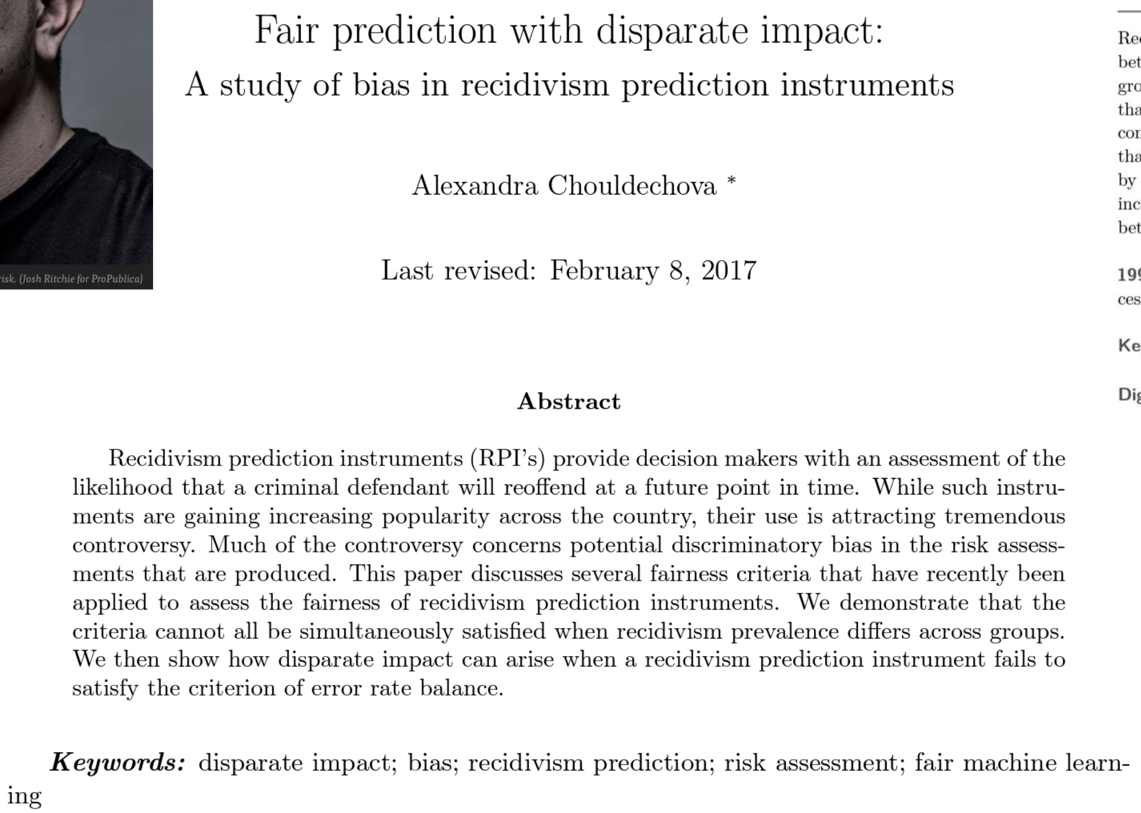
Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)



Fair prediction with disparate impact:
A study of bias in recidivism prediction instruments

Alexandra Chouldechova *

Last revised: February 8, 2017

Abstract

Recidivism prediction instruments (RPI's) provide decision makers with an assessment of the likelihood that a criminal defendant will reoffend at a future point in time. While such instruments are gaining increasing popularity across the country, their use is attracting tremendous controversy. Much of the controversy concerns potential discriminatory bias in the risk assessments that are produced. This paper discusses several fairness criteria that have recently been applied to assess the fairness of recidivism prediction instruments. We demonstrate that the criteria cannot all be simultaneously satisfied when recidivism prevalence differs across groups. We then show how disparate impact can arise when a recidivism prediction instrument fails to satisfy the criterion of error rate balance.

Keywords: disparate impact; bias; recidivism prediction; risk assessment; fair machine learning

[Chouldechova, BigData (2017)]

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.

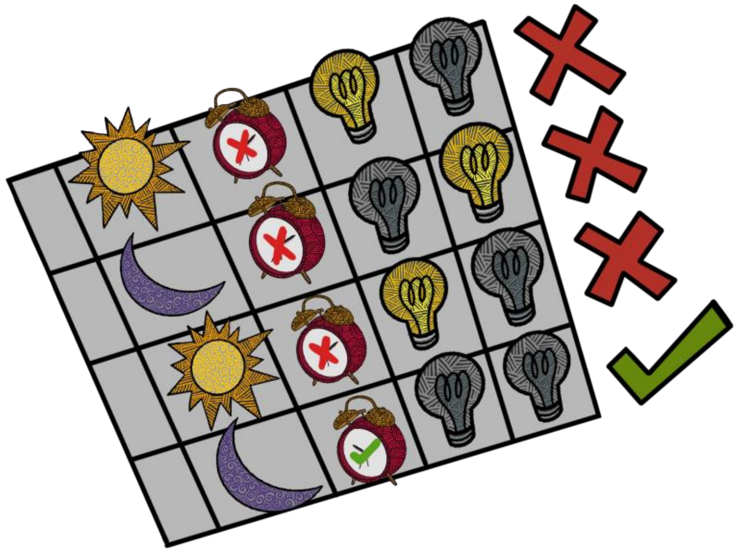
1998 ACM Subject Classification H.2.8 Database Applications, J.1 Administrative Data Processing

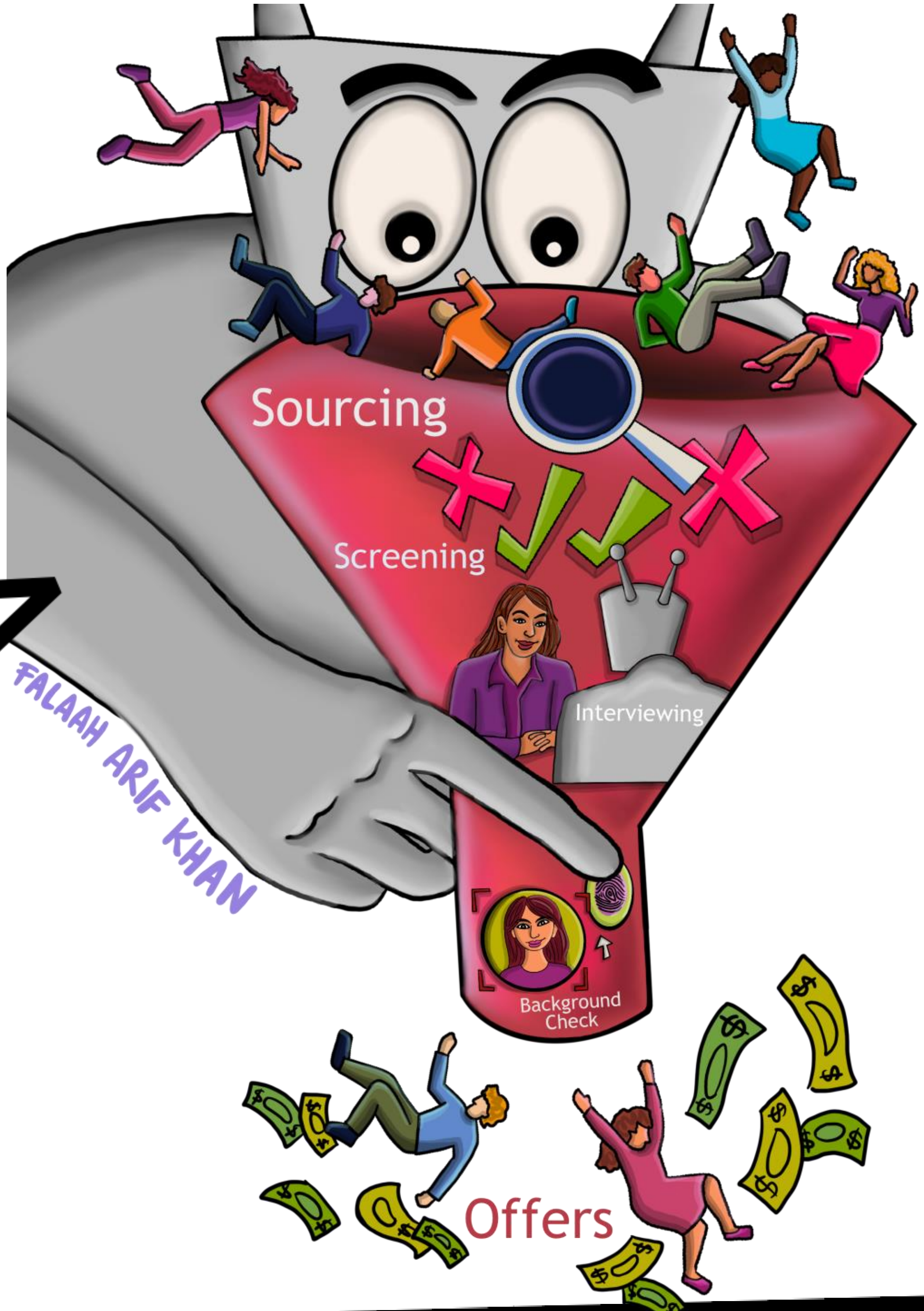
Keywords and phrases algorithmic fairness, risk tools, calibration

Digital Object Identifier 10.4230/LIPIcs.ITCS.2017.43

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

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**?

technical teaser *c. 2015*: fairness in classification

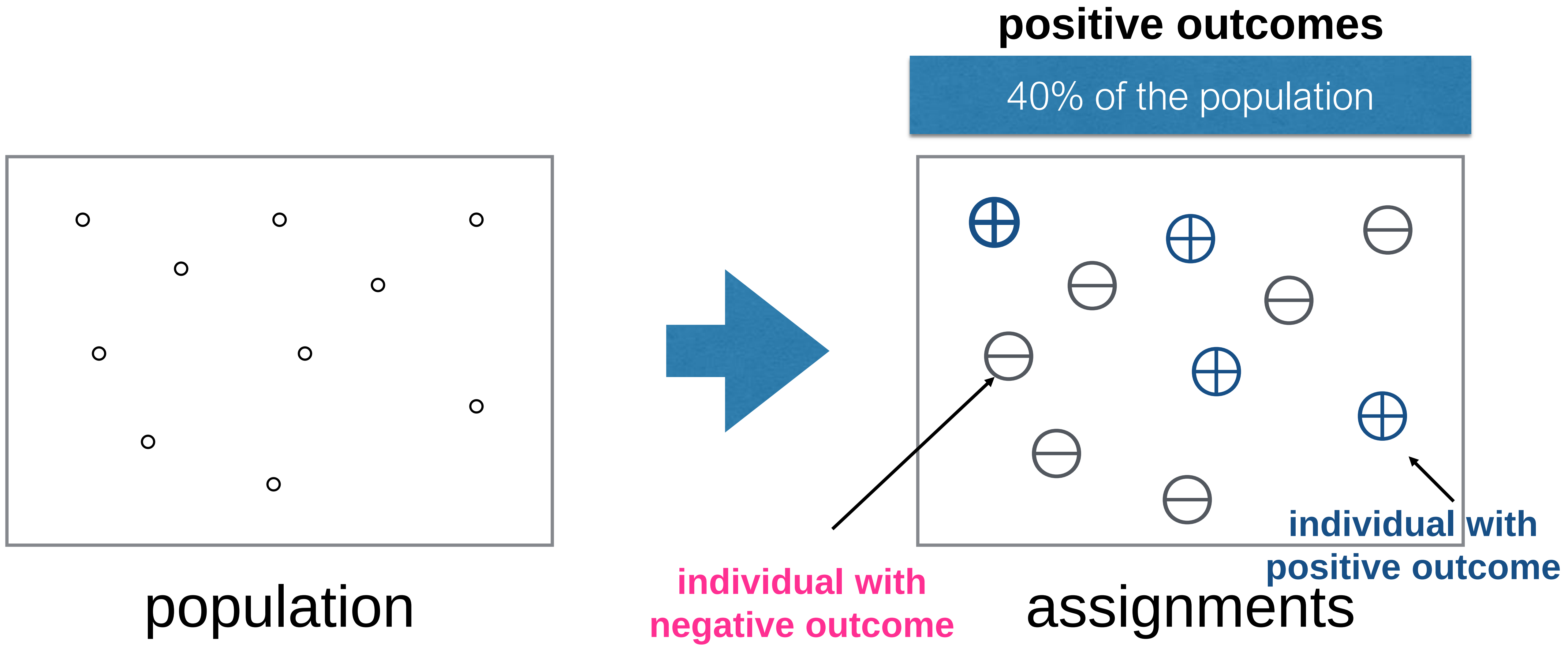
Vendors and outcomes

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~~

Fairness in classification

ss in classification is concerned with how outcomes are assigned to a po



positive outcomes

40% of the population

population

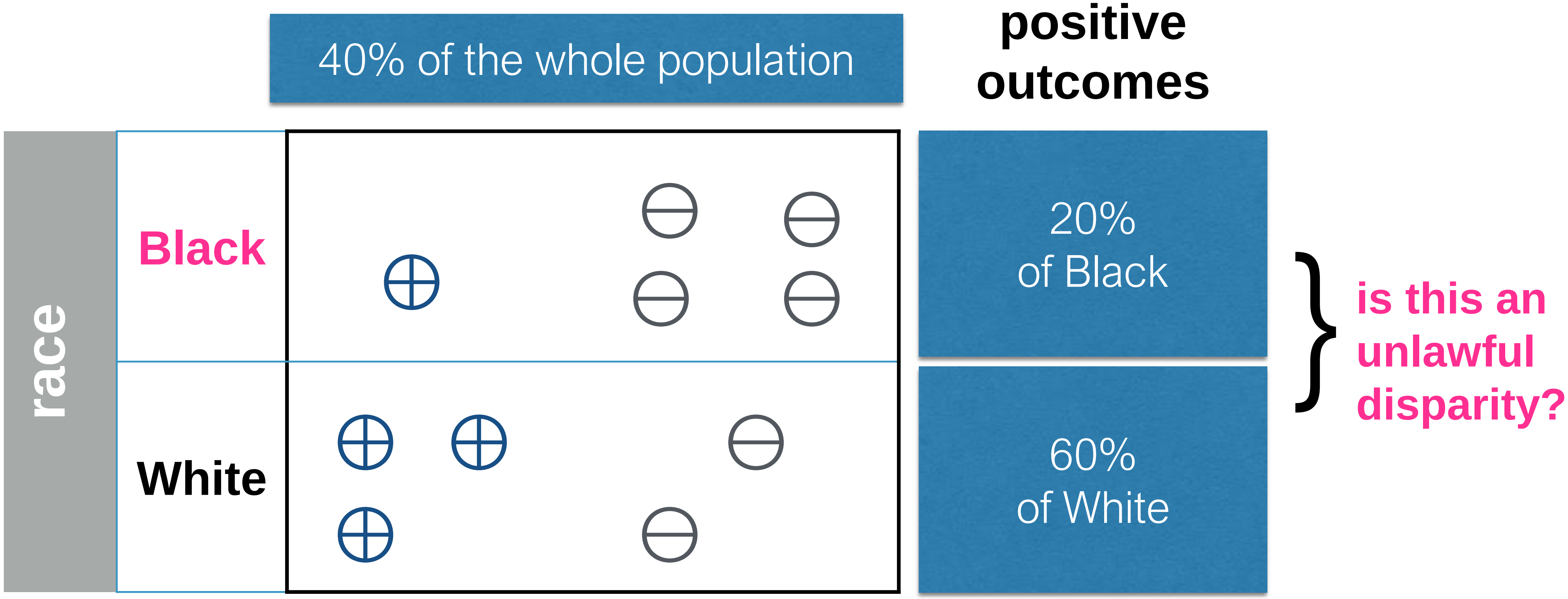
assignments

individual with negative outcome

individual with positive outcome

Fairness in classification

Sub-populations may be treated differently



40% of the whole population

positive outcomes

race

Black ⊕ ⊖ ⊖ ⊖ ⊖

White ⊕ ⊕ ⊕ ⊖ ⊖

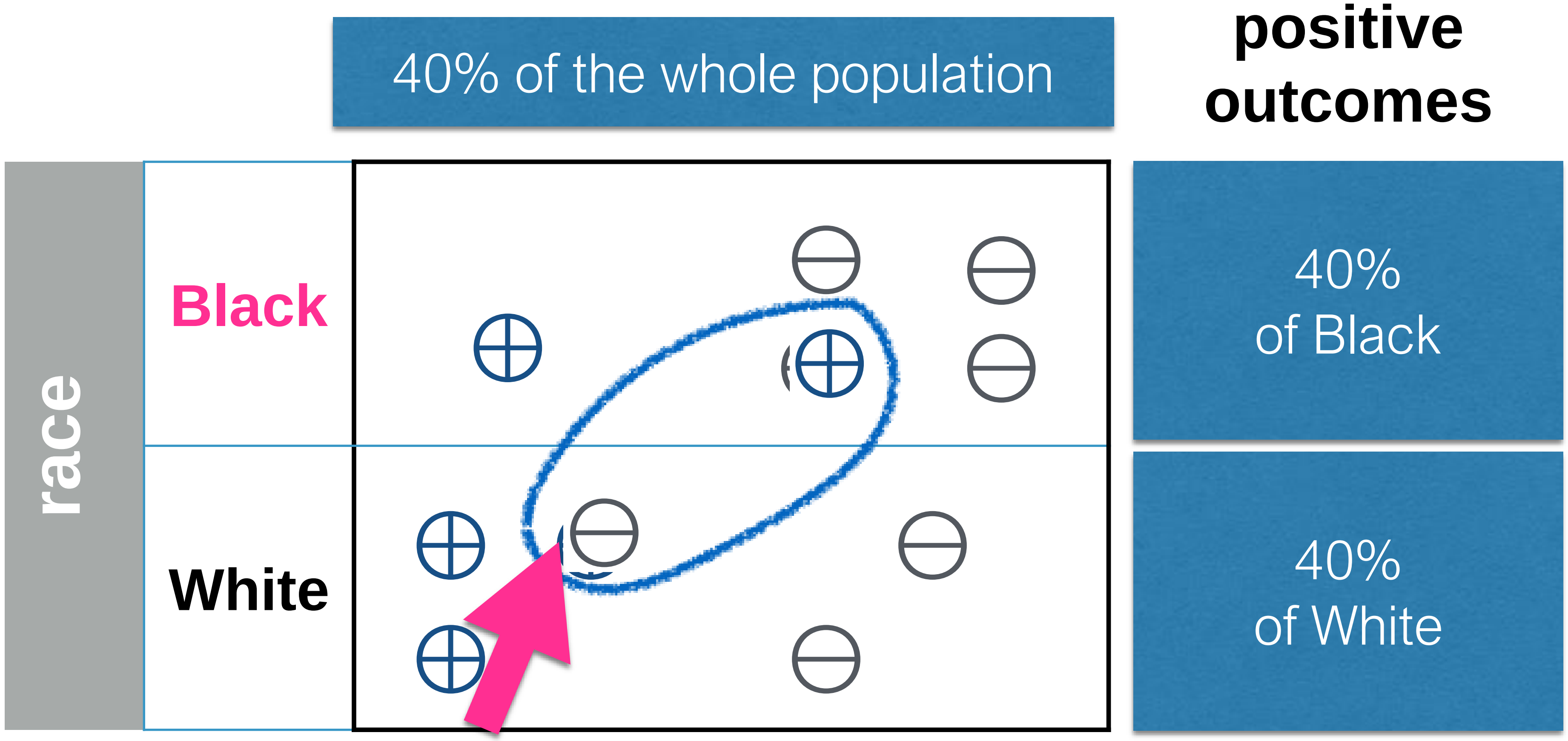
20% of Black

60% of White

is this an unlawful disparity?

Fairness in classification

Sub-populations may be treated differently



40% of the whole population

positive outcomes

race

Black

White

40% of Black

40% of White

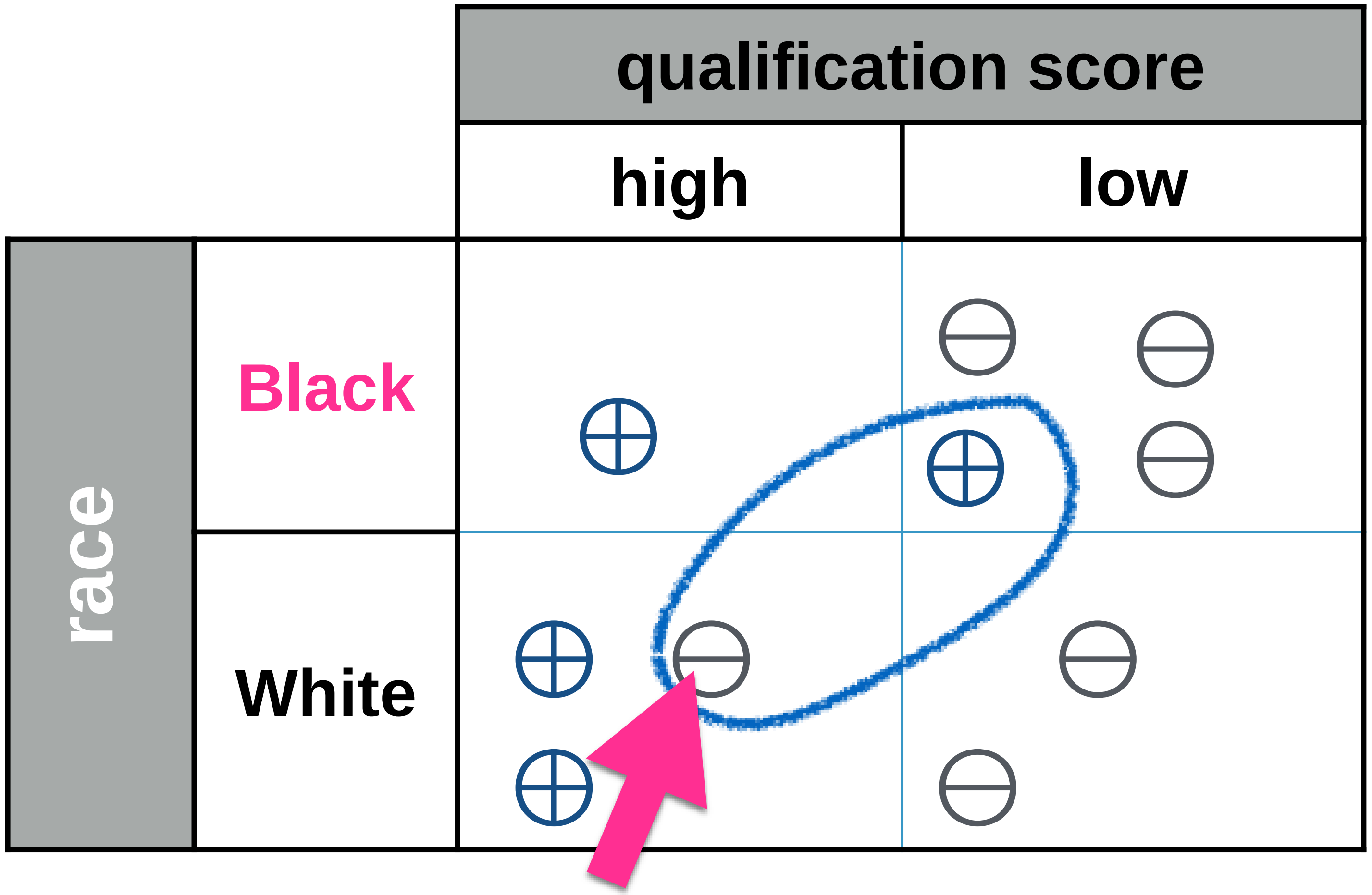
Fairness in classification

Explaining the disparity with proxy variables

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

quick discussion

Swapping outcomes



race	qualification score: high	qualification score: low
Black	⊕	⊖ ⊖ ⊕ ⊖
White	⊕ ⊖ ⊕	⊖ ⊖

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

Bias in computer systems

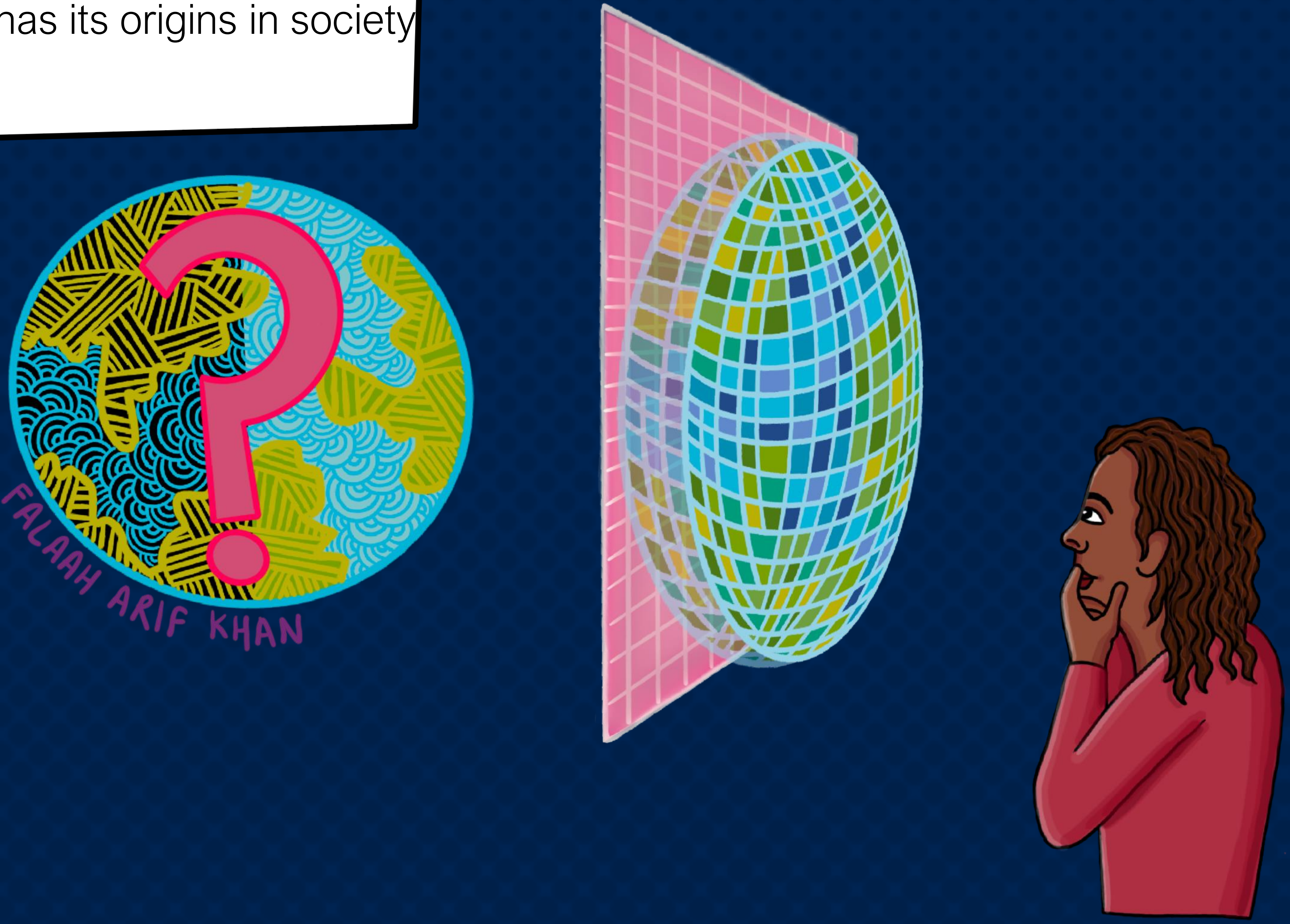
Pre-existing is independent of an algorithm and has origins in society

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

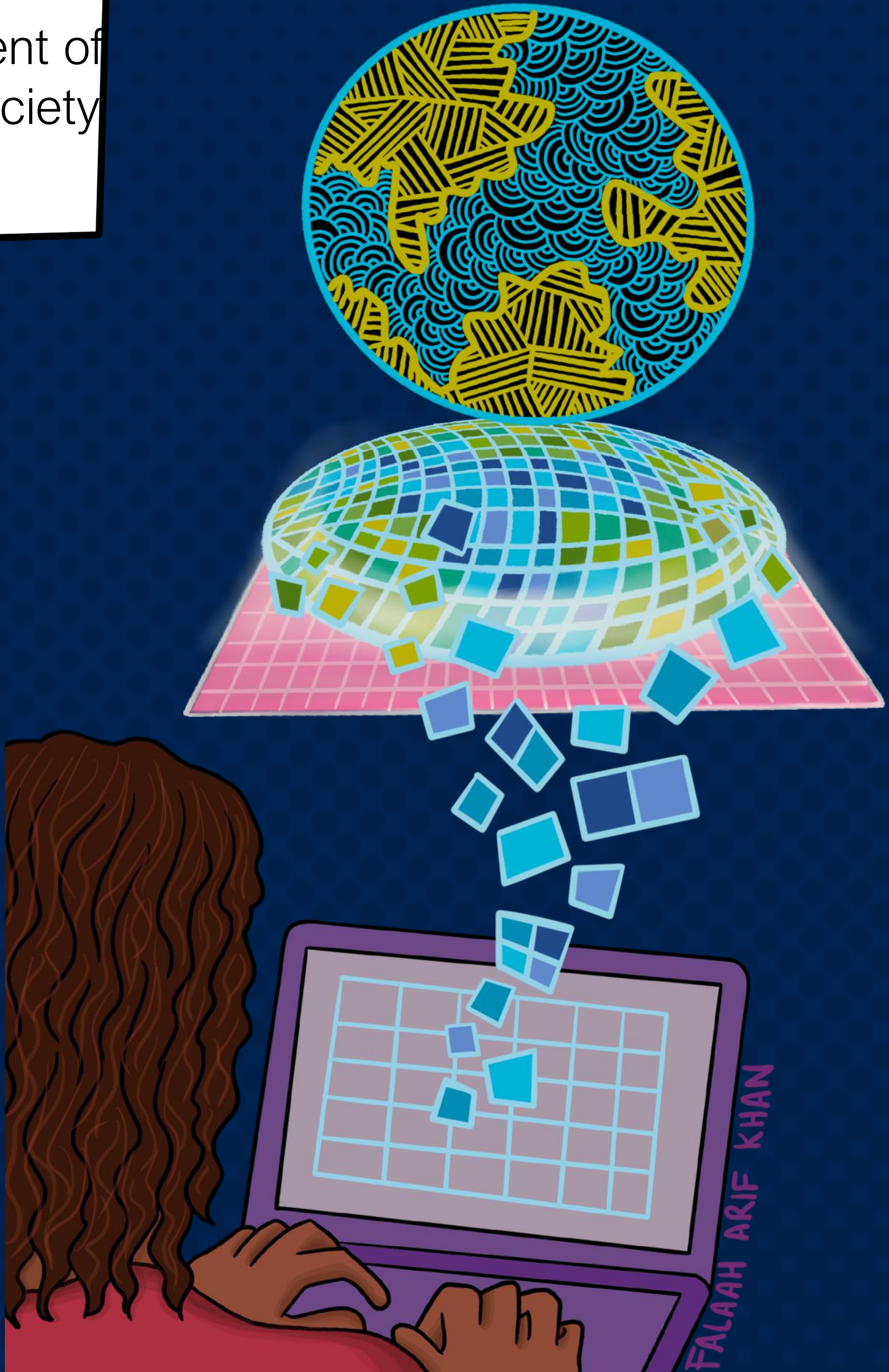
Emergent arises due to context of use

[Friedman & Nissenbaum (1996)]

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



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

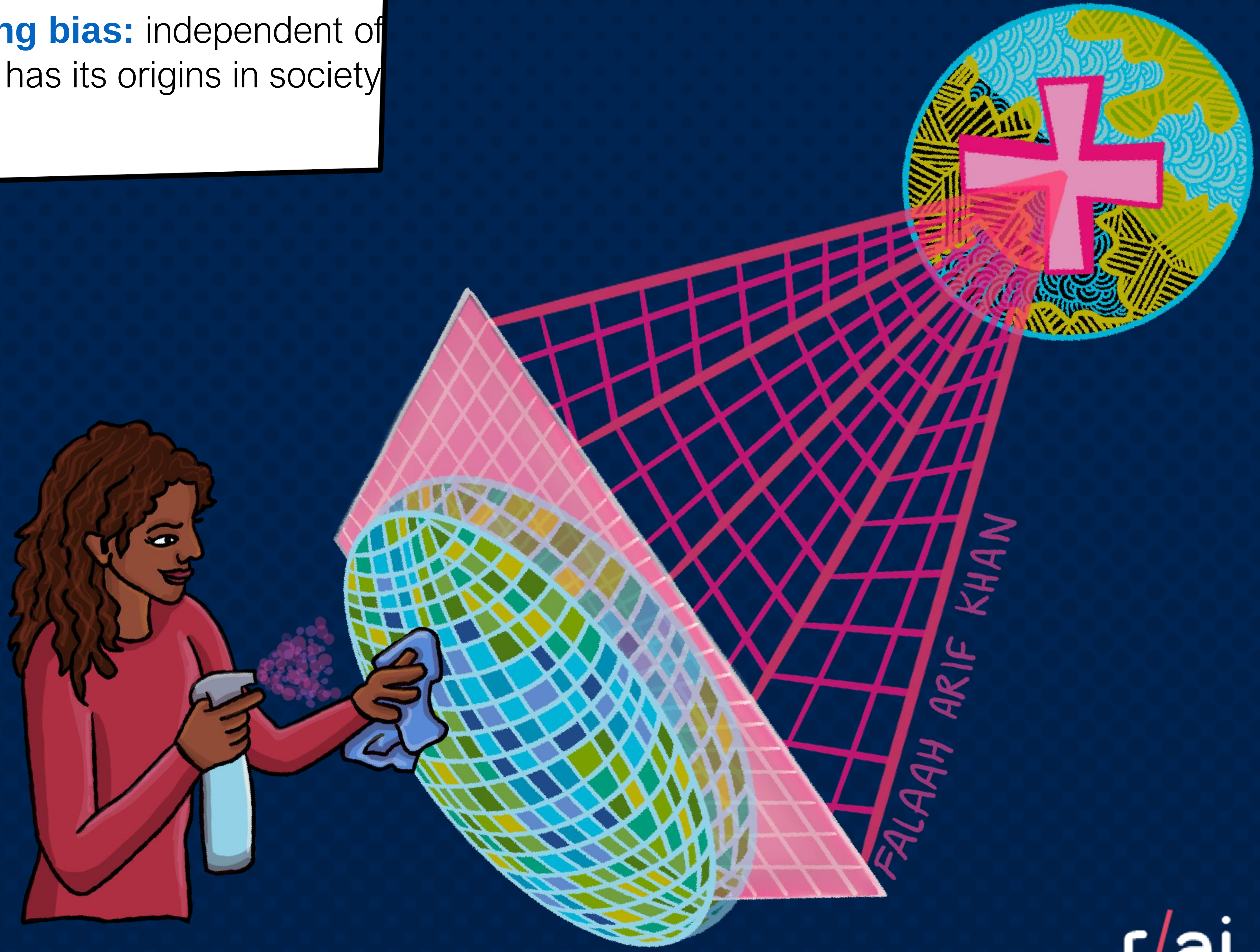


r/ai

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

r/ai

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



r/ai

bias can lead to discrimination

The evils of discrimination

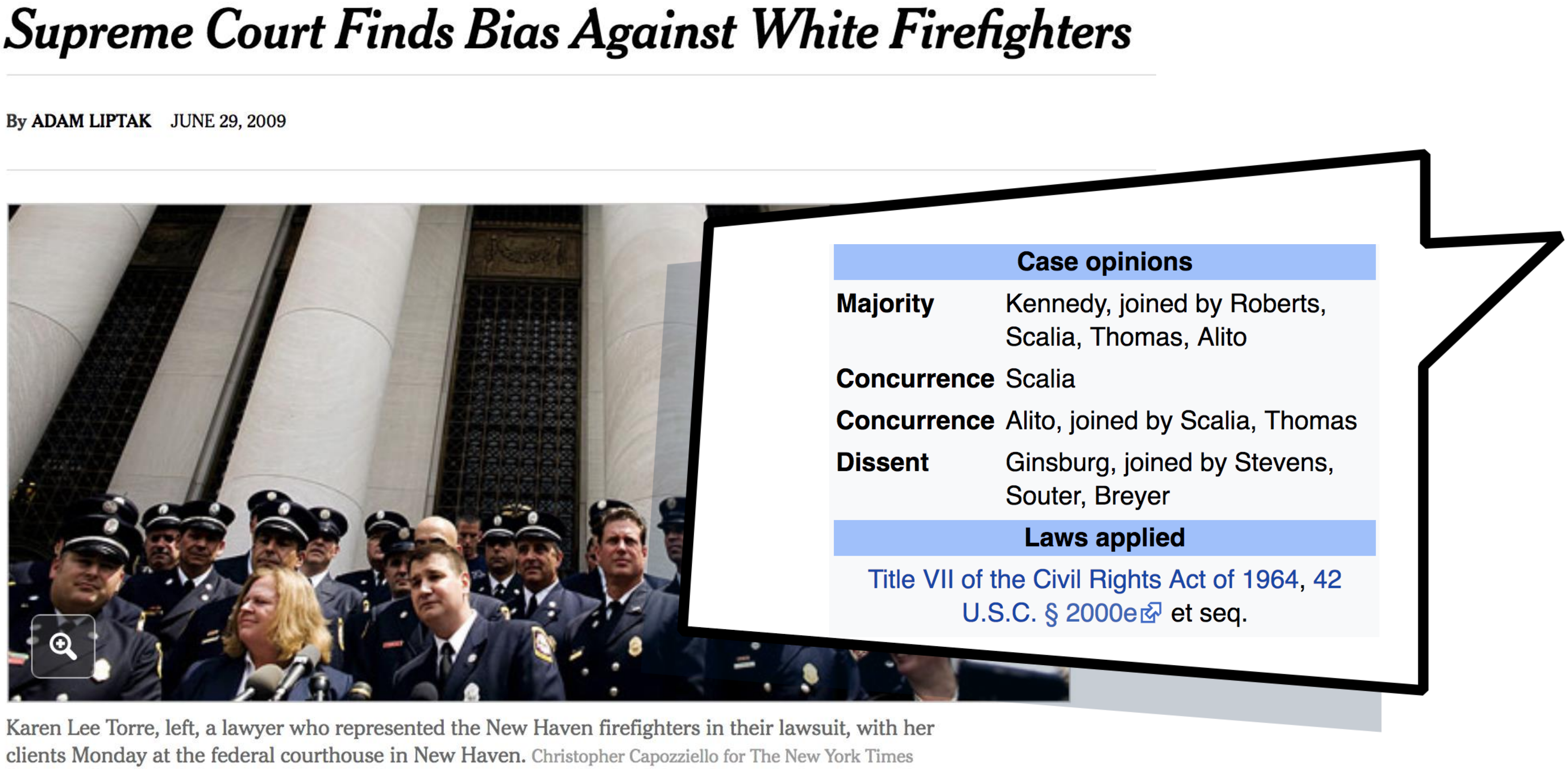
Disparate treatment

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**.

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

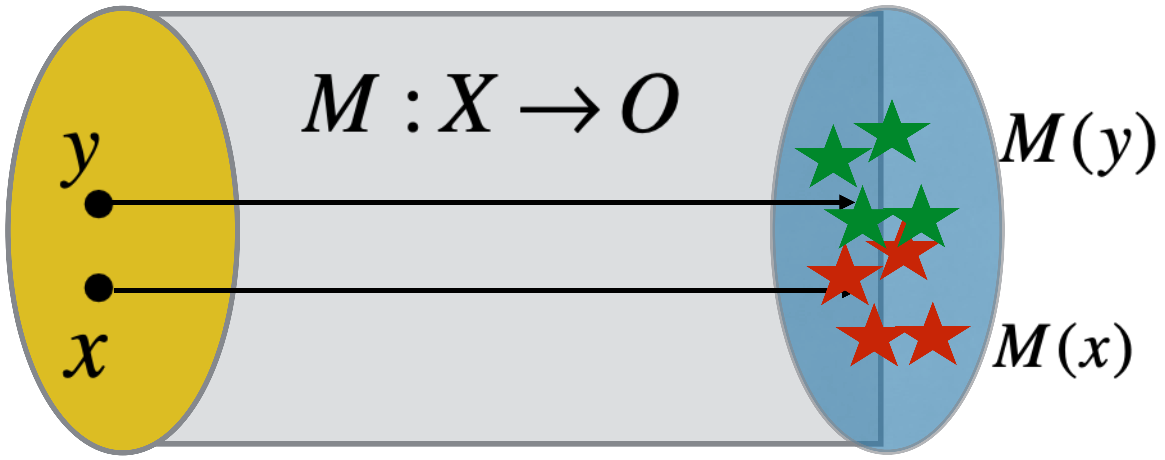
Case opinions	
Majority	Kennedy, joined by Roberts, Scalia, Thomas, Alito
Concurrence	Scalia
Concurrence	Alito, joined by Scalia, Thomas
Dissent	Ginsburg, joined by Stevens, Souter, Breyer
Laws applied	
Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e et seq.	

fairness through awareness

Fairness through awareness

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]

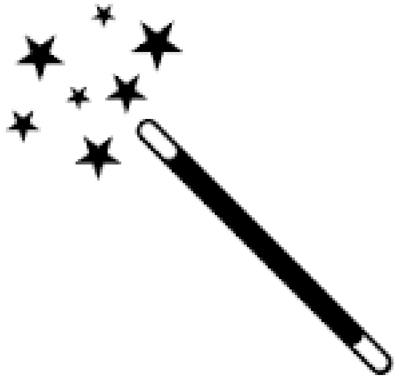
Fairness: Individuals who are **similar** for the purpose of classification task should be **treated similarly**.



X individuals

O outcomes

A task-specific distance metric is given $d(x,y)$

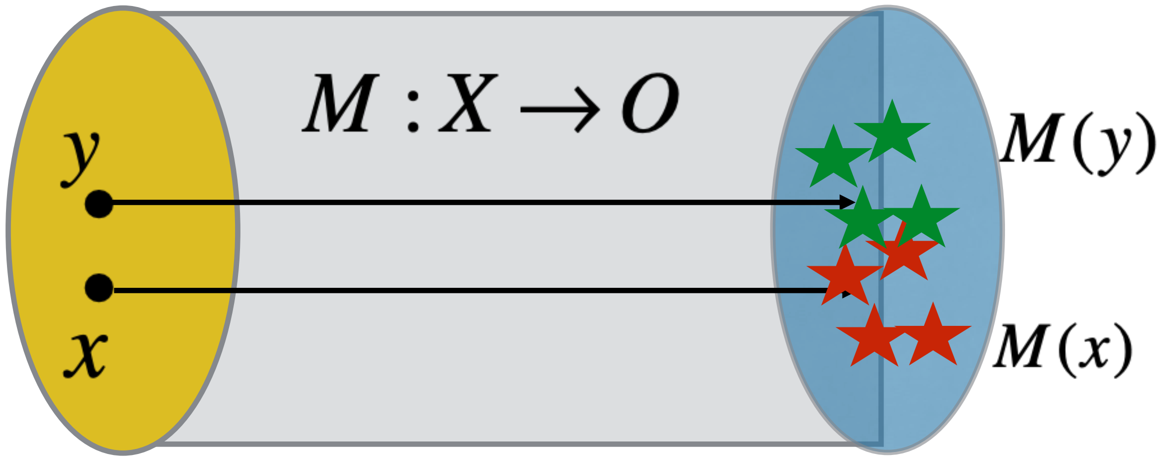


$M : X \to O$ is a **randomized mapping**: an individual is mapped to a distribution over outcomes

Fairness through a Lipschitz mapping

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]

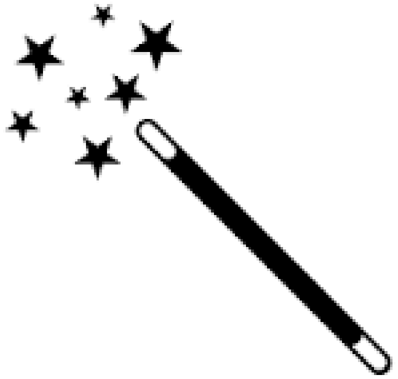
Fairness: Individuals who are **similar** for the purpose of classification task should be **treated similarly**.



X individuals

O outcomes

A task-specific distance metric is given $d(x,y)$



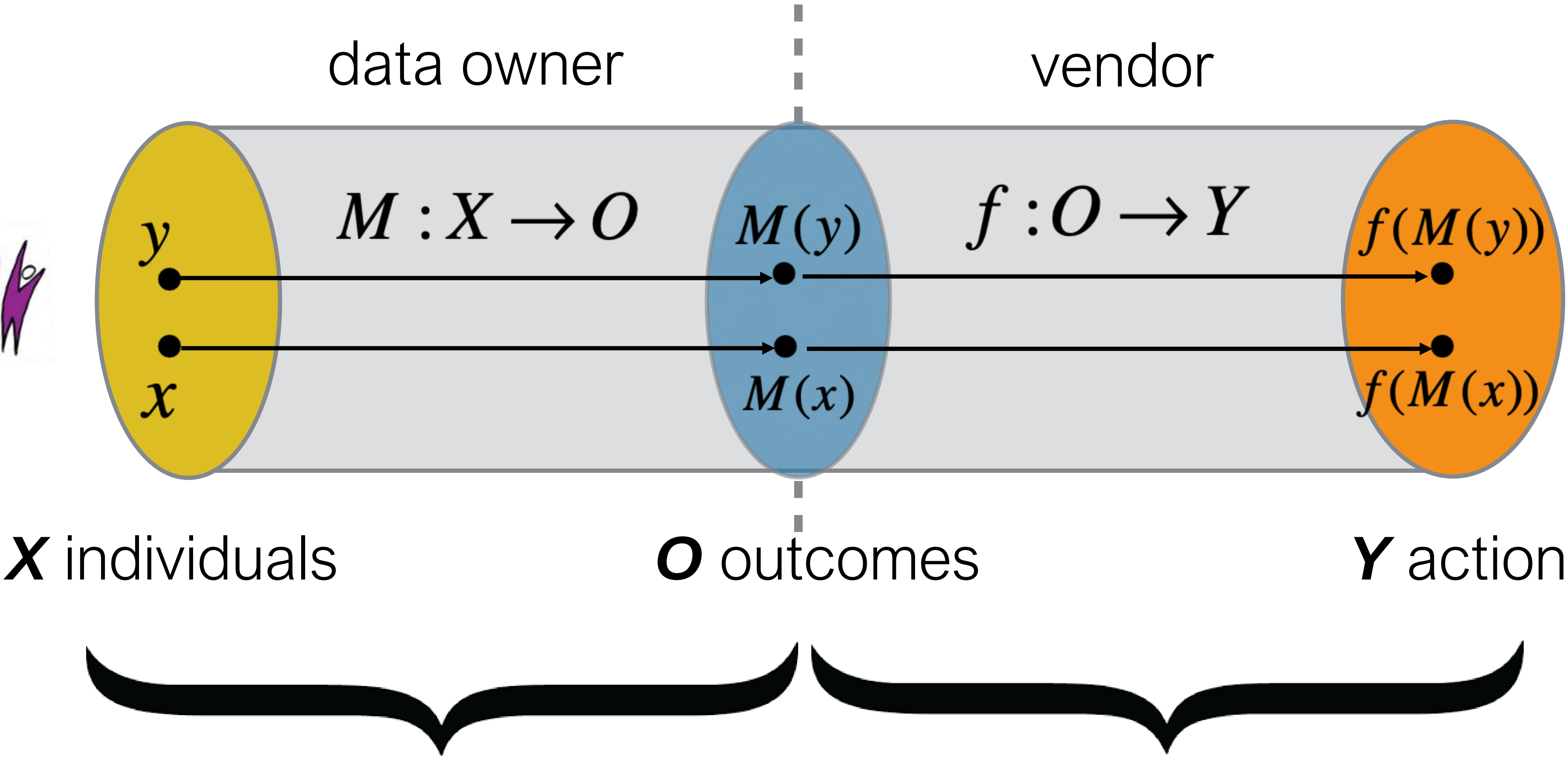
M is a Lipschitz mapping if $\forall x,y \in X \quad \|M(x),M(y)\| \le d(x,y)$

close individuals map to close distributions

there always exists a Lipschitz mapping - which?

Fairness through a Lipschitz mapping

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]



X individuals

O outcomes

Y actions

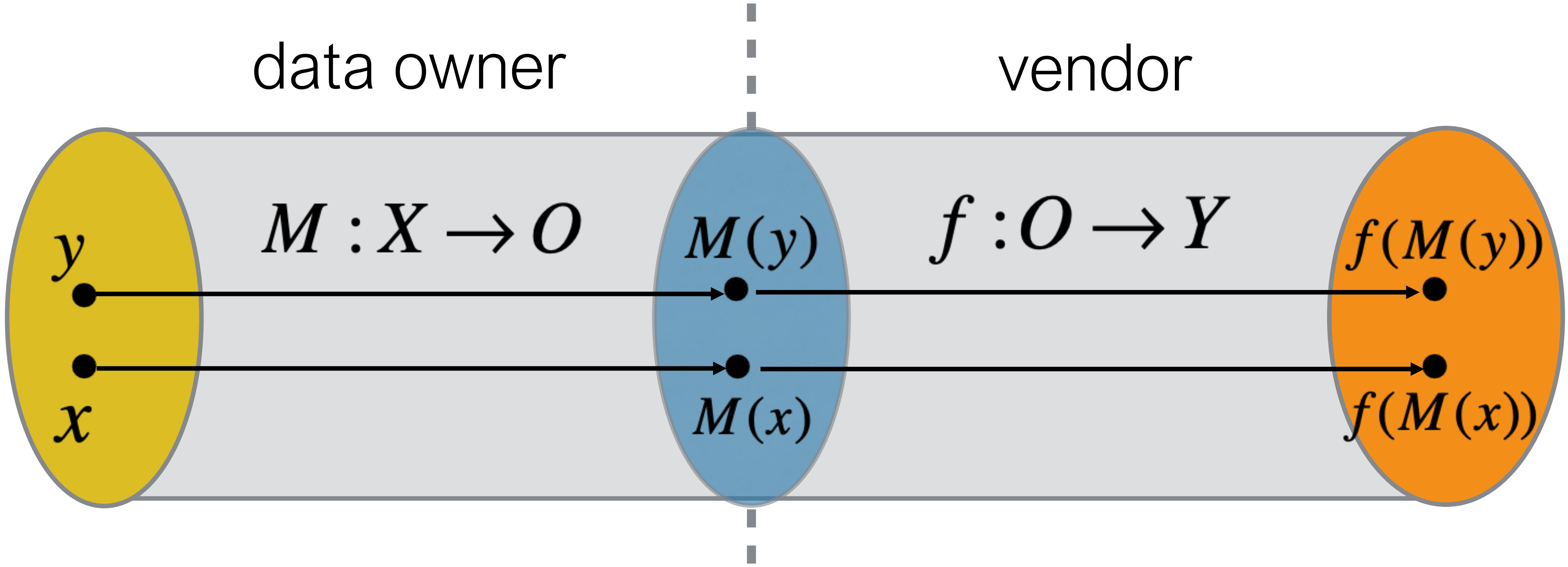
simpsons.wikia.com

fairness enforced at this step

vendor cannot introduce bias

Fairness through a Lipschitz mapping

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]



data owner

vendor

$M : X \to O$

$f : O \to Y$

y

x

$M(y)$

$M(x)$

$f(M(y))$

$f(M(x))$

simpsons.wikia.com

X individuals

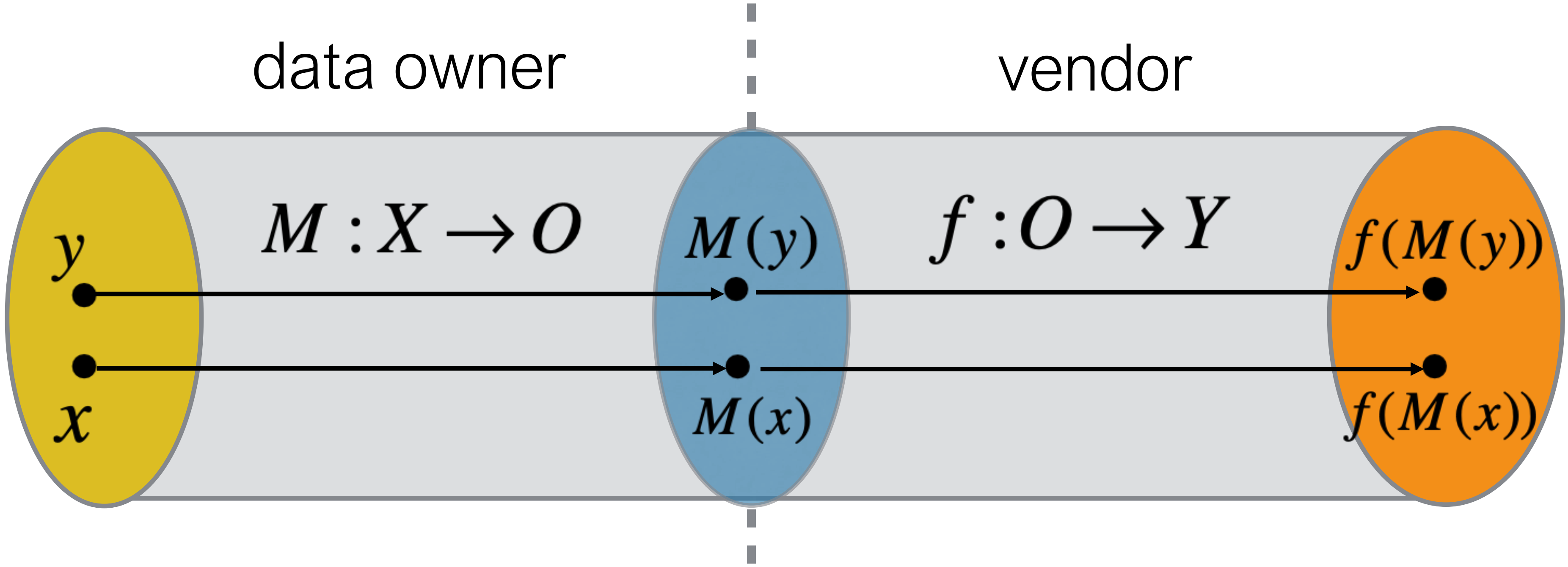
O outcomes

Y actions

Find a mapping from individuals to distributions over outcomes that minimizes expected loss, **subject to the Lipschitz condition**. Optimization problem: minimize an arbitrary loss function.

Fairness through a Lipschitz mapping

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]



simpsons.wikia.com

X individuals

O outcomes

Y actions

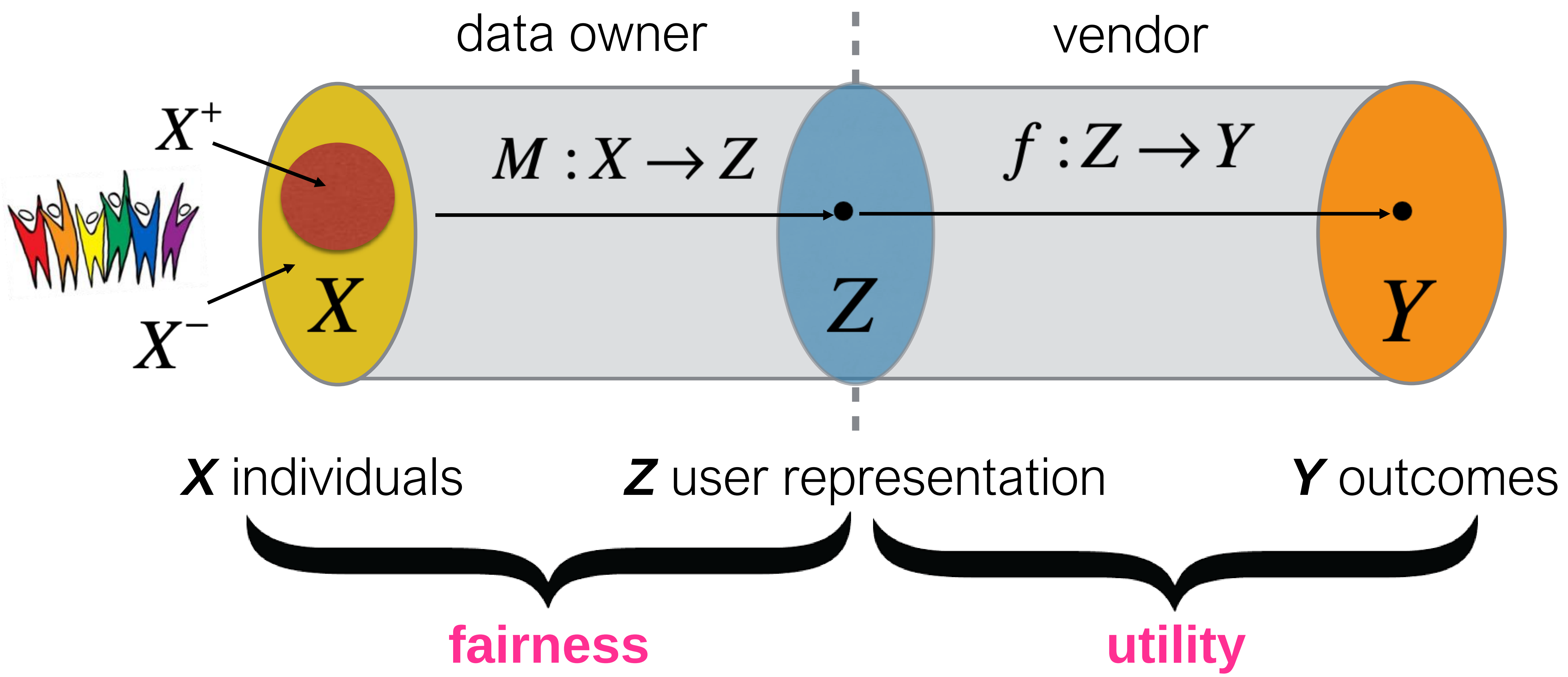
Computed with a linear program of size $poly(|X|,|Y|)$

the same mapping can be used by multiple vendors

learning fair representations

Learning fair representations

[R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork; *ICML 2013*]



data owner

vendor

X^+

X^-

X

$M : X \to Z$

Z

$f : Z \to Y$

Y

X individuals

Z user representation

Y outcomes

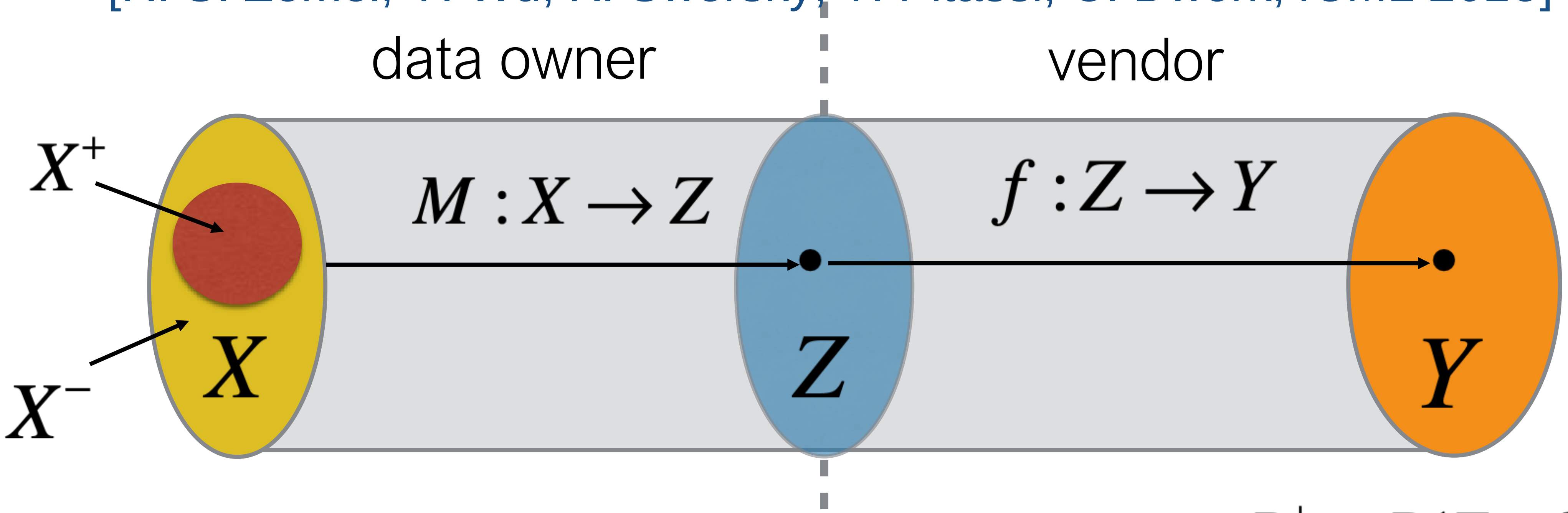
fairness

utility

Idea: remove reliance on a “fair” similarity measure, instead **learn** representations of individuals, distances

Fairness and utility

[R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork; *ICML 2013*]



earn a **randomized mapping** M(X) to a set of K prototypes Z

$$P_k^+ = P(Z = k \mid x \in X^+)$$

(X) should lose information about membership in S

$$P_k^- = P(Z = k \mid x \in X^-)$$

(X) should preserve other information so that vendor can maximize utility

$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

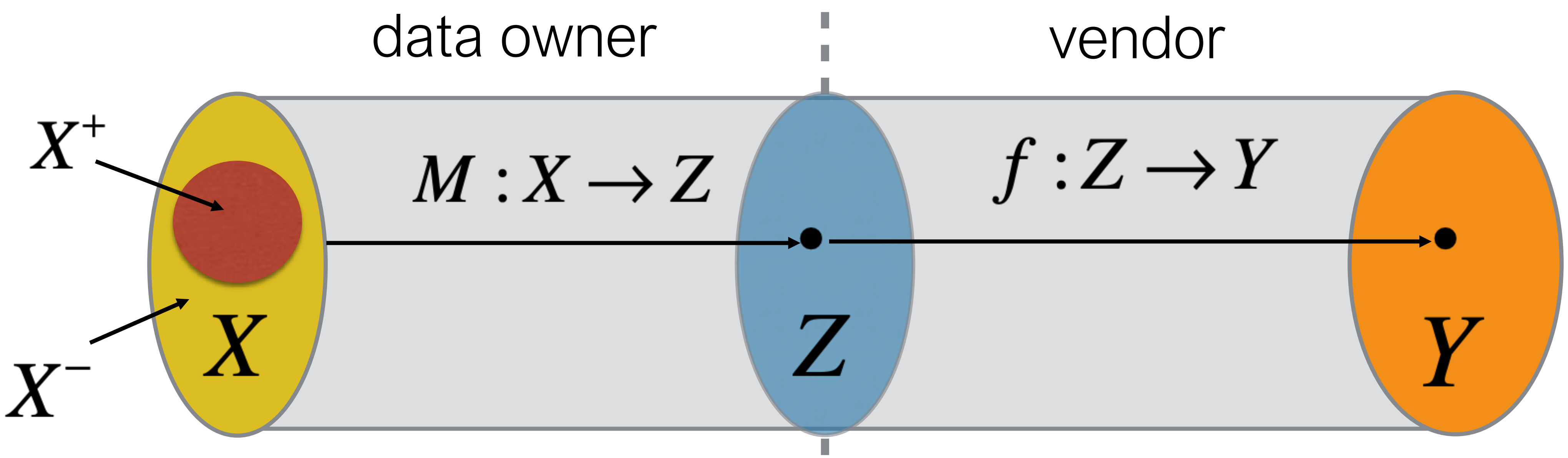
group fairness

individual fairness

utility

Fairness and utility

[R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork; *ICML 2013*]



$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

group fairness

individual fairness

utility

$$P_k^+ = P(Z = k \mid x \in X^+)$$

$$L_x = \sum_n (x_n - \widehat{x_n})^2$$

$$P_k^- = P(Z = k \mid x \in X^-)$$

$$L_y = \sum_n -y_n \log \widehat{y_n} - (1 - y_n) \log(1 - \widehat{y_n})$$

$$L_z = \sum_k \left| P_k^+ - P_k^- \right|$$

does this make sense?

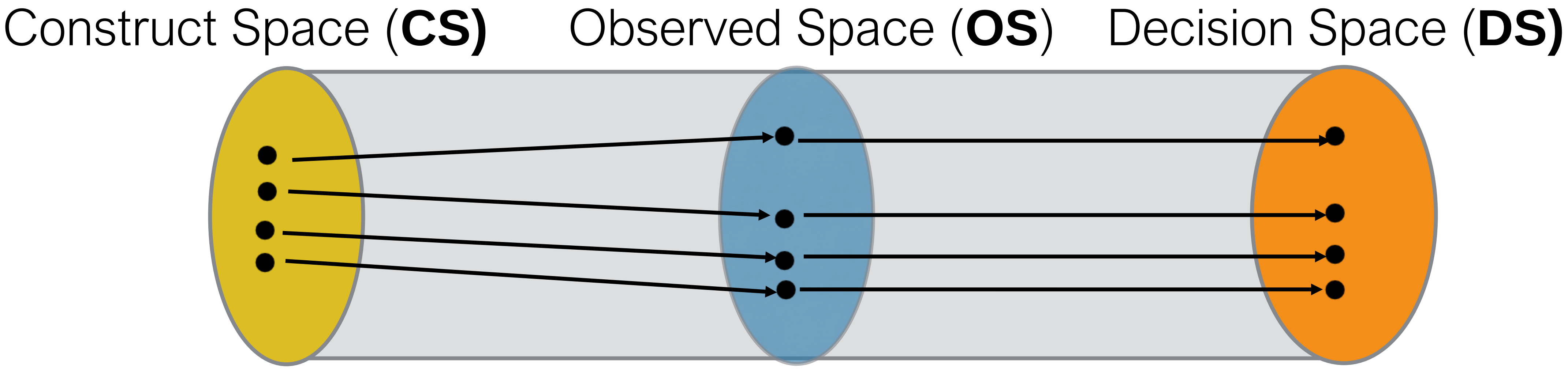
on the (im)possibility of fairness

On the (im)possibility of fairness

[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]

e out the difference between *beliefs* and *mechanisms* that logically follow from

hmic fairness is to study the interactions between different spaces that make u



Construct Space (**CS)**
Observed Space (**OS**)
Decision Space (**DS)**

On the (im)possibility of fairness

[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]

Construct Space	Observed Space	Decision Space
intelligence	SAT score	performance in college
grit	high-school GPA	
propensity to commit crime	family history	recidivism
risk-averseness	age	

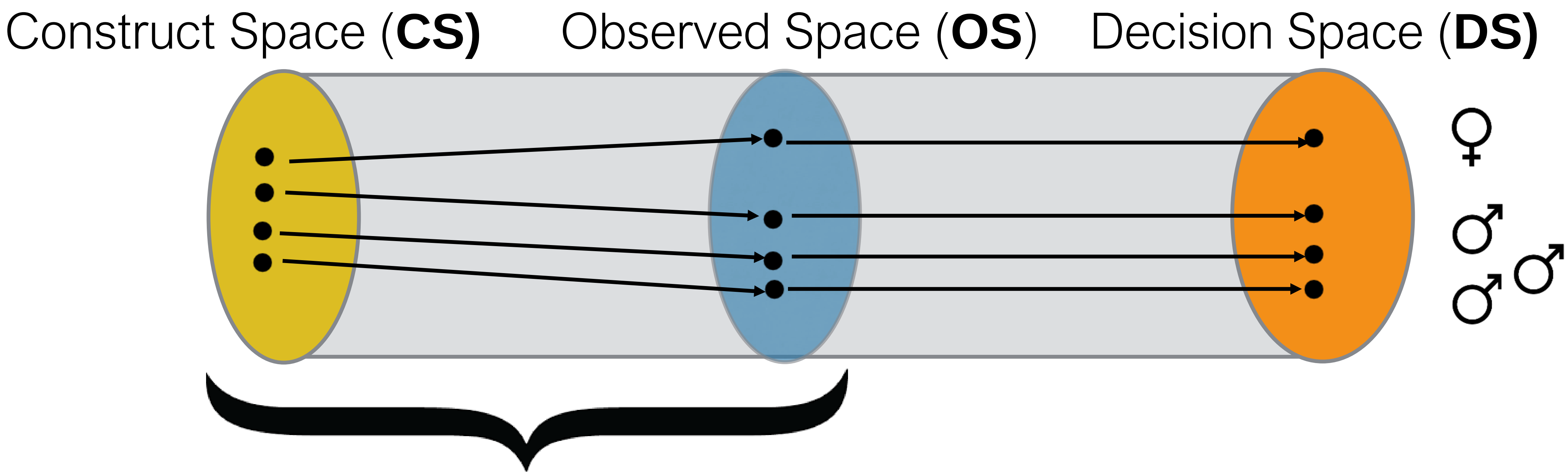
define fairness through properties of mappings

Fairness through mappings

[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]

Fairness: a mapping from CS to DS is (ε, ε')-fair if two objects that are no further than ε in CS map to objects that are no further than ε' in DS.

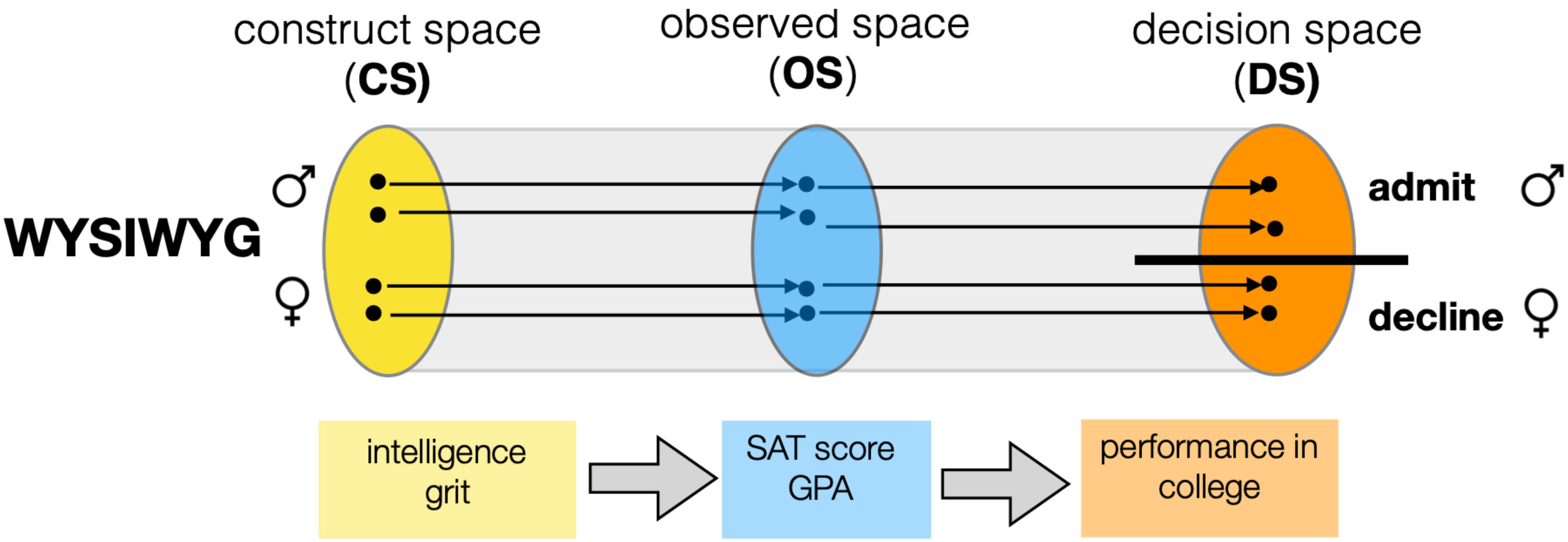
$$f : CS \rightarrow DS \qquad d_{CS}(x,y) < \varepsilon \Rightarrow d_{DS}(f(x), f(y)) < \varepsilon'$$



let's focus on this portion

WYSWYG

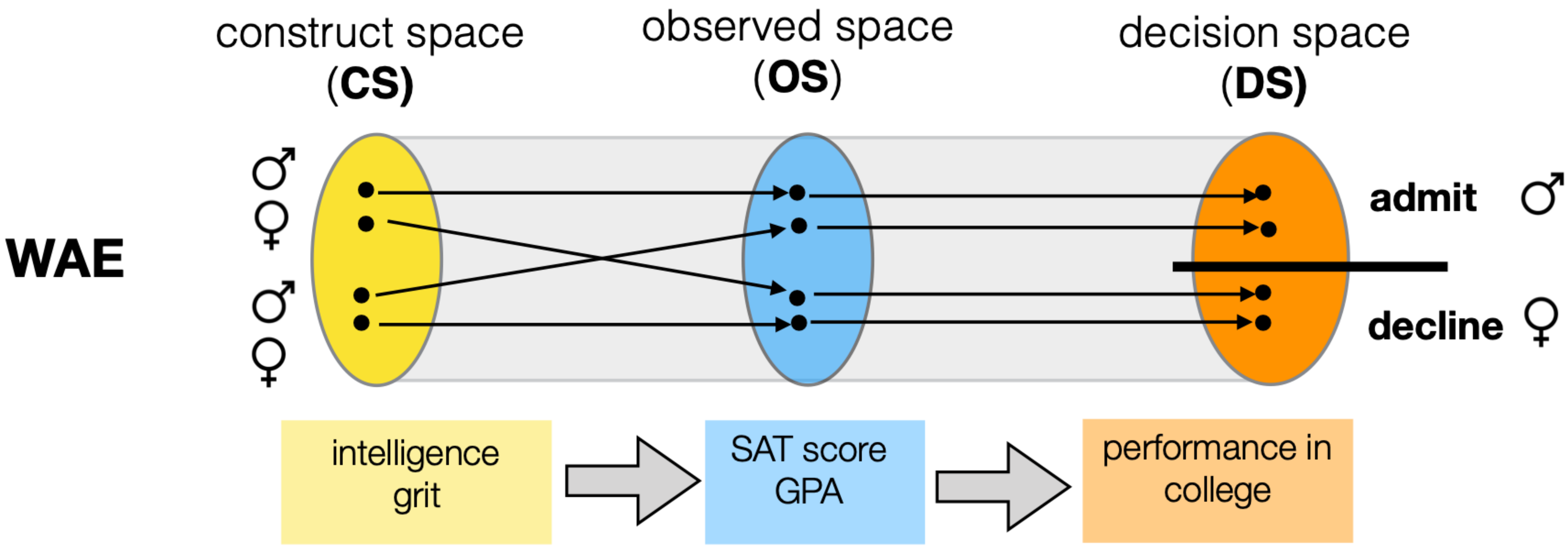
[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]



What you see is what you get (**WYSIWYG**): there exists a mapping from CS to OS that has low distortion. That is, we believe that OS faithfully represents CS. **This is the individual fairness world view.**

WAE

[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]

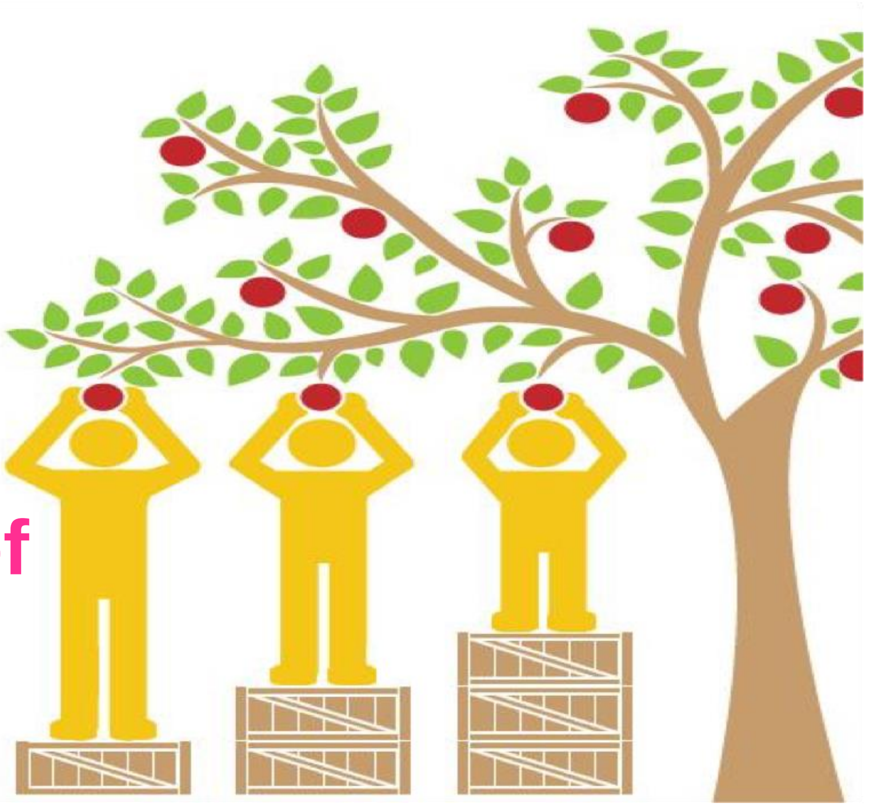


We are all equal (**WAE**): the mapping from CS to OS introduces **structural bias** - there is a distortion that aligns with the group structure of CS. **This is the group fairness world view.**

Structural bias examples: SAT verbal questions function differently in the African-American and in the Caucasian subgroups in the US. Other examples?

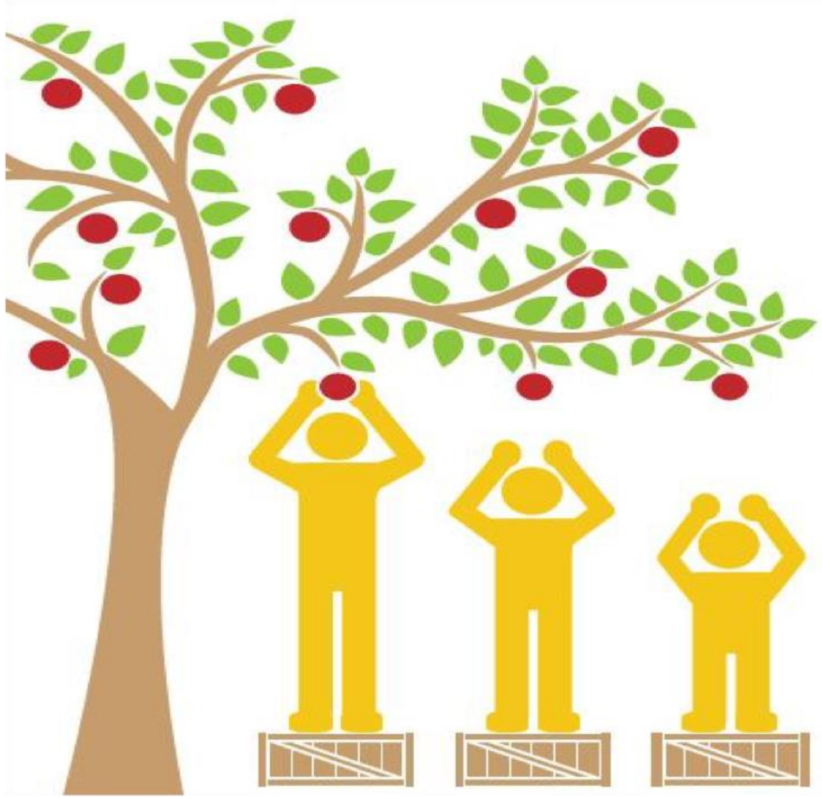
Fairness and worldviews

group fairness



equality of outcome

individual fairness



equality of treatment

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
- Recall “Learning Fair Representations”: a typical ML approach

$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

group fairness ↗ **individual fairness** ↖ **utility**

apples + oranges + fairness = ?

Responsible Data Science

Algorithmic Fairness II

February 6, 2025

Prof. Jonathan Colner

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

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai

The evils of discrimination

Disparate treatment

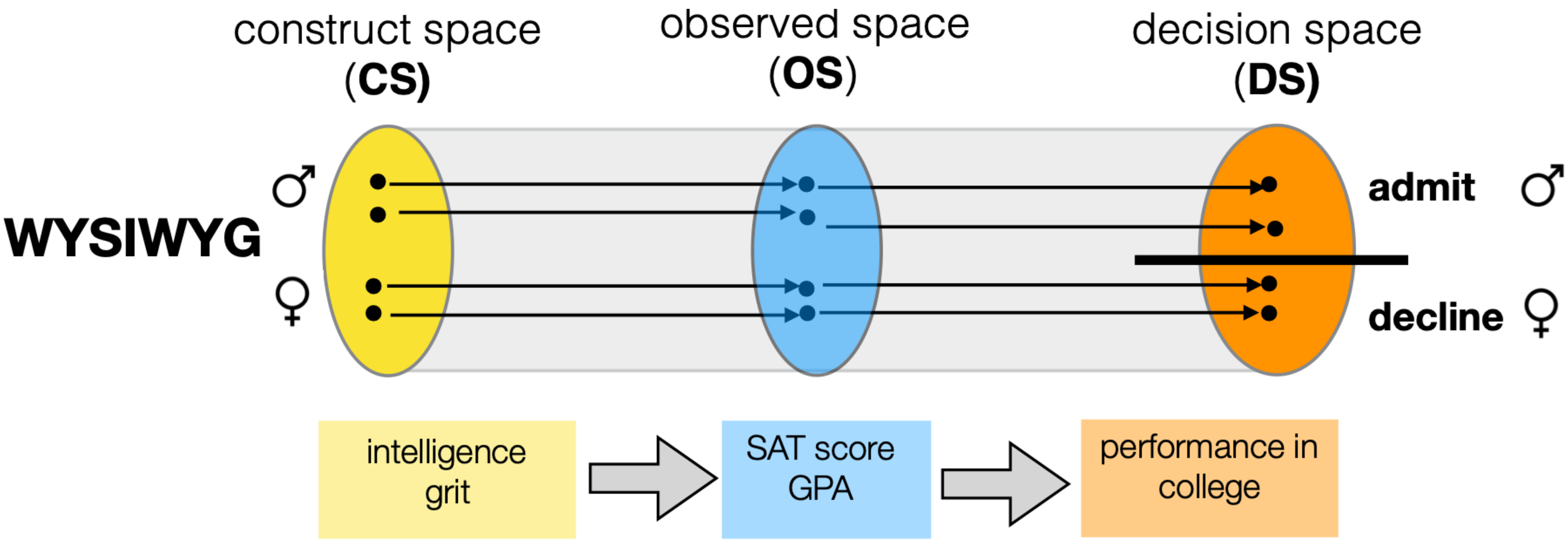
is the illegal practice of treating an entity, such as a job applicant or an employee, differently based on a **protected characteristic** such as race, gender, age, disability status, religion, sexual orientation, or national origin.

Disparate impact

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

WYSWYG

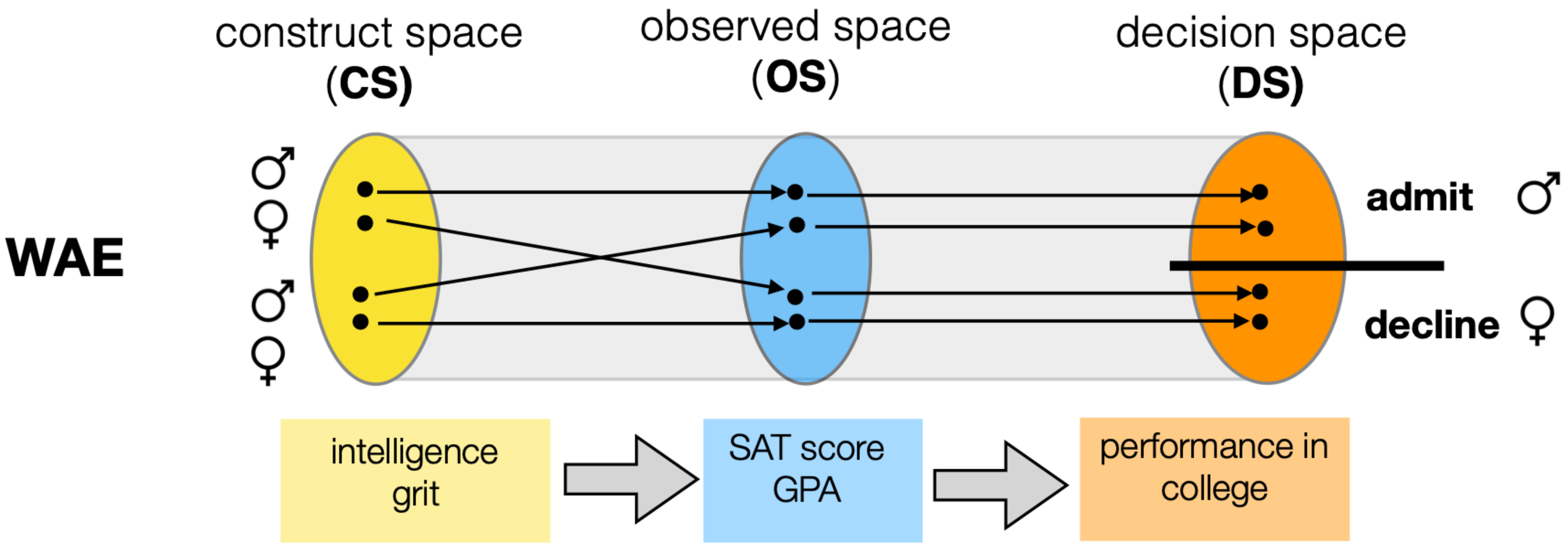
[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]



What you see is what you get (**WYSIWYG**): there exists a mapping from CS to OS that has low distortion. That is, we believe that OS faithfully represents CS. **This is the individual fairness world view.**

WAE

[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]



construct space (**CS**)

observed space (**OS**)

decision space (**DS**)

WAE

intelligence grit → SAT score GPA → performance in college

We are all equal (**WAE**): the mapping from CS to OS introduces **structural bias** - there is a distortion that aligns with the group structure of CS. **This is the group fairness world view.**

Structural bias examples: SAT verbal questions function differently in the African-American and in the Caucasian subgroups in the US. Other examples?

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
- Recall “Learning Fair Representations”: a typical ML approach

$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

group fairness **individual fairness** **utility**

apples + oranges + fairness = ?

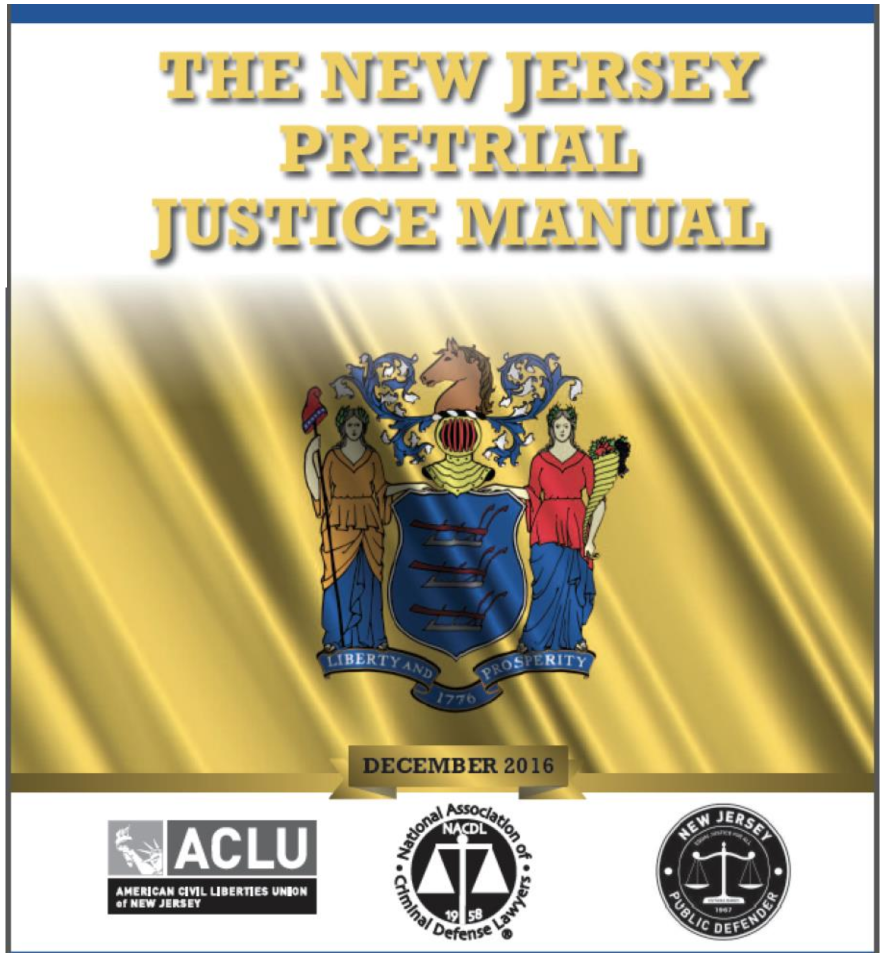
fairness in risk assessment

New Jersey bail reform

ACLU
AMERICAN CIVIL LIBERTIES UNION
of NEW JERSEY

National Association of Criminal Defense Lawyers®
NACDL
19 58

NEW JERSEY
EQUAL JUSTICE FOR ALL
ESTABLISHED
1967
PUBLIC DEFENDER



THE NEW JERSEY
PRETRIAL
JUSTICE MANUAL

LIBERTY AND PROSPERITY
1776

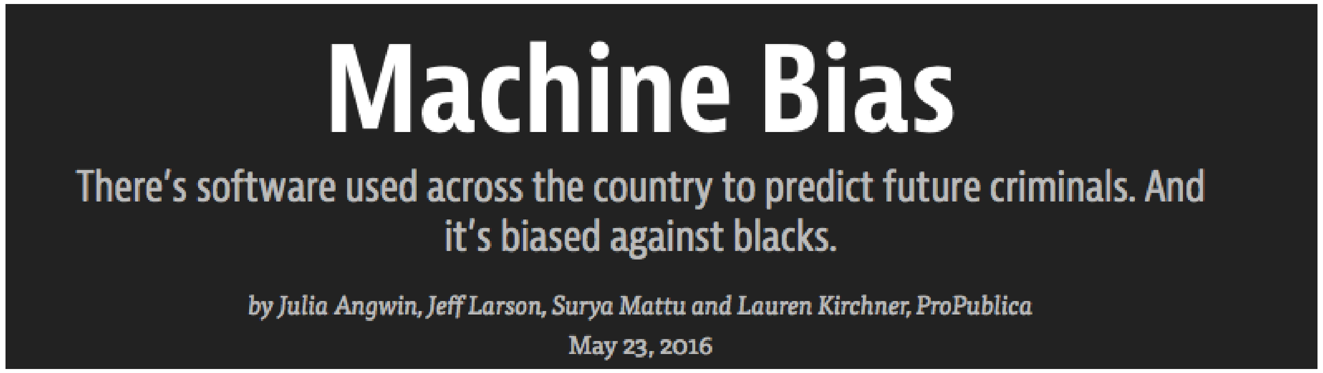
DECEMBER 2016

ACLU
AMERICAN CIVIL LIBERTIES UNION of NEW JERSEY
National Association of Criminal Defense Lawyers NACDL 19 58
NEW JERSEY PUBLIC DEFENDER

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.

https://www.nacdl.org/getattachment/50e0c53b-6641-4a79-8b49-c733def39e37/the-new-jersey-pretrial-justice-manual.pdf

ProPublica’s COMPAS study



Machine Bias

There’s software used across the country to predict future criminals. And it’s biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

May 2016

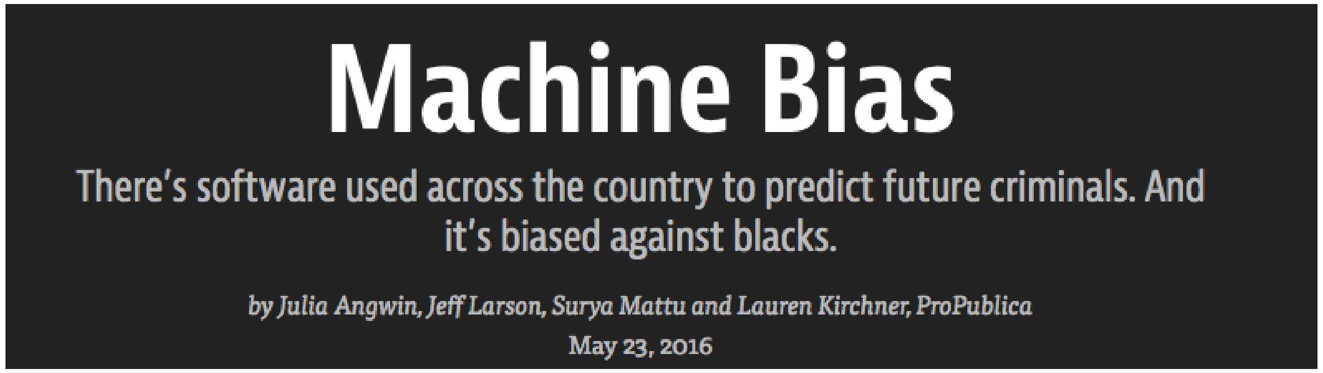
A commercial tool **COMPAS** automatically predicts some categories of future crime to assist in bail and sentencing decisions. It is used in courts in the US.

The tool correctly predicts recidivism **61% of the time.**

Blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend.
The tool 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.

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

Back to ProPublica's COMPAS study



Machine Bias

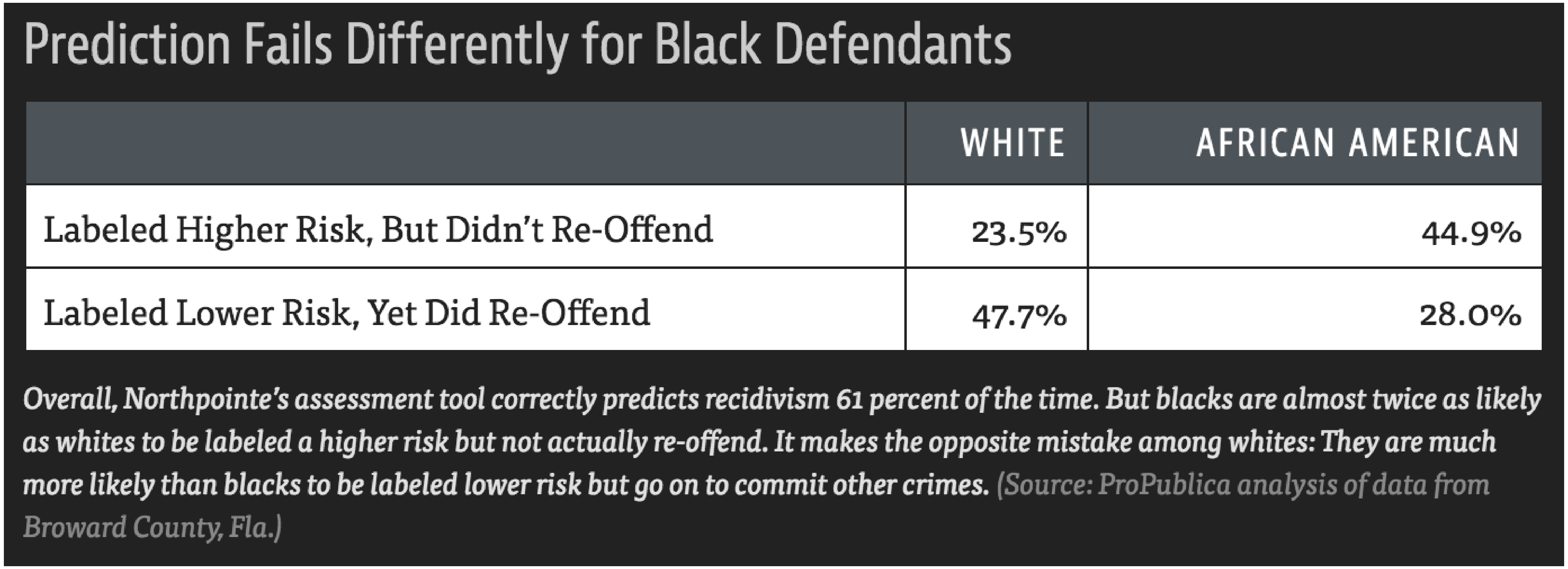
There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

May 2016

A commercial tool **COMPAS** automatically predicts some categories of future crime to assist in bail and sentencing decisions. COMPAS has been used by the U.S. states of NY, WI, CA, FL and other jurisdictions.



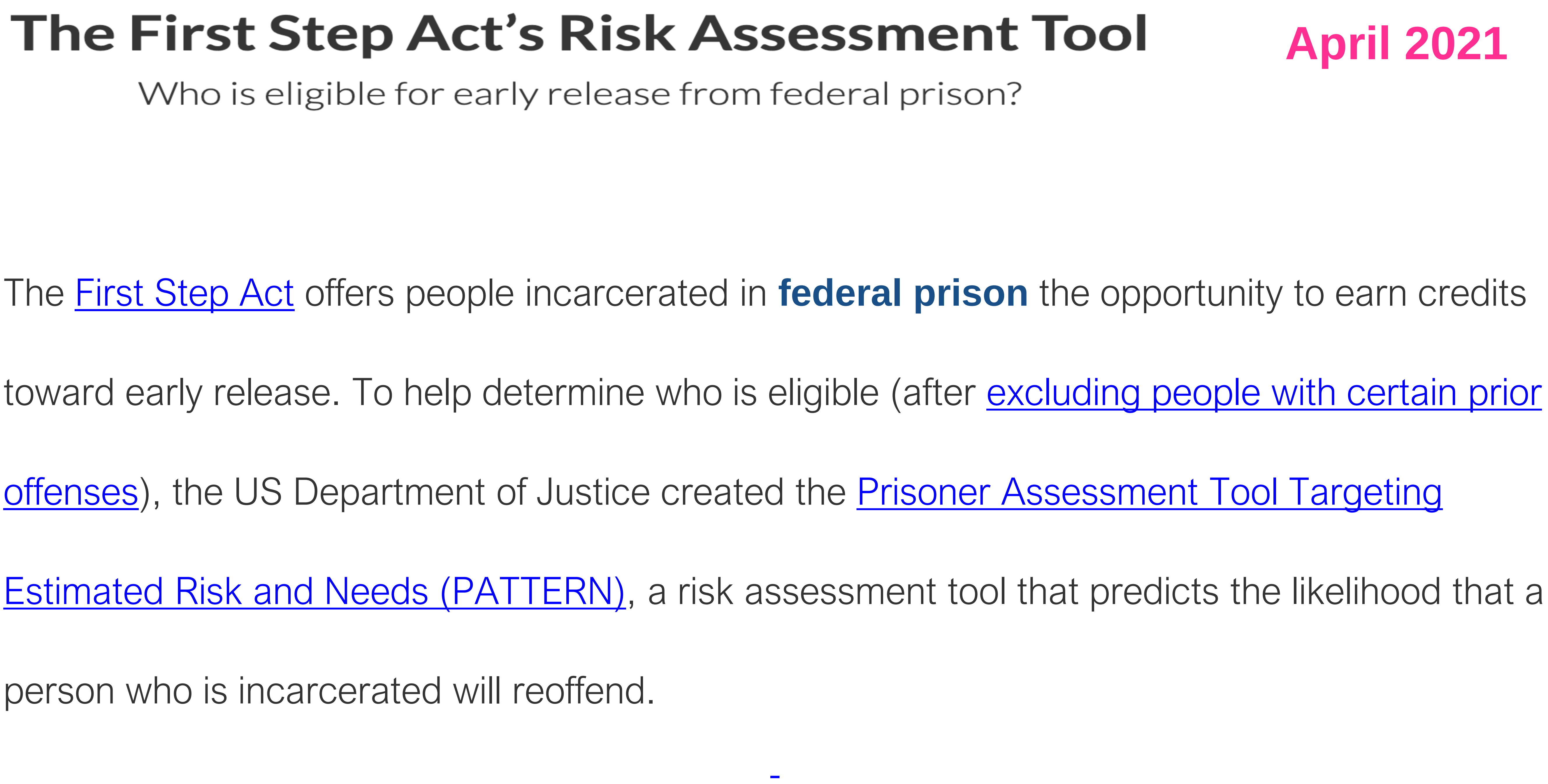
Prediction Fails Differently for Black Defendants

	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.)*

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

Similar tools are used today



The First Step Act's Risk Assessment Tool

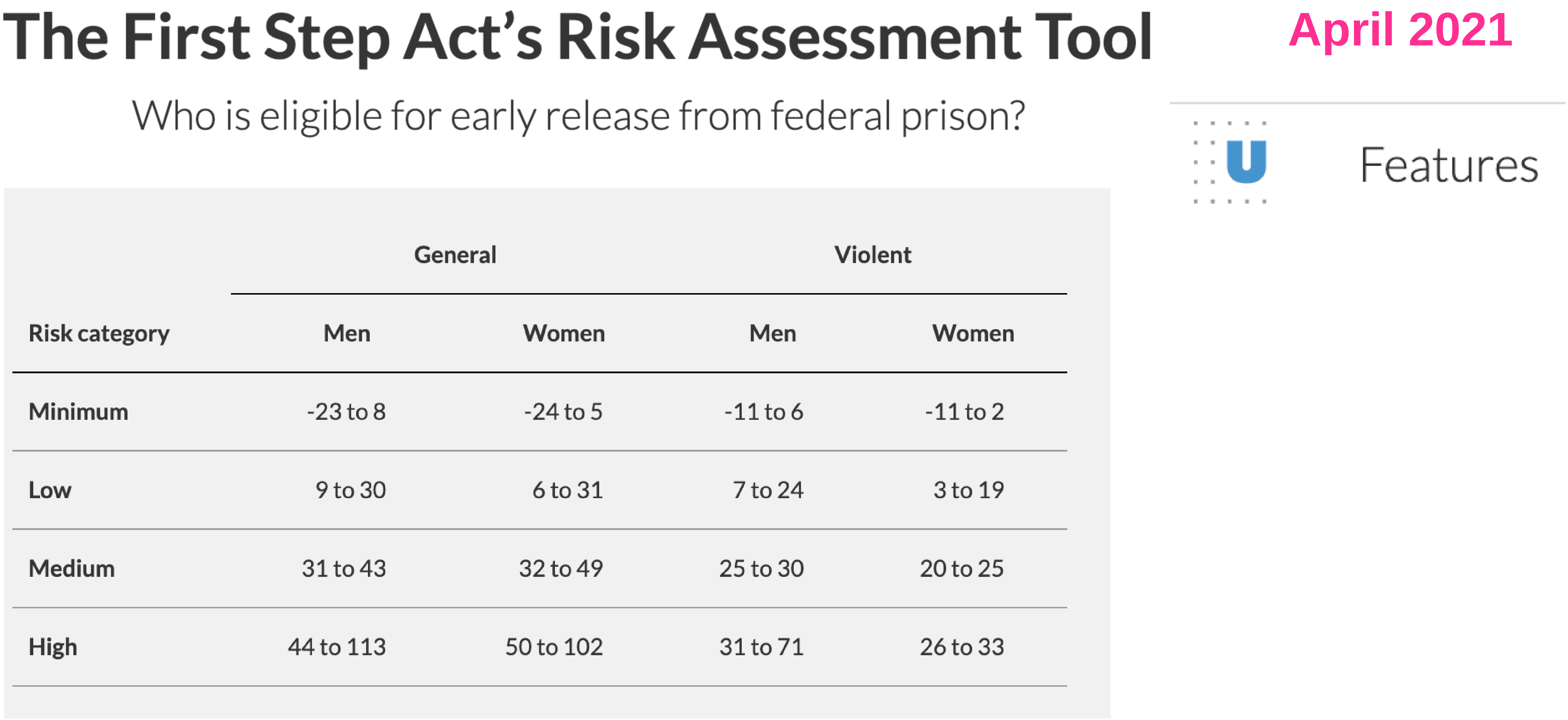
Who is eligible for early release from federal prison?

April 2021

The First Step Act offers people incarcerated in **federal prison** the opportunity to earn credits toward early release. To help determine who is eligible (after excluding people with certain prior offenses), the US Department of Justice created the Prisoner Assessment Tool Targeting Estimated Risk and Needs (PATTERN), a risk assessment tool that predicts the likelihood that a person who is incarcerated will reoffend.

https://apps.urban.org/features/risk-assessment/

These tools are used today



The First Step Act's Risk Assessment Tool

April 2021

Who is eligible for early release from federal prison?

U Features

	General		Violent	
Risk category	**Men**	**Women**	**Men**	**Women**
Minimum	-23 to 8	-24 to 5	-11 to 6	-11 to 2
Low	9 to 30	6 to 31	7 to 24	3 to 19
Medium	31 to 43	32 to 49	25 to 30	20 to 25
High	44 to 113	50 to 102	31 to 71	26 to 33

https://apps.urban.org/features/risk-assessment/

These tools are used today

LAW

Flaws plague a tool meant to help low-risk federal prisoners win early release

January 26, 2022 · 5:00 AM ET

Heard on Morning Edition

CARRIE JOHNSON

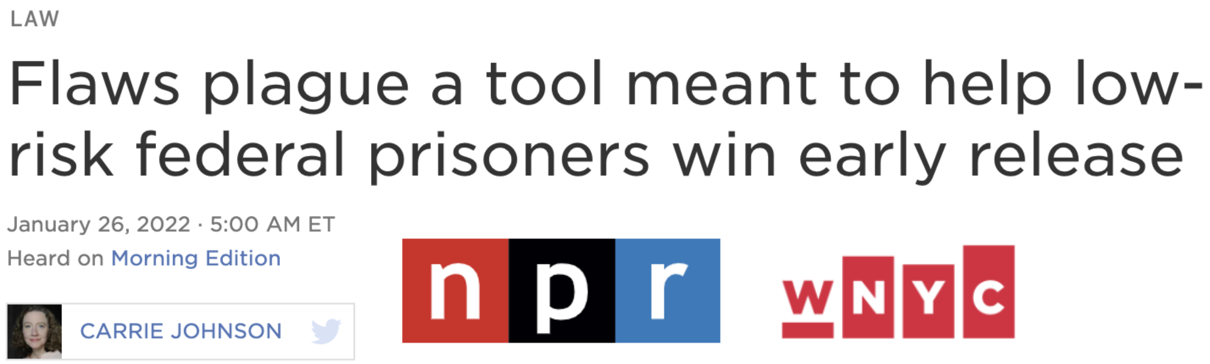
January 2022

Thousands of people are leaving federal prison this month thanks to a law called the First Step Act, which allowed them to win early release by participating in programs aimed at easing their return to society. But thousands of others may still remain behind bars because of fundamental flaws in the Justice Department's method for deciding who can take the early-release track. The biggest flaw: **persistent racial disparities that put Black and brown people at a disadvantage**.

[…] The algorithm, known as **Pattern**, **overpredicted the risk that many Black, Hispanic and Asian people** would commit new crimes or violate rules after leaving prison. At the same time, it also **underpredicted the risk for some inmates of color when it came to possible return to violent crime**.

https://www.npr.org/2022/01/26/1075509175/justice-department-algorithm-first-step-act

These tools are used today



LAW

Flaws plague a tool meant to help low-risk federal prisoners win early release

January 26, 2022 · 5:00 AM ET
Heard on Morning Edition

CARRIE JOHNSON

January 2022

Aamra Ahmad, senior policy counsel at the American Civil Liberties Union: ”The Justice Department found that **only 7% of Black people in the sample were classified as minimum level risk compared to 21% of white people**," she added. "This indicator alone should give the Department of Justice great pause in moving forward.”

Risk assessment tools are common in many states. But critics said Pattern is the first time the federal justice system is using an algorithm with such high stakes.

"**Especially when systems are high risk and affect people's liberty, we need much clearer and stronger oversight**," said Costanza-Chock [director of research & design for the Algorithmic Justice League]

https://www.npr.org/2022/01/26/1075509175/justice-department-algorithm-first-step-act

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

Calibration

positive
outcomes:
do recidivate

	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

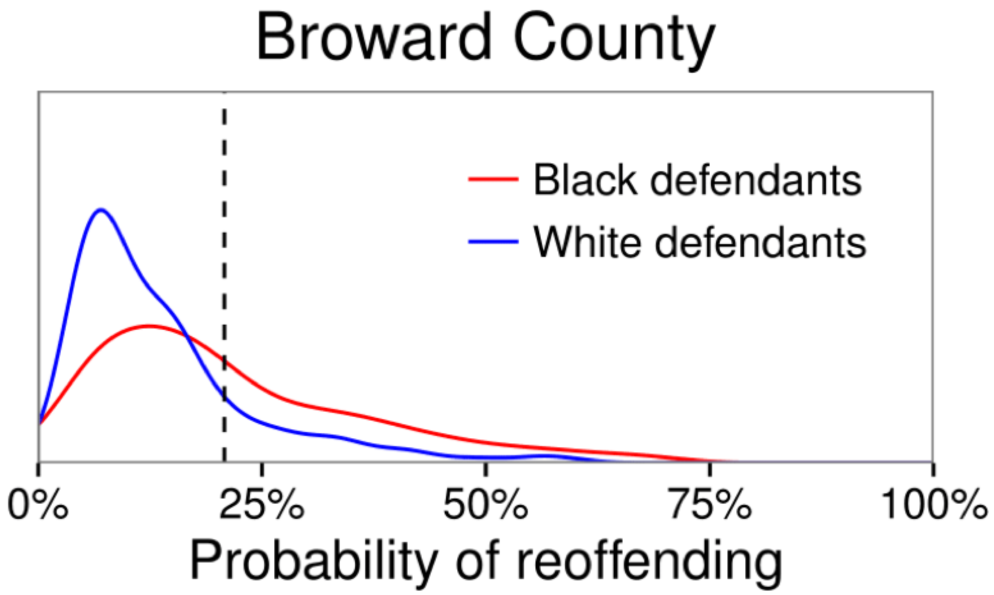
0.6 means 0.6 for any defendant - likelihood of recidivism

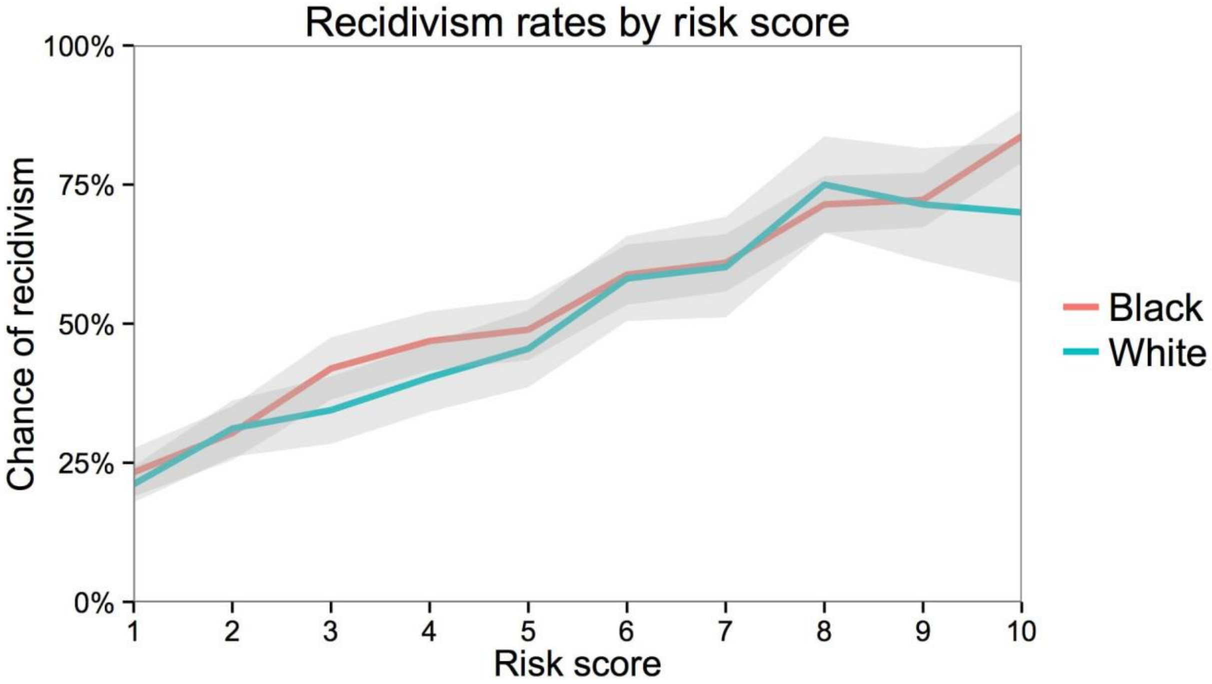
why do we want calibration?

COMPAS as a predictive instrument

Predictive parity (also called **calibration**)
an instrument identifies a set of instances as having probability *x* of constituting positive instances, then approximately an *x* fraction of this set are indeed positive instances, over-all and in sub-populations

COMPAS is well-calibrated: in the window around 40%, the fraction of defendants who were re-arrested is ~40%, both over-all and per group.



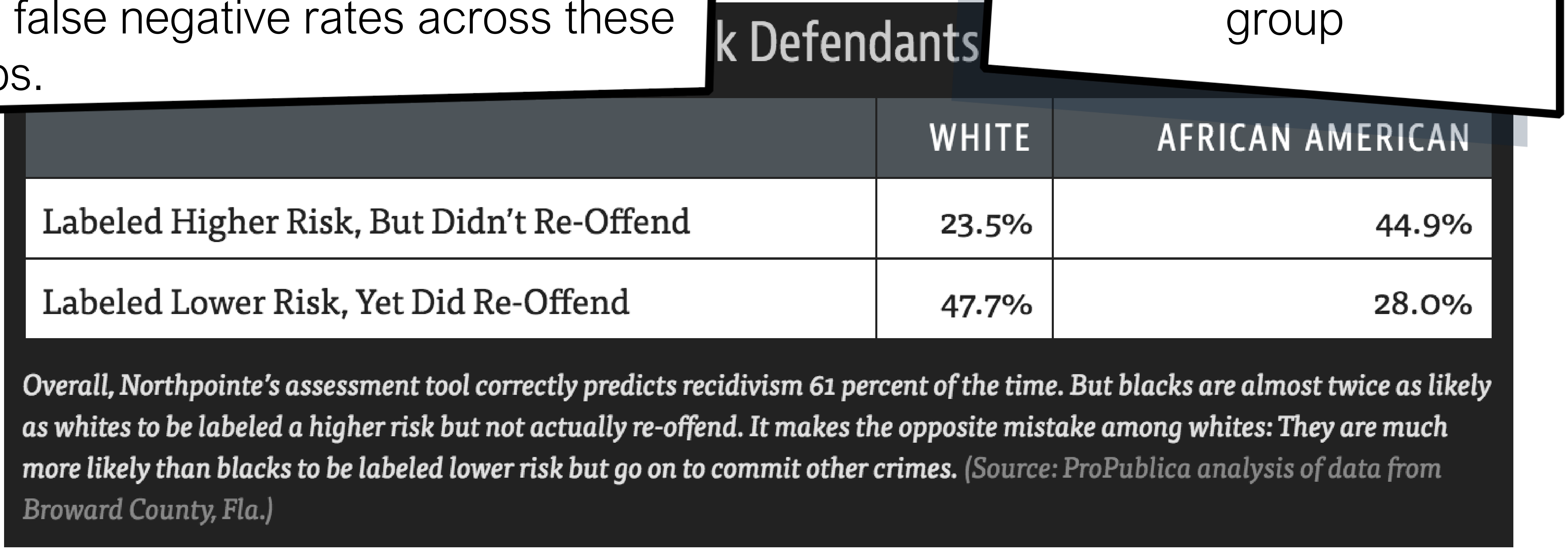


[plot from Corbett-Davies et al.; *KDD 2017*]

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



k 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.)

[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 of: 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 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 recidivates (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: of those labeled low-risk, how many will recidivate?

Defendant: how likely will I 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
- Recall “Learning Fair Representations”: a typical ML approach

$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

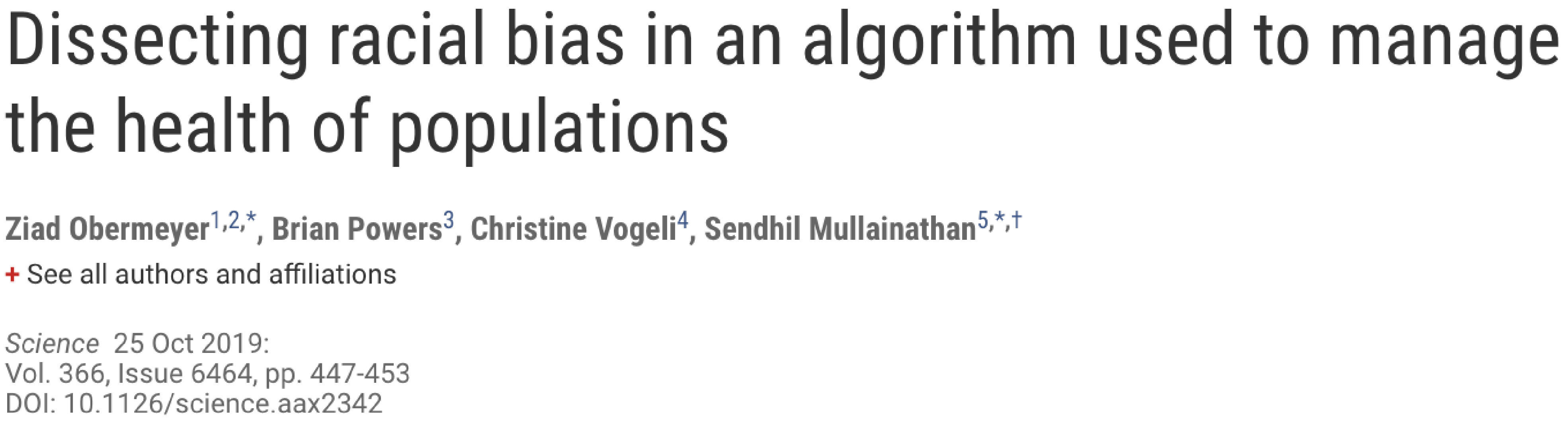
group fairness

individual fairness

utility

apples + oranges + fairness = ?

Racial bias in healthcare



Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2,*], Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5,*,†]

+ See all authors and affiliations

Science 25 Oct 2019:
Vol. 366, Issue 6464, pp. 447-453
DOI: 10.1126/science.aax2342

October 2019

Science

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and **affecting millions of patients**, exhibits significant **racial bias**: **At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses**. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm **predicts health care costs rather than illness**, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, **despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise**. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic
bias in many contexts.

r/ai

Racial bias in healthcare

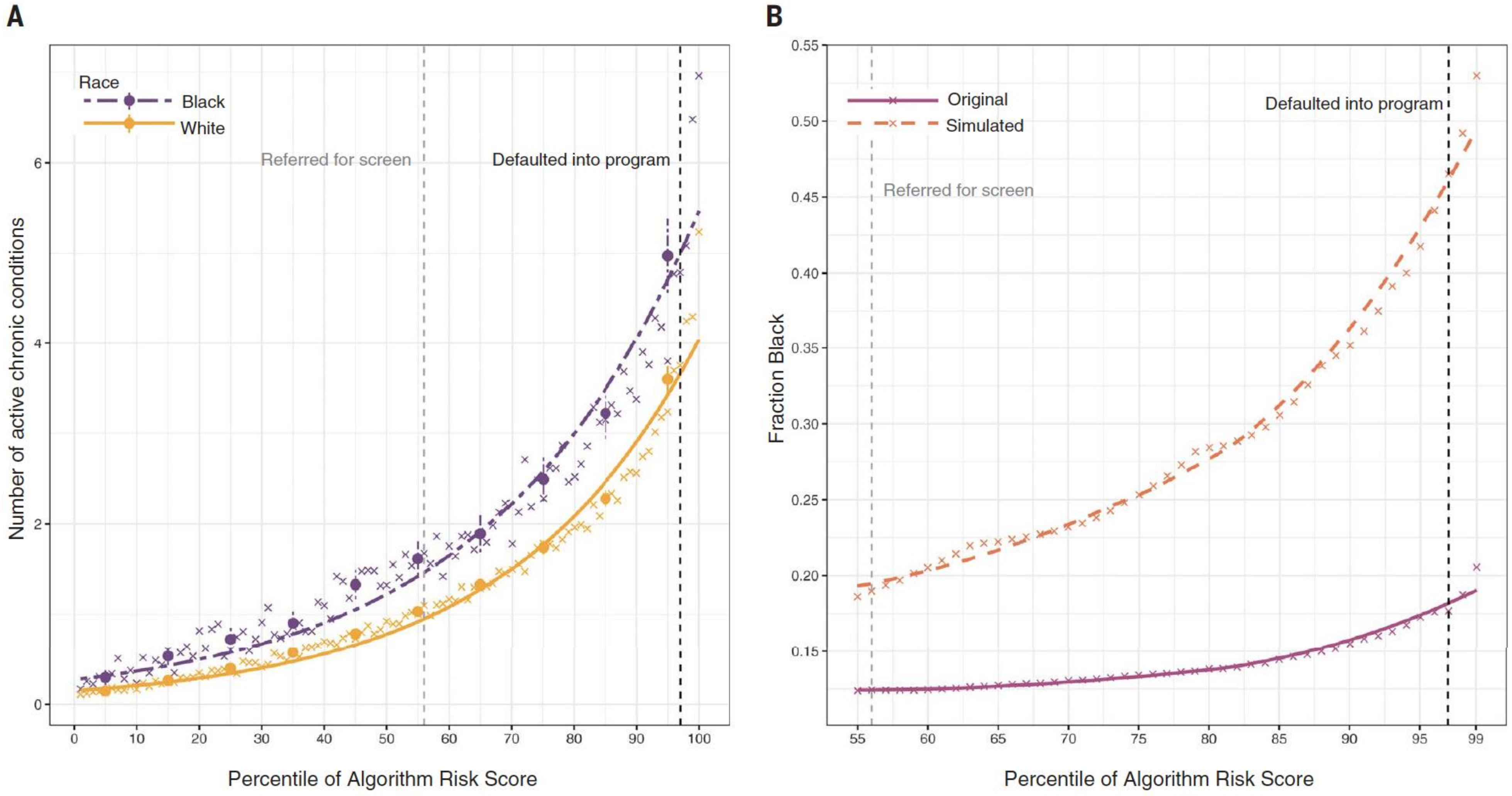


Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race. (**A**) Mean number of chronic conditions by race, plotted against algorithm risk score. (**B**) Fraction of Black patients at or above a given risk score for the original algorithm ("original") and for a simulated scenario that removes algorithmic bias ("simulated": at each threshold of risk, defined at a given percentile on the *x* axis, healthier Whites above the threshold are replaced with less healthy Blacks below the threshold, until the marginal patient is equally healthy). The × symbols show risk percentiles by race; circles show risk deciles with 95% confidence intervals clustered by patient. The dashed vertical lines show the auto-identification threshold (the black line, which denotes the 97th percentile) and the screening threshold (the gray line, which denotes the 55th percentile).

https://www.science.org/doi/10.1126/science.aax2342

Fixing bias in algorithms?

The New York Times

By Sendhil Mullainathan

ECONOMIC VIEW

Dec. 6, 2019

Biased Algorithms Are Easier to Fix Than Biased People

Racial discrimination by algorithms or by people is harmful — but that's where the similarities end.

Tim Cook

December 2019

In one study published 15 years ago, **two people applied for a job**. Their résumés were about as similar as two résumés can be. One person was named Jamal, the other Brendan.

In a study published this year, **two patients sought medical care**. Both were grappling with diabetes and high blood pressure. One patient was black, the other was white.

Both studies documented **racial injustice**: In the first, the applicant with a black-sounding name got fewer job interviews. In the second, the black patient received worse care.

But they differed in one crucial respect. In the first, hiring managers made biased decisions. In the second, the culprit was a computer program.

https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html

Fixing bias in algorithms?

The New York Times

By Sendhil Mullainathan

ECONOMIC VIEW

Dec. 6, 2019

Biased Algorithms Are Easier to Fix Than Biased People

Racial discrimination by algorithms or by people is harmful — but that's where the similarities end.

Tim Cook

Changing algorithms is easier than changing people: software on computers can be updated; the “wetware” in our brains has so far proven much less pliable.

[…] In a 2018 paper […], I took a cautiously optimistic perspective and argued that **with proper regulation, algorithms can help to reduce discrimination**.

But the key phrase here is “proper regulation,” which we do not currently have. We must ensure all the necessary inputs to the algorithm, including the data used to test and create it, are carefully stored. * […] We will need a well-funded regulatory agency with highly trained auditors to process this data.

https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html

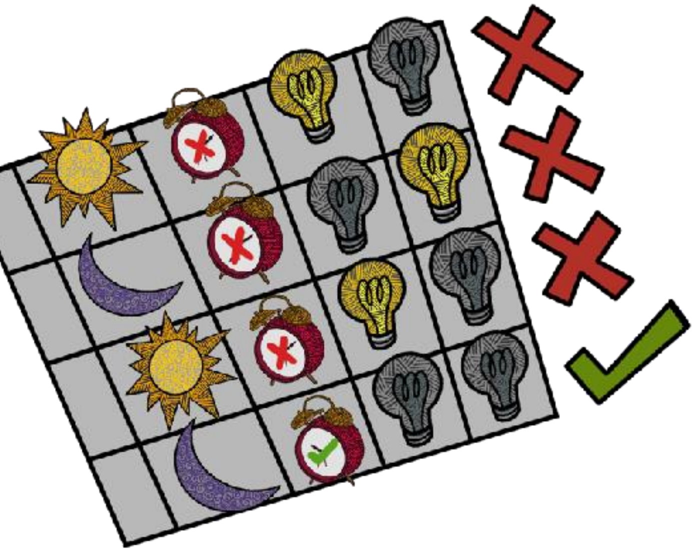
AI ethics teaser

This week's reading

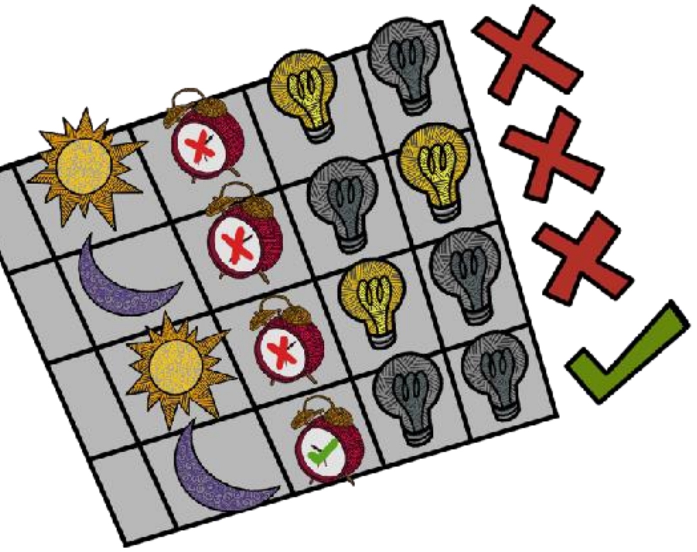
WE ARE AI
#3
Who lives. Who dies.
Who decides?
INEQUALITY
VALUES
UTILITARIANISM
UNCERTAINTY
ETHICS

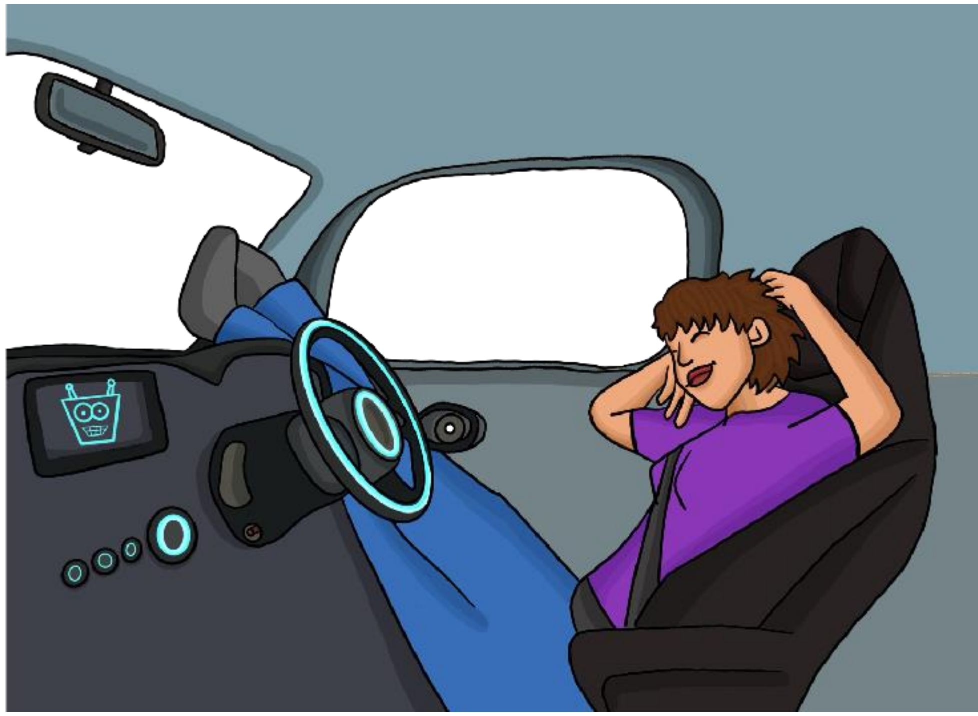
© Julia Stoyanovich, Mona Sloane and Falaah Arif Khan (2021)

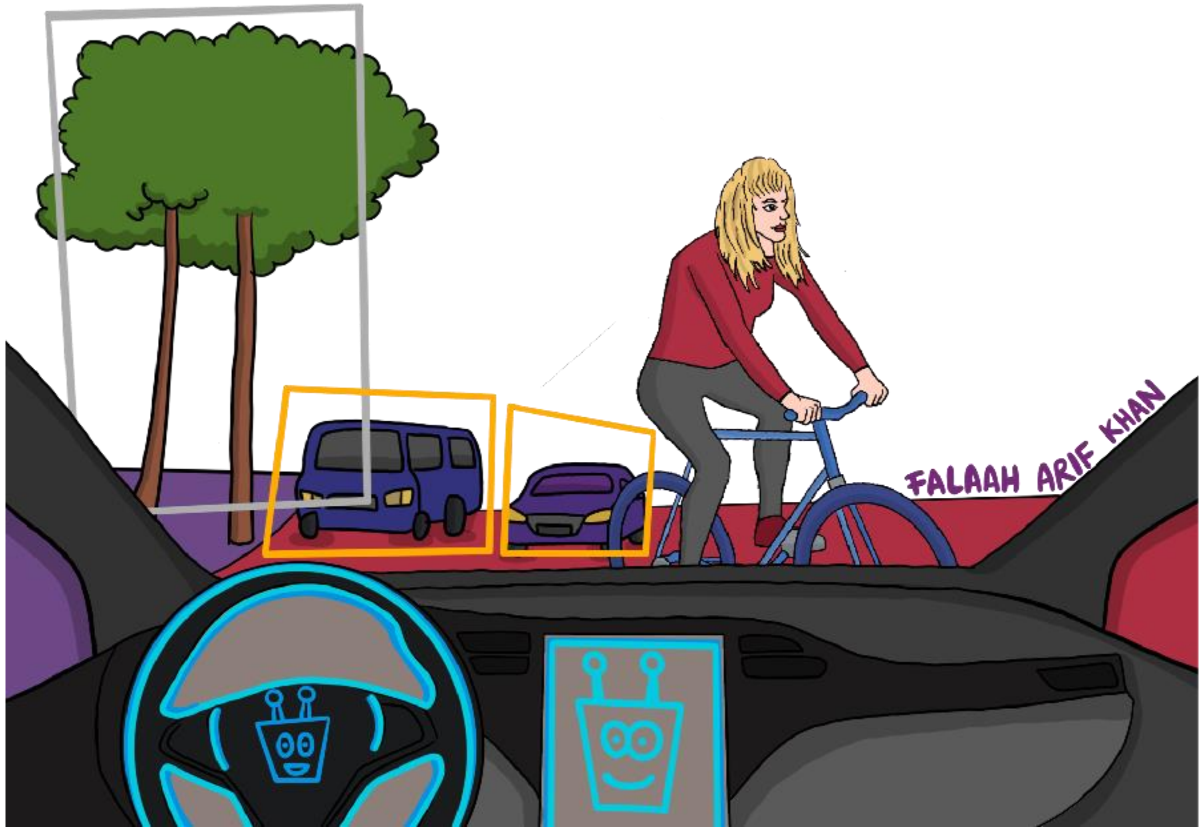
Mistakes lead to harms



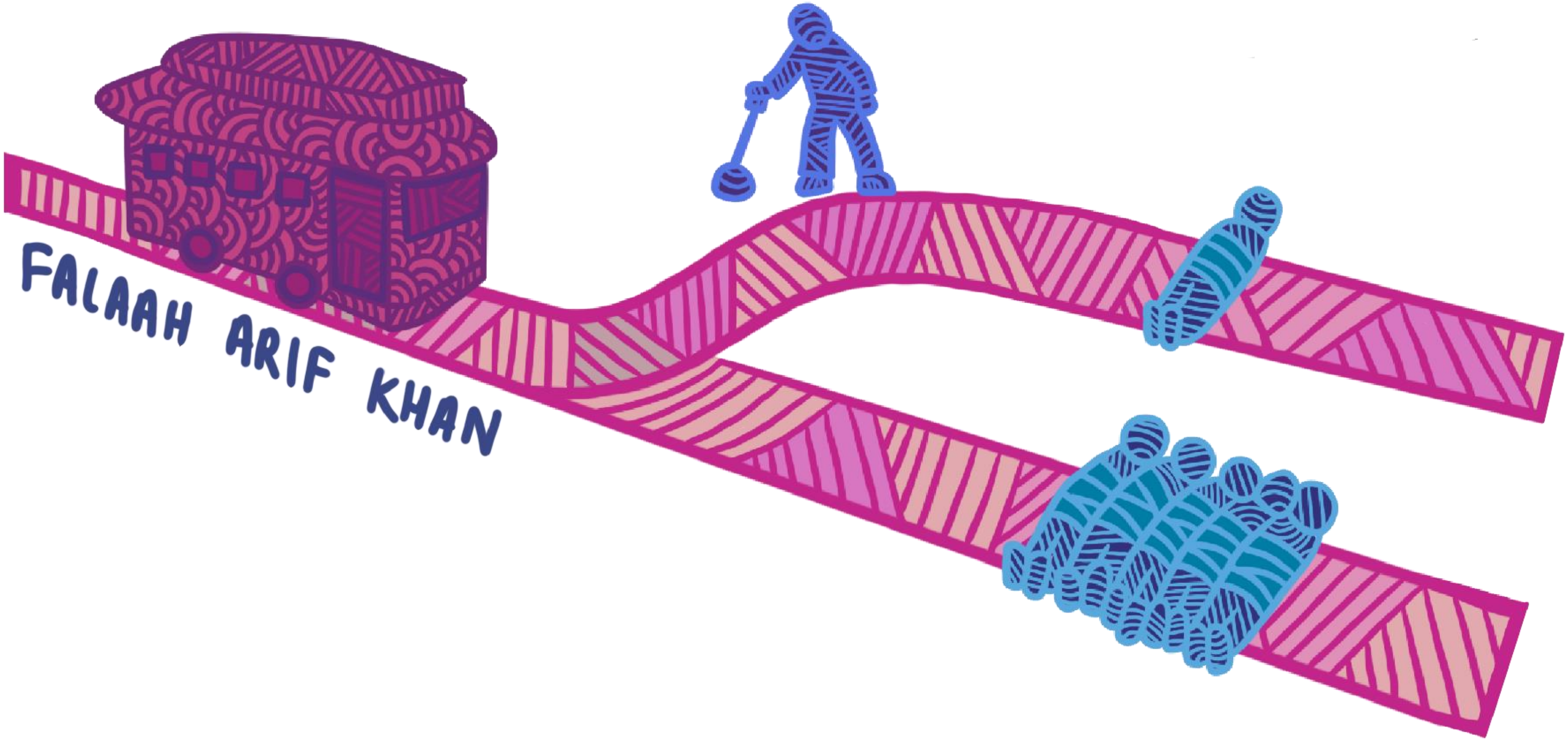
Mistakes lead to harms





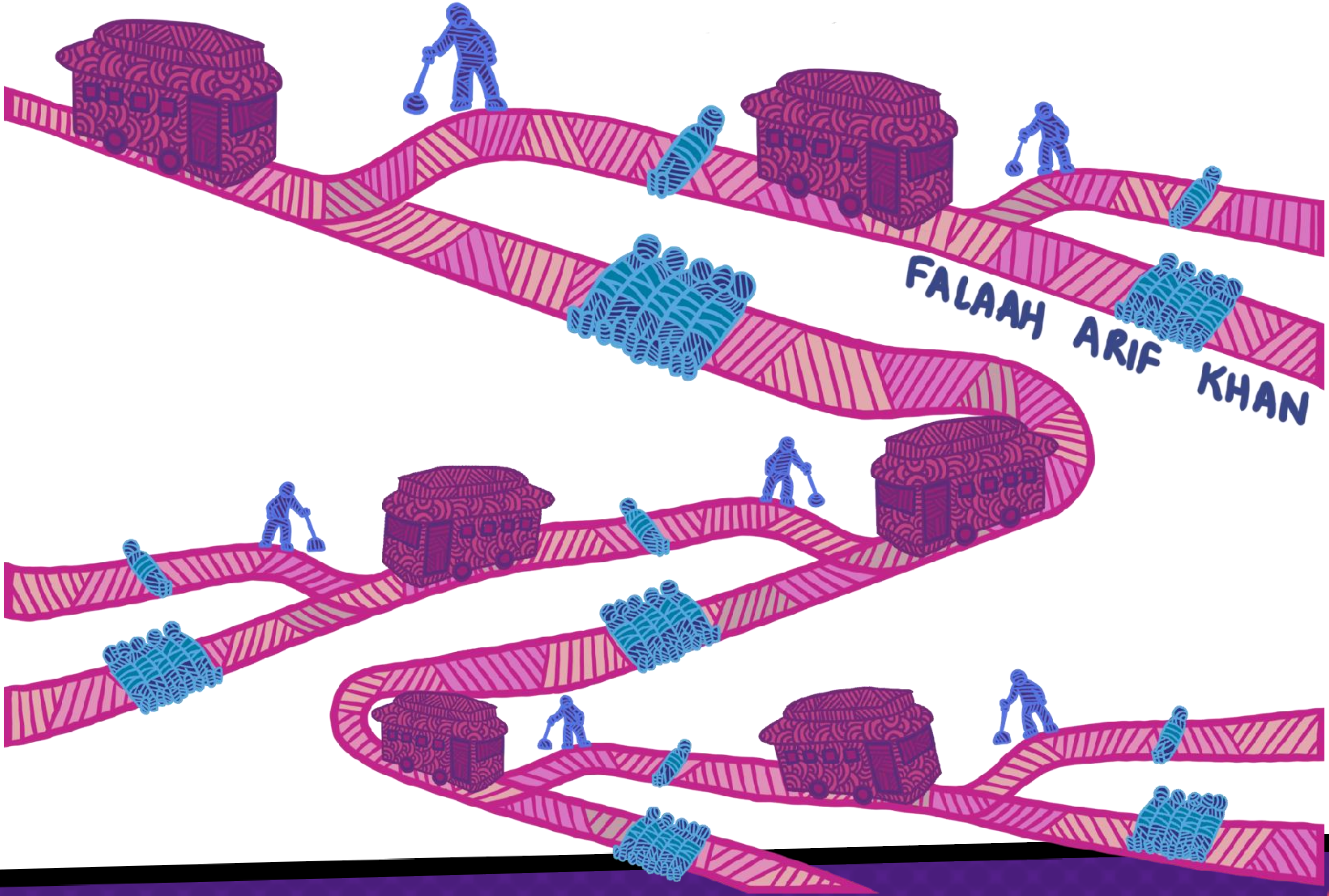


The trolley problem

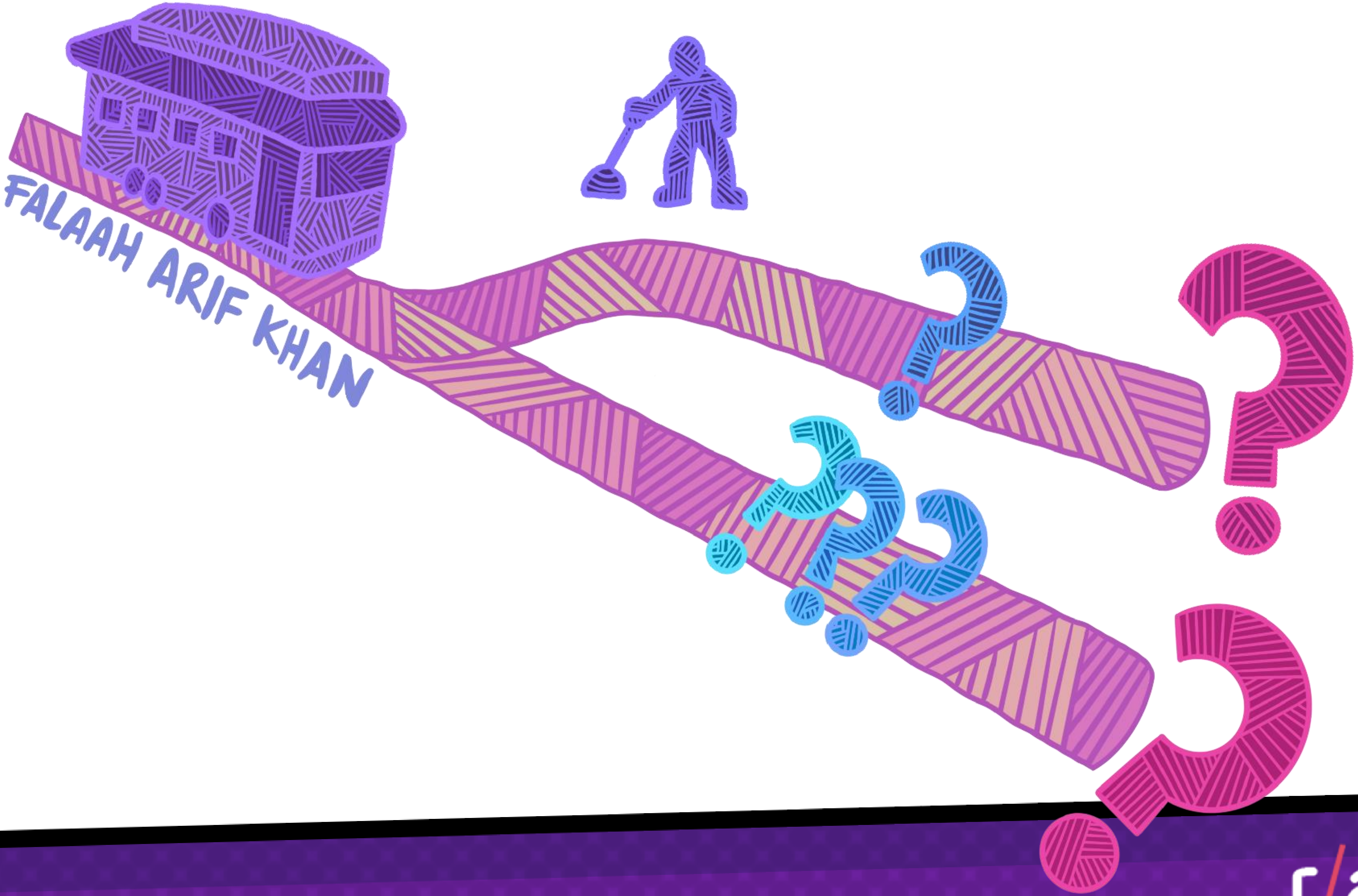


r/ai

The trolley problem



Dealing with uncertainty



Utilitarianism

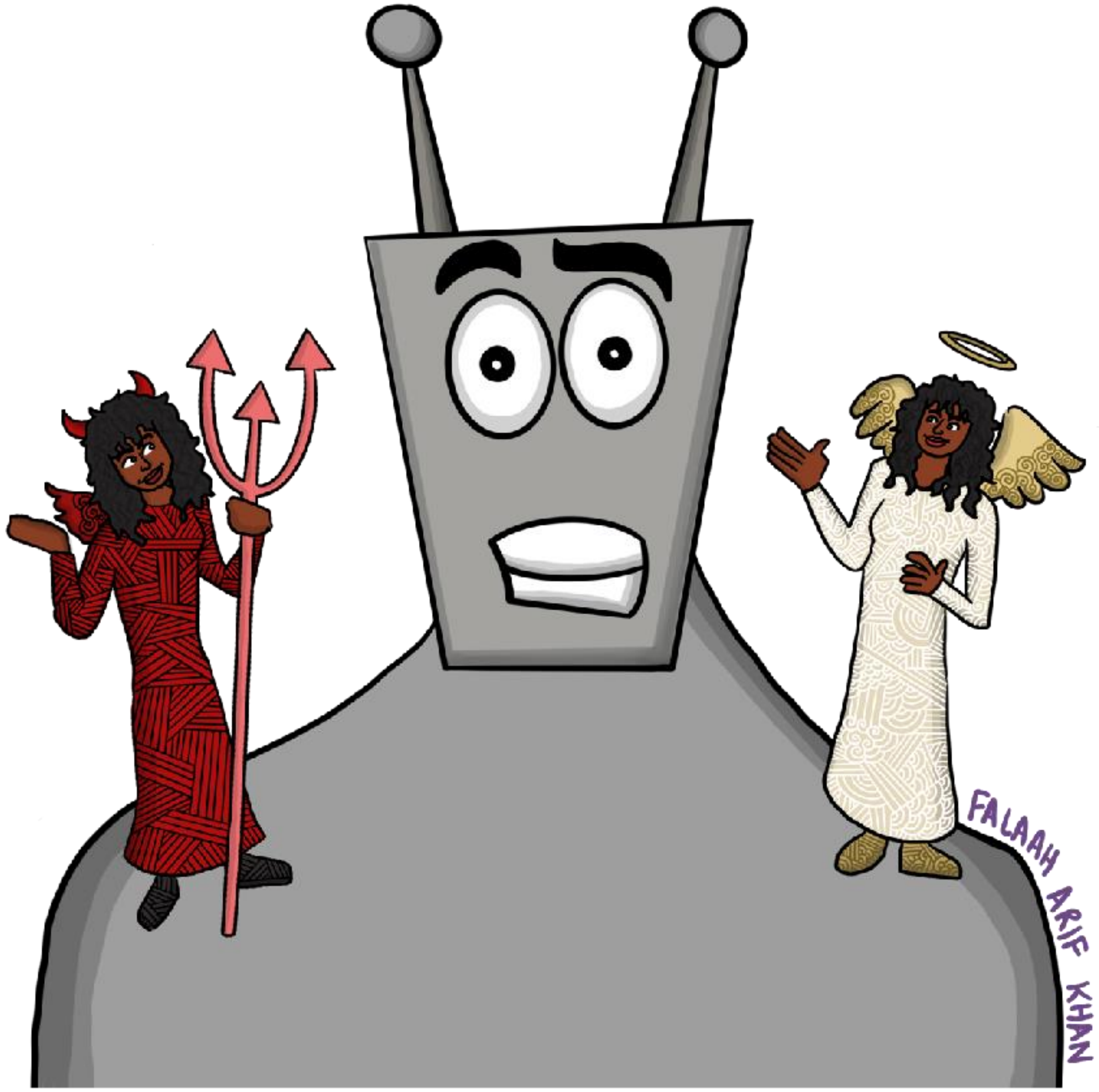
"It is the greatest happiness of the greatest number that is the measure of right and wrong."

Jeremy Bentham

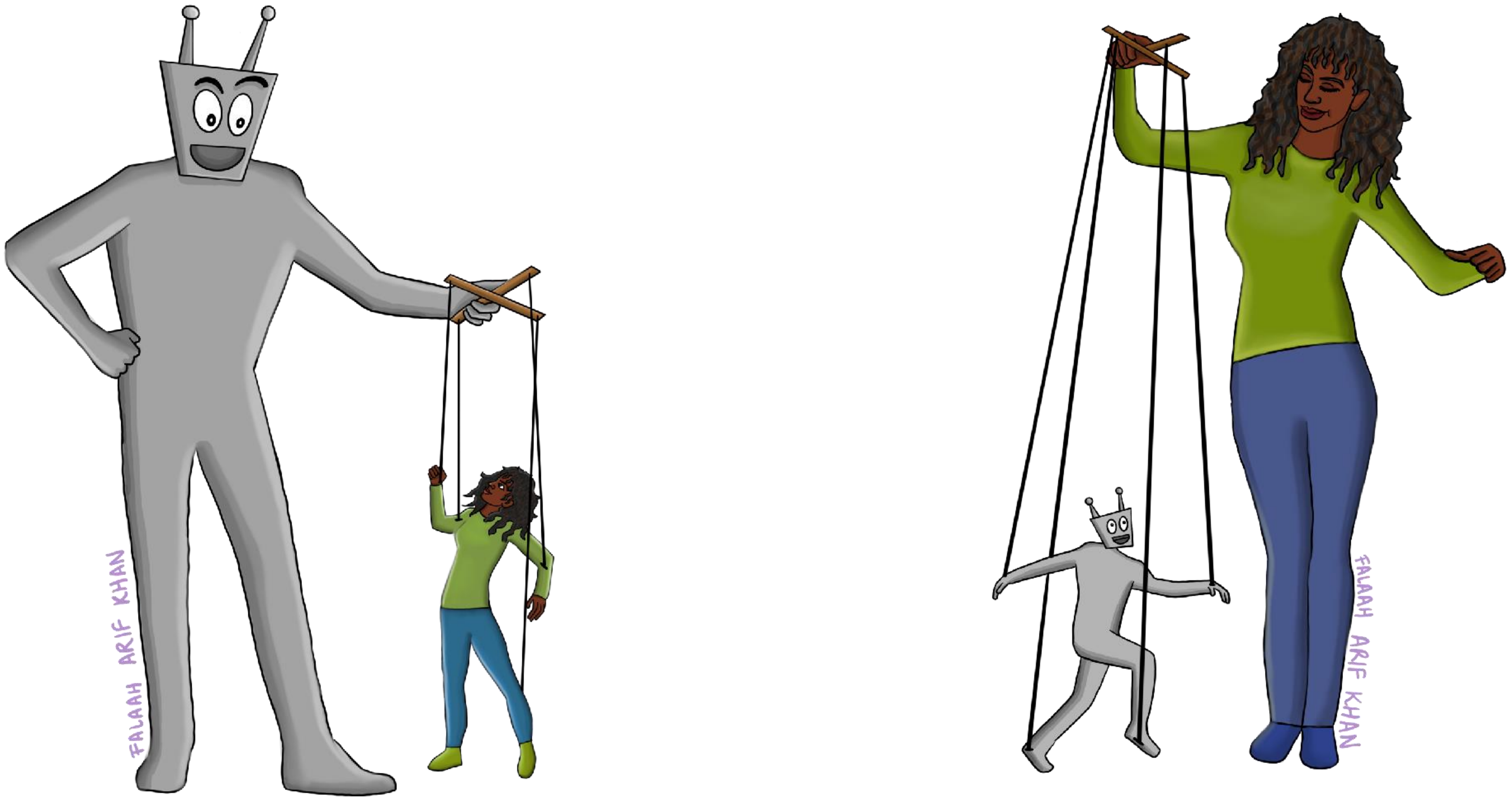
Algorithmic morality?

Algorithmic morality

is the act of attributing moral reasoning to algorithmic systems



Algorithmic morality?



Responsible Data Science

Algorithmic Fairness

Thank you!

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai