Fairness Measures in Machine Learning

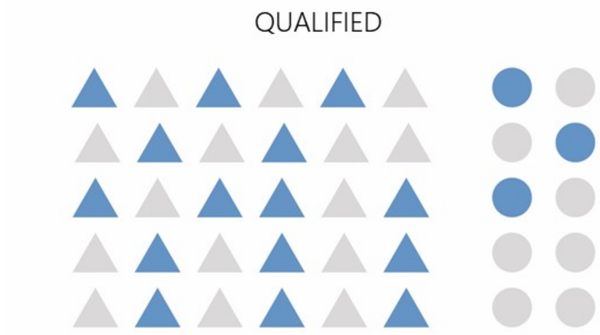
A Legal Perspective

Alesia Vallenas Coronel | Lehrstuhl für Rechtsphilosophie und Öffentliches Recht | Prof. Dr. Friederike Wapler

As the EU Commission's Artificial Intelligence Act is on its way, the pressure rises to ensure the legal compliance of algorithmic decision making systems. These systems are trained with data that often contain a historical bias. The question arises how to exclude or mitigate this bias. Promising approaches to implement ways to measure "fairness" have developed in the machine learning research field. **But how do we define and compute what is "fair"?** This problem may not be solved from a technical point of view alone but must consider a philosophical and legal perspective.

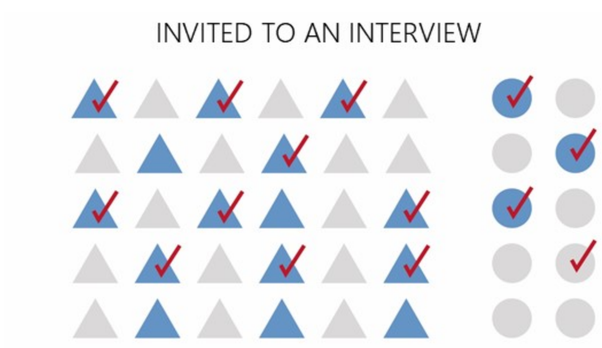
1. Who to hire?

A binary classification model in the context of hiring decisions is used to illustrate how fairness measures can be applied to algorithmic decision making systems. Assume 40 people apply for a vacant nursing position at a hospital, 30 women (triangles) and 10 men (circles). The model is to decide which of the 40 people are invited to an interview and which are not.



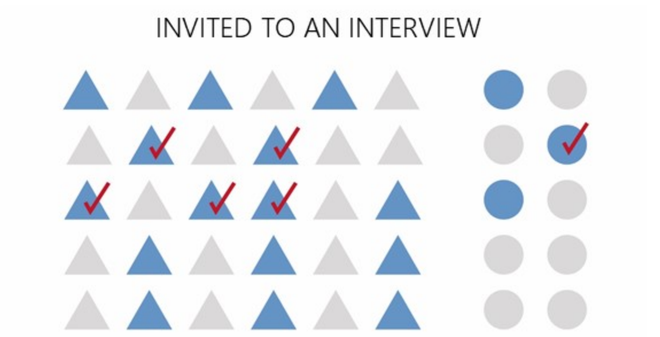
2. Group Fairness

The collection of characteristics of each applicant ("input") forms the basis for the classification ("output") of the model. Consider two classes, the desired outcome "should be invited to an interview" and the undesired outcome "should not be invited to an interview". The system's recommendations are based on a statistical analysis of historical data. One way to measure "fairness" is to evaluate whether groups were treated equally in the classification process (group fairness).



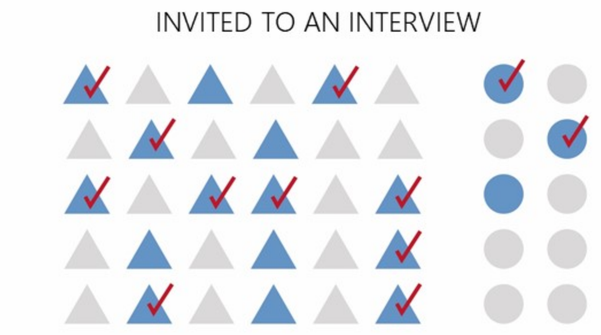
3. Independence

One subcategory of group fairness is to ensure a "fair" distribution of predictions within the two groups. Assume 50% of women and 30% of men are qualified for the job. If the system recommends the same percentage of women and men to be interviewed, it is considered fair. Here, the probability of acceptance of an applicant is 40% for both groups and therefore independent of their gender.



4. Conditional Independence

Assume individuals have to surpass a certain threshold of grade point average to be considered for the vacant position. The model is fair if the invited applicants from those who fulfil this conditional attribute is equal for both groups. Here, a third of both women and men exceeding the threshold are accepted.



5. Separation

The definition of separation is met if two equations are fulfilled. First, the possibility of accepting applicants who are qualified for the job has to be equal for both groups. Second, the possibility of accepting applicants who are actually unqualified for the job is also the same for both groups. Here, the chance of a qualified person being invited to be interviewed is 66% for both groups. While the chance of an unqualified person being invited is 0% for both groups as well.

... is this "fair"? These three examples of group fairness measures seem to connect the idea of fair treatment with some notion of equality. We want groups and individuals to be treated equally in a decision making process. How can this be done in a legally compliant way? This research project analyses fairness definitions and their implementation while considering fundamental rights such as anti-discrimination legislation, data protection law and the upcoming EU Artificial Intelligence Act.

This is part of the interdisciplinary research project "Trading off Non-Functional Properties in Machine Learning" (TOPML) in collaboration with several departments of the University of Mainz, such as the Institute of Computer Science, Mathematics, Philosophy and the University of Applied Sciences Mainz.

REFERENCES: J. Bernstein et al., How to prevent discriminatory outcomes in machine learning, Tech. rep., World Economic Forum (2018); M. Cerrato et al., Fair interpretable representation learning with correction vectors, CoRR, 2022; B. Dattner et al., The legal and ethical implications of using ai in hiring, Harvard Business Review 25 (2019); C. Dwork et al., Fairness through awareness, in: Proceedings of the 3rd innovations in theoretical computer science conference, ACM, 2012, pp. 214–226; M. P. Hauer et al., Computer Law and Security Review 00 (2021), pp. 1–26.