An Introduction to Fairness and Bias in Machine Learning

SCQ Summer School
July 24th-28th, 2023

Anna Sapienza, PhD
Senior Researcher - ISI Foundation, Italy
Assistant Professor - SODAS, University of Copenhagen

Contacts:
ansa@sodas.ku.dk
anna.sapienza@isi.it
www.annasapienza.it

An Introduction to Fairness and Bias in Machine Learning

SCQ Summer School
July 24th-28th, 2023

DTU

Germans Savcisens
PhD student in Computational Social Science
Technical University of Denmark

Algorithmic Fairness, Accountability and Ethics
Lecturer at IT University of Copenhagen

AI and ML are everywhere

Personal Assistants

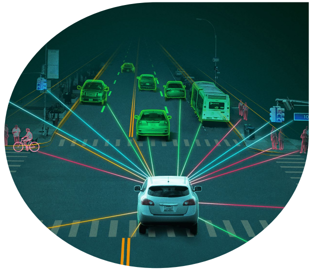
Social Media

YouTube

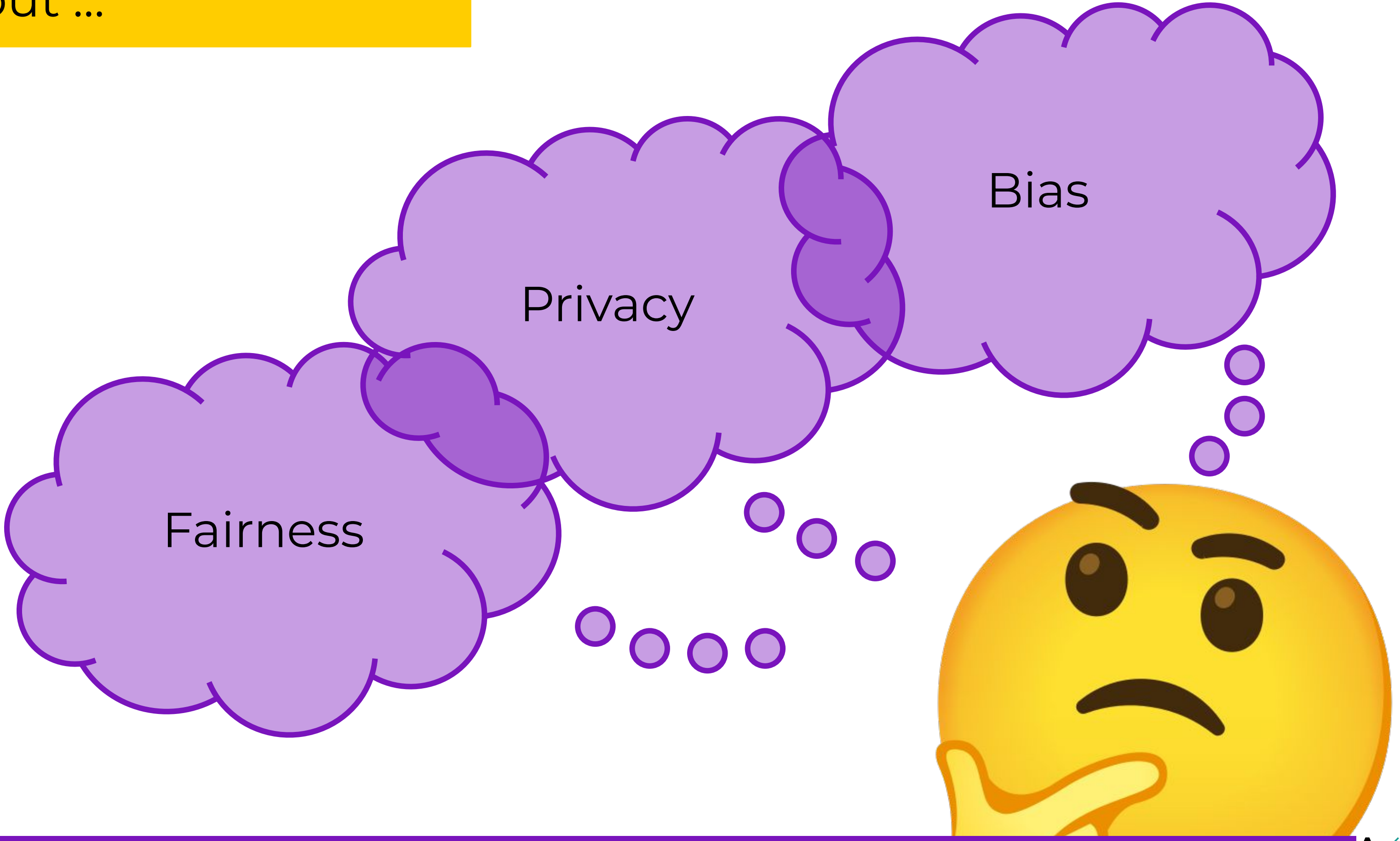
TikTok

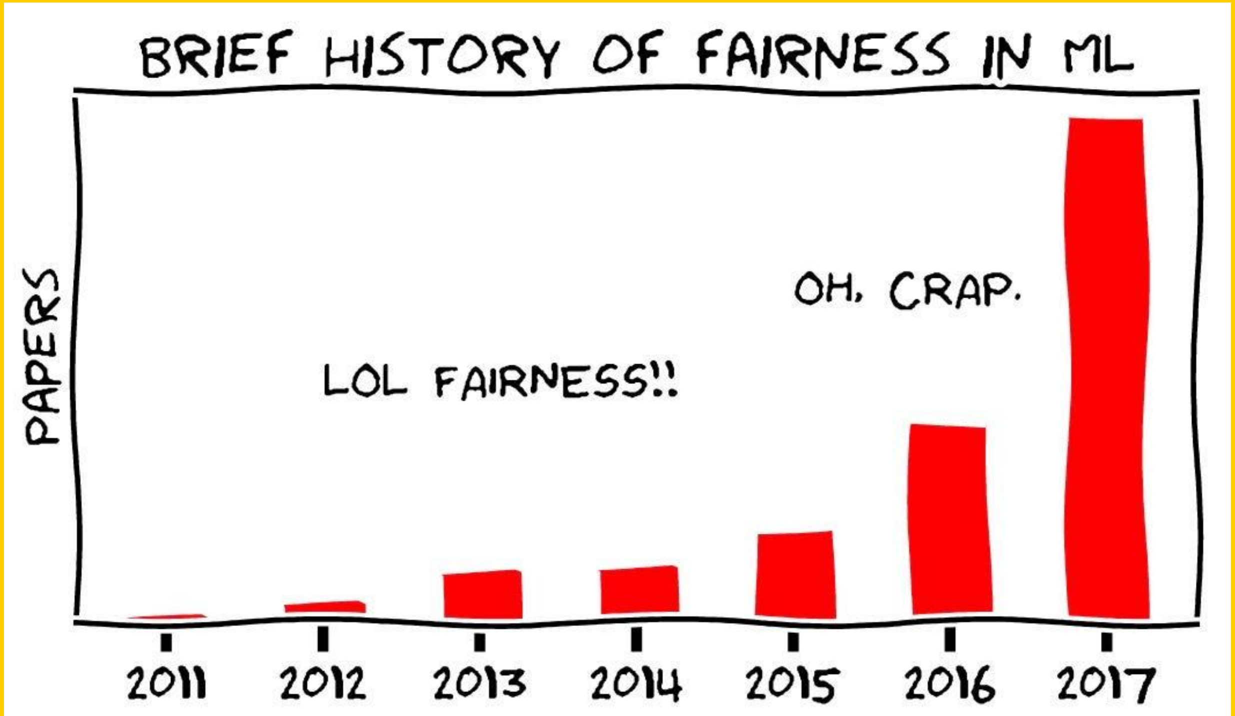
Meta

Other



But, what about …

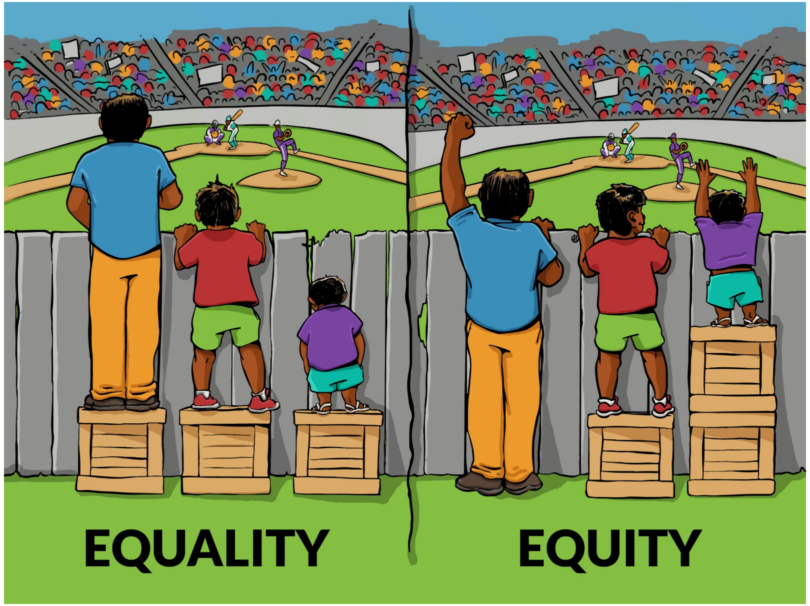




By Moritz Hardt

Disclaimer

- There are **many definitions** of fairness
- There is **no free lunch**
  - Fairness can **decrease accuracy**
  - Fairness definitions are **often incompatible**
- Fairness can be **achieved in different ways**



How would you define fairness?

Take 3 minutes to discuss with your group

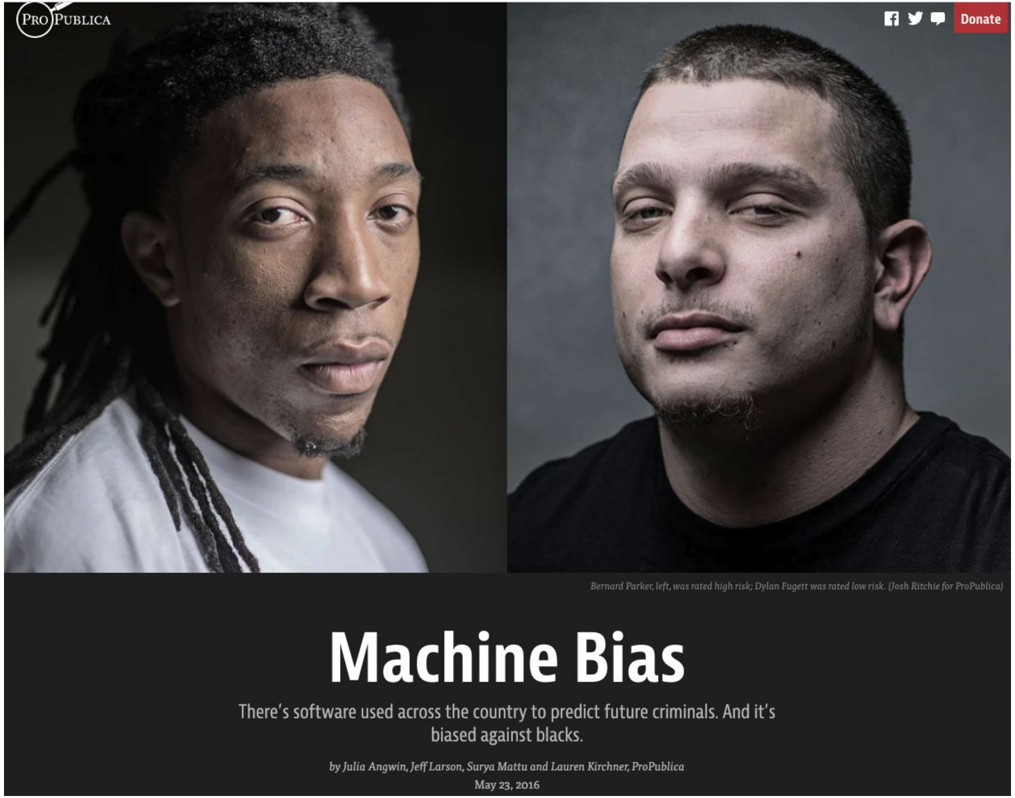
What is algorithmic fairness?

In the context of decision-making, fairness is the *absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics.* Thus, an unfair algorithm is one whose decisions are skewed toward a particular group of people.

A Survey on Bias and Fairness in Machine Learning

NINAREH MEHRABI, FRED MORSTATTER, NRIPSUTA SAXENA, KRISTINA LERMAN, and ARAM GALSTYAN, USC-ISI

Impact of algorithms

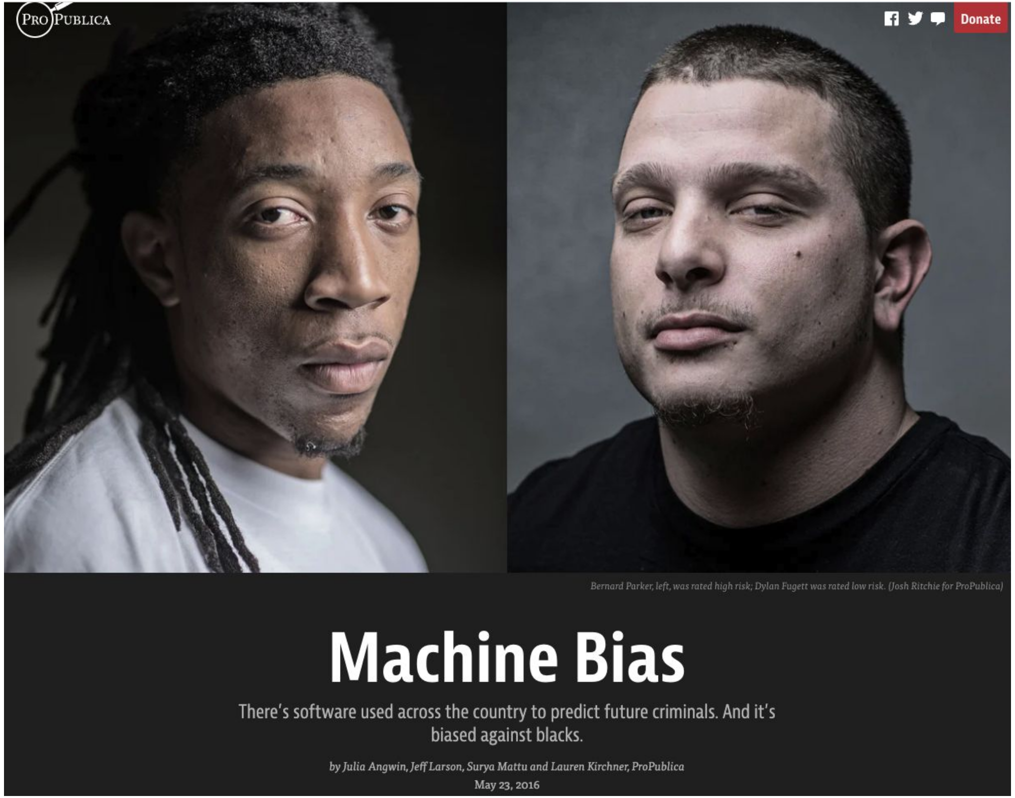


COMPAS

(Correctional Offender Management Profiling for Alternative Sanctions)

a popular commercial algorithm used by judges and parole officers for scoring criminal defendant's likelihood of reoffending (recidivism).

ProPublica Study



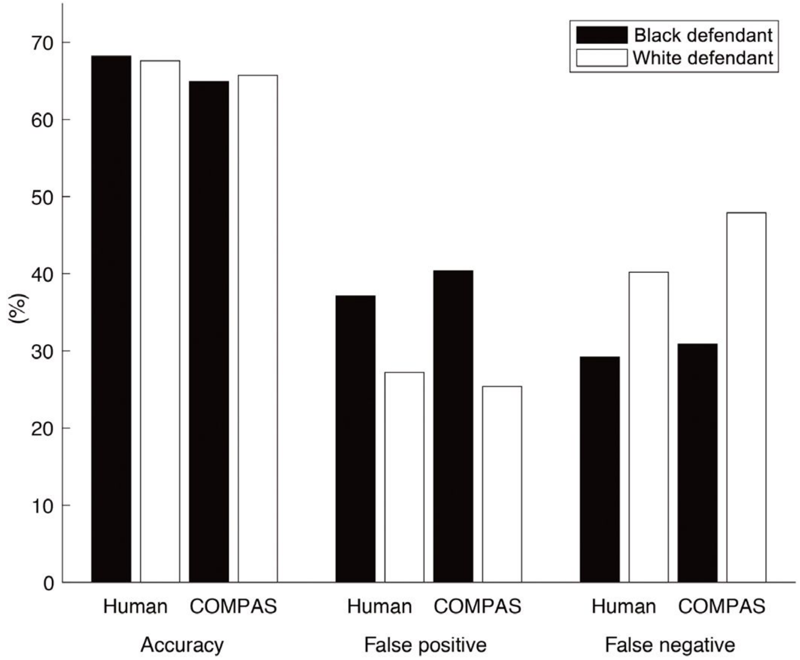
Key take away

COMPAS was found to be **biased against African-Americans:** it falsely predicts them to be at a **higher risk** of recommitting a crime or recidivism.

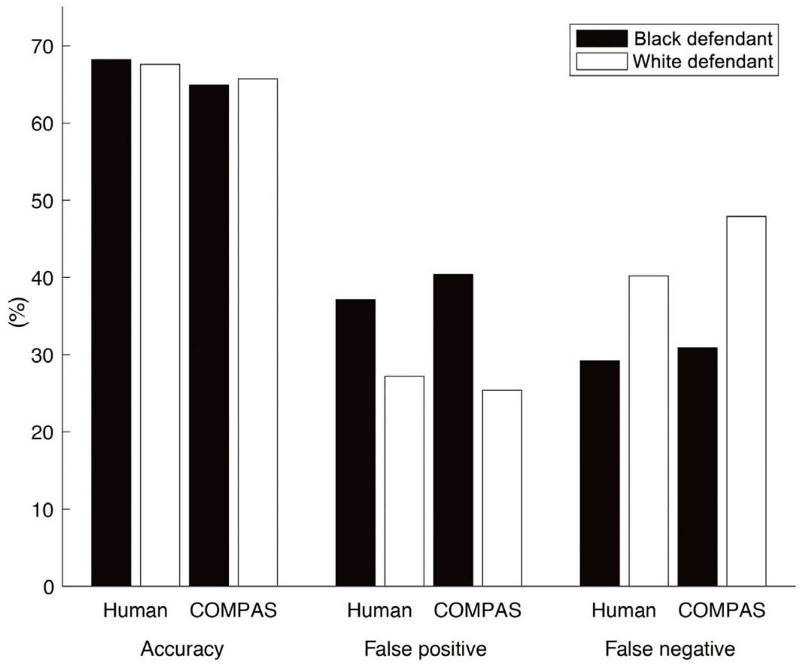
ProPublica: How we analyzed the COMPAS recidivism algorithm

MIT SERC: The dangers of risk prediction in the criminal justice system

COMPAS performance



COMPAS performance

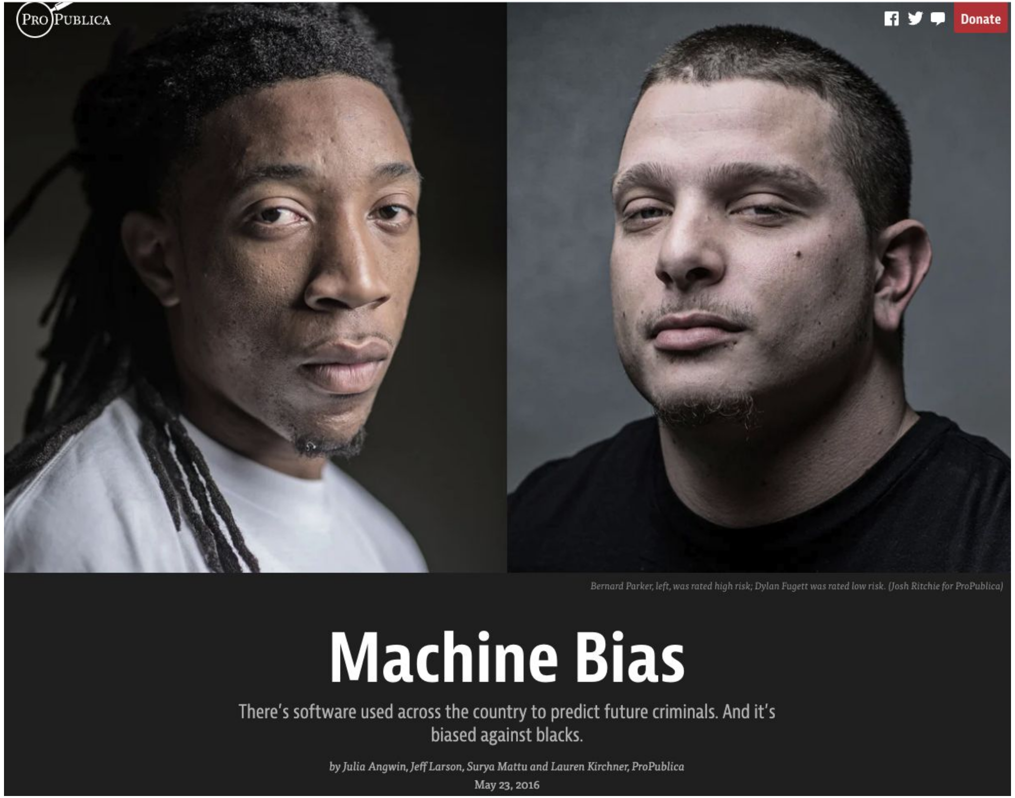


When considering using software such as COMPAS in making decisions that will significantly affect the lives and well-being of criminal defendants, it is valuable to ask whether we would put these decisions in the hands of random people who respond to an online survey because, in the end, the results from these two approaches appear to be indistinguishable.

The accuracy, fairness, and limits of predicting recidivism

Julia Dressel and Hany Farid*

Impact of algorithms



Failed due to biases …

… but what is bias?

What is bias?

Different definitions proposed

Some concepts can be vague

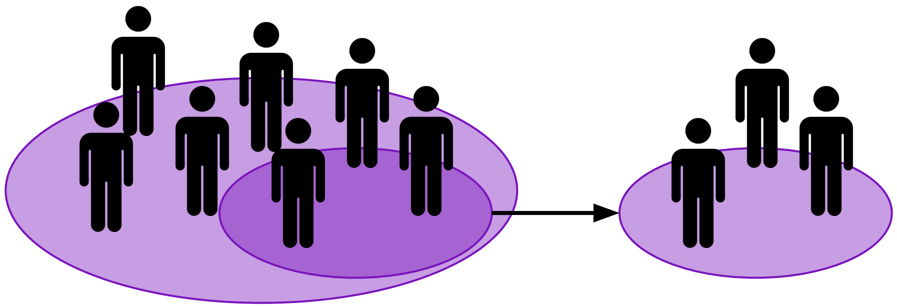
What is bias?

Defining bias in statistics

Statistical bias is anything that leads to a systematic difference between the **true parameters** of a population and the **statistics used to estimate** those parameters.

Population

Sample



Variable X*

Measurement / Estimation X

The measurement X is biased if $E[X^*] \neq E[X]$

What is bias?

Defining bias in sociology

A **tendency** (either known or unknown) to prefer a thing over another that **prevents objectivity** and influences understanding or outcomes in some way

Examples of Bias

> - A bias towards respecting male teachers more than female teachers.
> - Judging a group negatively because of their ethnicity.
> - Not accounting for students with disabilities when designing a test.
> - Framing a question on a survey to ensure a desired response.

What is bias?

Defining bias in Machine Learning and AI

There is no exact definition

What is bias?

Defining bias in Machine Learning and AI

The term bias is used to characterize the process leading to **prediction issues** and **possible unfairness**

What is algorithmic fairness?

In the context of decision-making, fairness is the *absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics.* Thus, an unfair algorithm is one whose decisions are skewed toward a particular group of people.

A Survey on Bias and Fairness in Machine Learning

NINAREH MEHRABI, FRED MORSTATTER, NRIPSUTA SAXENA, KRISTINA LERMAN, and ARAM GALSTYAN, USC-ISI

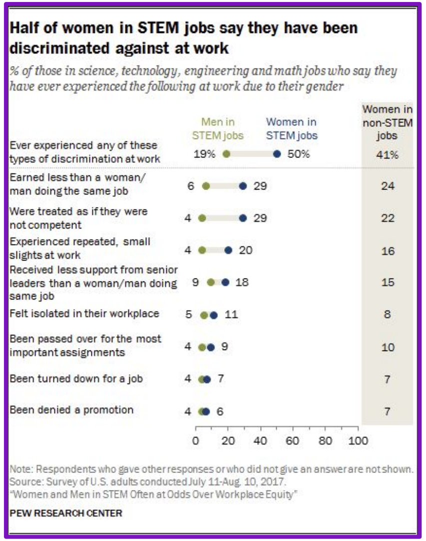
Does bias necessarily imply unfairness?

Take 3 minutes to discuss with your group

Bias vs Fairness

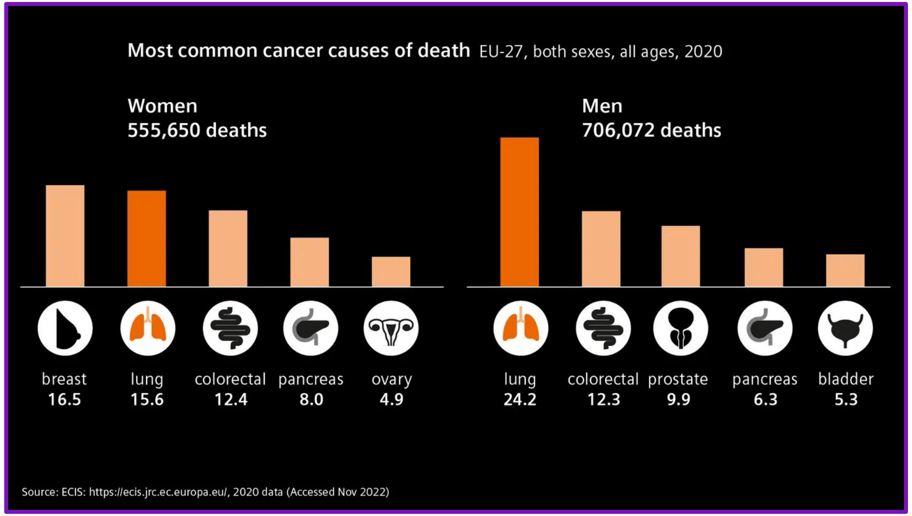
Bias **does not** necessarily imply unfairness

Gender and the workplace



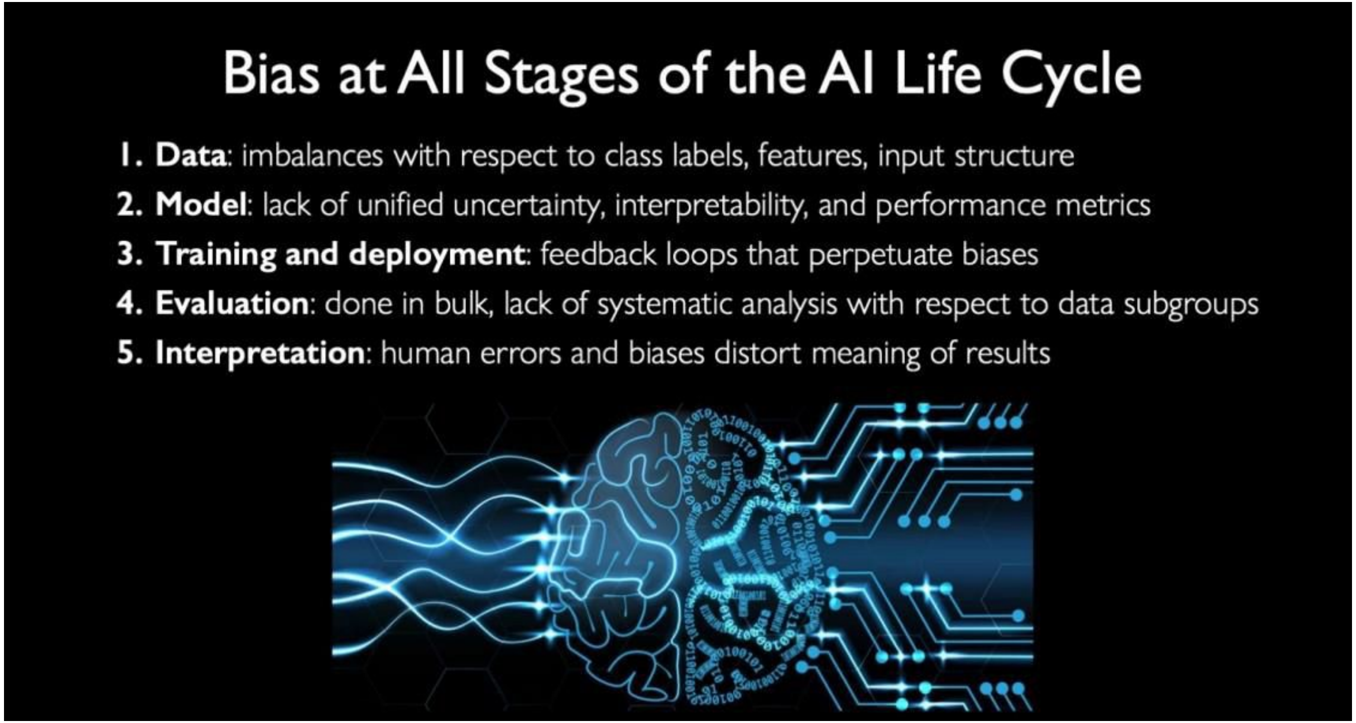
Gender discrimination is illegal

Gender in medical diagnosis



Gender specific medical diagnosis is desirable

Where is bias?



Slidefrom:©AlexanderAminiandAvaSoleimany
MIT6.S191:IntroductiontoDeepLearning,IntroToDeepLearning.com

There are many different types of bias

Manifestation of Bias

Bias can be manifested in data through:

1. Sensitive features and causal influences
2. Representativeness of data
3. Different data modalities (numerical, textual, etc.)

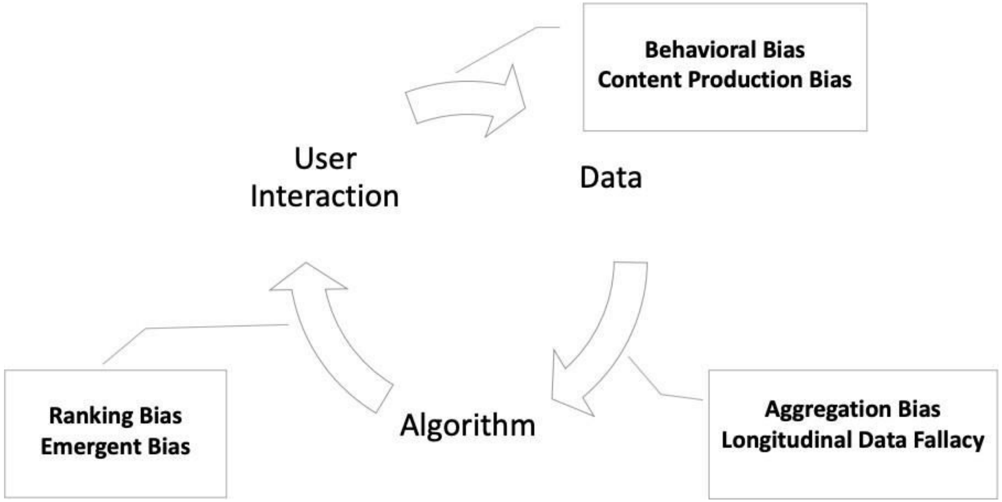
OVERVIEW

WIREs DATA MINING AND KNOWLEDGE DISCOVERY WILEY

Bias in data-driven artificial intelligence systems—An introductory survey

Eirini Ntoutsi[1] | Pavlos Fafalios[2] | Ujwal Gadiraju[1] | Vasileios Iosifidis[1] | Wolfgang Nejdl[1] | Maria-Esther Vidal[3] | Salvatore Ruggieri[4] | Franco Turini[4] | Symeon Papadopoulos[5] | Emmanouil Krasanakis[5] | Ioannis Kompatsiaris[5] | Katharina Kinder-Kurlanda[6] | Claudia Wagner[6] | Fariba Karimi[6] | Miriam Fernandez[7] | Harith Alani[7] | Bettina Berendt[8,9] | Tina Kruegel[10] | Christian Heinze[10] | Klaus Broelemann[11] | Gjergji Kasneci[11] | Thanassis Tiropanis[12] | Steffen Staab[1,12,13]

Sources of bias



A Survey on Bias and Fairness in Machine Learning

NINAREH MEHRABI, FRED MORSTATTER, NRIPSUTA SAXENA, KRISTINA LERMAN, and ARAM GALSTYAN, USC-ISI

Taxonomy of bias

Systematic distortions along different data properties:

1. Population biases
2. Behavioral biases
3. Content production biases
4. Linking biases
5. Temporal biases

Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries

Alexandra Olteanu [1,2], Carlos Castillo [3], Fernando Diaz [2] and Emre Kıcıman [4]*

Taxonomy of bias

1. **Population biases**
2. Behavioral biases
3. Content production biases
4. Linking biases
5. Temporal biases

Differences in demographics or other user characteristics between a user population represented in a dataset or platform and a target population

Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries

Alexandra Olteanu[1,2], Carlos Castillo[3], Fernando Diaz[2] and Emre Kıcıman[4]*

Taxonomy of bias

1. Population biases
2. **Behavioral biases**
3. Content production biases
4. Linking biases
5. Temporal biases

Differences in user behaviour across platforms or contexts, or across users represented in different datasets

Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries

Alexandra Olteanu [1,2], Carlos Castillo [3], Fernando Diaz [2] and Emre Kıcıman [4]*

Taxonomy of bias

1. Population biases
2. Behavioral biases
3. **Content production biases**
4. Linking biases
5. Temporal biases

Lexical, syntactic, semantic, and structural differences in the contents generated by users

Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries

Alexandra Olteanu[1,2], Carlos Castillo[3], Fernando Diaz[2] and Emre Kıcıman[4]*

Taxonomy of bias

1. Population biases
2. Behavioral biases
3. Content production biases
4. **Linking biases**
5. Temporal biases

Differences in the attributes of networks obtained from user connections, interactions and activity

Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries

Alexandra Olteanu[1,2], Carlos Castillo[3], Fernando Diaz[2] and Emre Kıcıman[4]*

Taxonomy of bias

1. Population biases
2. Behavioral biases
3. Content production biases
4. Linking biases
5. **Temporal biases**

Differences in populations and behaviours over time

Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries

Alexandra Olteanu[1,2], Carlos Castillo[3], Fernando Diaz[2] and Emre Kıcıman[4]*

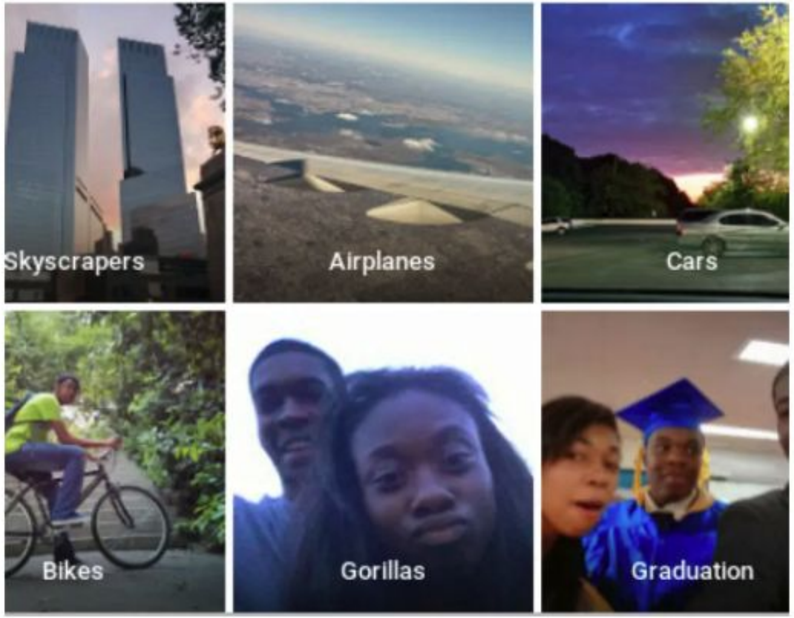
How can we handle biases?

Unfortunately

There is no standard approach or formula to debiasing algorithms and data

Let's start with some examples

Google Photos



Bias:
People with dark skin were labeledas *gorillas*

In 2015 Google Photos auto labels images uploaded to its site

Google Photos

TECH / GOOGLE / ARTIFICIAL INTELLIGENCE

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

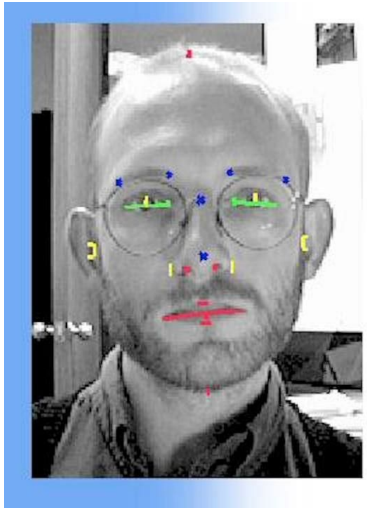
The AI algorithms in Google Photos sort images by a number of categories. Photo by Vjeran Pavic / The Verge

/ Nearly three years after the company was called out, it hasn't gone beyond a quick workaround

By James Vincent, a senior reporter who has covered AI, robotics, and more for eight years at The Verge.

Jan 12, 2018, 4:35 PM GMT+1 | 0 Comments / 0 New

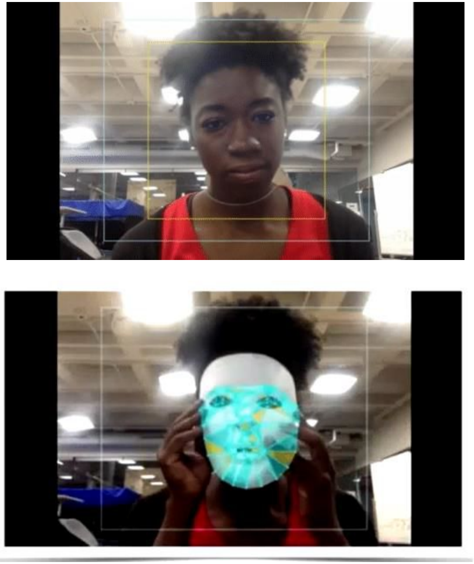
IBM Facial Recognition



In 2018 IBM sells software that detects faces and emotional reactions

Crawford, Kate. The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press, 2021.

IBM Facial Recognition



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female
Microsoft	94.0%	79.2%	100%	98.3%
FACE++	99.3%	65.5%	99.2%	94.0%
IBM	88.0%	65.3%	99.7%	92.9%

Bias:
Joy Buolamwiniet al. found that the software does not work equally well for all

IBM Facial Recognition

Tech

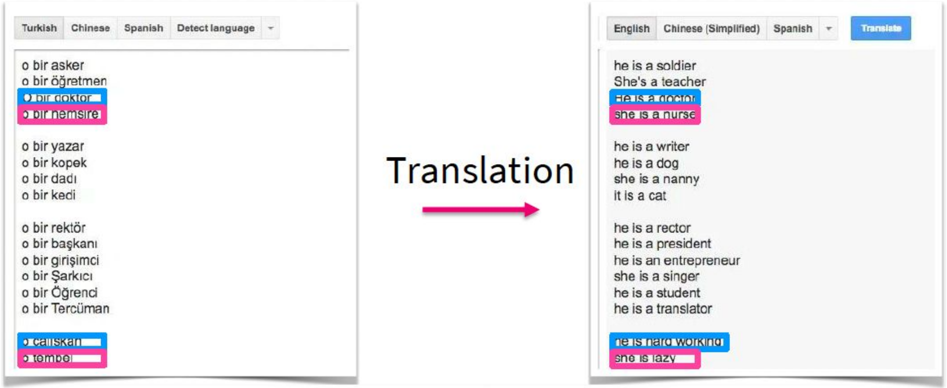
IBM abandons 'biased' facial recognition tech

BBC

9 June 2020

IBM added more pictures of the minority classes (2018) & in 2020 decided to stop providing general purpose facial recognition technologies

Google Translate

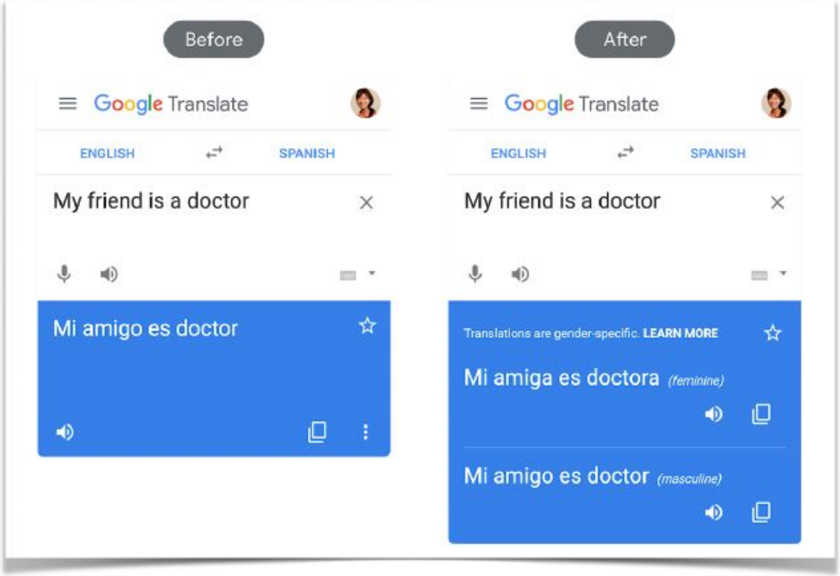


Google translate (2018) uses ML to translate from one language to others

Bias:

Reproduces gender and other stereotypes in a translated text

Google Translate



Built ML model to detect "gendered" translations and if thinks something is gendered it is hardcoded it to return multiple options

https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html

Unfortunately

There is no standard approach or formula to debiasing algorithms and data

Google Photos

Fixed by removing gorilla class

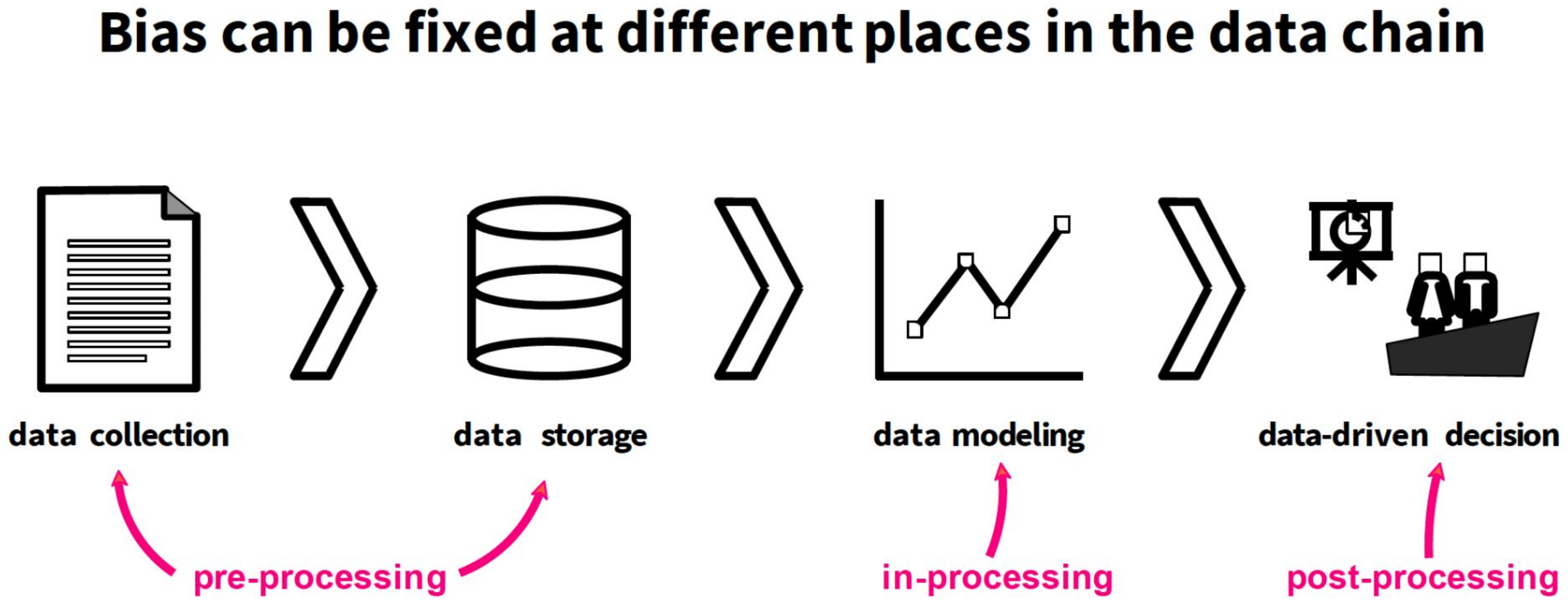
IBM

Fixed by adding more examples for unrepresented class

Google Translate

Added ML models & hardcoded response

Handling bias



Handling bias

Table 2. List of Papers Targeting and Talking about Bias and Fairness in Different Areas

Area	Reference(s)
Classification	[25, 49, 57, 63, 69, 73, 75, 78, 85, 102, 118, 143, 150, 151, 155]
Regression	[1, 14]
PCA	[133]
Community detection	[101]
Clustering	[8, 31]
Graph embedding	[22]
Causal inference	[82, 95, 111, 112, 123, 156, 160, 161]
Variational auto encoders	[5, 42, 96, 108]
Adversarial learning	[90, 152]
Word embedding	[20, 58, 165] [23, 162]
Coreference resolution	[130, 164]
Language model	[21]
Sentence embedding	[99]
Machine translation	[52]
Semantic role labeling	[163]
Named Entity Recognition	[100]

A Survey on Bias and Fairness in Machine Learning

NINAREH MEHRABI, FRED MORSTATTER, NRIPSUTA SAXENA, KRISTINA LERMAN, and ARAM GALSTYAN, USC-ISI

Handling bias

Table 1. Categorizing Different Fairness Notions into Group, Subgroup, and Individual Types

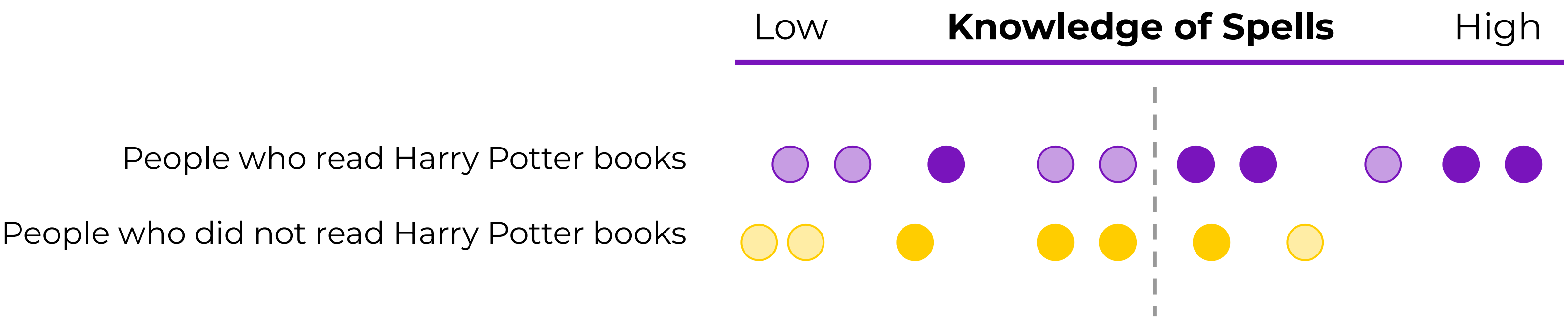
Name	Reference	Group	Subgroup	Individual
Demographic parity	[48, 87]	✓		
Conditional statistical parity	[41]	✓		
Equalized odds	[63]	✓		
Equal opportunity	[63]	✓		
Treatment equality	[15]	✓		
Test fairness	[34]	✓		
Subgroup fairness	[79, 80]		✓	
Fairness through unawareness	[61, 87]			✓
Fairness through awareness	[48]			✓
Counterfactual fairness	[87]			✓

A Survey on Bias and Fairness in Machine Learning

NINAREH MEHRABI, FRED MORSTATTER, NRIPSUTA SAXENA, KRISTINA LERMAN, and ARAM GALSTYAN, USC-ISI

Demographic parity

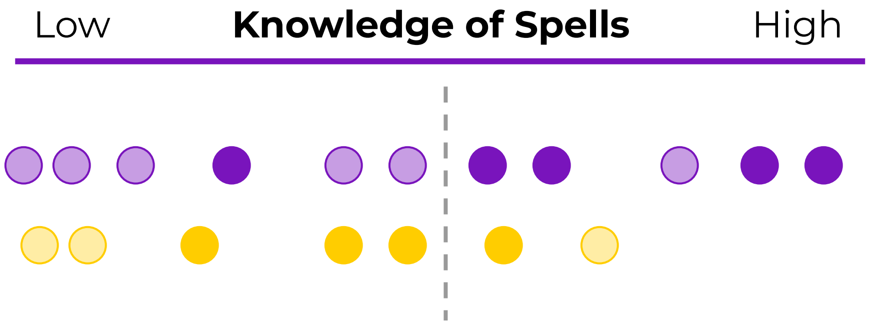
Admission to the Wizarding School



People indicated by full colors (and) will eventually become **Great Wizards**
We set a threshold for the admission (grey line)

Demographic parity

Number of applicants: 18
Number of purples: 11
Number of yellows: 7

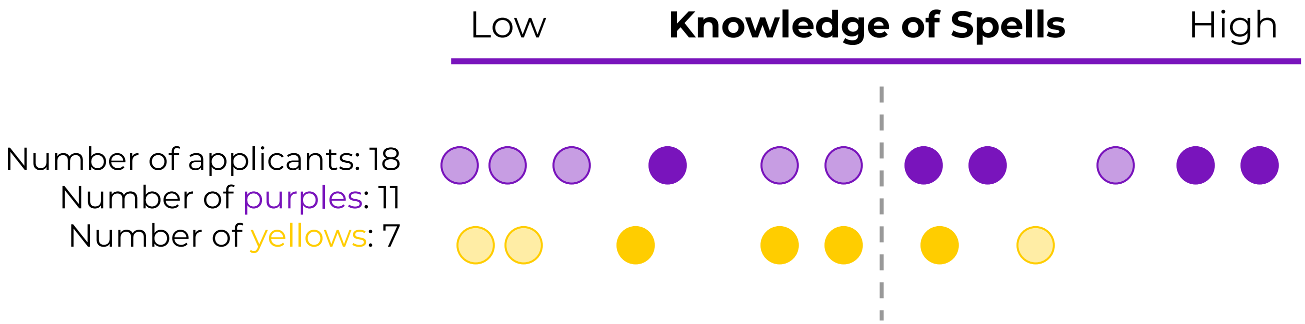


$$P(\text{Acceptance}) = \frac{\text{N. accepted}}{\text{Tot. applicants}} = 7/18 = 39\%$$

$$P(\text{Acceptance if purple}) = \frac{\text{N. accepted purple}}{\text{Tot. purple}} = 5/11 = 45\%$$

$$P(\text{Acceptance if yellow}) = \frac{\text{N. accepted yellow}}{\text{Tot. yellow}} = 2/7 = 29\%$$

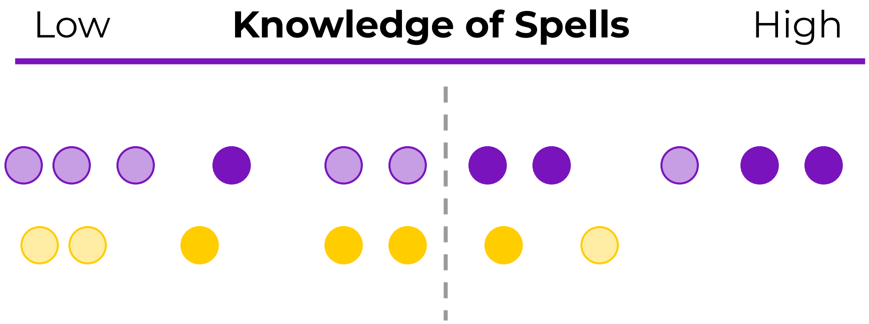
Demographic parity



P(Acceptance if purple) = P(Acceptance if yellow)

Demographic parity

Number of applicants: 18
Number of purples: 11
Number of yellows: 7



P(Acceptance if purple) = P(Acceptance if yellow)

Two options:

1. Admit less purple

$$P_{flip} = 1 - \frac{P(\text{Acceptance if yellow})}{P(\text{Acceptance if purple})} = 1 - 29/45 \approx 0.64$$

2. Admit more yellow

$$P_{flip} = \frac{P(\text{Acceptance if yellow})}{P(\text{Acceptance if purple})} = 29/45 \approx 0.36$$

Let's put it into practice!

In groups:

1. Go over the GitHub page (https://github.com/AnnaSapienza/CSS_SummerSchool/tree/main)
2. Try to solve the exercises
3. Discuss with your peers