(Un)fairness in ML

- ML algorithms increasingly used to make decisions about people
- Can lead to unfairness! (Related: privacy, transparency, diversity in ML, safe AI, future of work, ethics in ML/AI)
- Today: examples and discussion.
  Issues that arise. How? What to do?
- Mostly questions, not answers
- Slides credit: Alexandra Meliou, Gerome Miklau

Algorithms as pinnacle of fairness?

Big Data's Disparate Impact

104 California Law Review 671 (2016)

62 Pages • Posted: 11 Aug 2014 • Last revised: 30 Sep 2016

Solon Barocas
Cornell University

Andrew D. Selbst
Data & Society Research Institute; Yale Information Society Project

Date Written: 2016

Abstract

Advocates of algorithmic techniques like data mining argue that these techniques eliminate human biases from the decision-making process. But an algorithm is only as good as the data it works with. Data is frequently imperfect in ways that allow these algorithms to inherit the prejudices of prior decision makers. In other cases, data may simply reflect the widespread biases that persist in society at large. In still others, data mining can discover surprisingly useful regularities that are really just preexisting patterns of exclusion and inequality. Unthinking reliance on data mining can deny historically disadvantaged and vulnerable groups full participation in society. Worse still, because the resulting discrimination is almost always an unintentional emergent property of the algorithm's use rather than a conscious choice by its programmers, it can be unusually hard to identify the source of the problem or to explain it to a court.

TIME SUBSCRIBE

IDEAS • TECHNOLOGY

The Police Are Using Computer Algorithms to Tell If You're a Threat

Resilient cities Cities

Predicting crime, LAPD-style

Cutting edge data-driven analysis directs Los Angeles patrol officers to likely future crime scenes - but critics worry that decision-making by machine will bring 'tyranny of the algorithm'

● Join our live Q&A with Homicide Watch this Friday

▲ PredPol co-developer P Jeffrey Brantingham at the Unified Command Post in Los Angeles. 'This is not Minority Report,' he said. Photograph: Damian Dovarganes/AP

ACLU GET UPDATES / DONATE

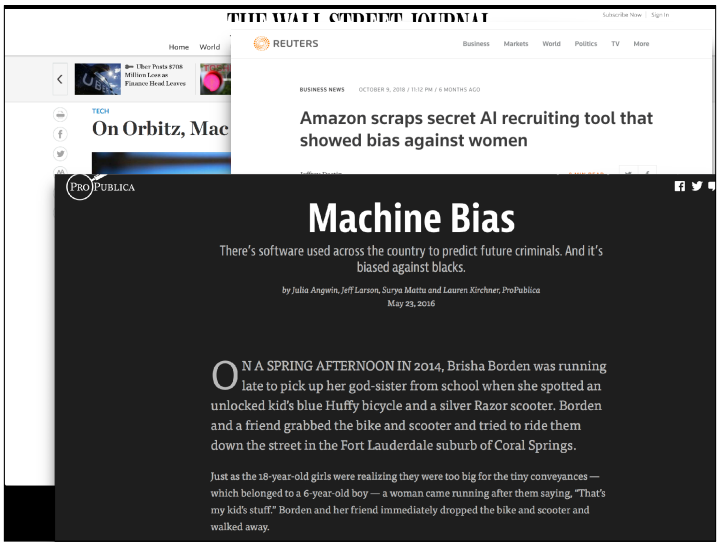
The Government Is Blacklisting People Based on Predictions of Future Crimes

By Hina Shamsi, Director, ACLU National Security Project
OCTOBER 8, 2018 | 9:15 AM

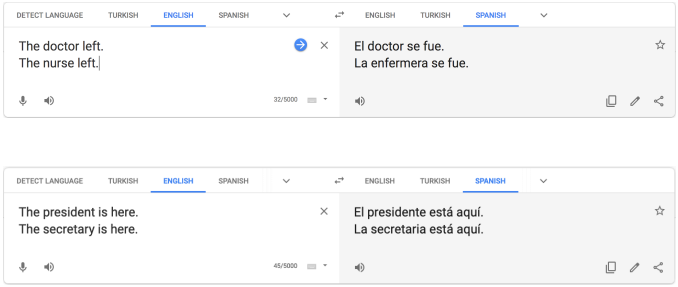
TAGS: Discriminatory Profiling, National Security

Imagine: You've never been charged with any crime, yet the government blacklists you as a terrorism threat and bans you from flying indefinitely. You're separated from family members, can't get to weddings or funerals or religious obligations, and lose jobs because you can't travel or your employer finds out you're blacklisted.

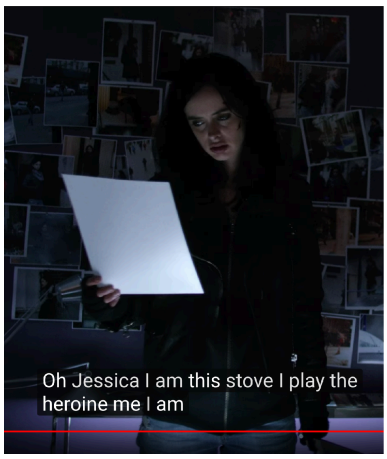
You know what the government has done violates your constitutionally



algorithms can exacerbate societal biases



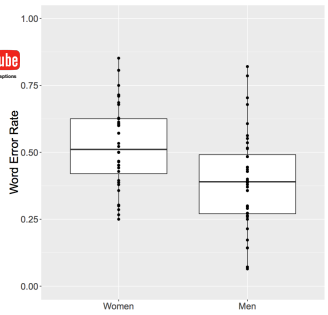
algorithms don't provide the same service to all



YouTube automatic captions

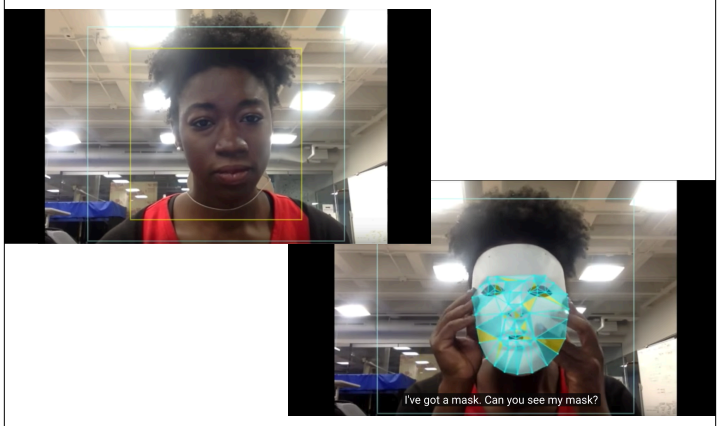
algorithms don't provide the same service to all

YouTube automatic captions



Rachael Tatman, "Gender and Dialect Bias in YouTube's Automatic Captions" in 2017 Workshop on Ethics in Natural Language Processing

algorithms don't provide the same service to all



Joy Buolamwini
https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms

Staples online pricing

THE WALL STREET JOURNAL.

WHAT THEY KNOW

Websites Vary Prices, Deals Based on Users' Information

By JENNIFER VALENTINO-DEVRIES, JEREMY SINGER-VINE and ASHKAN SOLTANI
December 24, 2012

It was the same Swingline stapler, on the same Staples.com website. But for Kim Wamble, the price was $15.79, while the price on Trude Frizzell's screen, just a few miles away, was $14.29.

A key difference: where Staples seemed to think they were located.

WHAT PRICE WOULD YOU SEE?

lower prices offered to buyers who live in more affluent neighborhoods

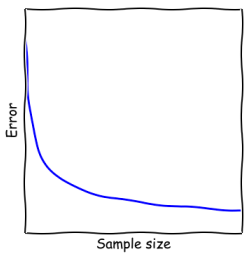
Any many more examples…

How does this happen?

"Social Mirror"

- **"Social Mirror" = biased training data**
- Replicate biased judgments of people
  - Thought experiment: college admissions
- Discover preexisting patterns of exclusion / inequality
  - Google translate

Sample size disparity



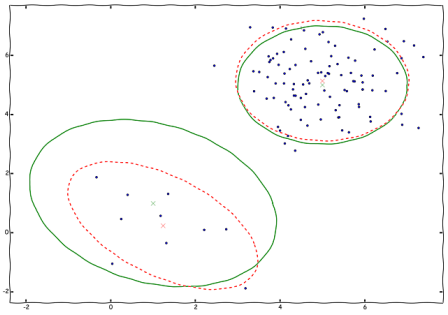
- more data —> lower error
- less data —> more error
- "by definition that there is always proportionately less data available about minorities"
- different error rates among groups
- what does it mean to "be 95% accurate"?

Example: faces

http://vis-www.cs.umass.edu/lfw/devTrain.html

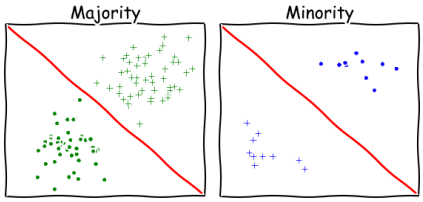
- Labeled faces in the wild: "more than 13,000 images of faces collected from the web"
  - Who appears on the web
- Long-term effects of data / responsible use

Technical Example: Gaussian Mixture Models



Cultural Differences

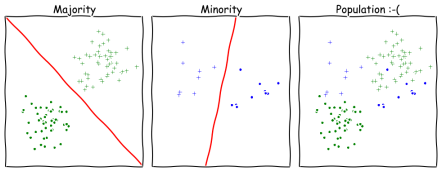
- Social network: is username fake?
  Majority: names short, common
  Minority: names long, rare



- Minority patterns may be overwhelmed by majority

Possible Solutions?

- Learn a separate classifiers for each group?
  Thorny issues: how to define minority groups?
  Now taking explicit action on protected attribute.
- Learn one classifier? Increased complexity



What can we do?

- Legal notions
  - Regulated domains: credit, education, housing, employment, etc. (includes marketing, advertising)
  - Protected classes: race, color, sex, religion, national origin, citizenship, age, disability, …
  - Disparate treatment: formal or intentional
  - Disparate impact: unjustified or avoidable
- Try to design ML systems to be fair
  - Easier said than done. Fairness is domain-specific. Pitfalls in simple approaches.
  - Nascent field of research

Example: Proxies

Data, Responsibly
fairness, neutrality and transparency in data analysis

Julia Stoyanovich
Drexel University, USA

Serge Abiteboul
INRIA, France

Gerome Miklau
UMass Amherst, USA

The evils of discrimination

Disparate treatment is the illegal practice of treating an entity, such as a creditor or employer, differently based on a **protected characteristic** such as race, gender, age, religion, sexual orientation, or national origin.

http://www.allenovery.com/publications/en-gb/Pages/Protected-characteristics-and-the-perception-reality-gap.aspx

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

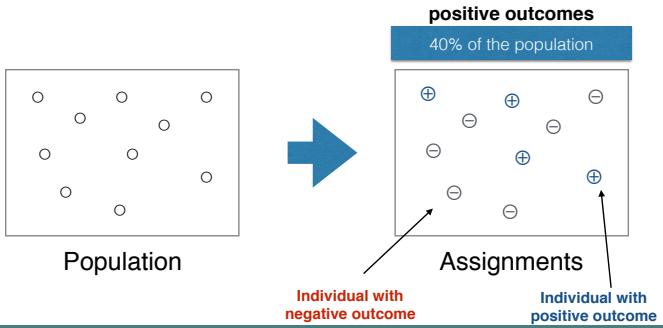
Outcomes

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

Positive Outcomes	Negative Outcomes
offered employment	denied employment
accepted to school	rejected from school
offered a loan	denied a loan
offered a discount	not offered a discount

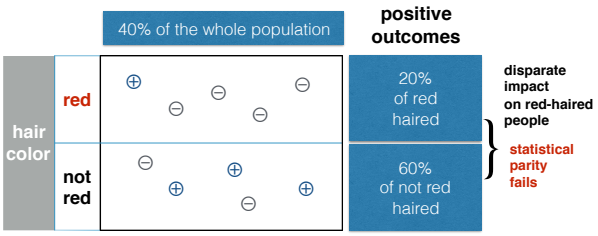
Assigning outcomes to populations

Fairness is concerned with how outcomes are assigned to a population



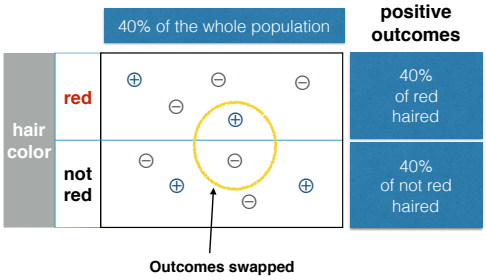
Sub-populations may be treated differently

Sub-population: those with red hair (under the same assignment of outcomes)



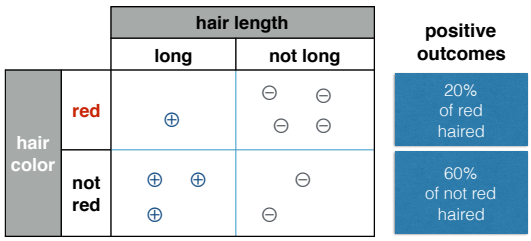
Enforcing statistical parity

Statistical parity (aka **group fairness**)
demographics of the individuals receiving any outcome are the same as demographics of the underlying population



Redundant encoding

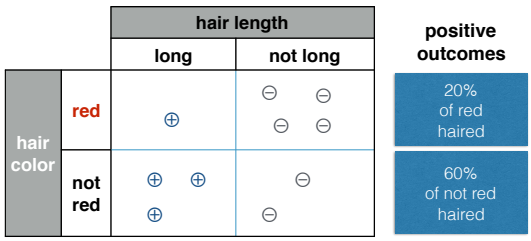
Now consider the assignments under both **hair color** (protected) and **hair length** (innocuous)



Deniability
The vendor has adversely impacted red-haired people, but claims that outcomes are assigned according to hair length.

Blinding does not imply fairness

Removing **hair color** from the vendor's assignment process does not prevent discrimination



Assessing disparate impact
Discrimination is assessed by the effect on the protected sub-population, not by the input or by the process that lead to the effect.

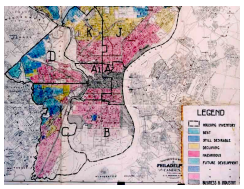
Redundant encoding

Let's replace hair color with **race** (protected), hair length with **zip code** (innocuous)

		zip code: 10025	zip code: 10027	positive outcomes
race	black	⊕	⊖ ⊖ ⊖ ⊖	20% of black
	white	⊕ ⊕ ⊕	⊖ ⊖	60% of white

The evils of discrimination

Substituting hair color (protected) with hair length (innocuous) or race (protected) with zip code (innocuous) are examples of **redundant encoding**.



Redlining is the practice of arbitrarily denying or limiting financial services to specific neighborhoods, generally because its residents are people of color or are poor.

Discrimination may be unintended

Staples website estimated user's location, **offering discounts** to those near rival stores, leading to discrimination w.r.t. to average income.

		rival store proximity: close	rival store proximity: far	positive outcomes
income	low	⊕	⊖ ⊖ ⊖ ⊖	20% of low income
	high	⊕ ⊕ ⊕	⊖ ⊖	60% of high income

Discrimination
Whether intentional or not, discrimination is unethical and, in many countries, illegal.

Many Other Issues

- Transparency
- Privacy
- Safe AI
- Future of work
- Ethics in big tech: your role