# Safe and Explainable Machine Learning for High-Impact Decisions

**Khadija Zanna**

Rice University
khzanna@rice.edu

## Introduction and Background

High-stakes decision-making domains such as healthcare, finance, and education demand machine learning (ML) models that are not only accurate, but also fair and transparent. Bias in these models can lead to discriminatory outcomes, undermining trust and causing real harm (Hanna et al. 2025). Early research in algorithmic fairness introduced statistical metrics such as demographic parity, equalized odds to quantify group disparities in model outcomes (Dwork et al. 2012; Hardt, Price, and Srebro 2016). These metrics are simple to compute and enforce, but they examine associations among variables, rather than causal mechanisms, and therefore, can fail to capture the underlying reasons for bias (Hinnefeld et al. 2018). Additionally, enforcing such metrics naively may ignore legitimate group differences, particularly in sensitive domains like healthcare, where outcome differences may reflect medically justified factors rather than unfair bias (Barale, Rovatsos, and Bhuta 2025). This limitation has prompted a shift toward causal fairness, which examine data-generating mechanisms to explain why disparities occur (Makhlouf, Zhioua, and Palamidessi 2022). For instance, counterfactual fairness requires that an individual's prediction remains the same in a counterfactual world where a protected attribute (e.g., racial background or gender) is changed (Kusner et al. 2017). Similarly, path-specific fairness distinguishes "fair" causal pathways from "unfair" ones. For example, gender might influence a hiring decision through relevant skills or level of education, but not directly, or through biased proxy variables (Pan et al. 2021). At the same time, legal and policy frameworks define fairness at a high level, for example, by prohibiting the explicit use of protected attributes (disparate treatment) in decision-making, yet translating such principles into ML model design terms remains challenging (Wachter, Mittelstadt, and Russell 2021). The central question driving my thesis therefore is: "How can we design ML systems for high-stake decisions that ensure fairness and accountability in alignment with legal/policy standards, while maintaining high performance and interpretability?". To address this, my research has explored a combination of technical approaches in bias mitigation, and explainability.



## Research Contributions to Date

My research began by addressing bias in a high-impact domain: mental health prediction (Zanna et al. 2022). In my first study, we identified that anxiety prediction models using physiological sensor data exhibited performance disparities across demographic subgroups (e.g., age, ethnicity, and income). We developed a multi-task learning (MTL) framework that integrates uncertainty estimation via Monte Carlo dropout (Gal and Ghahramani 2016) to guide fairness mitigation in this model. We use uncertainty signals to adjust model reliance on potentially bias-inducing features during training, improving performance parity across subgroups. Intuitively, when the model is less certain, its predictions might rely on spurious correlations (related to the sensitive attribute). We trained the MTL model to reduce uncertainty on the main task while preserving higher uncertainty on the auxiliary task, jointly minimizing outcome loss to encourage more robust and fair decision boundaries. In the Anxiety Dataset, this MTL method reduced bias against several demographic groups compared to a baseline model. For example, it narrowed the performance gap between younger and older patients, and mitigated disparate error rates across ethnic groups. Our approach outperformed a standard reweighting technique for fairness (which adjusts data sample weights) in reducing these disparities (Kamiran and Calders 2012). This work demonstrated that incorporating fairness objectives into the training process (via multi-task uncertainty modeling) can yield more equitable models. By analyzing feature importance, we also uncovered meaningful relationships (e.g., certain heart rate variability features were strongly associated with both anxiety and specific demographic factors), highlighting areas where domain experts might need to be cautious about bias.

One key challenge is that improving fairness often degrades accuracy. Building on this notion and the findings from the first study, we applied a Pareto optimization approach to jointly optimize fairness and accuracy (Zanna and Sano 2024). This yielded a range of models representing fairness-performance trade-offs, from which one can choose a balance that meets application requirements. We achieved up to a 20% reduction in false positive rate gaps on datasets like COMPAS, with minimal accuracy loss. Notably, at certain Pareto-optimal points, our model even improved accuracy while increasing fairness, by virtue of better general-

ization through MTL. We also incorporated saliency map explanations to ensure that model behavior is interpretable, visualizing the influence of input features, so that stakeholders can verify that the model's decisions are based on appropriate factors rather than sensitive attributes. This work highlights that with careful modeling of uncertainty while considering performance objectives, it is possible to enhance fairness and transparency in ML models without sacrificing performance.

As my research progressed, it became clear that identifying the source of bias in a model's decision process is crucial to effective mitigation and building trust. We developed a framework that leverages Large Language Models (LLMs) to assist causal discovery (CD) for bias analysis. Traditional CD algorithms (e.g., PC, GES) can infer a causal graph, but struggle with high-dimensional data and often ignore domain knowledge, leading to spurious or missed connections. Our approach used LLM-guided reasoning to evaluate causal plausibility for edges linking protected attributes to outcomes, combining these cues with statistical tests (mutual information, correlation) in a dynamic scoring scheme. The system performed a breadth-first search over possible causal edges, querying the LLM for plausibility of relationships (e.g., "Could gender influence income through education in this context?"). This hybrid method narrowed the search space and improves causal graph accuracy. An active learning loop prioritized the most uncertain or influential edge decisions to query next, reducing query complexity from quadratic to linear. To evaluate fairness sensitivity, we constructed a semi-synthetic benchmark from the UCI Adult dataset, embedding a domain-informed causal graph with injected noise, label corruption, and latent confounding. We assessed how well CD methods including ours, recover both graph structure and fairness-critical paths (e.g., sex $\rightarrow$ education $\rightarrow$ income). Our results from this work showed that LLM-guided methods, including the proposed method, demonstrated competitive or superior performance in recovering such pathways under noisy conditions. We discussed the implications for fairness auditing in real-world datasets. One of such is that CD methods, particularly when applied naively, can introduce spurious pathways or erroneous connections involving sensitive attributes. Across baseline methods, we observed frequent over- or under-attribution of causal influence to variables like sex or age, leading to inflated estimates of structural bias. This is especially problematic in fairness contexts, where false positives (e.g., wrongly inferring a direct path from a sensitive attribute to an outcome) may trigger unjustified policy interventions or obscure true sources of inequity.

## Ongoing and Future Work

The final strand of my thesis addresses operationalizing high-level fairness policies in ML systems. Organizations often declare policies like "the algorithm shall not discriminate based on X," but enforcing such statements in an ML pipeline is non-trivial. My aim is to develop a policy-to-ML pipeline translator that uses causal inference to impose these requirements on model behavior. Policies such as "no direct use of race" map to constraints eliminating direct causal paths, while more nuanced policies specify allowed mediators (e.g., qualifications) for sensitive attributes (Nabi and Shpitser 2018). My system will parse these policies into formal constraints, apply CD, and enforce fairness, drawing on methods like path-specific effect removal or counterfactual data augmentation (Nabi and Shpitser 2018). The modular pipeline includes a natural language parser for policy text to formal constraints (using keywords like "not use," "only through," etc.), a causal modeler to define graph, and an enforcement module that applies interventions at the data, algorithm, or outcome level to satisfy the constraints. The system can take a policy such as "ensure equal opportunity for all groups" and automatically tune a classifier to have balanced true positive rates across those groups (a parity constraint, as per Hardt et al. 2016 (Hardt, Price, and Srebro 2016)), or take "prevent disparate impact" and perform a causal analysis to remove any unjustified indirect bias effects. I will evaluate the system by measuring fairness metrics such as disparate impact, controlled direct effects, and path-specific effects, alongside standard model performance. This evaluation will include real-world datasets and controlled synthetic benchmarks to assess robustness under varying conditions. If undesirable results such as significant accuracy degradation or unresolved proxy influences emerge, the system will generate a report detailing the nature of the trade-offs observed, the fairness constraints applied, and their impact on model utility. This allows stakeholders to make informed decisions about model deployment and cautions against use when fairness objectives cannot be achieved without unacceptable performance loss.

This approach is highly interdisciplinary, connecting legal definitions of discrimination with technical measures, an area highlighted by scholars as needing attention (Alvarez et al. 2024). By the end of this project, I aim to demonstrate a case study where a real-world policy is translated into a model constraint and successfully prevents prohibited bias while preserving model performance. Such a pipeline would provide a blueprint for regulators and industry to audit and align AI systems with societal values in a rigorous, systematic way.

## Conclusion and Significance

In summary, my thesis presents a comprehensive framework for safe and explainable ML in high-impact applications, with a special focus on fairness. I address bias at multiple levels: algorithm design (MTL for bias mitigation), model inference (uncertainty-based fairness optimization), and model interpretability (CD of bias pathways). Collectively, these efforts lay the groundwork for the ultimate goal of my research, which is bridging the gap between high-level fairness principles and practical ML implementation. By using causal inference as the unifying framework between policy and code, I seek to ensure that future AI systems can be explicitly constrained to behave fairly and explain their decisions in terms of cause and effect. This brings us closer to AI that is not just experimentally fair on certain datasets, but compliant with ethical and legal standards in deployment, along with explainability.

## References


Alvarez, J. M.; Colmenarejo, A. B.; Elobaid, A.; Fabbrizzi, S.; Fahimi, M.; Ferrara, A.; Ghodsi, S.; Mougan, C.; Papageorgiou, I.; Reyero, P.; et al. 2024. Policy advice and best practices on bias and fairness in AI. *Ethics and Information Technology*, 26(2): 31.

Barale, C.; Rovatsos, M.; and Bhuta, N. 2025. When Fairness Isn't Statistical: The Limits of Machine Learning in Evaluating Legal Reasoning. *arXiv preprint arXiv:2506.03913*.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

Hanna, M. G.; Pantanowitz, L.; Jackson, B.; Palmer, O.; Visweswaran, S.; Pantanowitz, J.; Deebajah, M.; and Rashidi, H. H. 2025. Ethical and bias considerations in artificial intelligence/machine learning. *Modern Pathology*, 38(3): 100686.

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

Hinnefeld, J. H.; Cooman, P.; Mammo, N.; and Deese, R. 2018. Evaluating fairness metrics in the presence of dataset bias. *arXiv preprint arXiv:1809.09245*.

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1): 1–33.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Makhlouf, K.; Zhioua, S.; and Palamidessi, C. 2022. Survey on causal-based machine learning fairness notions. arXiv. *arXiv preprint arXiv:2010.09553*, 10.

Nabi, R.; and Shpitser, I. 2018. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Pan, W.; Cui, S.; Bian, J.; Zhang, C.; and Wang, F. 2021. Explaining algorithmic fairness through fairness-aware causal path decomposition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1287–1297.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2021. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, 41: 105567.

Zanna, K.; and Sano, A. 2024. Enhancing Fairness and Performance in Machine Learning Models: A Multi-Task Learning Approach with Monte-Carlo Dropout and Pareto Optimality. *arXiv preprint arXiv:2404.08230*.

Zanna, K.; Sridhar, K.; Yu, H.; and Sano, A. 2022. Bias reducing multitask learning on mental health prediction. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 1–8. IEEE.