Responsible Data Science

Algorithmic Fairness

January 31 & February 2, 2022

Prof. George Wood
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**.

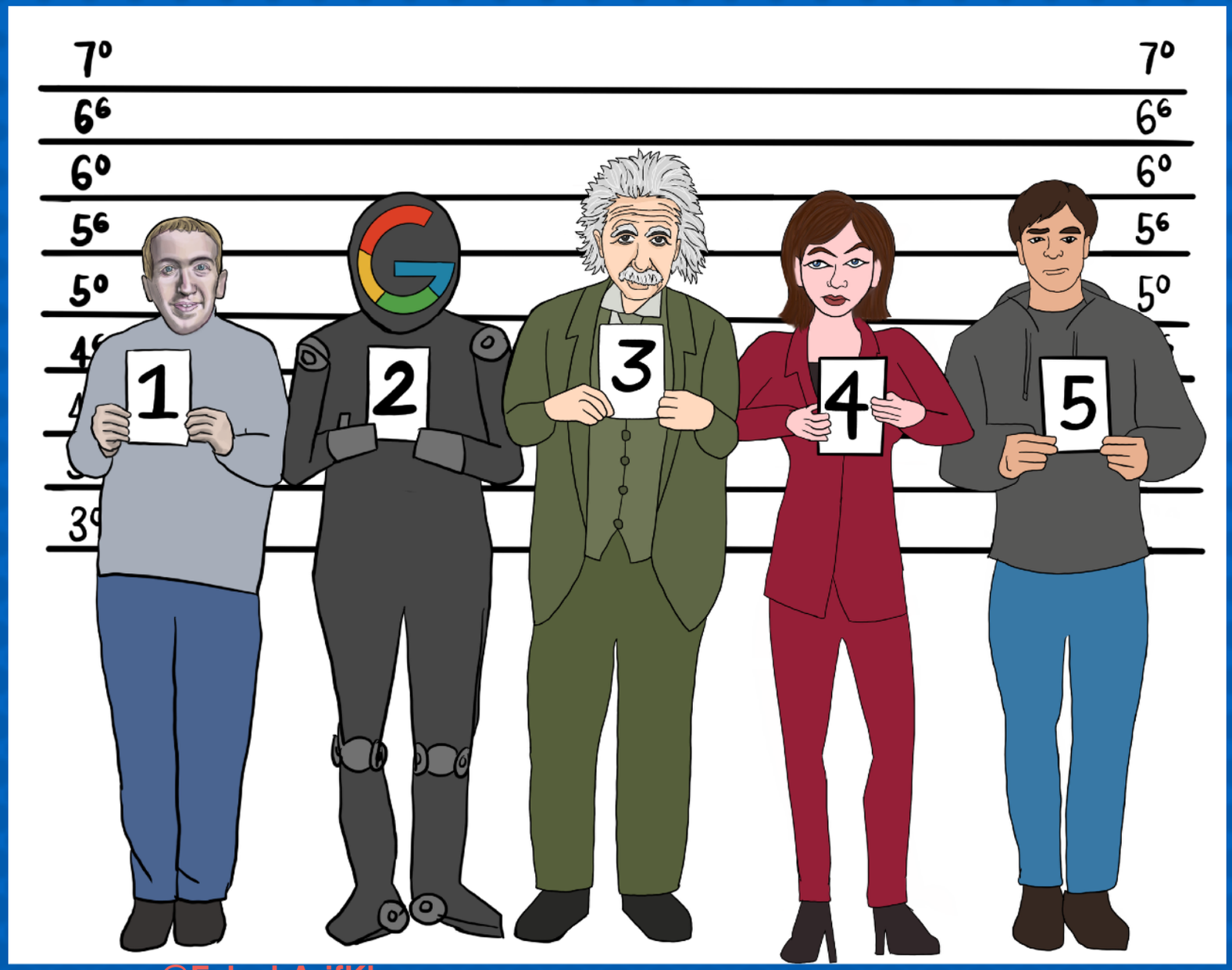
Bottom line: we will learn that many of the problems are **socio-technical**, and so cannot be “solved” with technology alone.

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

Nuance, please!

r/ai

We all are responsible



@FalaahArifKhan

Reading: Algorithmic bias

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)]

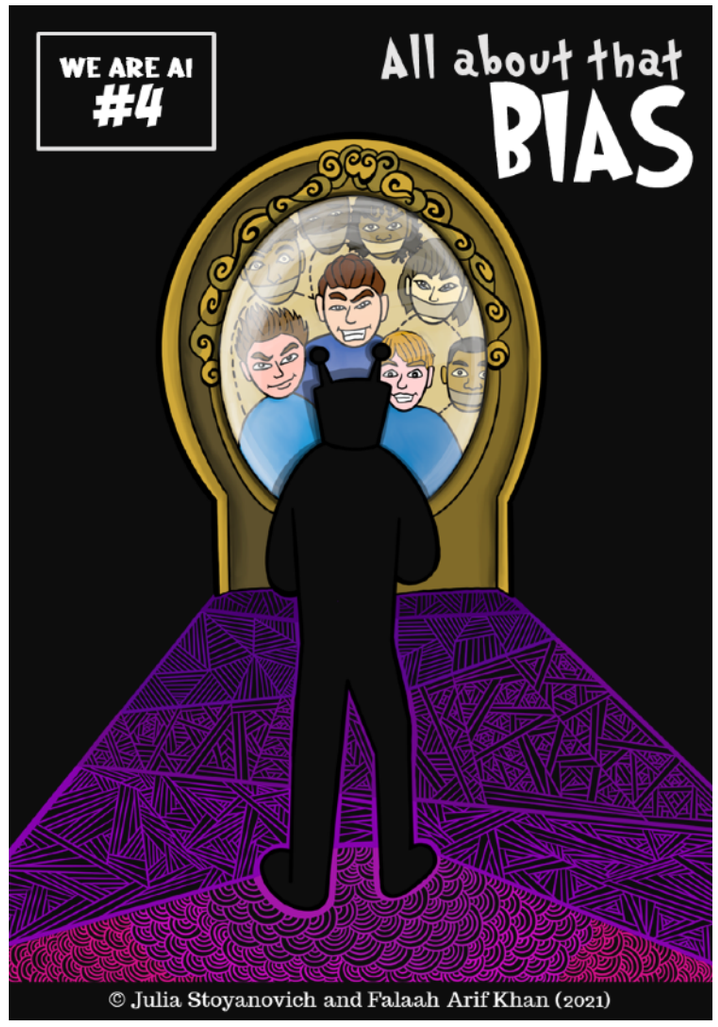
DOI:10.1145/3376898

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)]



r/ai

Reading: Fairness in risk assessment

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 learn-

[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

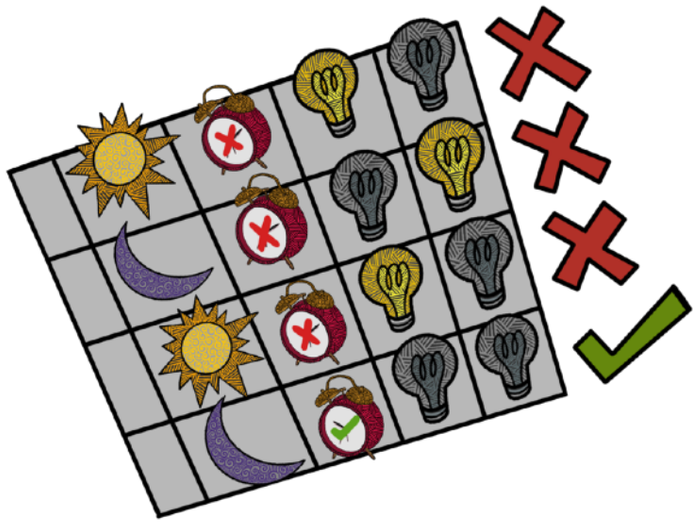
Keywords and phrases algorithmic fairness, risk tools, calibration

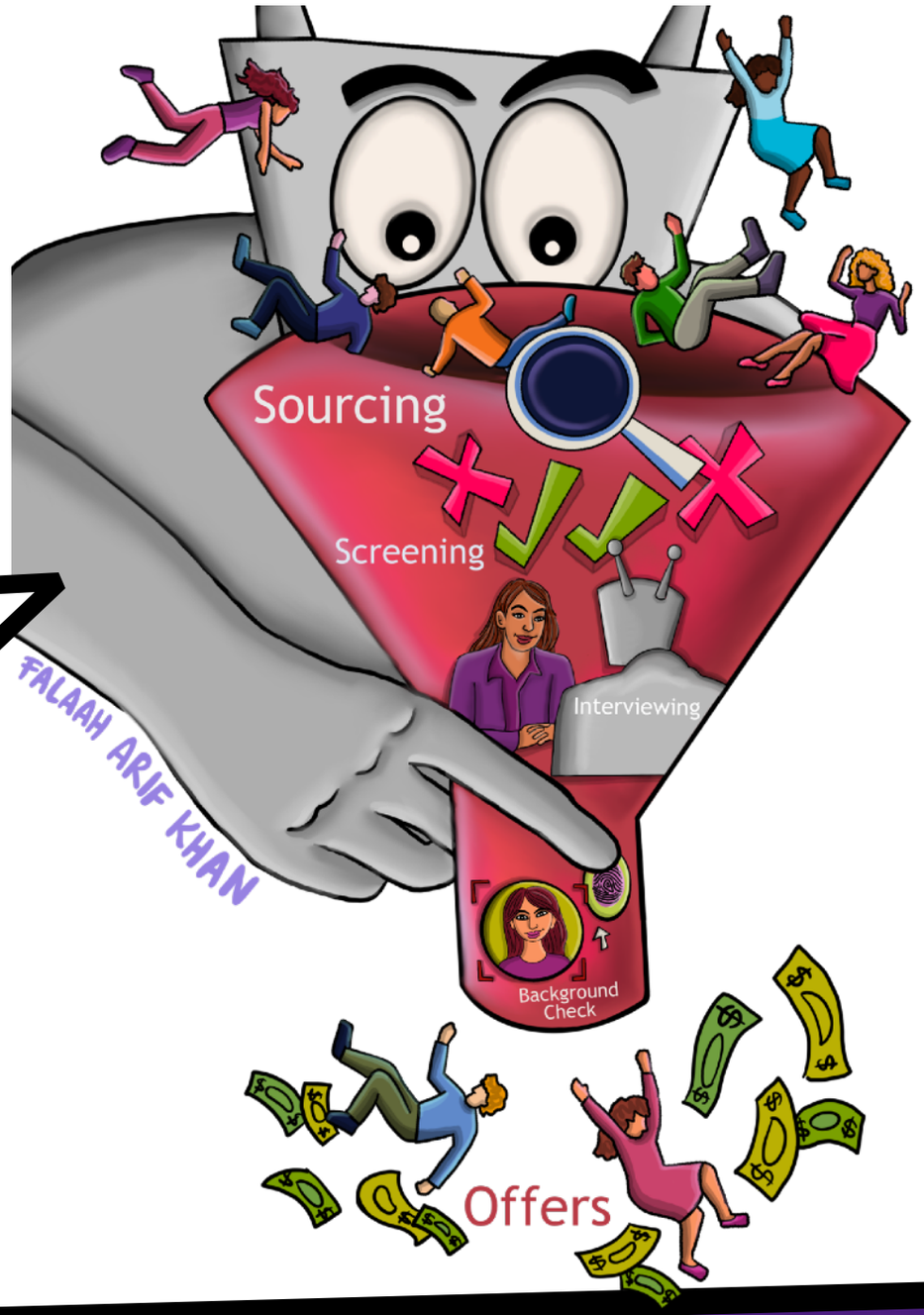
Digital Object Identifier 10.4230/LIPIcs.ITCS.2017.43

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

r/ai

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

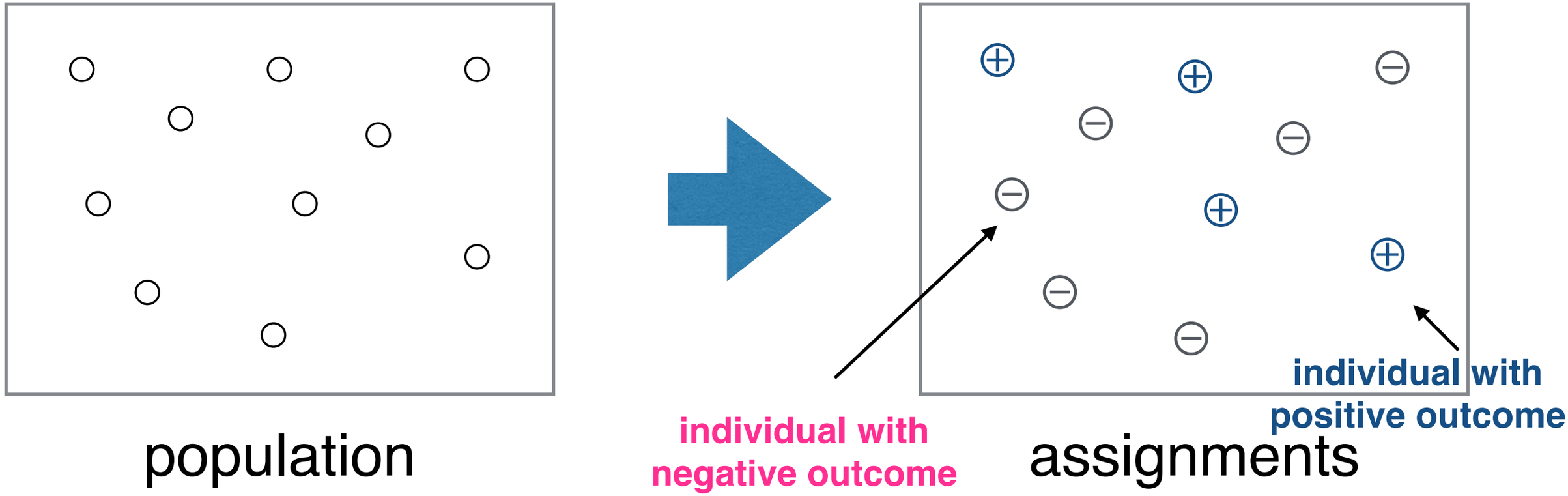
r/ai

Fairness in classification

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

positive outcomes

40% of the population

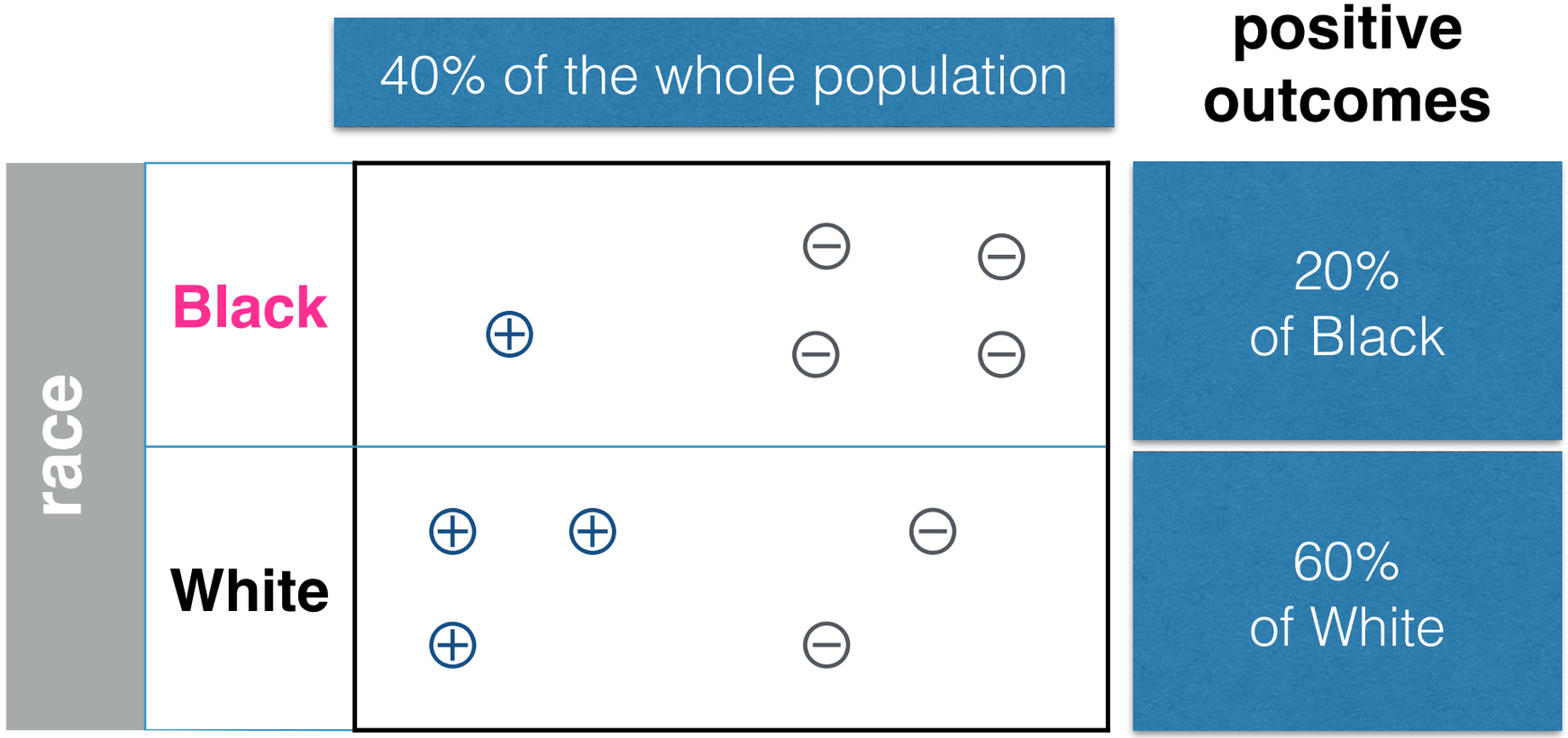


population

assignments

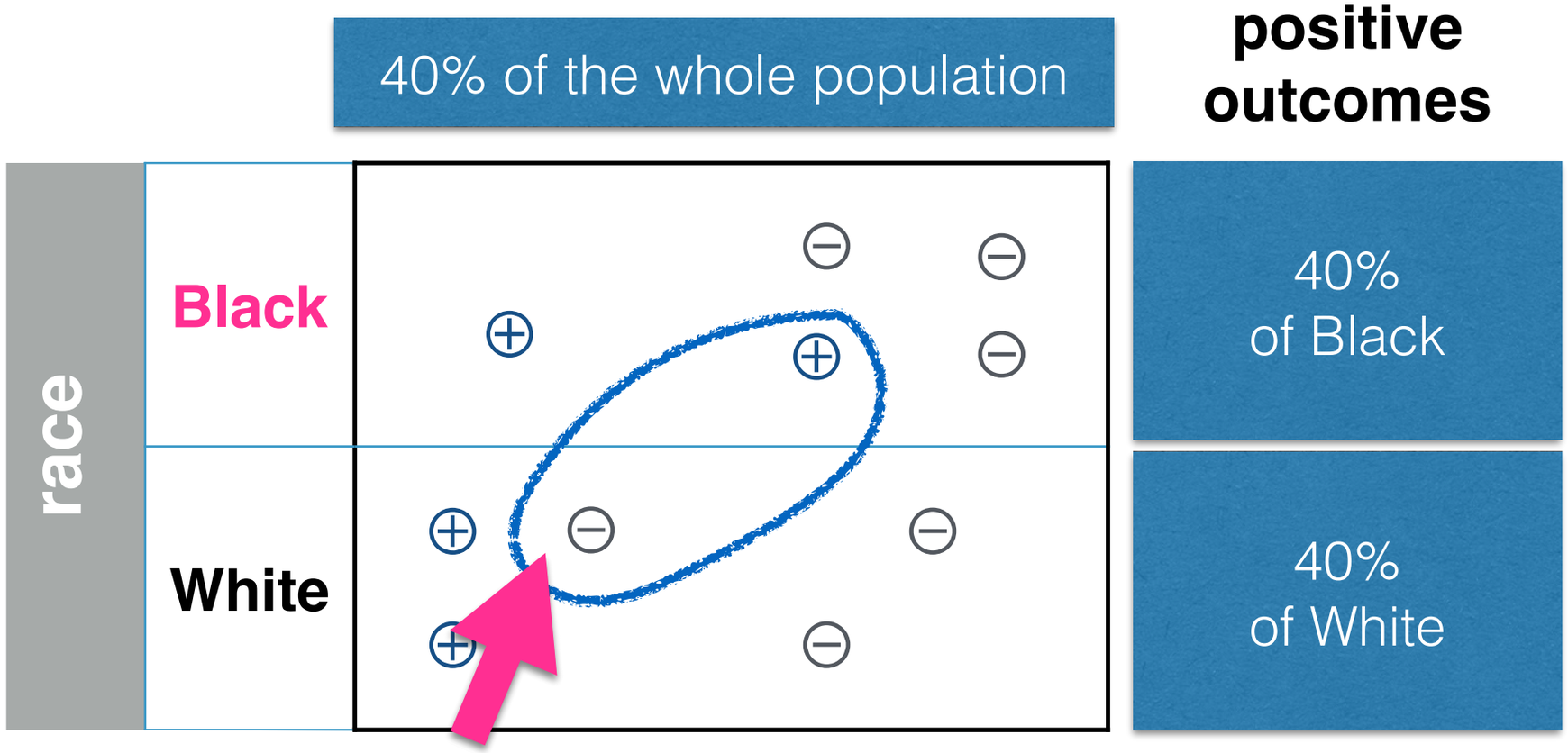
Fairness in classification

Sub-populations may be treated differently



Fairness in classification

Sub-populations may be treated differently



Fairness in classification

Explaining the disparity with proxy variables

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

positive outcomes

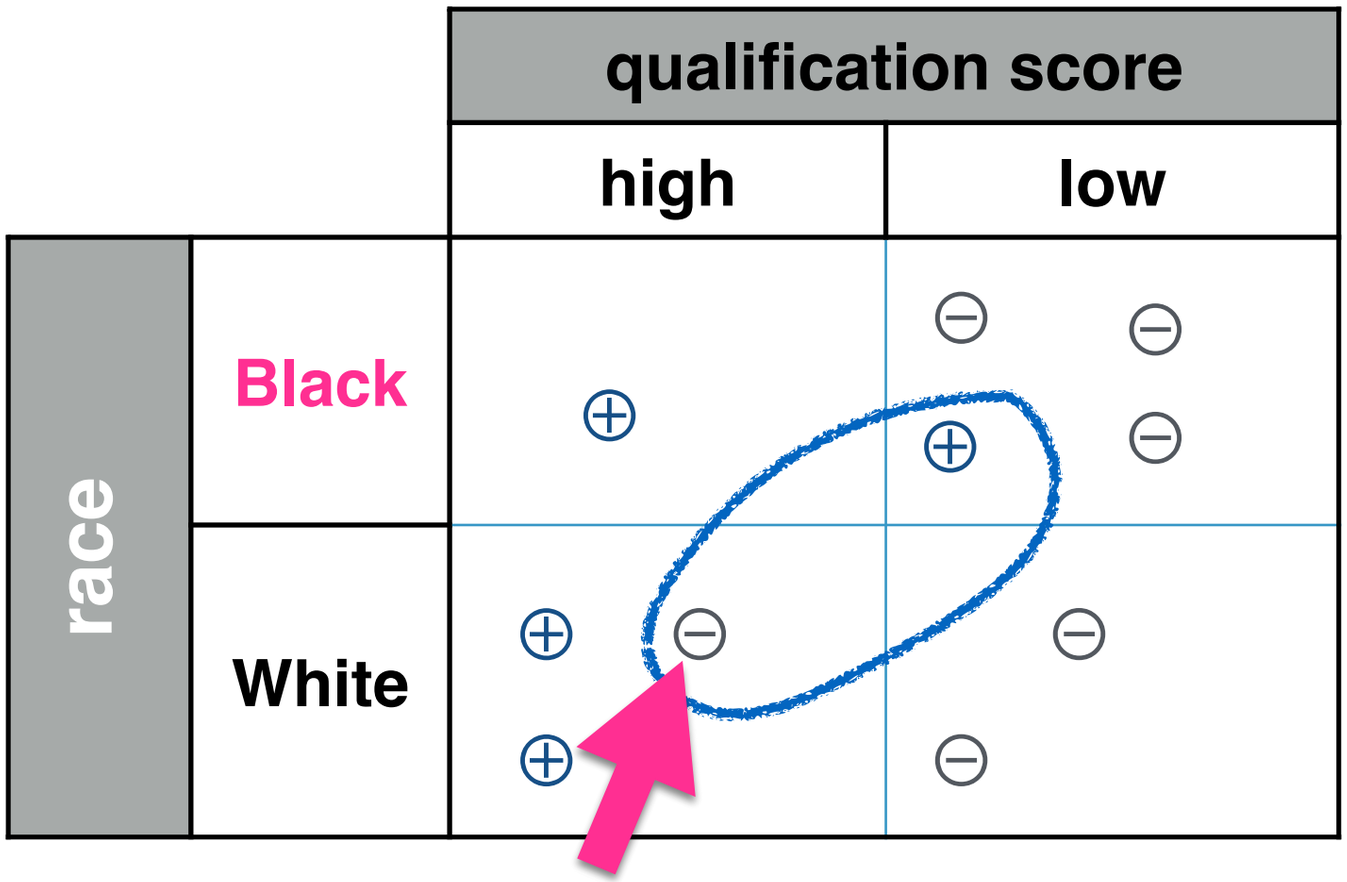
20% of Black

60% of White

discussion

r/ai

Swapping outcomes



positive outcomes

40%
of Black

40%
of White

r/ai

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

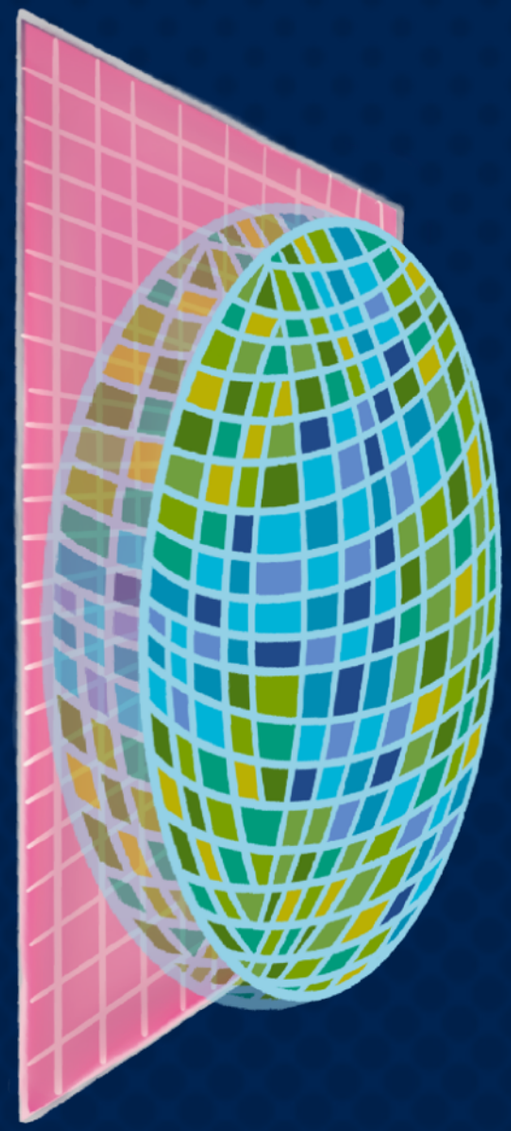
Technical is introduced or exacerbated by the technical properties of an ADS

Emergent arises due to context of use

[Friedman & Nissenbaum (1996)]

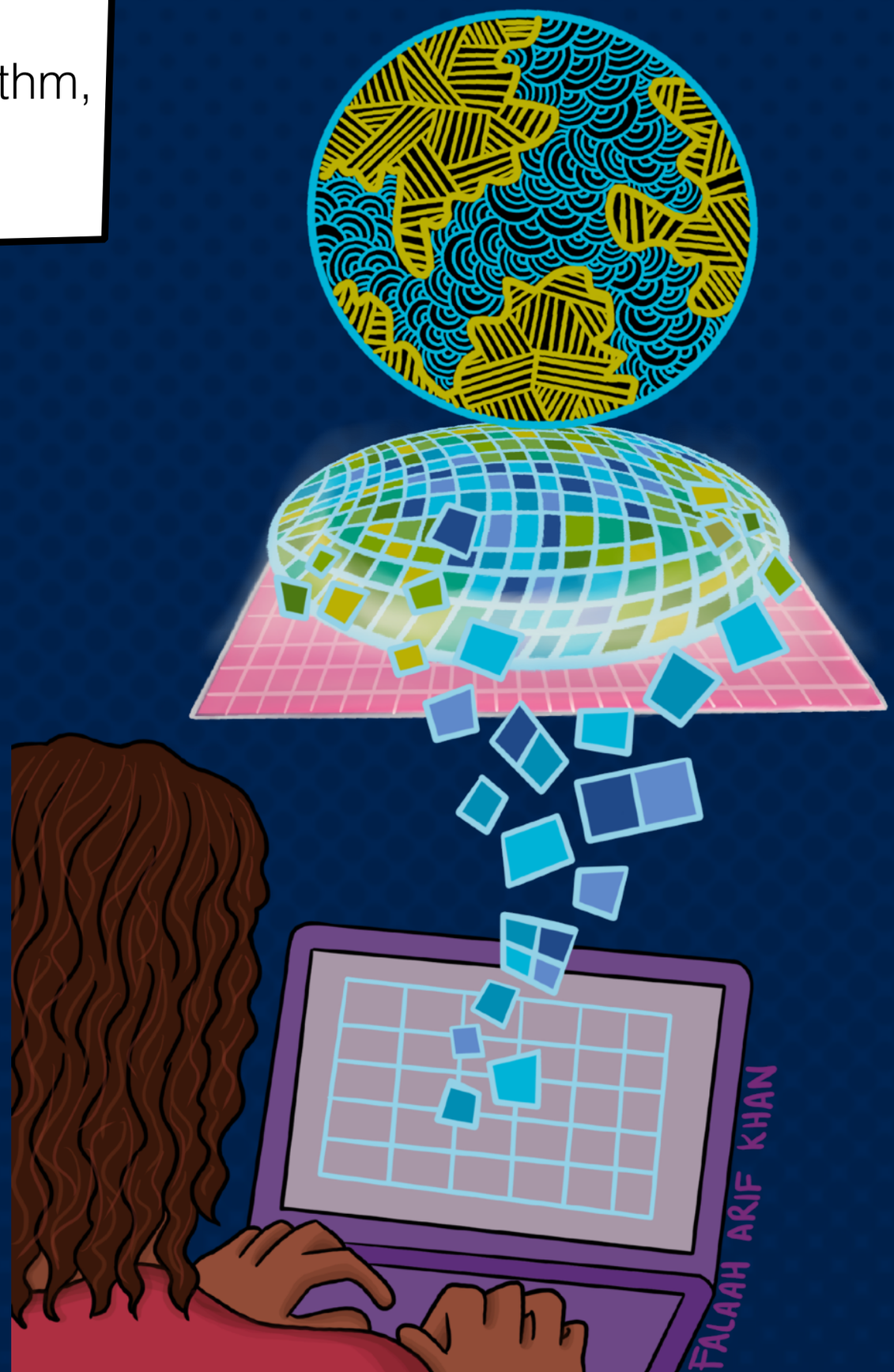
r/ai

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

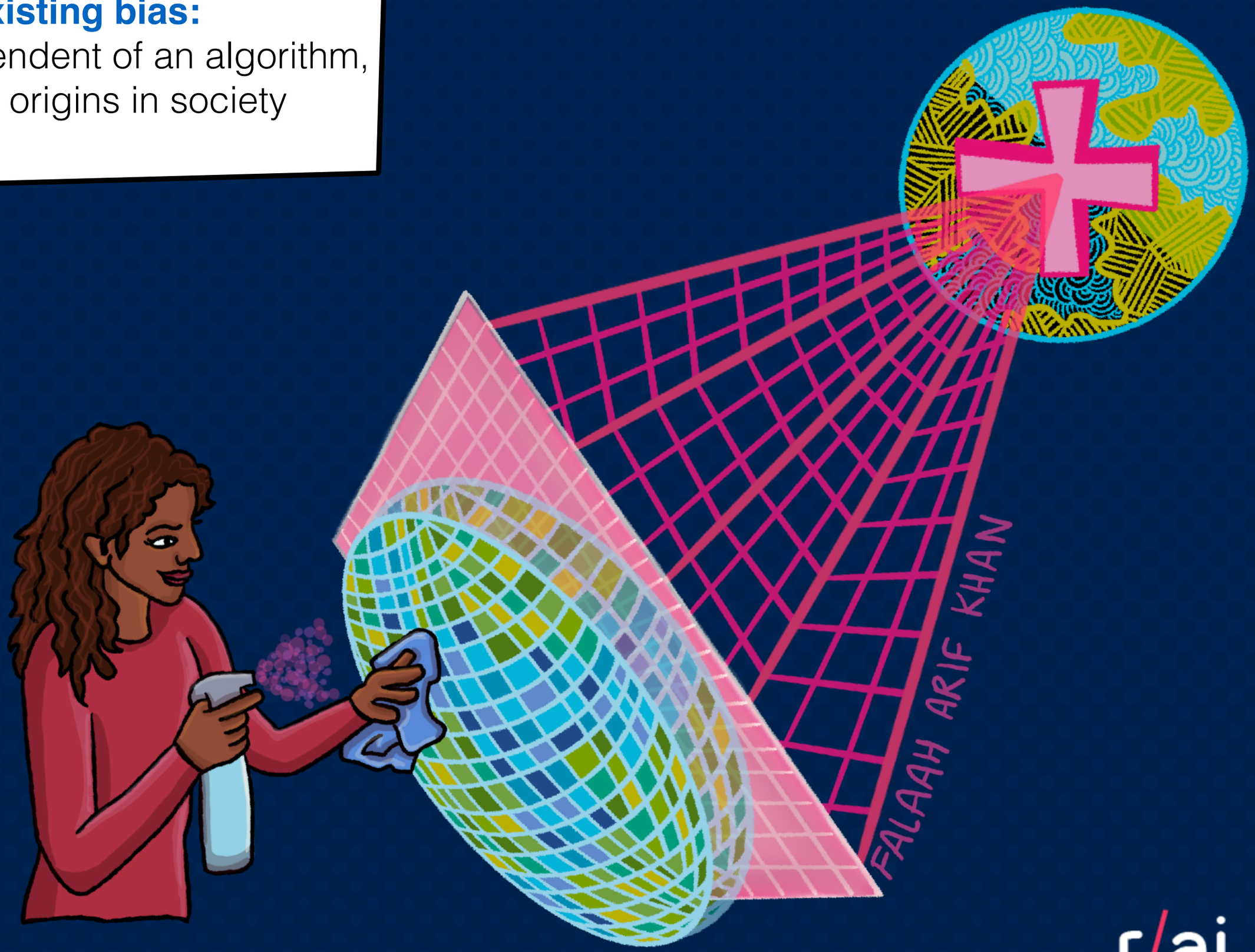


r/ai

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

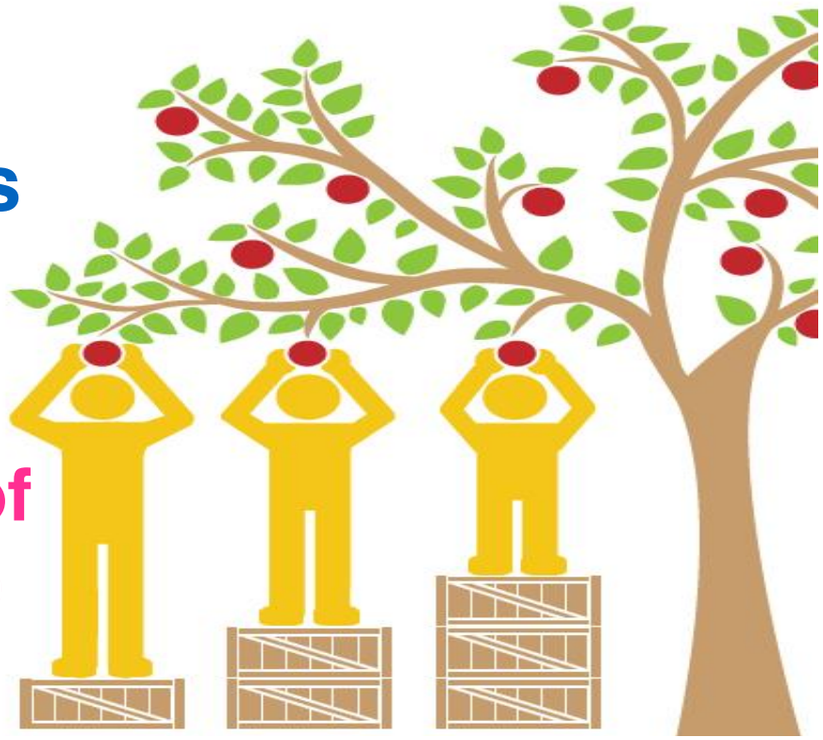
Karen Lee Torre, left, a lawyer who represented the New Haven firefighters in their lawsuit, with her clients Monday at the federal courthouse in New Haven. Christopher Capozziello for The New York Times

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

group fairness

equality of outcome



individual fairness

equality of treatment

r/ai