Fairness in AI: a practitioner's perspective

MSc Artificial Intelligence

Hinda Haned

January 7th, 2020

h.haned@uva.nl

About me

2015

Lead Data Scienist

Ahold Delhaize

A.R
ARTIFICIAL INTELLIGENCE FOR RETAIL

2018

Chair Data Science
Special appointment

ILPS

Best practices for safe and responsible applications of machine learning

Lecture goals

This lecture will cover important aspects of fairness in AI from the practitioner's point of view:

- What does it mean to be fair?
- How do we avoid bias in practice?
- What are the challenges of making AI fair in practice?

Fairness in 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.

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.

"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?

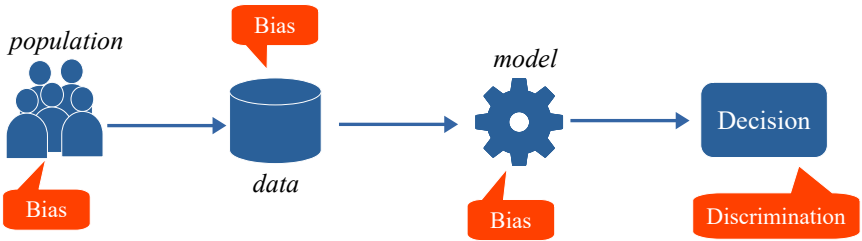
- Fairness is concerned with how outcomes are assigned to particular group of individuals
- Core principles: avoid bias even if it is supported by data, as to avoid the perpetuation of existing discrimination (distributive justice)
- Fairness is a political construct: someone decides to avoid (direct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

Source: Kate Crawford's NIPS 2017 Keynote presentation: the trouble with Bias.

Bias occurs throughout the algorithmic pipeline

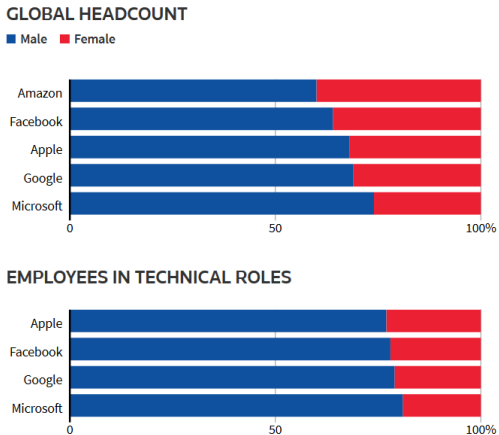


Types of bias

- **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.

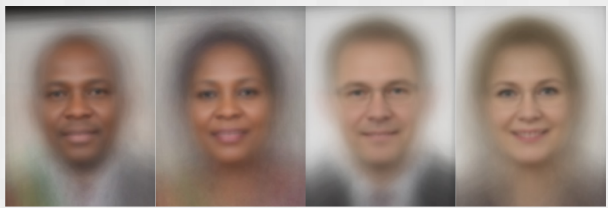
Historical bias



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

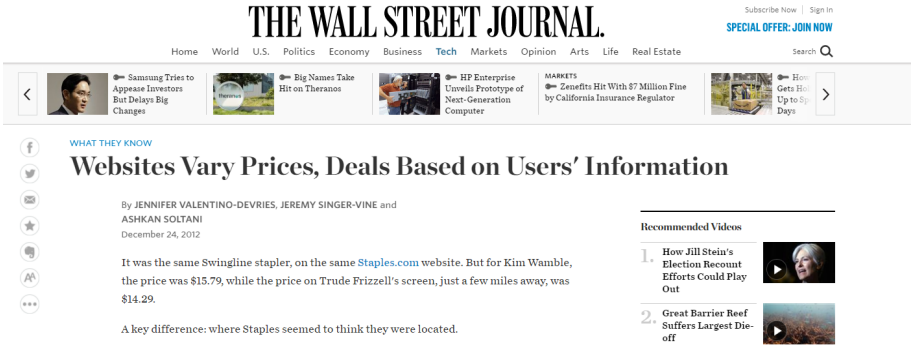
Representation bias

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



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.

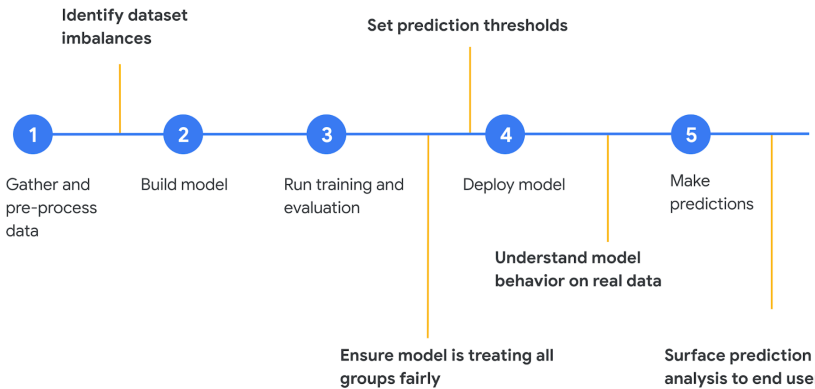
Recommended Videos

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



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

Regulation: GDPR

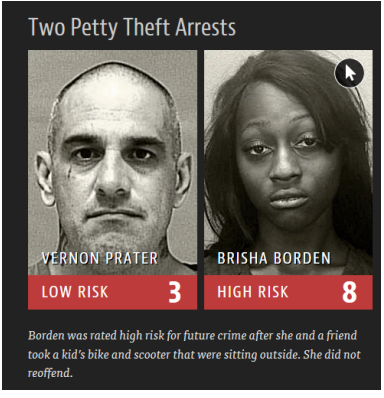
"Data subjects have a right to **meaningful information** about the **logic involved** and to the significance and the **envisaged consequence** of automated decision-making"

Ethics


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

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."



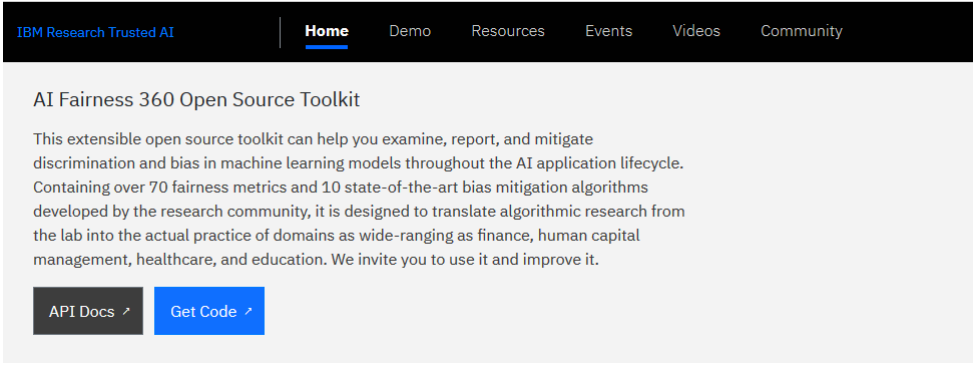
FAT AI: Fairness Accountability & Transparency in AI

https://www.fatml.org/

Mitigation algorithms

- **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/

Mitigating 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

Mitigating algorithmic bias

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

Define fairness objective

There are more than 20 (statistical) definitions for fairness:

- Defined based on predicted outcomes, e.g. credit score predicted as good/bad
- Defined on predicted and actual outcomes, e.g, sentencing based on high or low risk of recidivism vs. actual observations after prison term

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

Fairness metrics

- Defined on predicted and actual outcomes
  - Predictive parity
  - False positive error rate balance
  - False negative error rate balance
  - Equalised odds
  - Conditional use accuracy equality
  - Overall accuracy equality
  - Treatment equality
- Defined based on predicted outcomes only
  - Group fairness or statistical parity
  - Conditional statistical parity

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

Mitigation or fairness algorithms have only been developed for classification tasks:

- 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

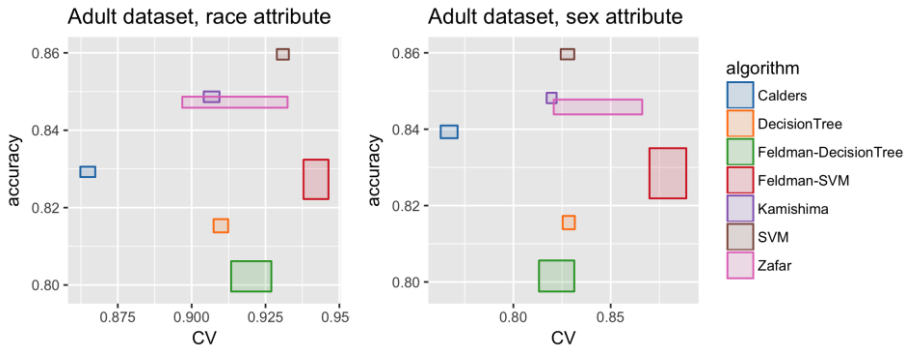
Mitigation algorithms

Pre-processing	Re-weighing (Kamiran & Calders, 2012)
	Optimized pre-processing (Calmon et al., 2017)
	Learning fair representations (Zemel et al., 2013)
	Disparate impact remover (Feldman et al., 2015)
In-processing	Adversarial debiasing (Zhang et al., 2018)
	Prejudice remover (Kamishima et al., 2012)
Post-processing	Equalized odds post-processing (Hardt et al., 2016)
	Calibrated eq. odds postprocessing (Pleiss et al., 2017)
	Reject option classification (Kamiran et al., 2012)
	Fairness-focused regularization (Kamishima et al, 2019)
	Two Naive Bayes (Calders & Verwer, 2010)

Limitations of mitigation algorithms

- Sensitive attributes are known
- Ground truth or observable outcomes are available
- For all methods, there is a trade-off to be found between utility and a desired measure of fairness
- Fairness interventions might cause harm (Liu et al, 2017)
- Fairness-preserving algorithms tend to be sensitive to fluctuations in dataset composition, and to different forms of pre-processing (Friedler et al, 2019)

Illustration: limitations of mitigation algorithms



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.

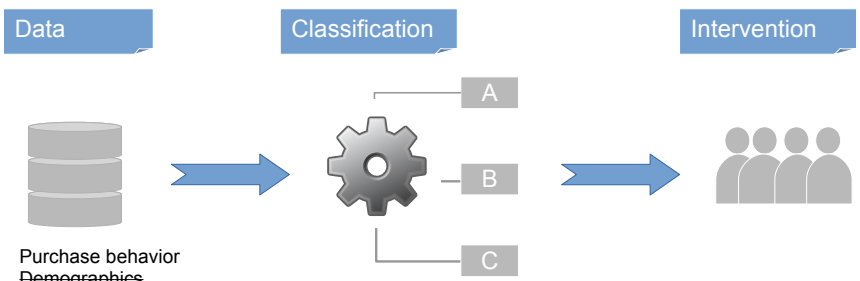
Illustration: mitigating bias in retail

Background

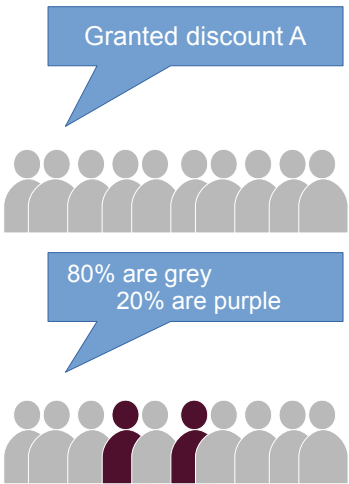
- Businesses increasingly aware of the potential for bias and its impact
- Objective: develop a non-biased decision-making process while preserving as much as possible the quality of the decision

Example: customer segmentation

- Challenge: demographic data cannot be used explicitly (GDPR)
- Objective: avoid harm of allocation



Business objective: avoid allocation harm



Avoid harm of allocation: withholding opportunity from a given group

Reminder: 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

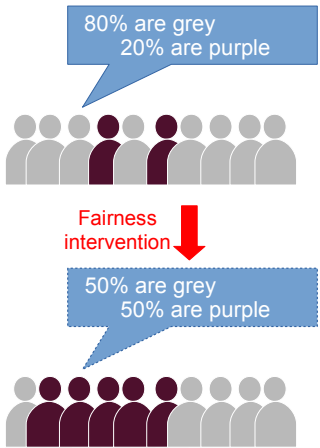
2. Define fairness objectives

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



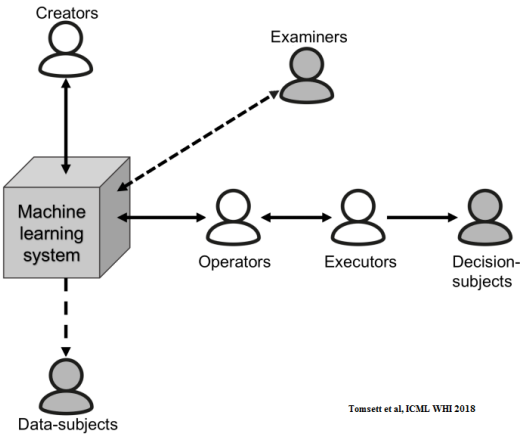
- Choice of method is constrained by availability of sensitive features
- Satisfying fairness objective is not enough
- What is the cost of the intervention?

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: analysing customer behavior over time to analyse effect of intervention (opportunity)

Lessons learned

Manage your stakeholders



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?

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

h.haned@uva.nl
github.com/hindantation