Identifying, Mitigating, and Anticipating Bias in Algorithmic Decisions

Joachim Baumann

University of Zurich, Switzerland
Zurich University of Applied Sciences, Switzerland
baumann@ifi.uzh.ch

Abstract

Today's machine learning (ML) applications predominantly adhere to a standard paradigm: the decision maker designs the algorithm by optimizing a model for some objective function. While this has proven to be a powerful approach in many domains, it comes with inherent side effects: the power over the algorithmic outcomes lies solely in the hands of the algorithm designer, and alternative objectives, such as *fairness*, are often disregarded. This is particularly problematic if the algorithm is used to make consequential decisions that affect peoples lives. My research focuses on developing principled methods to characterize and address the mismatch between these different objectives.

Problem Statement

ML is used in consequential decision-making across various domains, such as recidivism reduction, hiring, and credit lending. However, social biases can be learned and reinforced by algorithms that use historical human data. Different types of bias exist, which can be introduced at every step of the ML pipeline. Fairness has become an essential consideration in algorithmic decision-making due to recent cases highlighting the importance of mitigating unjustified disadvantages toward certain demographic groups (Barocas, Hardt, and Narayanan 2023).

One line of work in ML strives for fairness across different groups, which resulted in many different so-called group fairness criteria. Group fairness criteria are concerned with the impact of algorithmic decision-making on individuals' lives (Mehrabi et al. 2021). The goal is to avoid systematic disadvantages across sociodemographic groups.

Despite the promising progress of ML, there is a caveat: good prediction models do not automatically lead to fairness. Even a perfectly accurate predictor does not guarantee fair outcomes for the affected individuals. However, quantifying the fairness of decision-making systems is not straightforward as there is no universally-accepted definition of what it means for an algorithmic decision-making system to be fair. Any morally appropriate notion of fairness heavily depends on the context and even within a given context reasonable people can disagree on what is "fair". Making matters even more difficult, previous studies have demonstrated that such systems are vulnerable to runaway feedback loops, e.g., when police are repeatedly sent back to the same neighborhoods regardless of the actual rate of criminal activity, which exacerbate existing biases. Understanding those feedback loops is crucial to be able to anticipate biases before they even emerge. Hence, identifying, mitigating, and anticipating bias is essential to build fairer algorithmic systems.

- **Identifying bias:** To identify whether a decision-making system is biased towards a protected group, ML researchers have brought forward various group fairness metrics. The choice of a metric requires a moral assessment (i.e., what does "fairness" mean in a given context?) to find an appropriate mathematical representation of fairness – a so-called fairness metric.
- **Mitigating bias:** Once a morally appropriate metric has been specified to quantify the fairness of an algorithmic decision-making system, one can use this metric to improve the fairness of the system. Thereby, the idea is to use so-called bias mitigation methods as part of the ML development to avoid potential negative impacts on people and society at large.
- **Anticipating bias:** In practice, the automated decisions often have dynamic feedback effects on the system itself that can perpetuate biases over time. Therefore, it is crucial to not only develop short-sighted solutions aimed at identifying and fixing existing biases but to focus on a more long-term-oriented view that strives to anticipate and prevent biases.

Contributions

In Baumann et al. (2023b), we propose a comprehensive bias *identification* framework that links group fairness metrics to more theories of distributive justice. The framework reveals the normative choices associated with standard group fairness metrics and allows for an interpretation of their moral substance. Additionally, the framework expands on standard fairness metrics to address criticisms, including their parity-based nature, the lack of comparison of resulting consequences for different groups, and the insufficient representation of distributive justice literature. We also present an approach to balance the perspective of decision makers and decision subjects in decision-making systems by eliciting and formalizing values from both parties, using well-known

theories of distributive justice to evaluate fairness and the concept of Pareto efficiency to compare decision rules. Our proposed framework can guide decision-making system implementation and aid in audits by highlighting the values embedded in the system.

In Baumann et al. (2023a), we highlight the importance of understanding (and *identifying*) bias in data and its impact on ML based decision-making systems. The primary contribution of the paper is a modeling framework for generating synthetic datasets with different forms of biases, which are used to showcase the interconnection between biases and their effect on performance and fairness evaluations. Additionally, the paper provides insights into mitigating specific types of bias through post-processing techniques. This framework is critical to understanding the impact of bias on ML-based decision-making systems and can aid in developing more fair and equitable systems.

In Baumann, Hannák, and Heitz (2022), we investigate optimal bias *mitigation* techniques to satisfy the fairness concept of predictive parity. We formulate algorithmic fairness as a constrained optimization problem and provide a counter-intuitive result: For a decision maker, it may be optimal to leave out the most promising individuals of a group to generate predictive parity in a globally optimal way. This is in contrast to the analogous solutions for statistical parity or equality of opportunity, where optimal decisions always take the form of threshold rules (Corbett-Davies et al. 2017; Hardt, Price, and Srebro 2016).

In Pagan and Baumann et al. (2023), we investigate feedback loops, such as recurring police deployment to identical neighborhoods irrespective of the actual crime rates, to *anticipate* potential biases. In reality, automated decision-making often induces dynamic feedback impacts on the system that can reinforce over time. This longevity can challenge the effectiveness of myopic design decisions in managing the system's progression. We contribute a categorization of various types of feedback loops that may emerge in ML-based decision-making systems.

Future Work

Beyond short-term fairness My research agenda in fair machine learning primarily focuses on transitioning from a static viewpoint to a more dynamic framework, where addressing long-term fairness becomes essential. This entails expanding and refining our recent work on feedback loops (Pagan and Baumann et al. 2023), by conducting a more thorough and systematic review of the literature.

Shifting power away from the algorithm designer I am also invested in exploring tools designed to resist AI and evaluating their effectiveness in situations where enforcing fairness on algorithm designers is not feasible, often due to the absence of regulatory frameworks. Algorithmic collective action in machine learning is a promising starting point for research that aims at redirecting an algorithm designer's objective towards an alternative solution through pooled resources (Hardt et al. 2023). However, it remains to be seen whether algorithmic collective action can effectively shift power from platforms to users or content creators, especially in contexts involving recommender systems.

Positive social impact The overarching objective of this research goes beyond making theoretical contributions to the field of fair ML. The primary focus is on ensuring a positive societal impact. Therefore, it is crucial to bridge the gap between theoretical advancements in ML and their practical implementation. Our recent paper, which utilizes ML to distribute rental assistance more equitably to individuals at imminent risk of homelessness, underscores several challenges in ethically designing predictive decision support tools (Vajiac et al. 2024). Further research is needed to ensure that these sociotechnical systems are deployed fairly and without disadvantaging certain members of society.

References

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Baumann, J.; Castelnovo, A.; Crupi, R.; Inverardi, N.; and Regoli, D. 2023a. Bias on Demand: A Modelling Framework That Generates Synthetic Data With Bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1002–1013.

Baumann, J.; Hannák, A.; and Heitz, C. 2022. Enforcing Group Fairness in Algorithmic Decision Making: Utility Maximization Under Sufficiency. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2315–2326.

Baumann, J.; Hertweck, C.; Loi, M.; and Heitz, C. 2023b. Distributive Justice as the Foundational Premise of Fair ML: Unification, Extension, and Interpretation of Group Fairness Metrics. arXiv:2206.02897.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806.

Hardt, M.; Mazumdar, E.; Mendler-Dünner, C.; and Zrnic, T. 2023. Algorithmic Collective Action in Machine Learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, 12570–12586.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 3323–3331.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6).

Pagan, N.; Baumann, J.; Elokda, E.; De Pasquale, G.; Bolognani, S.; and Hannák, A. 2023. A Classification of Feedback Loops and Their Relation to Biases in Automated Decision-Making Systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*.

Vajiac, C.; Frey, A.; Baumann, J.; Smith, A.; Amarasinghe, K.; Lai, A.; Rodolfa, K.; and Ghani, R. 2024. Preventing Eviction-Caused Homelessness through ML-Informed Distribution of Rental Assistance. Forthcoming in *Proceedings of the AAAI Conference on Artificial Intelligence*.