Fair and accountable AI

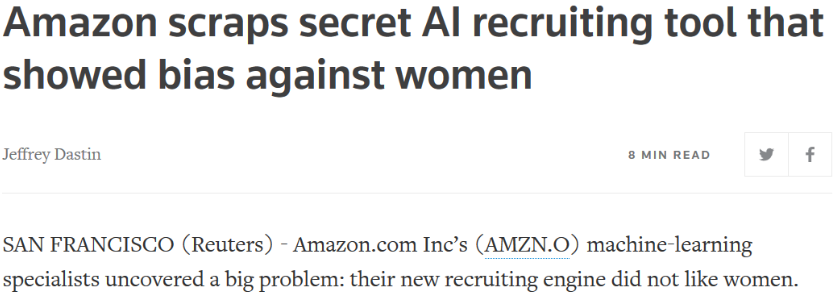
Super Day 2019

Hinda Haned

December 17th, 2019

h.haned@uva.nl

Fair and accountable AI: why bother?

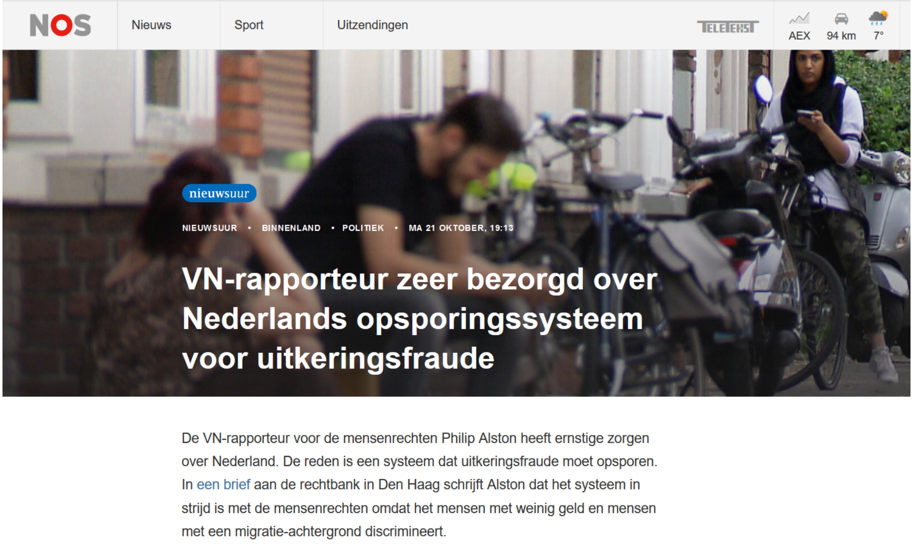


Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

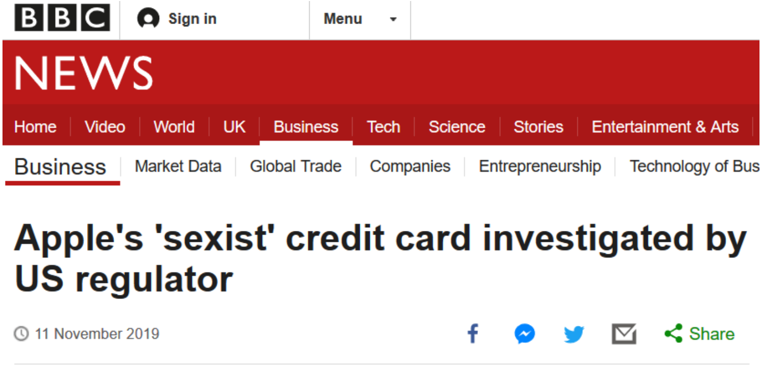
SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.



NIEUWSUUR • BINNENLAND • POLITIEK • MA 21 OKTOBER, 19:13

VN-rapporteur zeer bezorgd over Nederlands opsporingssysteem voor uitkeringsfraude

De VN-rapporteur voor de mensenrechten Philip Alston heeft ernstige zorgen over Nederland. De reden is een systeem dat uitkeringsfraude moet opsporen. In een brief aan de rechtbank in Den Haag schrijft Alston dat het systeem in strijd is met de mensenrechten omdat het mensen met weinig geld en mensen met een migratie-achtergrond discrimineert.



Apple's 'sexist' credit card investigated by US regulator

11 November 2019

Complex systems raise concern

- Why this ad?
- Why this discount?
- Why this recommendation?

- Why was I rejected?
- Can I change the outcome?
- When will the system fail?

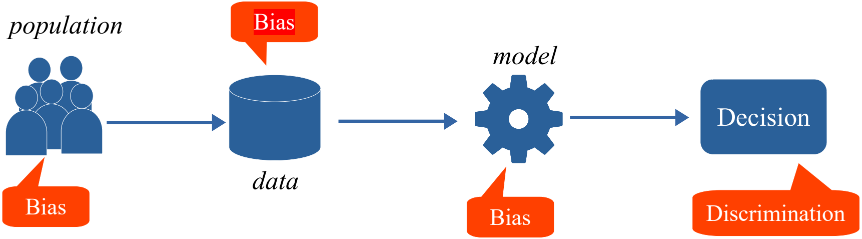
> “Algorithms are written and maintained by people, and machine learning algorithms adjust what they do based on people’s behavior. As a result algorithms can reinforce human prejudices.”

C.C. Miller. When algorithms discriminate, NYT, 2019.

What is bias?

- Systematic errors that create unfair outcomes
- Sources: algorithm design, biased data collection or selection
- Algorithms learn and perpetuate bias

Bias occurs throughout the algorithmic pipeline



population

Bias

Bias

data

model

Bias

Decision

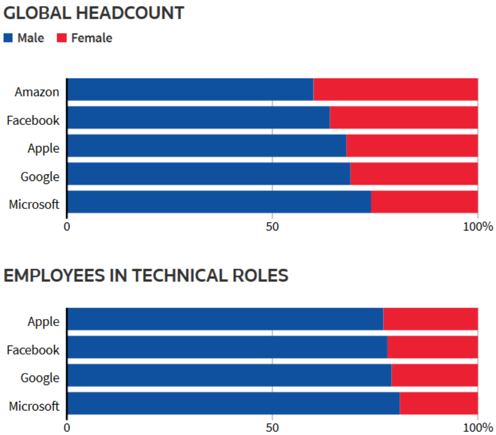
Discrimination

Types of bias

- **Historical bias** reflects structural societal issues
- **Representation bias** certain groups are under-represented in the training data
- **Measurement bias** training data are proxies for some ideal features and labels

simplifed from Suresh & Guttag. A Framework for understanding unintended consequences of machine learning, 2019.

Historical bias

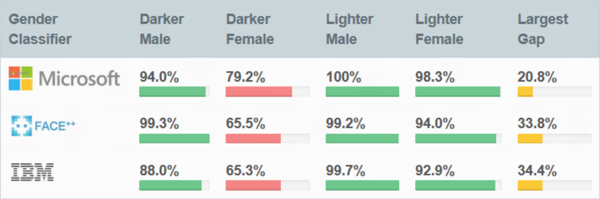


GLOBAL HEADCOUNT

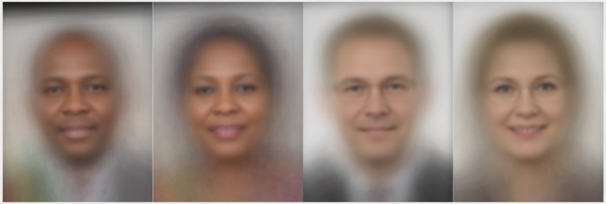
EMPLOYEES IN TECHNICAL ROLES

Note: Amazon does not disclose the gender breakdown of its technical workforce.
Source: Latest data available from the companies, since 2017.
By Han Huang | REUTERS GRAPHICS

Representation bias

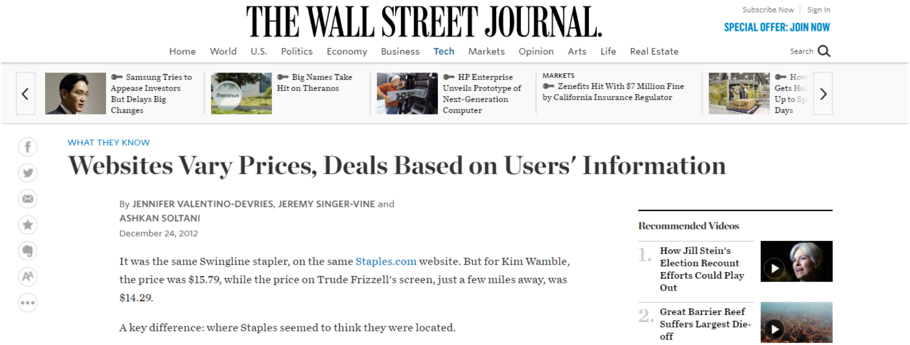


Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



http://gendershades.org/overview.html

Measurement bias



THE WALL STREET JOURNAL.
Subscribe Now | Sign In
SPECIAL OFFER: JOIN NOW
Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Estate
Search

Samsung Tries to Appease Investors But Delays Big Changes

Big Names Take Hit on Theranos

HP Enterprise Unveils Prototype of Next-Generation Computer

MARKETS

Zenefits Hit With $7 Million Fine by California Insurance Regulator

How Gets Ho Up to Sp Days

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.

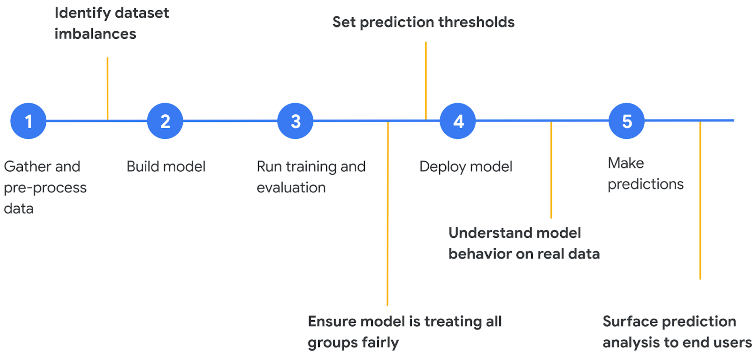
Recommended Videos

1. How Jill Stein's Election Recount Efforts Could Play Out

2. Great Barrier Reef Suffers Largest Die-off

How can we avoid bias?

Best practice



Identify dataset imbalances

Set prediction thresholds

1 Gather and pre-process data

2 Build model

3 Run training and evaluation

4 Deploy model

5 Make predictions

Understand model behavior on real data

Ensure model is treating all groups fairly

Surface prediction analysis to end users

Source: `https://ai.google/`

Regulation: GDPR

"Data subjects have a right to **meaningful information** about the **logic involved** and to the significance and the **envisaged consequence** of automated decision-making"

Ethics

European Commission

INDEPENDENT

HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE

SET UP BY THE EUROPEAN COMMISSION

ETHICS GUIDELINES FOR TRUSTWORTHY AI

FAT AI community

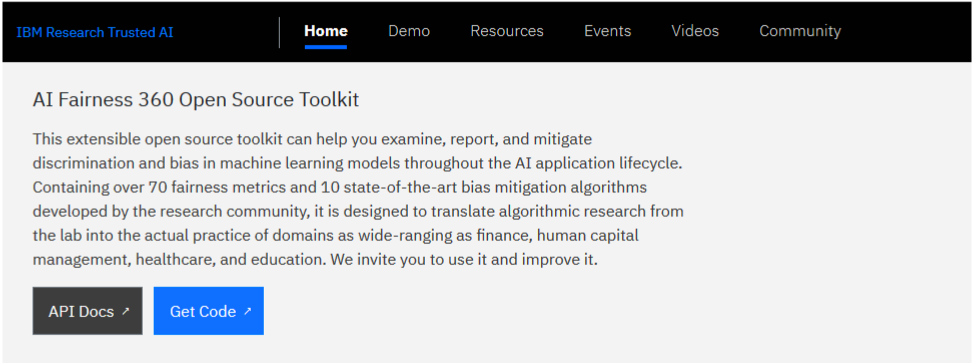
FAT AI: Fairness Accountability & Transparency in AI

ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*)

A computer science conference with a cross-disciplinary focus that brings together researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems.

https://www.fatml.org/

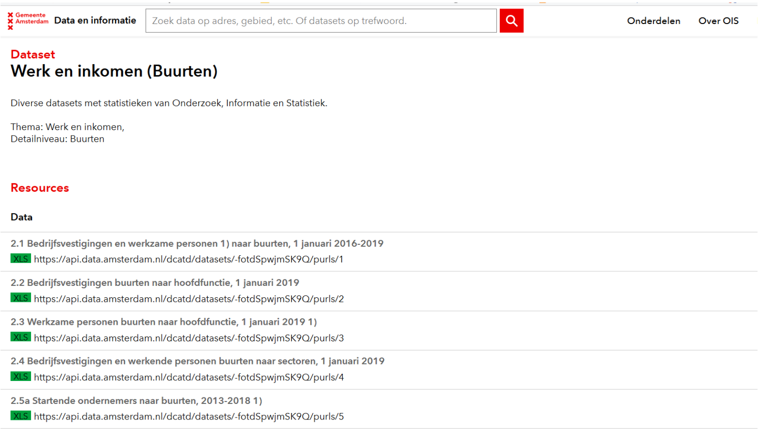
Tools: bias testing and mitigation



AI Fairness 360 Open Source Toolkit

This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. Containing over 70 fairness metrics and 10 state-of-the-art bias mitigation algorithms developed by the research community, it is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education. We invite you to use it and improve it.

Open Data



Dataset

Werk en inkomen (Buurten)

Diverse datasets met statistieken van Onderzoek, Informatie en Statistiek.

Thema: Werk en inkomen,
Detailniveau: Buurten

Resources

Data

2.1 Bedrijfsvestigingen en werkzame personen 1) naar buurten, 1 januari 2016-2019
XLS https://api.data.amsterdam.nl/dcatd/datasets/-fotdSpwjmSK9Q/purls/1

2.2 Bedrijfsvestigingen buurten naar hoofdfunctie, 1 januari 2019
XLS https://api.data.amsterdam.nl/dcatd/datasets/-fotdSpwjmSK9Q/purls/2

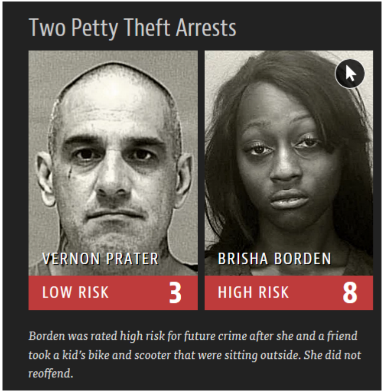
2.3 Werkzame personen buurten naar hoofdfunctie, 1 januari 2019 1)
XLS https://api.data.amsterdam.nl/dcatd/datasets/-fotdSpwjmSK9Q/purls/3

2.4 Bedrijfsvestigingen en werkende personen buurten naar sectoren, 1 januari 2019
XLS https://api.data.amsterdam.nl/dcatd/datasets/-fotdSpwjmSK9Q/purls/4

2.5a Startende ondernemers naar buurten, 2013-2018 1)
XLS https://api.data.amsterdam.nl/dcatd/datasets/-fotdSpwjmSK9Q/purls/5

Open Data

“Through a public records request, ProPublica obtained two years worth of COMPAS scores from the Broward County Sheriff’s Office in Florida. We received data for all 18,610 people who were scored in 2013 and 2014.”



Two Petty Theft Arrests

Borden was rated high risk for future crime after she and a friend took a kid’s bike and scooter that were sitting outside. She did not reoffend.

Challenges

Practical limitations

- Sensitive attributes are unknown
- Regulation constraints
- Realised outcomes are unavailable
- Fairness intervention impact is not monitored over time
- Stakeholders goals vs. fairness goals misaligned

Mitigation algorithms

Methods for fair classification

- Pre-processing: modify the train data
- In-processing: modify the algorithm's objective function to incorporate fairness constraints/penalty
- Post-processing: modifies the predictions produced by the algorithm

Mitigation algorithms

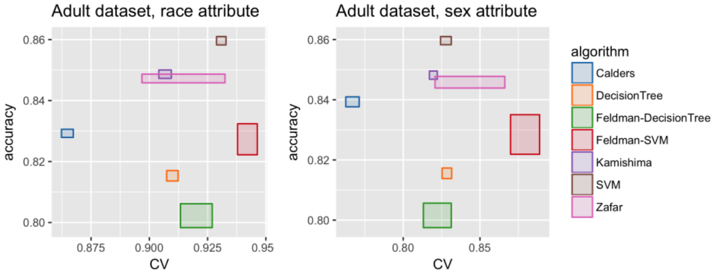
Fairness-preserving algorithms tend to be sensitive to fluctuations in dataset composition, and to different forms of pre-processing

- Adult data set: prediction task is predicting whether an individual makes more or less than $50,000 per year
- Fairness goal: group fairness/statistical parity (equal probability of positive outcomes across groups, 1 is prefect parity)

Friedler et al. A comparative study of fairness-enhancing interventions in machine learning, FAT* 2019

Mitigation algorithms: no easy fix

Each algorithm is tested on 10 random train/test splits and a rectangle centered on the mean, width/height equal to the standard deviation



Friedler et al. A comparative study of fairness-enhancing interventions in machine learning, FAT* 2019.

The way forward

Ask fundamental questions

1. Why do you need AI for this task?
2. Is the system transparent?
3. When and how does the system fail?
4. Who is responsible for the errors?

Algorithmic bias: no quick fix

> "Any real machine-learning system seeks to make some change in the world. To understand its effects, then, we have to consider it in the context of the larger socio-technical system in which it is embedded."

Barocas et al. Fairness and machine learning, fairmlbook.org, 2019.

Interdisciplinary research

Diverse Domains
e.g. aviation, justice

Diverse Users
e.g. pilots, judges

Diverse Criteria
e.g. fairness, privacy

social science – mathematics – computer science – law – ethics

Thank you

`h.haned@uva.nl`

`github.com/hindantation`