Ethical issues regarding fairness in prediction algorithms

Universität Hildesheim

Summer Semester 2022-23

Teachers : Prof. Dr. Thomas Mandl, Stefan Dreisiebner

Submitted by: Sruthy Annie Santhosh

International Master's in Data Analytics program

Matr. Nr. : 312213

E-Mail: santhosh@uni-hildesheim.de

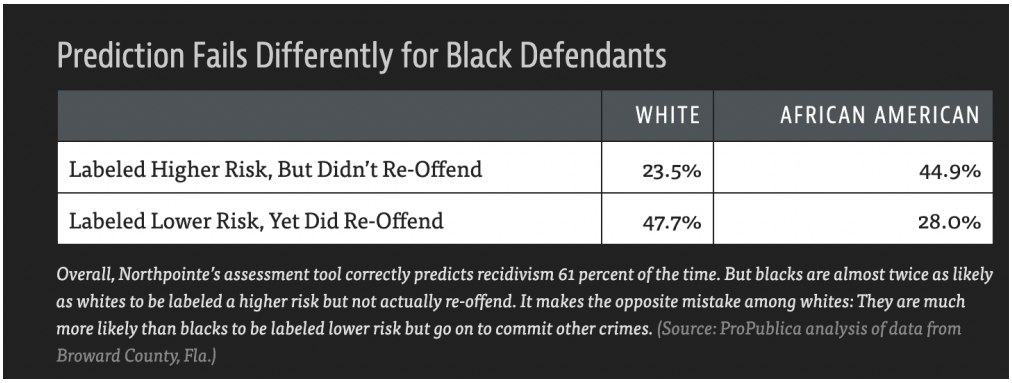
Table of content

1. Introduction

Today, prediction algorithms are being used in many different sectors of life. They are used to provide recommendations, predict weather, predict the changes in stock market and for predicting the demand and sales of products. They are also becoming widely popular in situations which can directly affect an individual's life like in pre-trial release or detention assessment, hiring decisions, loan-lending approval decisions and so on. Thus with prediction algorithms being prevalent in such high stakes decisions, it is important that they provide fair results. We expect algorithms to be objective, fair and mitigate the issues that can arise due to human misconceptions. But there have been many reports that machine learning algorithms can also be biased and unfair. This can be due to various reasons(Mehrabi, 2021).

The term fairness means " *the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics"* (Mehrabi, 2021). But the results of an algorithm is largely dependent on the inherent bias in the data. This can lead to predictions being discriminating towards specific sectors of society. This bias in data can occur during sampling or can be the societal bias which exists due to historical practices. The assessment tool Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS is widely used by the courts in many states of the USA. This tool provides a score for every arrested individual based on which it indicates how much risk is posed by the release of that individual to society. It also shows their chances of re-arrest. Based on this tool, the decision to release before the trial can be

made. But this tool was found to be increasingly discriminating towards Black people. It was found to be twice more likely to flag a Black defendant wrongly for future criminal activities than a white defendant. Also, white defendants were more commonly mislabelled as of low risk (Angwin,2016).



Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Fig 1: The above table shows that the false positive rates of Black people are very much higher than the false positive rates of white. While the false negative rates are very much higher for white people. This table was published by (Angwin, 2016)

Unfairness can also occur from the models, decision space and evaluation metrics being used. Since the field of ' *Fairness in Machine Learning* ' is a relatively new field, these biases can also occur due to an unawareness by the creators of the algorithm. But nowadays more and more importance is given to this concept and hence learning algorithms are being made to correct and regulate these discriminatory and prejudical notions (Stewart, 2020). The purpose of this paper is to delve more into the benefits and drawbacks of the usage of prediction algorithms and the ethical issues surrounding them.

2. Con Arguments

2.1 Bias in Data and Algorithms

One of the main reasons of an algorithm producing a discriminating result is due to the inherent bias in the data population on which it was trained on. Since algorithms are data-driven, it can perpetuate and amplify the biases in the data. These biased outcomes lead to biased decisions which can in turn increase the bias in the data for future learning. For example, for a loan granting selection tool which produces slightly biased results, the model further learns from only those to whom loan was granted. It will not know whether the other section which was excluded may repay the loans or not. Hence the bias will be further perpetuated.

There are mainly two types of bias which occur in data: statistical bias and societal bias. Statistical bias occurs due to non-representation in the data samples collected. It also accounts for any measurement errors which arise during data collection. If there is missing data from one set of population, then the training will produce inaccurate results as algorithms favour majority samples. In the case of pretrial release or detention, the information regarding whether those who were detained will be re-arrested is not present as they were never released(Mitchell, 2018).

Societal bias arises due to existing unfairness in the social structures. These can be due to historical human decisions. In the tool COMPAS, it takes into consideration

whether your parents or relatives have been arrested to calculate your risk score. Historically, Black people were more prone to get arrested due to human misconceptions. Hence due to this they may still have higher scores, even though reality of their actions may be far from true (Angwin, 2016).

Algorithmic bias can also give unfair predictions. Though algorithms do not favour any set of data, they have a tendency for the majority group due to statistical basis. Supervised learning algorithms are by name discriminative in nature because the objective is to minimize the error which favors the majority. Since by definition, there will be lesser data on minority groups, results produced will be worse(Stewart, 2020).

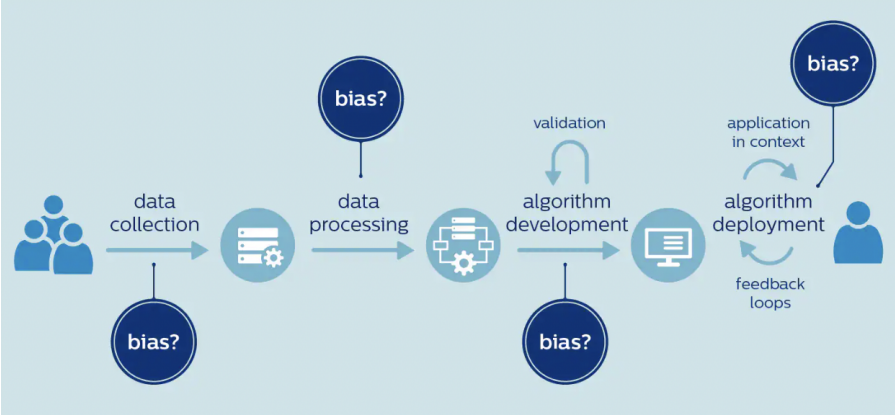


Fig 2: This figure showcases the different biases which occur throughout the pipeline of an algorithm (Henk van Houten, 2020).

2.2 Lack of Transparency and Indirect Relationships

Another main drawback of prediction algorithms is their lack of transparency. Transparency and interpretability are crucial to understand the algorithm and to enhance the trust in the predictions of the algorithm. This will also help keep the

creators and companies behind the algorithm accountable. Statistical validation of the datasets can help mitigate some issues of missing or non-representation of data. The COMPAS tool was developed by a profit organisation and was adopted by many judicial courtrooms before completing proper testing and validation. Transparency at all points in the pipeline of the algorithm like development, training, testing and predictions will help the different stakeholders to understand the outputs (McCradden, 2020). In the case of the COMPAS tool, the stakeholders include the developers and the company behind the tool, the users of tools like the judicial system and the police and the individuals on which the tool is used upon. Hence if tools are not tested rigorously to make sure their outcomes are not biased, then it cannot be fair in nature.

Sometimes prediction algorithms can produce unfair results from indirect relationships with sensitive attributes. Sensitive attributes consist of those variables which can be used to discern between privileged and underprivileged sections of people. These include factors like gender, race, ethnicity and age which are not legally allowed to be used by algorithms. Though they may not be directly influencing the decision, sometimes a proxy attribute can be used which is derived from the sensitive attribute. Hence even though the method seems neutral, it can favour protected members (Pessach, 2022).

An example would be the CV screening tool which was used for hiring applicants in Amazon. The attributes used were educational qualifications and historical data which both seem to be non-sensitive attributes. But it was found to be preferential towards men as historically more men were hired before (MacCarthy, 2019). Another case would be usage of the zip code of the applicants for granting the loans. Even though

this seems to be legitimate, zipcode can lead to racial discrimination as some areas maybe populated like that (Köchling, 2020).

2.3 Equality Vs Equity

There have been many definitions proposed for the notion of 'Fairness' in prediction algorithms. One definition defines a fair algorithm as one that "*provides equal rates of true positives and false positives for protected and unprotected groups* "(equalized odds), while another definition says that "*an algorithm is fair if the protected and unprotected groups have the same positive true rates*" (equal opportunity) (Mehrabi, 2021). These definitions cover a wide range of situations and hence are quite different in their view of fairness. Hence synthesising a unified fairness problem and solution which can be used across different use cases is an issue that needs to be tackled. This is kind of similar to the case of discourse ethics, where the notions of ethics evolve through time and can change depending on the validity of the arguments. Here the different definitions of fairness need to be thoroughly studied to get a compatible solution for all cases (Mehrabi, 2021).

While these definitions revolve around the idea of equality, there is very little research done on the idea of equity. Equity is defined as a '*concept by which each individual or group is given enough resources or opportunities to succeed*'. Equity also needs to be formulated and compared with the current definitions of equality to achieve true fairness (Mehrabi, 2021). For example, denying someone a loan for education can have a very

different impact to denying someone the same amount of loan for a vacation home (Mitchell, 2018).

Another drawback of pursuing a higher degree of fairness will be the trade off in accuracy. Traditionally a model that does not take into account fairness produces better accuracy. Therefore in many cases developers and profit organizations may hesitate in adding additional constraints that can reduce the performance of the model. But new research is being done to mitigate this issue(Pessach, 2022).

3. Pro Arguments

3.1 Control Human Prejudices

Using fairness measures in prediction algorithms human prejudices can be controlled and avoided. While some bias can exist due to data, even human predictions can have bias. But unlike human decisions, machine learning algorithms will provide a consistent prediction across similar groups. Hence rejecting prediction algorithms does not serve the purpose. There are ways in which the data bias can be controlled to an extent, which can help provide similar predictions for similar groups of data. The algorithms will not be prejudiced by any individual's experiences or perceptions (Mitchell, 2018).

For example, sometimes judges can let their personal experiences cloud their judgements. But with the help of an algorithm, scores can be made based on certain

sets of attributes. If the algorithm and data are made transparent, it can ensure a more fair decision making. Another example can be the pre-screening for loan granting tools. No individual can influence the decision of the algorithm. The bankers, people who lend the money or the applicants themselves play no role in the prediction of the algorithm. This can help reduce the amount of corruption and personal favouritism in the system.

3.2 Realistic Distribution of Data

When data is collected thoroughly and properly without any non-representation, it yields a more realistic distribution. As biases do exist in the real-world, it is important to take prior data into consideration. Doing this can help improve fairness in prediction algorithms. For example, while providing recommendations in an e-commerce site, it is beneficial to take into account the age, gender and financial status of the user as it can filter products more effectively. Also in the case of breast cancer detection for females, it was found that the reason for the tool to be discriminating towards Black women was due to the lack of data for that group. By increasing their representation in the data by using race as one of the inputs a more accurate model can be made (Stewart, 2020). Also while some algorithms can reveal an inherent historical bias, it can help make the stakeholders of the situation become aware of the same and steps can be taken to mitigate the same.

In some cases, the discrimination between outcomes maybe explainable. In such cases the differences in results based on a sensitive attribute is legitimate. Let us take the

example of a search engine for finding the images of CEOs. It gives as output more male images than female. But this is because, historically, there are more male than female CEOs. Hence the question remains whether this is fair or unfair. Another example would be the case of predicting the average salary of male and female employees for a company A. It was found that male employees had higher salaries than female employees which can seem to be disparate based on gender. But in reality, average salary is based on the number of hours worked and statistically men have worked more hours than women in company A. Thus predicting same salary would have led to an unfair prediction here which can cause reverse discrimination. Hence in some cases it is beneficial to obtain the realistic distribution of data (Mehrabi, 2021).

3.3 Interventions at Individual and Aggregate levels

Though prediction algorithms can be unfair due to data and algorithmic bias, fairness can be kept in check by providing interventions at individual and aggregate levels. Expanding the decision space to include more supportive interventions can mitigate fairness concerns. Interventions remove disparity and bring about more fairness in the results. The main methods used for the same are (Mehrabi, 2021 & Pessach 2022):

- Pre-processing : Here the inherent bias in data is removed by the algorithm.
- In-processing : These techniques make changes to the algorithm by modifying the objective function or by adding constraints.
- Post-processing : In this technique, some parts of data will be held out for testing purposes. These data will not be used in training. This can help identify the biases in the program.

Also as discusses in the above section, in some cases sensitive attributes can be used for discriminatory results. In such cases, more importance can be given to those variables and training can be done in a way to regulate the problem. They can be used to generalize the algorithm and fairness metrics can prevent it from being discriminatory (Stewart, 2020).

Prediction algorithms also aid in faster decision making. They can be used as a tool to help speed up processes instead of solely relying on them for decision making. By adding interventions checks fairness can be ensured, unlike the case where decisions are solely depended on a group of individuals.

4. Conclusion

In (Stewart, 2020) it is being mentioned that the bias in the prediction algorithms are in majority cases due to ignorance regarding the notions of fairness. Since this is a fairly new area of research, developers were not adequately prepared with the constraints needed to achieve fairness. This can be related to the ethical theory of virtue ethics that explains that humans will naturally do what is good, if they know that it is the right thing to do. Nowadays, ethics has become an integral part of machine learning and many courses are also being held so as to spread more awareness.

As mentioned in the above section, to enhance the trust in the predictions of algorithms, it is necessary for them to be transparent and understandable. Transparency can be

achieved in all stages of the pipeline. If the data collection process is transparent, it can help avoid the non representational bias and understand the societal bias. Transparency can also help improve the accountability of the creators or owners of the algorithm. Transparency is one of the major key points to achieve fairness. Though for proprietary softwares, it cannot be fully achieved as trade secrets or data privacy can be violated, algorithms used in the public sector (pre-trial risk assessment tool) can be made transparent. But care must be taken so as to not violate the data protection and privacy laws (McCradden, 2020).

Like in the case of new medicines and drugs, an ethical review board or certifications can be made mandatory before real life implementation of tools. This can enforce vigorous testing of the tool which can uncover any hidden biases (Stewart, 2020). A major factor which led to the COMPAS tool having a disparate impact was its lack of testing before being sold. If there are regulations to follow and a regulatory body to enforce them, then unfairness can be controlled to an extent. Andrew Tutt talks about this in his paper "*An FDA for Algorithms* "(Tutt, 2017).

Anti-discriminatory laws also include the decisions made by prediction algorithms. Thus these laws prevent disparity in sensitive attributes. They include both direct and indirect discriminations. The **Algorithmic Accountability Act** of 2022, in the US enforces companies to evaluate their automated prediction algorithms for " *inaccurate, unfair, biased, or discriminatory decisions*" (MacCarthy, 2019). Laws like this help prevent and control unfairness in algorithms. The **Artificial Intelligence Act** (2021) proposed by the

EU also subjects high risk assessment tools like CV scanners to legal regulations. This Act can become a global standard which can make AI have a positive rather than negative impact on our lives(The AI Act, 2021).

As discussed in the above sections, many solutions of biased predictions are not transferable across use-cases. This is because each use case may use a different definition of Fairness. Hence it is crucial to design a unified definition so that different solutions can be unified and be transferable across domains. This requires further research and analysis in the field of " *fairness in algorithms*" (Mitchell, 2018).

Thus we can reach the conclusion that prediction algorithms are useful techniques which when exercised under certain regulations and rules can produce fair results.

5. References

1. Angwin, J., Larson, J., Mattu, S., and Kirchner, L.(2016). Machine Bias : There's software used across the country to predict future criminals. and it's biased against blacks. Propublica.org.Bias in Criminal Risk Scores Is Mathematically Inevitable …
2. Houten, H. v. (2020). For fair and equal healthcare, we need fair and bias-free AI . Retrieved from a blog handled by the Philips Company on 27.08.2022 . https://www.philips.com/a-w/about/news/archive/blogs/innovation-matters/2020/20201116-for-fair-and-equal-healthcare-we-need-fair-and-bias-free-ai.html
3. Köchling, A., Wehner, M.C.(2020) Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR

recruitment and HR development. Bus Res 13, 795–848 . https://doi.org/10.1007/s40685-020-00134-w

4. MacCarthy, M. (2019). Fairness in Algorithmic Decision Making. Report published and produced by the Brookings Institution's Artificial Intelligence and Emerging Technology (AIET) Initiative, The Centre of Technology Innovation. https://www.brookings.edu/research/fairness-in-algorithmic-decision-making/

5. McCradden, M. D., Joshi, S., Mazwi, M., Anderson, J. A., (2020) .Ethical limitations of algorithmic fairness solutions in health care machine learning. The Lancet Digital Health Journal, Volume 2, Issue 5,E221-E223. https://doi.org/10.1016/S2589-7500(20)30065-0

6. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1-35. https://arxiv.org/pdf/1908.09635.pdf

7. Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2018). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*.

8. Pessach, D. and Shmueli, E., 2022. A Review on Fairness in Machine Learning. *ACM Computing Surveys (CSUR)*, *55*(3), pp.1-44. https://doi.org/10.1145/3494672

9. Stewart, M. (2020). Programming Fairness in Algorithms. Published by Towards Data Science, a Medium publication. https://towardsdatascience.com/programming-fairness-in-algorithms-4943a13dd9f8

10. Tutt, A.,(2017) An FDA for Algorithms. Administrative Law Review 69, pg 83 . Available at SSRN: https://ssrn.com/abstract=2747994 or http://dx.doi.org/10.2139/ssrn.2747994

11. The Artificial Intelligence Act (2021) by the European Commission, retrieved from the website, is maintained by the Future of Life Institute (FLI) on 27.08.2022. https://artificialintelligenceact.eu/