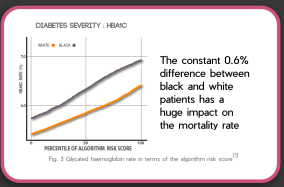
DEFINITION : by machine learnt scoring systems, we describe models using machine learning (ML) frameworks to assign a score to entities (here individuals)

BIAS IN MACHINE LEARNT SCORING SYSTEMS

- HOW IT CREEPS IN AND WHY IT PERPETUATES RACIAL DISCRIMINATION -

An analysis through two characteristic examples :
The Health and Judicial Systems in the United States



The constant 0.6% difference between black and white patients has a huge impact on the mortality rate

Fig. 3 Glycated haemoglobin rate in terms of the algorithm risk score[7]

I. HOW BIAS CREEPS IN THE MODEL

Data collection
- The data is often unrepresentative of the context in which it will be applied, marginalising minorities
- Often reflects existing prejudices

Choosing what to optimise for
- Harder than we'd think as optimizing for results and fairness actually leads to worse scoring systems

Data preparation and choosing the right variables
- Inputs highly influence the model's prediction accuracy
- Really difficult to measure impacts on the model's bias
- We can obviously choose not to add inputs such as someone's origins but it isn't sufficient !

→ Many metrics such as wage are not race-blind metrics[9] due to historical or socio-economic causes and can be a proxy for race

II. WHY BIAS IN ML IS HARD TO CORRECT

Unfitting models
- Normally, the data is split randomly between training data and testing data before training → both have the same biases → the model will hardly be able to flag biased data

AI black box problem
- Contrary to "normal algorithms" machine learning algorithms are hard to interpret → we can't know where the biases come from
- To use a system in important situations (e.g. judging people) it should at least be able to explain its reasoning, be "provably democratic"
- Nowadays the scoring systems function more as "autocracies"[3] giving out scores with little to no explanation

Defining fairness
- Centuries long debated subject by philosophers and lawmakers alike → Many definitions (statistical parity, equality of false negatives..) some being mutually exclusive[10]
→ Even harder to define fairness mathematically (required in ML)

A social context problem : the "portability trap" [10]
- Reusable ML scoring systems ignore social context → can't be developed and implemented in two places (ethnic backgrounds vary)

Model Optimality versus Personal Fairness
- The most important metric is often the error rate and including debiasing in the models would reduce their effectiveness (pareto curve)[15]

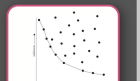


Fig. 1 Pareto graph for «unfairness, error»[15]

Feedback loop
- The algorithms aren't dynamically implemented : they set the statistics in stone → are only a representation of society at one point in time and perpetuate these statistics[13]

JUDICIAL SYSTEM : DEEP ROOTED RACISM

- Systemic racism : racial discrimination rooted in an institution as the fruit of a historical and political process
- 400 days longer sentences for African Americans with similar crimes comparison[13]
- Biased sentencing decisions
- Profiling in the police force and judge's biases : discriminating factors that recidivism prediction algorithms perpetuate

COMPAS : decision support tool using a machine learnt algorithm to assess the probability of a defendant to reoffend

Risk score computed using a form comprised of 137 questions[7]

NORTHPOINTE
COMPAS developer

HOW DATASET BIASES ARE REFLECTED THROUGH WELL-CALIBRATED ALGORITHMS

By defining the score criteria as the recidivism risk, NORTHPOINTE's analysis of COMPAS' outcomes attests of its accuracy compared to human sentencing decisions → no statistical bias.[4]
The African American recidivism score is higher[4] according to COMPAS : why such a disparity ?
- Police profiling, systemic racism and socio-economic inequalities induce dataset bias
→ Systemic racism has to be reduced in the datasets before using them in accurate scoring algorithms

"Powerful algorithms can be harmful and unfair, even when they're unbiased in a strictly technical sense." [4]

"Once you understand the bias in the algorithm, not only do you understand the bias in the humans that shaped the algorithms, but you also have a roadmap to fixing it"
Ziad Obermeyer

HEALTH SYSTEM : COST AND RACIAL DISCRIMINATION

African Americans often face discrimination regarding healthcare services. The US healthcare system is assisted by commercial decision-making algorithms predicting the patient's needs by computing a risk score.

At a given algorithm risk score :
- African American patients tend to be way sicker
- Having a 0.6% higher[7] (Fig. 3) concentration in glycated haemoglobin, African Americans with diabetes have a 30% higher all-cause mortality rate[7]
- They spend 1800$ less[11] than white patients a year

OPTUM : main healthcare decision-making algorithm in the US, applied to 200M [5] **americans a year**
- Doesn't include ethnicity as a parameter but racial biases can still be observed [5]
- The cost being the main variable in the process produces discriminating outcomes[11]

REDUCING TECHNICAL BIAS IN OPTUM

The cost variable inevitably induces social and racial discrimination

Combining health prediction along with cost prediction would reduce 84%[11] of the excess in active chronic conditions in the set of African American patients

Fig. 2 Representation of how a model perpetuates the state of the world through feedback loops [13]

TECHNICAL SOLUTIONS

General advice
- Datasets more representative of the target population are required
- Algorithms to mitigate bias → divided up in 3 types tackling respectively[3] :
1. Pre-processing 2. In-processing 3. Post-processing

One of the solutions : a semi-supervised VAE [16]
- A Variational Autoencoder learns the latent structure of the dataset without manual input → no further human bias (it reduces each data point to a restrained number of defining attributes)
- With this latent representation, the training data is resampled to increase the probability of rarer data and diminishing that of over-represented attributes → "our improvement" : the model could use a disentangled-VAE[16] to further understandability and accuracy.

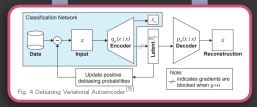


Fig. 4 Debiasing Variational Autoencoder[16]

SOCIAL UNDERSTANDING OF FAIR DECISION-MAKING ALGORITHMS

- Fairness doesn't come with an absolute definition as multiple interpretations can be given on the same already socially conditioned data computed by an algorithm
- It is thus difficult to translate a social reality into a reliable and absolutely fair mathematical model
- However, reducing racial discrimination within social infrastructure would enhance the integrity of the datasets, thus allowing well-calibrated algorithms to achieve accurate results regarding a chosen conception of fairness

TOWARDS A GLOBAL ACKNOWLEDGEMENT : MACHINE LEARNING IN THE EUROPEAN JURISDICTION

- The European Union considers decision-making algorithms as a relevant issue to deal with
- The ethical european charter[8] for the use of AI in Justice suggests 5 principles that public and private actors have to respect, in particular, the non-discrimination principle that warns about the use of "sensible data" in sentencing decisions
- Since 2018 more rights for users regarding automated or intelligent algorithms have been legally mandated by the GDPR

BIBLIOGRAPHY

[1] Obermeyer, Ziad, et al. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." Science, vol. 366, no. 6464, Oct. 2019, p. 447, doi:10.1126/science.aax2342.
[2] Khan, Kar-Tan, et al. "Glycated Haemoglobin, Diabetes, and Mortality in Men in Norfolk Cohort of European Prospective Investigation of Cancer and Nutrition (EPIC-Norfolk)." BMJ, vol. 322, no. 7277, Jan. 2001, p. 3. doi:10.1136/bmj.322.7277.15.
[3] Nawesh Mishrabi, et al. "A Survey on Bias and Fairness in Machine Learning". ArXiv: abs/1908.09635, 2019.
[4] Gong, Alex. "Ethics for Powerful Algorithms (1 of 4)". Medium, 3 Aug. 2016.
[5] https://medium.com/@AlexGong/ethics-for-powerful-algorithms-1-of-3-a060054c2bb0
[6] Julia Angwin, "Machine Bias: Investigating Algorithms Used across America Are Charged with Racial Bias", The Guardian, 23 Oct. 2016. www.theguardian.com. https://www.theguardian.com/society/2016/oct/25/healthcare-algorithm-racial-bias-optum
[7] Corbett-Davies, Sam, et al. "A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks. It's Actually Not That Clear." Washington Post, Oct. 2016. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublica/
[8] Angwin, Julia, et al. "Machine Bias". ProPublica, May. 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[9] Cowx, Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaire". CCE déc. 2018
[10] Knight, Will. "The Financial World Wants to Open AI's Black Boxes." MIT Technology Review, 2 Apr. 2020, www.technologyreview.com/2017/04/13/152980/the-financial-world-wants-to-open-ais-black-boxes/.
[11] Sellist, Andrew D., et al. "Fairness and Abstraction in Sociotechnical Systems." Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19, 2019, pp. 59–68., doi:10.1145/3287560.3287598.
[12] Verma, Sahil, and Julia Rubin. "Fairness Definitions Explained." FairWare '18, 2018, doi:10.1145/3194770.3194776.
[13] Wachter-Boettcher, Sara. "The Ethical Algorithm: the Science of Socially Aware Algorithm Design, by Kearns, Michael and Roth, Aaron." Oxford University Press, 2020, pp. 15–19.
[14] Solon Barocas, Moritz Hardt, & Arvind Narayanan (2019). Fairness and Machine Learning. fairmlbook.org. chapter 1 to 5.
[15] Butrin, Traci. "Skin Color and the Criminal Justice System: Beyond Black-White Disparities in Sentencing." Journal of Empirical Legal Studies, vol. 12, no. 3, John Wiley & Sons, Ltd, Sept. 2015, p. 290–349, doi:10.1111/jels.12077.
[16] Jones, Alexander, et al. "Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure." Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019.
[16] Irina Higgins, et al. "Early Visual Concept Learning with Unsupervised Deep Learning." (2016).

RISK ASSESSMENT


Alexandre Hayderi


Victor Nazianzeno


Léo Theytaz


Javier Garcia

EPFL HUM-122 b.