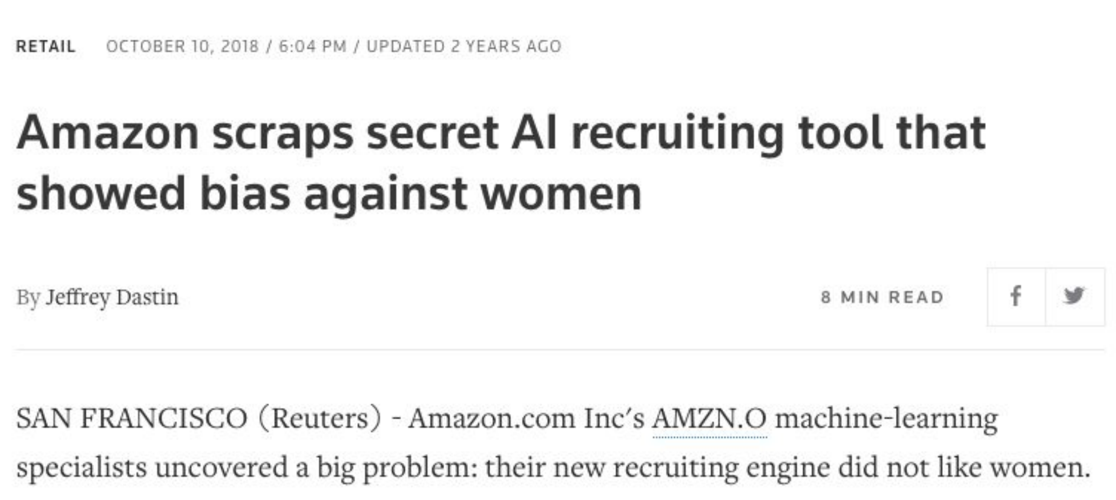
Fairness

Issues while blindly shipping ML-based apps.

Examples of algorithmic Bias

- Amazon’s computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry.
- Amazon’s system taught itself that male candidates were preferable. It penalized resumes that included the word “women’s,” as in “women’s chess club captain.”



RETAIL OCTOBER 10, 2018 / 6:04 PM / UPDATED 2 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

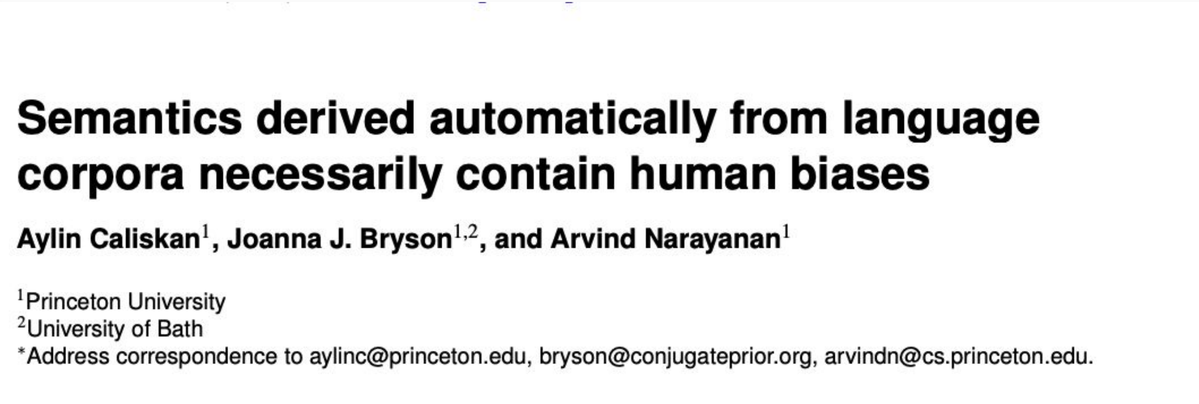
By Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc’s AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

- European American names are more likely than African American names to be closer to pleasant than to unpleasant.
- Female words associated with family and males associated with career.
- Female words associated with arts while men to sciences.



Semantics derived automatically from language corpora necessarily contain human biases

Aylin Caliskan[1], **Joanna J. Bryson**[1,2], **and Arvind Narayanan**[1]

[1]Princeton University
[2]University of Bath
*Address correspondence to aylinc@princeton.edu, bryson@conjugateprior.org, arvindn@cs.princeton.edu.

https://arxiv.org/pdf/1608.07187.pdf

- Facial recognition systems fail more with people of darker skin.

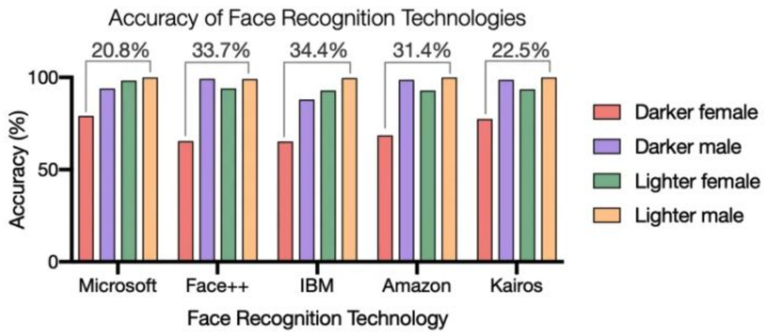
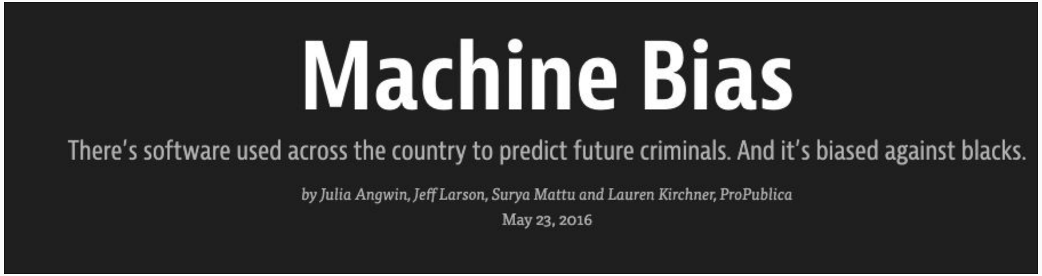


Figure 1: Auditing five face recognition technologies. *The* ***Gender Shades project*** *revealed* ***discrepancies*** *in the classification accuracy of face recognition technologies for different skin tones and sexes. These algorithms consistently demonstrated the poorest accuracy for darker-skinned females and the highest for lighter-skinned males.*

https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

- The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.
- White defendants were mislabeled as low risk more often than black defendants.



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

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

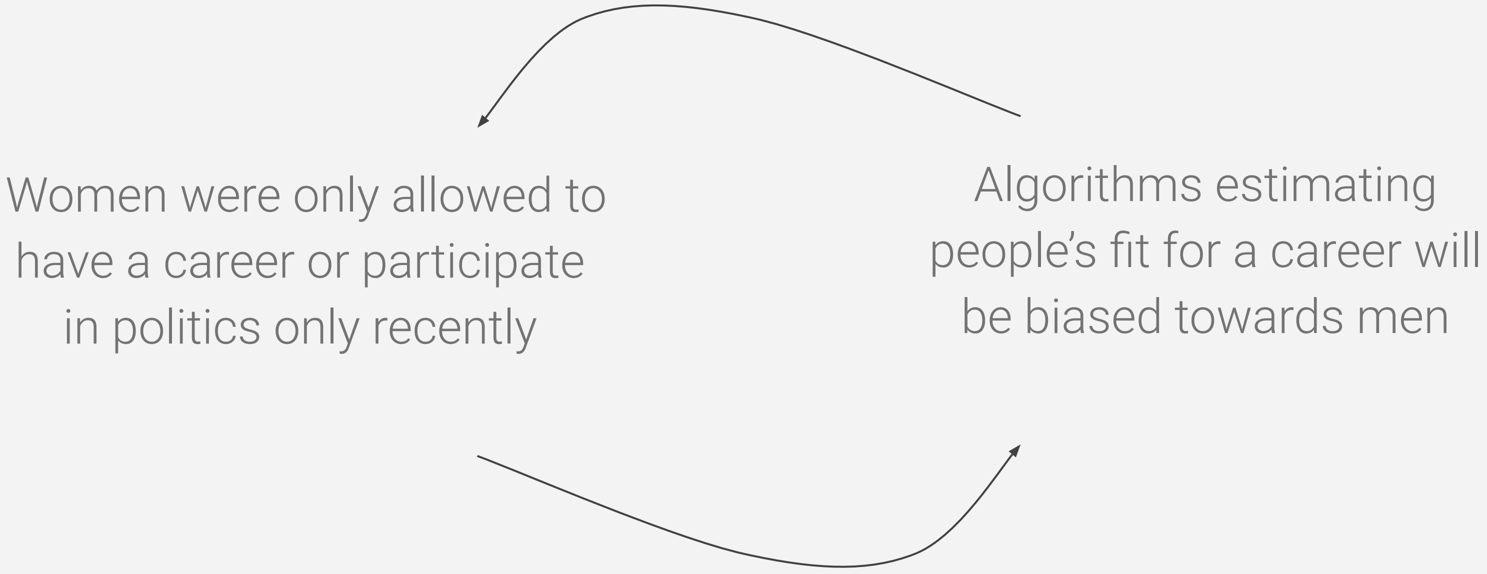
Causes

Historical Human Biases

- Woman are meant to stay at home instead of having a career.
- Black men are more likely criminals
- Poor people do not know how to work
- People from poor neighbors should get a low paying job instead of having a career
- Women should do arts instead of science or engineering

Underrepresented groups

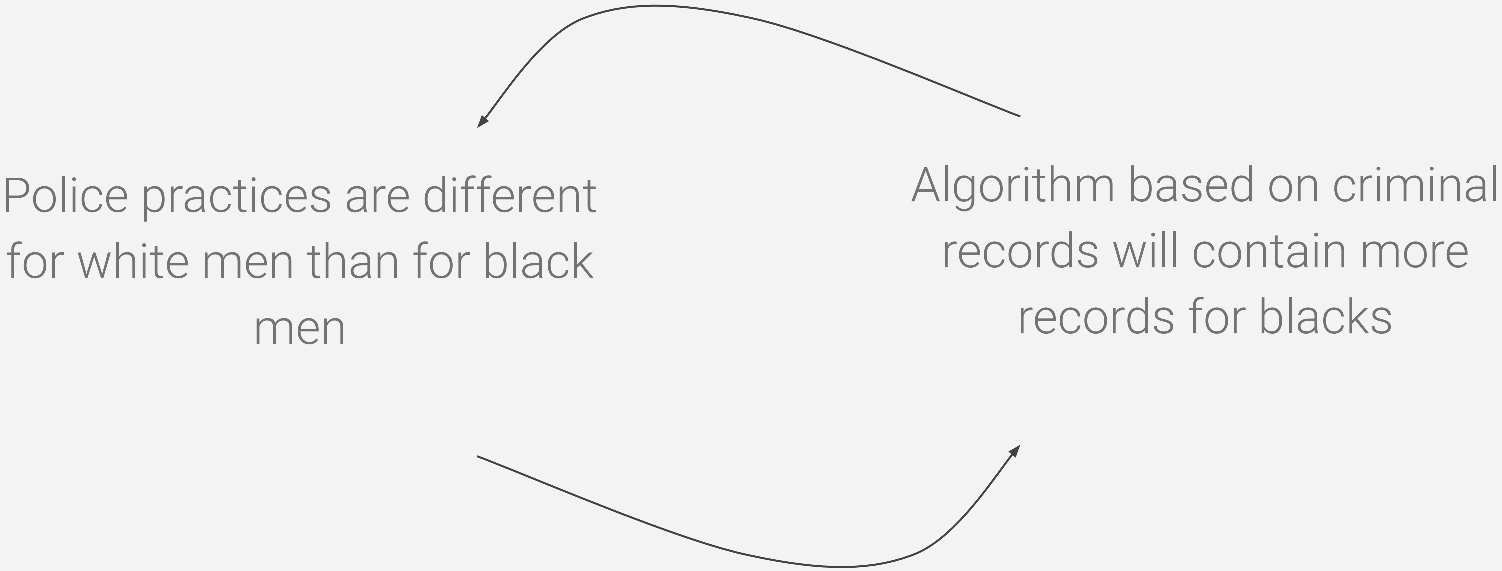
- Less black people go to university
- Less women are currently in an IT job
- Less black people with smartphones
- Less women leading successful business



Women were only allowed to have a career or participate in politics only recently

Algorithms estimating people's fit for a career will be biased towards men

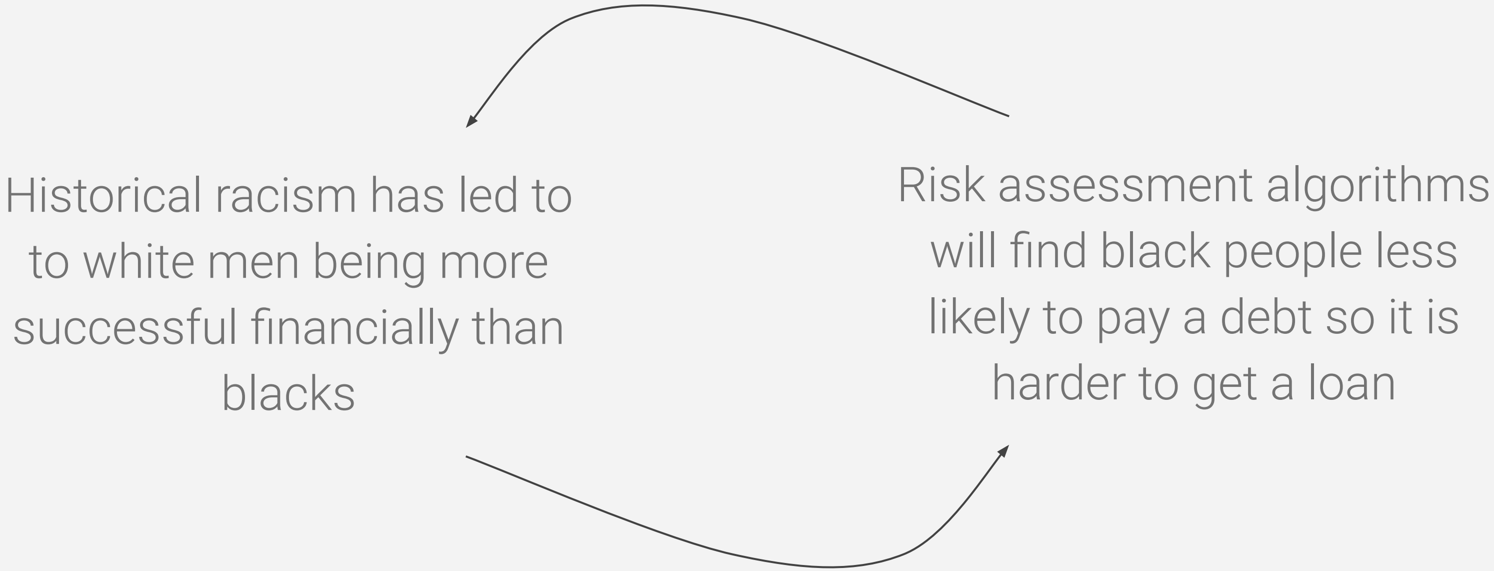
Historical Human Bias | Underrepresented groups



Police practices are different for white men than for black men

Algorithm based on criminal records will contain more records for blacks

Historical Human Bias | Underrepresented groups



Historical racism has led to to white men being more successful financially than blacks

Risk assessment algorithms will find black people less likely to pay a debt so it is harder to get a loan

Historical Human Bias | Underrepresented groups

“African-Americans who are primarily the target for high-interest credit card options might find themselves clicking on this type of ad without realizing that they will continue to receive such predatory online suggestions. In this and other cases, the algorithm may never accumulate counter-factual ad suggestions (e.g., lower-interest credit options) that the consumer could be eligible for and prefer.”

- Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms

“Researchers at Georgetown Law School found that an estimated 117 million American adults are in facial recognition networks used by law enforcement, and that African-Americans were more likely to be singled out primarily because of their over-representation in mug-shot databases [even if they were innocent]”

- Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms

Metrics

Algorithmic decision making and the cost of fairness

Sam Corbett-Davies
Stanford University
scorbett@stanford.edu

Emma Pierson
Stanford University
emmap1@stanford.edu

Avi Feller
Univ. of California, Berkeley
afeller@berkeley.edu

Sharad Goel
Stanford University
scgoel@stanford.edu

Aziz Huq
University of Chicago
huq@uchicago.edu

https://arxiv.org/pdf/1701.08230.pdf

Notation

$x_i \in \mathbb{R}^p$	Visual attributes of an individual (e.g. gender, race, etc.).
$X : \Omega \rightarrow \mathbb{R}^p$	Random variable that takes the values of individuals visual attributes.
$d : \mathbb{R}^p \rightarrow [0, 1]$	Decision rule, the probability of taking an action for an individual
$g : \mathbb{R}^p \rightarrow \{g_1, g_2, ..., g_k\}$	Membership of individuals visual attributes to a group (e.g. black men).
$y_i \in \{0, 1\}$	Benefit of an individual
$Y : \Omega \rightarrow \{0, 1\}$	Random variable that takes the values of the benefit of an individual

Statistical Parity

Equal rates for decision function regardless of membership to a group. For example, white and black defendants are detained at equal rates.

$$\mathbb{E}[d(X)|g(X)] = \mathbb{E}[d(X)]$$

Conditional Statistical Parity

Controlling for a limited set of “legitimate” risk factors, an equal rates of decision functions are computed within each group. For example, among defendants who have the same number of prior convictions, black and white defendants are detained at equal rates.

$$\mathbb{E}[d(X)|l(X), g(X)] = \mathbb{E}[d(X)|l(X)]$$

$$l : \mathbb{R}^p \to \mathbb{R}^m$$

Predictive equality

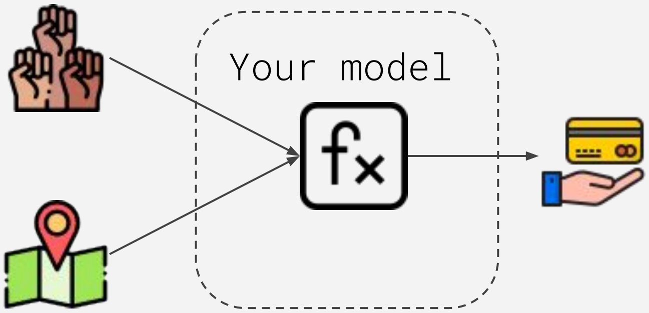
Accuracy of decisions is equal across membership groups, as measured by false positive rate (or any other metric). For example, among defendants who have not gone on to commit a violent crime if release, detention rates are equal across race groups.

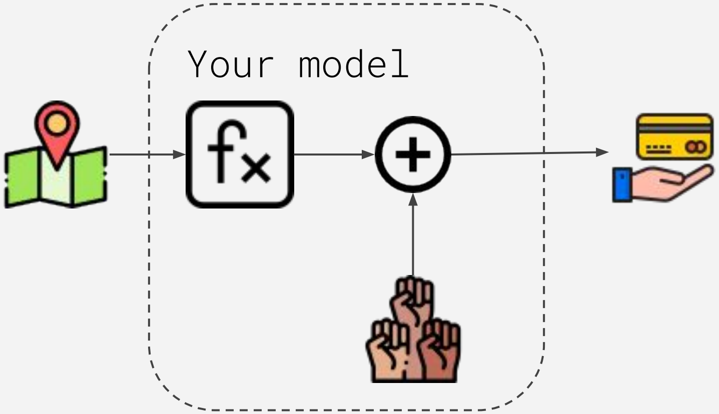
$$\mathbb{E}[d(X)|Y = 0, g(X)] = \mathbb{E}[d(X)|Y = 0]$$

Mitigation

Whether we are defining the target variable, labelling, collecting training data, using feature selection or making decision on the basis of the resulting model we could make room for our final result to have disproportionately adverse impact on protected classes, whether by specifying the problem to be solved in ways that affect classes differently, failing to recognize or address statistical biases, reproducing past prejudice, or considering an insufficiently rich set of factors.

- https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899





Ignoring a protected attribute may not be enough

Build a dataset so that it equally represent groups with respect to protected attributes

Consider the algorithm you are using, including preprocessing and postprocessing steps.