UNIVERSITY OF AMSTERDAM

janssen | PHARMACEUTICAL COMPANIES OF Johnson & Johnson

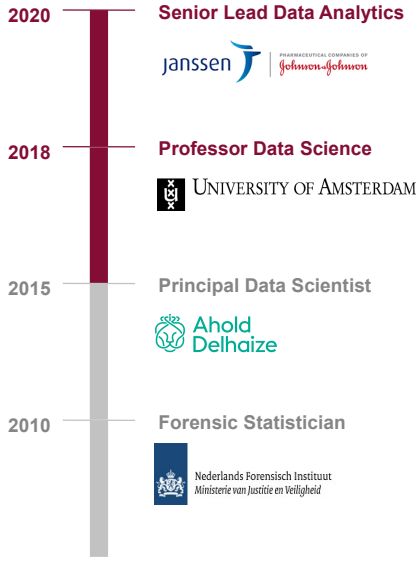
AI algorithms: defining and mitigating bias

Hinda Haned

Januray 20th, 2021

h.haned@uva.nl

About me


2020
Senior Lead Data Analytics
janssen
PHARMACEUTICAL COMPANIES OF
Johnson & Johnson
2018
Professor Data Science
UNIVERSITY OF AMSTERDAM
2015
Principal Data Scientist
Ahold
Delhaize
2010
Forensic Statistician
Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

The Civic AI Lab: partners

UNIVERSITEIT VAN AMSTERDAM

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

An engaging Society where all citizens have equal opportunity to participate and engage in a fair, transparent manner

VISION

Civic AI Lab: mission

To develop AI that promotes economic & social human rights

To develop AI that advances inclusive socio-cultural systems

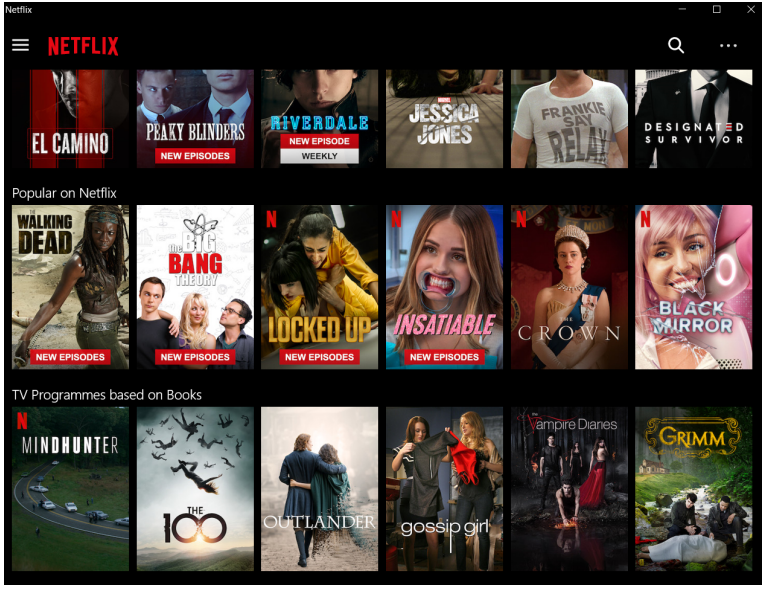
MISSION

Representative Data

Socially Grounded Fairness Measures

Custom-fit Algorithms

How do AI algorithms help us?


Netflix
NETFLIX
EL CAMINO
PEAKY BLINDERS
NEW EPISODES
RIVERDALE
NEW EPISODE
WEEKLY
JESSICA JONES
FRANKIE SAY RELAX
DESIGNATED SURVIVOR
Popular on Netflix
THE WALKING DEAD
NEW EPISODES
THE BIG BANG THEORY
NEW EPISODES
LOCKED UP
NEW EPISODES
INSATIABLE
NEW EPISODES
THE CROWN
BLACK MIRROR
TV Programmes based on Books
MINDHUNTER
THE 100
OUTLANDER
gossip girl
Vampire Diaries
GRIMM

van Hinda | Producten | Bonus | Allerhande Box | Recepten | Winkels | Acties | Meer | Zoeken naar..

Voor jou in de Bonus

30 van je eerdere aankopen zijn in de Bonus

BONUS

1.49

500 g

AH Pitloze witte druiven

BONUS

0.59

per stuk

AH Courgette

1.29

BONUS

1.09

per stuk

AH Avocado

2 VOOR 2.99

1.69

per stuk

AH Mango

BONUS

2.29

140 g

AH Sweet eve frambozen

BONUS

0.59

per stuk

AH Aubergine

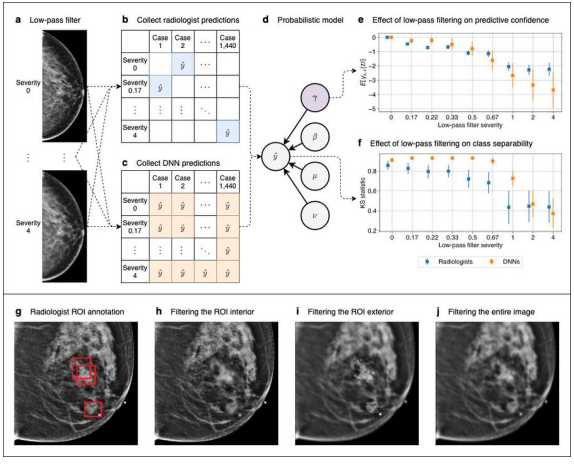
4.08

25% KORTING

3.06

6 x 0,33 l

Coca-Cola Regular



Breast cancer screening

Source: Makino et al, 2020, `https://arxiv.org/abs/2011.14036`

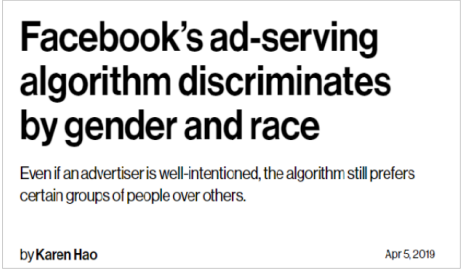

https://gcn.com/articles/2019/07/10/ai-disaster-relief.aspx

Can AI algorithms fail?


NOS
Nieuws
Sport
Uitzendingen
nieuwsuur
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.


2,780 views | Jan 28, 2019, 07:20pm
Forbes
Amazon Refuses To Quit Selling 'Flawed' And 'Racially Biased' Facial Recognition
Zak Doffman Contributor


BBC
Sign in
News
Sport
Reel
Worklife
Travel
Future
NEWS
Home
Video
World
UK
Business
Tech
Science
Stories
Entertainment & Arts
Amazon scrapped 'sexist AI' tool
10 October 2018
Share


Facebook's ad-serving algorithm discriminates by gender and race
Even if an advertiser is well-intentioned, the algorithm still prefers certain groups of people over others.
by Karen Hao
Apr 5, 2019

Algorithmic fairness: why bother?

Compliance

Compliance with General Data Protection Regulation (GDPR – 2018) – article 22

The data subject shall have the **right not to be subject to a decision based solely on automated processing**, including **profiling**, which produces legal effects concerning him or her or similarly significantly affects him or her

GDPR fines & notices

GDPR fines & notices

Date	Organisation	Amount	Issued by	Reason(s)
2020-07-14	Google LLC (Google Belgium)	€600,000	Belgium (GBA/APD)	Failure to respect a citizen's right to be forgotten.
2020-07-06	BKR	€840,000	The Netherlands (AP)	Failing to give access to personal data free of charge, failing to provide easy means of accessing the data, putting unreasonable limits on the number of requests per individual [43]
2019-12-17	Doorstep Dispensaree	£275,000	UK (ICO)	"cavalier attitude to data protection", having left 500,000 patient records in an unsecured location [42]
2019-12-09	1&1 Ionos	€9,550,000	Germany (BfDI)	Insufficient protection of personal data, failing to put "sufficient technical and organizational measures" in place to protect customer data in its call centers. Violation of article 32 of GDPR [41]
2019-10-17	Vueling Airlines	€30,000	Spain (AEPD)	Failing to obtain valid consent to process customer cookies, as per privacy notice. [40]
2019-09-20	Online retailer Morele.net	€645,000	Poland (UODO)	Insufficient protection of personal data, leading to the exposure of data of about 2.2 million people[39]
2019-09-19	Unnamed retailer	€10,000	Belgium (GBA/APD)	Demanding an electronic identity card to create a customer loyalty card.[38]

https://en.wikipedia.org/wiki/GDPR_fines_and_notices

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

What is fairness?

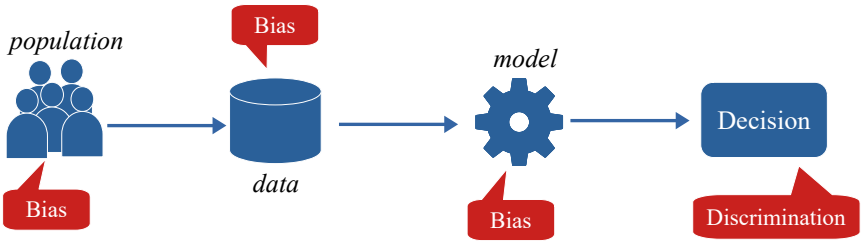
- Fairness is concerned with how outcomes are assigned to particular group of individuals
- Core principles: avoid bias even if it is supported by data, as to avoid the perpetuation of existing discrimination (distributive justice)
- Fairness is a political construct: someone decides to avoid (direct or indirect) harm

Types of harm

- **Harms of allocation** when a system allocates or withholds certain groups, an opportunity or resource. Economically oriented view (e.g. who gets a discount, who gets hired),
- **Harms of representation** systems reinforce the subordination of certain groups along the lines of identity like race, class, gender etc. (e.g. search results biased against a group),
- **Reduced quality of service** when a system produces degraded user/customer experience based on their personal characteristics (e.g stereotyping and denigration).

adapted from: Kate Crawford's NIPS 2017 Keynote presentation: the trouble with Bias & Challenges of incorporating algorithmic 'fairness' into practice: FAccT 2019 tutorial

Bias occurs throughout the algorithmic pipeline

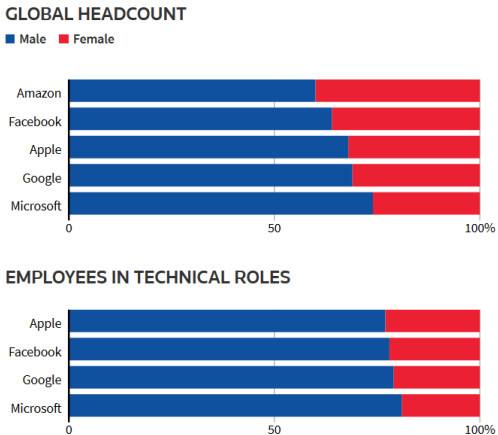


Types of bias

- **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.

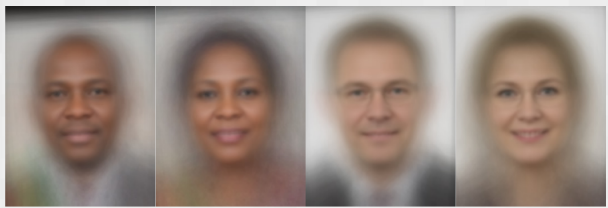
Historical bias



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%

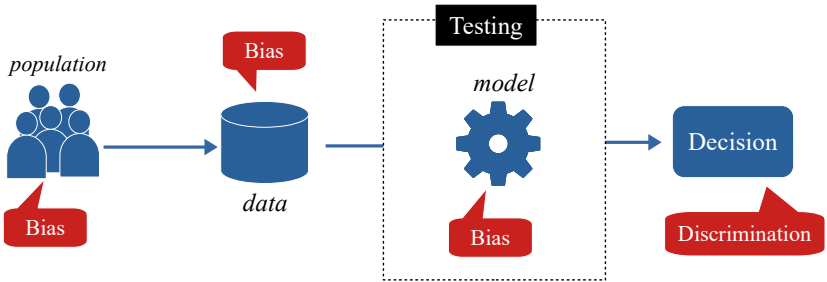


http://gendershades.org/overview.html

Measurement bias

How can we avoid bias?

Best practice



- check datasets imbalances
- ensure model treats all groups fairly
- understand model behavior
- monitor prediction

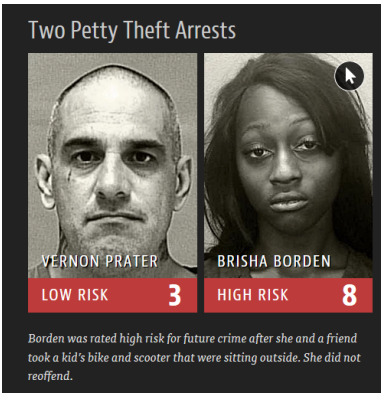
Regulation: GDPR

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

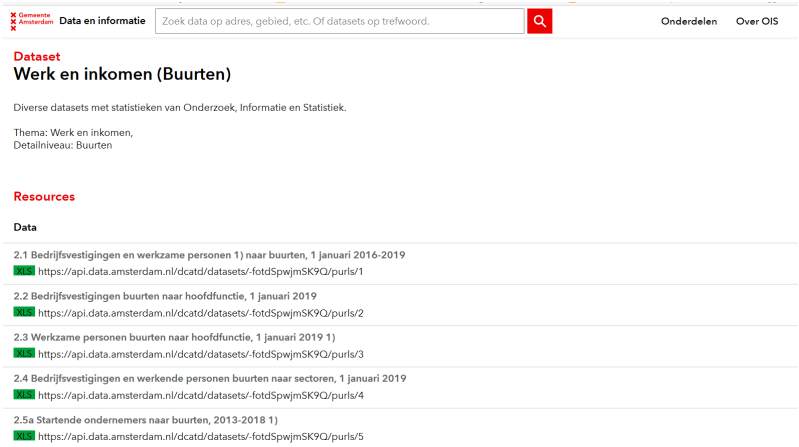
Ethics


European Commission
INDEPENDENT
HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE
SET UP BY THE EUROPEAN COMMISSION
AI
ETHICS GUIDELINES FOR TRUSTWORTHY AI

"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."



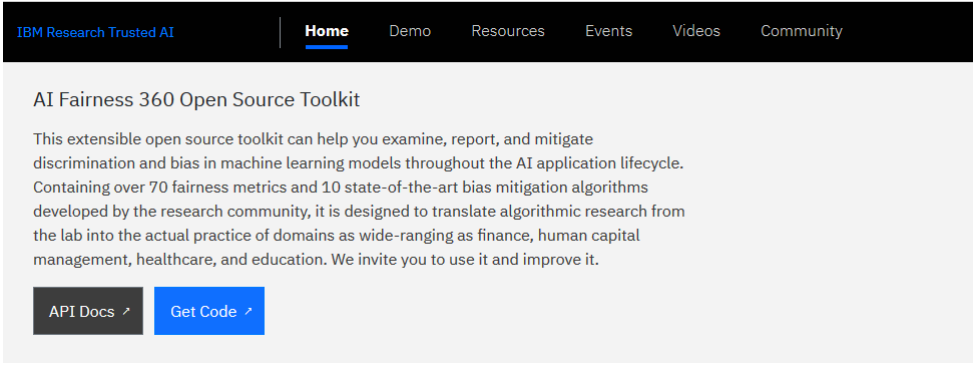
Open Data


Gemeente Amsterdam
Data en informatie
Zoek data op adres, gebied, etc. Of datasets op trefwoord.
Onderdelen
Over OIS
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

https://facctconference.org/

- **Mitigation** The action of reducing the severity, seriousness, or painfulness of something
- **Mitigation algorithms** Algorithms to remove or reduce bias in data and model outputs

Mitigation algorithms



https://aif360.mybluemix.net/

How do we mitigate algorithmic bias in practice?

Mitigating algorithmic bias

- There is no unifying framework to tackle algorithmic bias testing and mitigation
- In most use cases, mitigation is performed after a system is built and decisions have been made based on this system

- **Model:** classifier trained on historical data to predict when a check is relevant (yes/no)
- **Utility:** reveal mistakes/thefts while preserving customer experience

How can we build a model that is useful to the business and fair to its customers?

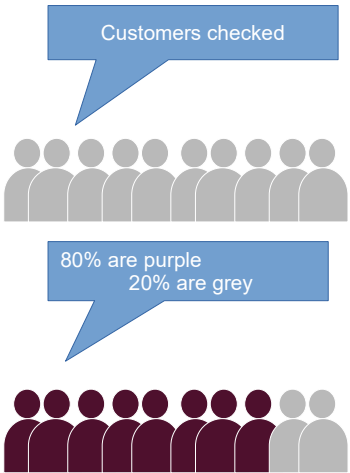
Questions

- What are the potential harms that could occur in this case?
- What is a fair outcome? What is an unfair outcome?
- How can we ensure fair outcomes?
- Who is responsible for ensuring fairness?
- What are the responsibilities: system owner, store manger, data scientist, customer, legislator?

Fundamental questions

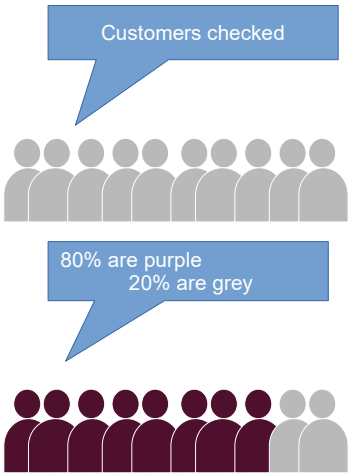
1. Why do you need AI for this task?
2. Is the system transparent?
3. When and how does the system fail?
4. Who is responsible for the errors?

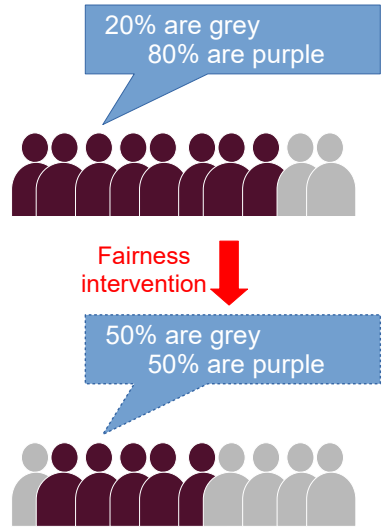
Business objective: avoid harm



Avoid a degraded customer experience for a given group

Business objective: avoid harm





Bias mitigation steps

1. Identify sensitive attribute
2. Define fairness objective
3. Choose fairness intervention
4. Monitor fairness intervention

1. Identify sensitive attribute

- Sensitive attributes such as age, gender or ethnicity are not directly encoded in the data but they can strongly correlate with other features, such as purchase behaviors
- Working assumption: sensitive attributes are unknown but underlying distributions are accessible for fairness testing

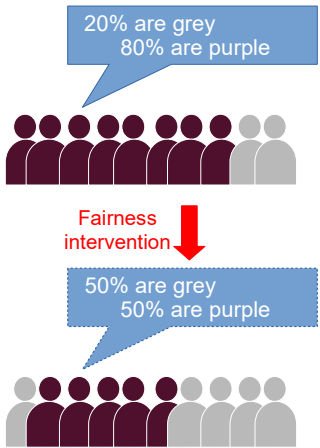
2. Define fairness objectives

Average predictions compared across groups:

- **Statistical parity** subjects in both subgroups have equal probabilities of being assigned to the positive predictive class
- **Conditional statistical parity** extends statistical parity by allowing additional legitimate attributes to affect the outcomes

Verma & Rubin. Fairness definitions explained, FairWare, 2018.

3. Choose fairness intervention: requirements



- Choice of method is constrained by availability of sensitive features
- Satisfying fairness objective is not enough
- What is the cost of the intervention?

4. Monitor fairness intervention: requirements

- Monitoring fairness interventions as part of the product pipeline might not align with business objectives
- Product pipelines often require simplicity and maintainability
- Added complexity: analyzing customer behavior over time to analyze effect of intervention (opportunity)

Practical challenges

In practice, there are many limitations to testing and correcting for fairness in this case

- Sensitive attributes are unknown and cannot be integrated in the model
- No ground truth available
- Potential fairness intervention impact cannot be monitored
- Stakeholders goals vs. fairness goals are misaligned

Evaluating & mitigating algorithmic bias requires navigating uncertainty.

Going further

- A curated list of fairness in AI resources.
- Challenges of incorporating algorithmic 'fairness' into practice: tutorial at FAccT 2019, Video.
- A detailed overview of statistical definitions in available from this paper.
- Open source toolkit to examine, report, and mitigate algorithmic bias: AI Fairness 360.

Beyond mitigation

> "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.

Thank you

https://hindantation.github.io