Auditing and Validating Fairness and Ethics in Machine Learning Systems

Disa Sariola

Computer Science, Tulane University
New Orleans, LA USA
dsariola@tulane.edu

Introduction

Ethics are embedded in the choices we make every day, whether they are in our personal interactions or when designing computational systems. Ethics should be seen as not just a philosophical pursuit or a legal construct of jurisprudence, instead it is a framework for operating and decision-making. In the context of machine learning, ethical concerns emerge in subtle but rife ways: from how data is collected and labeled, to who defines the contextual fairness. After all, the word 'fair' can mean different things; conforming with the established rules or marked by impartiality and honesty. But, what does it mean in the context of computational frameworks? Despite the frequent framing of ethical concerns and fairness as technical problems, issues like fairness are rooted in value judgments made by real humans. Decisions about inclusion, categorization, and weighting reflect the assumptions and priorities of the people behind the algorithms. At scale, these decisions don't disappear into abstraction. Instead, they shape outcomes, influencing who benefits from, or is harmed by, automated systems. While ethical theory and legal scholarship have long tackled questions of justice and fairness, computer science often treats these topics as secondary to performance metrics or efficiency. These disconnects risk minimizing the real-world stakes of algorithmic decision-making. My research aims to bridge this gap by examining how fairness interventions can be examined from a broader perspective through the lens of algorithmic auditing. I aim to study how fairness interventions efforts align with the stated goals of the audit framework, and whether they meaningfully address the inequities they aim to solve. The goal should not be to reduce fairness to a checkbox, but to understand it in the context of the data and the broader society the system is embedded within.

Central Research Question

Causal machine learning (Schölkopf 2022) offers a framework for understanding not just 'what' a model predicts, but 'why' it arrives at those predictions – whether the explanation lies in the data, the model architecture, or the interactions between the two. This orientation is especially important in fields like algorithmic fairness, where understanding causal relationships is essential for making ethical interventions. These decisions shape data, its curation, preprocessing, and its labeling. These also deeply influence model behavior, the choices are not actually technically neutral (Frické 2015). They are decisions that eventually impact real peoples lives and outcomes. They are value-laden, and often reproduce unexamined assumptions. In this sense, the epistemology of machine learning is tightly linked to ethics. Economists have long used audits to probe for systemic bias. An audit study is a specific type of field experiment primarily used to test for discriminatory behavior when survey and interview questions induce social desirability bias (Gaddis 2018). Originating around the civil rights movement, such studies were instrumental in exposing discriminatory practices and continue to inform legal and policy debates (Reed 2021).

In computer science, however, the use of audit studies and techniques remain underdeveloped when compared to the field of social sciences. Though the language of 'auditing' appears in fairness literature (Barocas, Hardt, and Narayanan 2023; Kearns et al. 2018), its use is often much more superficial, at points divorced from the methodologies that characterize its application in the social sciences. Likewise, much of the work on fairness interventions in machine learning remains focused on outcomes (Lee et al. 2019; Wang, Harper, and Zhu 2020), without focusing as deeply on the causal structure behind them. This emphasis on surface-level outcomes tends to obscure the more fundamental questions:

- What biases are embedded in the data, and which, if any, might be justified?
- Can fairness be reduced to group-level parity, or must it also account for individual-level treatment?

Many algorithmic interventions optimize for statistical fairness metrics under the assumption that large-scale corrections will suffice, what might constitute as the 'law of large numbers'. But this sidesteps the very real tension between group fairness and individual fairness, a tension that becomes especially important when interventions begin to flop under more complex conditions. Our work addresses these gaps through causal inference (Hernán and Robins 2010) tools to audit fairness treatments at the individual level. In our paper "The Illusion of Fairness: Auditing Fairness Inter-

ventions with Audit Studies", we introduce a method based on counterfactual reasoning, i.e. 'What outcome would this individual have received had they belonged to a different subgroup?' Using age-based audit study (Neumark, Burn, and Button 2019) data on hiring callbacks in the United States, we know that older applicants received callbacks at a rate 4.66% lower than their younger counterparts. Crucially, because this audit study controls for all variables except age, we are able to treat these comparisons as causally meaningful.

Leveraging the audit study dataset allows us not only to quantify existing discrimination but also to evaluate the effectiveness, and the unintended consequences, of common fairness interventions. To test these methods, we introduced what we could call 'justified bias' into the data by making a subgroup more qualified w.r.t. the baseline in the audit study data. This intervention was done to simulate a situation of what if you truly have a pool of applicants that are more qualified, in other words – a situation where the bias is fair and desired in the learning paradigm. What we observed was the following: popular fairness treatments, designed to equalize outcomes, began to create discrimination when differences in qualification were real and meaningful. These methods failed to account for individual-level variation, treating fairness as an aggregate property rather than a relational one. By combining causal reasoning with audit study design, we aim to make the case for more context-aware and ethically grounded fairness interventions, ones that do not just smooth statistical disparities but address questions about the origins and legitimacy of those disparities in the first place.

In the current landscape of machine learning, the integration of fairness and ethics remains fragmented and inconsistent. One central issue lies in the treatment of ethical considerations as post-hoc additions to system design. Rather than embedding them into the architecture and objectives of the automated decision making pipeline from the outset, the concerns are often introduced as an afterthought. These retroactive considerations limit the depth and efficacy of ethical intervention, reducing these interventions to a surface-level glaze rather than a foundational component of the system logic. A second critique concerns the dominance of formal fairness metrics as stand-ins for ethical soundness. While statistical measures, such as demographic parity (Feldman et al. 2015) or equalized odds (Hardt, Price, and Srebro 2016), are useful for benchmarking, they frequently abstract away from the socio-technical realities in which algorithms operate. These metrics often reinforce reductive assumptions by relying on protected attributes as the primary lens through which disparity is measured, overlooking the individual experience. Moreover, they do little to account for structural inequalities or to illuminate mechanisms of exclusion embedded within data-generating processes.

A further limitation is the field's insufficient engagement with causality and interpretability when evaluating fairness interventions. Most existing treatments like Equalized Base Rate (Kleinberg, Mullainathan, and Raghavan 2016; Li, Goel, and Ash 2022) approach fairness by centering on distributional outcomes, without justifying the causal pathways that produce them. As a result, the reasons behind observed disparities are ignored, and interventions may inadvertently obscure or even worsen underlying biases. A more effective approach would be to frame fairness through the lens of causal inference, by formalizing and examining the relationships between input variables and predicted outcomes. This shift enables targeted auditing, not merely of the model behavior, but of the data itself, the assumptions, and the contextual factors that shape behavior. Ultimately, addressing these shortcomings requires a more rigorous theoretical and methodological alignment between the goals of fairness and the tools currently used to implement them.

Work To Date and Future Plans

In addition to the work presented in the submitted AIES paper, which addresses foundational challenges in fairness auditing, I have also pursued research on modeling human decision-making under ethical constraints. This includes leveraging Markov Decision Processes (MDPs) (Puterman 1990) and Multi-Alternative Decision Field Theory (MDFT) (Busemeyer and Townsend 1993; Glazier et al. 2022) to simulate and understand decision-making dynamics when agents must navigate trade-offs under normative constraints. Subsequently, my research will build on insights from audit studies, notably their ability to expose bias through a counterfactual structure. One of the central challenges is the lack of audit study data in many real-world scenarios. Due to that, it becomes difficult to determine the extent and distribution of true discrimination. To address this challenge, I am going to investigate how tools from causal inference can be used to approximate audit-like insights in the absence of controlled experiments. My next project explores this in the context of judicial decision-making, within the domain of bond issuance. I intend to examine cases where bond amounts were modified after a request for review, using these instances as a quasi-audit. This data can reveal judgments that were deemed as requiring mitigation. By comparing these cases with those in which bond amounts remained unchanged following appeal, I aim to infer patterns of disparity. The underlying assumption in this analysis is that initial bond determinations are not inherently discriminatory, given the time-constrained nature of judicial hearings, as a judge has approximately one to three minutes for each bond decision (Nworah, Joki, and Farrell 2017). Nonetheless, by treating adjustments in bond outcomes as signals of perceived unfairness, this line of inquiry may offer a middle ground between audit studies and conventional observational data. Through this I hope to refine techniques for evaluating fairness when traditional audit designs are inaccessible.

References

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

Busemeyer, J. R.; and Townsend, J. T. 1993. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, 100(3): 432.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.

Frické, M. 2015. Big data and its epistemology. *Journal of the association for information science and technology*, 66(4): 651–661.

Gaddis, S. M. 2018. An introduction to audit studies in the social sciences. In *Audit studies: Behind the scenes with theory, method, and nuance*, 3–44. Springer.

Glazier, A.; Loreggia, A.; Mattei, N.; Rahgooy, T.; Rossi, F.; and Venable, B. 2022. Learning Behavioral Soft Constraints from Demonstrations. arXiv:2202.10407.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Hernán, M. A.; and Robins, J. M. 2010. Causal inference.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, 2564–2572. PMLR.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Lee, M. K.; Jain, A.; Cha, H. J.; Ojha, S.; and Kusbit, D. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–26.

Li, N.; Goel, N.; and Ash, E. 2022. Data-centric factors in algorithmic fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 396–410.

Neumark, D.; Burn, I.; and Button, P. 2019. Is It Harder for Older Workers to Find Jobs? New and Improved Evidence from a Field Experiment. *Journal of Political Economy*, vol. 127, no. 2].

Nworah, A.; Joki, P.; and Farrell, J. 2017. The cost of buying freedom: strategies for cash bail reform and eliminating systemic injustice. *The Sheller Center for Social Justice*.

Puterman, M. L. 1990. Markov decision processes. *Handbooks in operations research and management science*, 2: 331–434.

Reed, V. M. 2021. Civil rights legislation and the housing status of black Americans: Evidence from fair housing audits and segregation indices. In *The housing status of black Americans*, 29–41. Routledge.

Schölkopf, B. 2022. Causality for machine learning. In *Probabilistic and causal inference: The works of Judea Pearl*, 765–804.

Wang, R.; Harper, F. M.; and Zhu, H. 2020. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–14.