

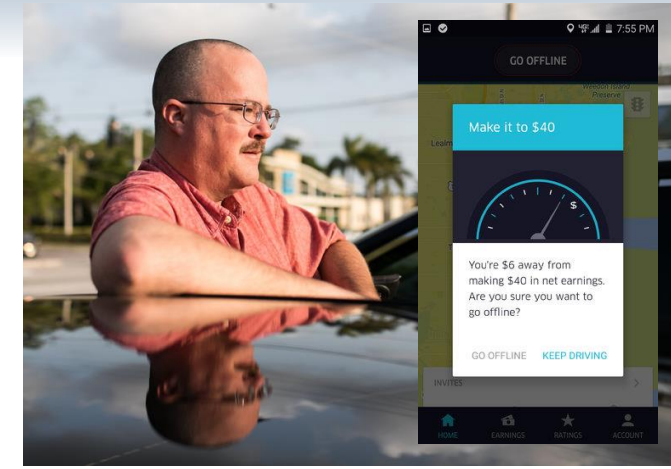


# **Ethics of Artificial Intelligence**

# Ethics of AI

## Will a new technology:

- disempower **individuals vs corporations?**
  - ⇒ user modeling; data mining; fostering addictive behaviors; developmental effects on children
- disempower **individuals vs governments?**
  - ⇒ facilitate disinformation (deep fakes; bots masquerading as people; filter bubbles); enable qualitatively new military or security tactics
- take **autonomous actions** in a way that obscures responsibility
  - ⇒ autonomous weapons; self-driving cars; loan approval systems
- disproportionately affect **vulnerable/marginalized groups**
  - ⇒ automated decision making tools trained in ways that may encode existing biases



# **BIAS AND FAIRNESS**

# Human bias

Bias in people refers to our tendency to take quick decisions based on little information

Published online 11 April 2011 | Nature | doi:10.1038/news.2011.227

**News**

## Hungry judges dispense rough justice

When they need a break, decision-makers gravitate towards the easy option.

Zoë Corbyn

*Journal of Economic Perspectives—Volume 12, Number 2—Spring 1998—Pages 41–62*

## Evidence on Discrimination in Mortgage Lending

Helen F. Ladd

## Science faculty's subtle gender biases favor male students

Corinne A. Moss-Racusin<sup>a,b</sup>, John F. Dovidio<sup>b</sup>, Victoria L. Brescoll<sup>c</sup>, Mark J. Graham<sup>a,d</sup>, and Jo Handelsman<sup>a,1</sup>


<sup>a</sup>Department of Molecular, Cellular and Developmental Biology, <sup>b</sup>Department of Psychology, <sup>c</sup>School of Management, and <sup>d</sup>Department of Psychiatry, Yale University, New Haven, CT 06520

Edited\* by Shirley Tilghman, Princeton University, Princeton, NJ, and approved August 21, 2012 (received for review July 2, 2012)

**The Observer**  
Stop and search

## Racial bias in police stop and search getting worse, report reveals

Despite reforms, black people are nine times more likely than white people to be checked for drugs




Mark Townsend  
Home Affairs Editor  
@marktownsenduk  
Sat, 10 Dec 2010 10:02 BST

1716  
This article is over 8 months old

## Is it easier to get a job if you're Adam or Mohamed?

By Zack Adkins and Clara Mancini  
BBC Inside Out

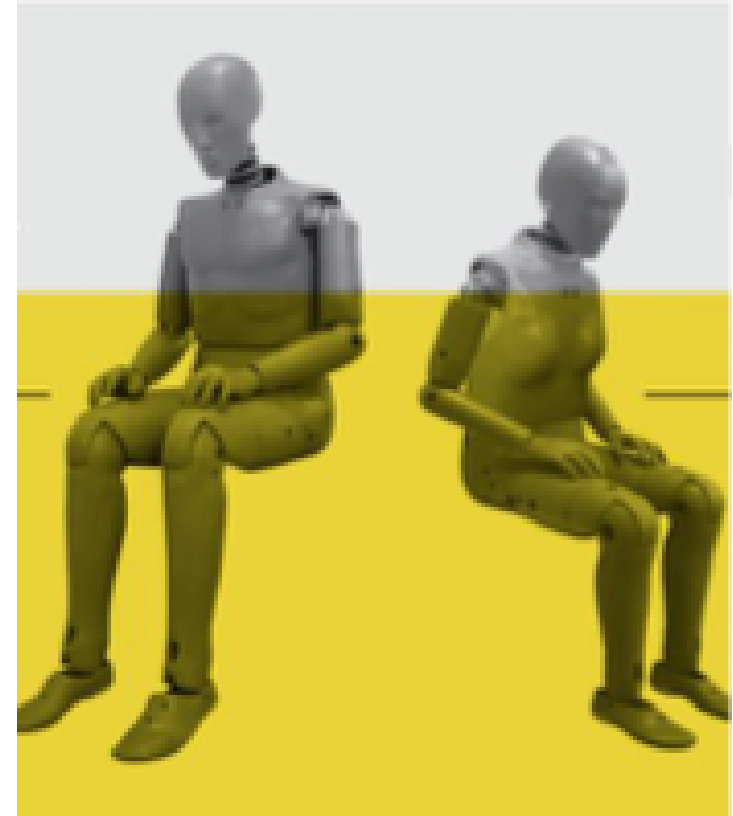
6 February 2017



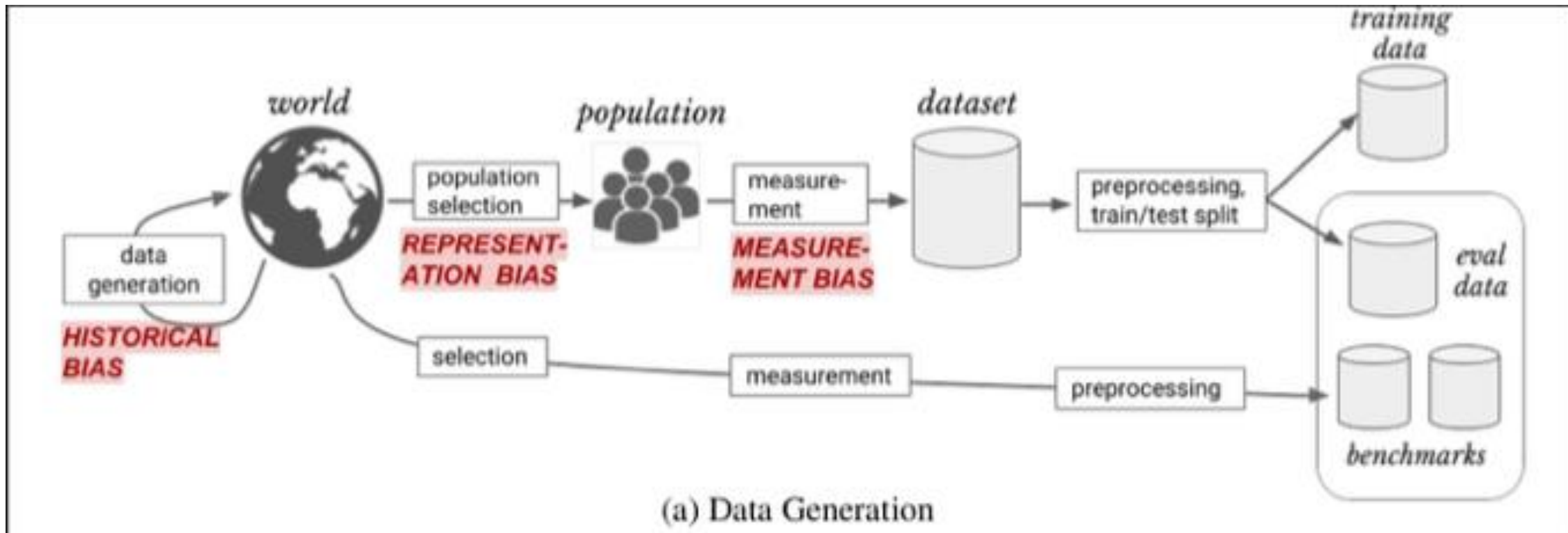
The two CVs sent out obtained the same level of qualifications and experience.

A job seeker with an English-sounding name was offered three times the number of interviews than an applicant with a Muslim-sounding name, BBC test found.

# Can technology have bias?



# Sources of bias in ML algorithms



# Historical bias

**Historical bias** arises when there is a misalignment between world as it is and the values or objectives to be encoded and propagated in a model. It is a normative concern with the state of the world, and exists even given perfect sampling and feature selection.

†

**Example: image search** In 2018, 5% of Fortune 500 CEOs were women (Zarya, 2018). Should image search results for “CEO” reflect that number? Ultimately, a variety of stakeholders, including affected members of society, should evaluate the particular harms that this result could cause and make a judgment. This decision may be at odds with the available data even if that data is a perfect reflection of the world. Indeed, Google has recently changed their Image Search results for “CEO” to display a higher proportion of women.



# Representation bias

**Representation bias** arises while defining and sampling a development population. It occurs when the development population under-represents, and subsequently fails to generalize well, for some part of the use population. 1

1. **The sampling methods only reach a portion of the population.** For example, datasets collected through smartphone apps can under-represent lower-income or older groups, who are less likely to own smartphones. Similarly, medical data for a particular condition may be available only for the population of patients who were considered serious enough to bring in for further screening.
2. **The population of interest has changed or is distinct from the population used during model training.** Data that is representative of Boston, for example, may not be representative if used to analyze the population of Indianapolis. Similarly, data representative of Boston 30 years ago will likely not reflect today's population.



# Measurement bias

**Measurement Bias** arises when choosing and measuring features and labels to use; these are often proxies for the desired quantities. The chosen set of features and labels may leave out important factors or introduce group- or input-dependent noise that leads to differential performance.

1

3. **The defined classification task is an oversimplification.**

In order to build a supervised ML model, some label to predict must be chosen. Reducing a decision to a single attribute can create a biased proxy label because it only captures a particular aspect of what we really want to measure. Consider the prediction problem of deciding whether a student will be successful (e.g., in a college admissions context). Fully capturing the outcome of ‘successful student’ in terms of a single measurable attribute is impossible because of its complexity. In cases such as these, algorithm designers resort to some available label such as ‘GPA’ (Kleinberg et al. 2018), which ignores different indicators of success achieved by parts of the population.

1. **The measurement process varies across groups.** For example, if a group of factory workers is more stringently or frequently monitored, more errors will be observed in that group. This can also lead to a feedback loop wherein the group is subject to further monitoring because of the apparent higher rate of mistakes (Barocas and Selbst 2016).
2. **The quality of data varies across groups.** Structural discrimination can lead to systematically higher error rates in a certain group. For example, women are more likely to be misdiagnosed or not diagnosed for conditions where self-reported pain is a symptom (Calderone 1990). In this case, “*diagnosed* with condition X” is a biased proxy for “has condition X.”

# Why worrying about bias in algorithms

Decisions made by a ML algorithm are:

- Cheap
- Scalable
- Automated
- Self-reinforcing
- Seemingly objective
- Often lacking appeals processes
- Not just predicting but also causing the future

# Fairness in algorithms

- Nowadays, more attention is placed on algorithms being **fair**, and not just accurate.
- Fairness can be measured as:
  - demographic (or statistical) parity: population percentage should be reflected in the output classes
  - Equality of false negatives or equalized odds: constant false-negative (or both false-negative and true-negative) rates across groups.
  - Equal opportunity: equal True Positive Rate for all groups
  - Other metrics...
- Accuracy and fairness tend to be at odds with each other.
- Algorithms can be audited to test their fairness.
- *Are we ethically required to sacrifice accuracy for fairness?*

# Algorithms to promote engagement

- Large, popular social media platforms use algorithms to increase user engagement
- Proposed content is designed to keep the user on the website longer
  - It also often becomes more extreme as the user follows the suggestions
  - Sometimes with very disturbing results: <https://www.npr.org/sections/thetwo-way/2017/11/27/566769570/youtube-faces-increased-criticism-that-its-unsafe-for-kids>
- They also tend to promote content that the user will agree/engage with, creating echo-chambers
  - Some theorize that echo-chambers can push people towards more extreme opinions
  - Do you agree? Should social media be required to change their recommendation algorithms to avoid these issues?
  - <https://www.pnas.org/doi/10.1073/pnas.2023301118>

# When the metric becomes the target (Goodhart's Law)

*"When a measure becomes a target it ceases to be a good measure"*

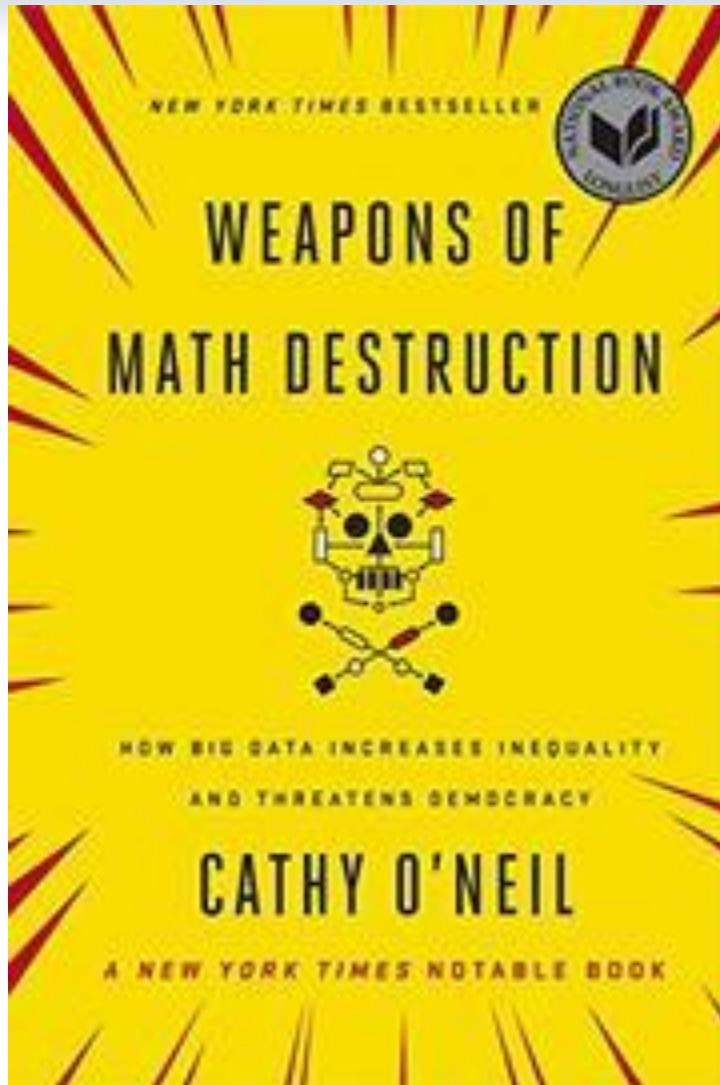
- Metrics introduced in the [British public healthcare system](#) (e.g. waiting time in ER) caused people to game it:
  - Cancelled scheduled operations to draft extra staff to ER
  - Required patients to wait outside the ER, e.g. in ambulances
  - Put stretchers in hallways and classified them as "beds"
  - Hospital and patients reported different wait times
- Big Data is significantly changing college applications (not in a good way)
  - Universities are given higher ranking for things such as receiving more applications, being more selective, and having more students accept their offers (while tuition is not considered)
  - This even pushed some mid-tier universities to reduce the number of offer letter sent out, especially to good students who they think would not accept. Students are losing their safety options
- Is this always undesirable? Can you think of ways to avoid this trap?

# Ethics of algorithmic pricing

- Algorithms are currently used to adjust prices based on:
  - Willingness of buyer
  - Availability
- You are probably familiar with car-sharing apps, like Uber.
- Unlike cabs, which work with fixed rates, the cost of a Uber ride is determined by an algorithm, based on supply and demand.
- This can introduce unfairness in several ways:
  - Underserved (poorer) neighbors get higher fares than more served ones
  - Surges in time of crisis (hostage siege in Sidney in 2014)
  - Algorithmic wage discrimination
    - [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=4331080](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4331080)
    - <https://www.latimes.com/business/technology/story/2023-04-11/algorithmic-wage-discrimination>



# Interesting reads





If you'd like to know more

**DSCI 430 - FAIRNESS, ACCOUNTABILITY,  
TRANSPARENCY AND ETHICS (FATE) IN DATA  
SCIENCE**