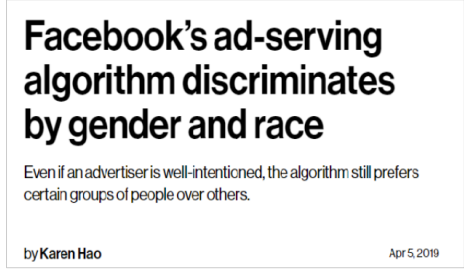
Defining and mitigating algorithmic bias: a practitioner's perspective

Hinda Haned

April 13th 2021

University of Amsterdam
h.haned@uva.nl

Complex systems raise concern



What is bias?

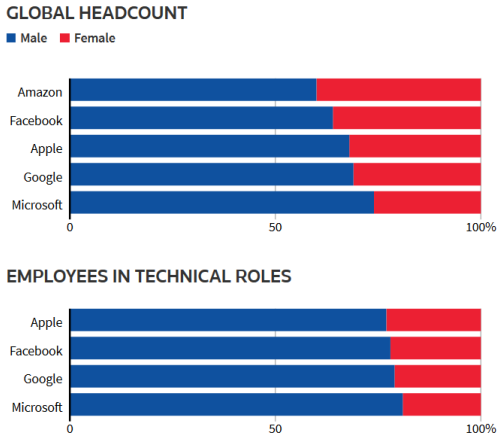
- Systematic errors that create unfair outcomes
- Sources: algorithm design, biased data collection or selection
- Algorithms learn and perpetuate bias

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.

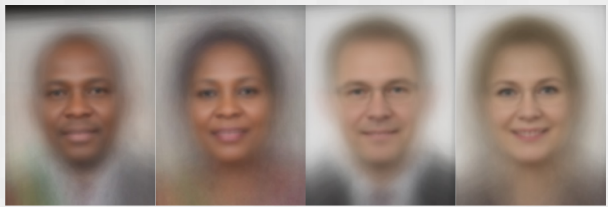
Historical bias



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

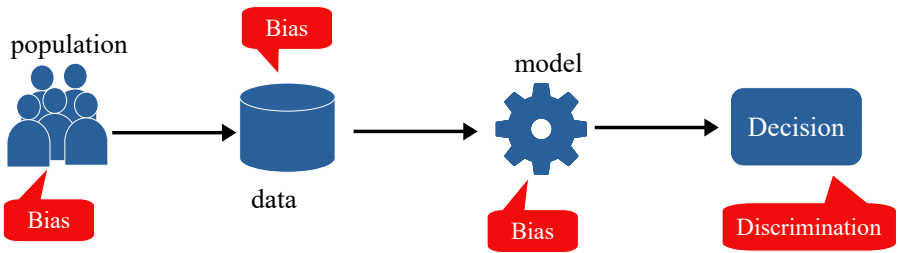
Measurement bias

How 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

Bias occurs throughout the data science pipeline



Mitigation algorithms

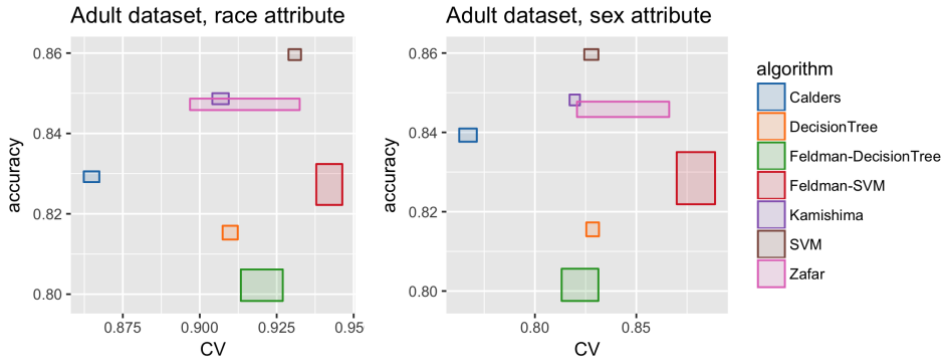
Mitigation or fairness algorithms have only been developed for classification tasks:

- 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

Limitations of mitigation algorithms

- Unrealistic assumptions: sensitive attributes are known & ground truth or observable outcomes are available
- Trade-off: utility vs. desired (quantifiable) measure of fairness
- Reliability: sensitivity to fluctuations in dataset composition, and to different forms of pre-processing (Friedler et al, 2019)

Illustration: limitations of mitigation algorithms



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

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.

Mitigation vs. Fundamental Questions

- Why do you need an automated system for this task?
- Is the system transparent?
- What are the potential harms that could occur?
- What is a fair outcome? What is an unfair outcome?
- When and how does the system fail?
- Who is responsible for the errors?

Thank you

https://hindantation.github.io

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)