Fair and Trustworthy AI for Modern Algorithmic Systems

The central goal of my PhD research is to **construct** and **audit** fair and trustworthy AI systems. I will discover, implement, and proliferate algorithms which not only achieve good performance, but also address fairness, reliability, and safety while simultaneously building auditing tools and understanding fundamental limits of access levels. To achieve these goals, my research is focused on the following key areas broadly related to trustworthy and fair AI.

Proposed Research Questions

(1) In the context of **black-box API systems**, what is the power and limit of algorithmic auditing and model post-processing for fairness, robustness, or trustworthiness? [4]

(2) How do we investigate the fairness, trustworthiness, and transparency of algorithmic systems **utilizing machine learned predictors**, such as recommender/ranking systems as well as two-sided marketplaces? [1, 2]

Methodology Summary. In order to understand the limits of black-box post-processing of models, I plan on extending recent black-box optimization methods such as [6, 5] to closed-source language models. In addition, I also plan to utilize recent theoretical work within *interactive proof systems* [3] in order to build novel auditing techniques for this same setting. To study algorithmic systems *utilizing* machine learned predictions, I will mainly rely on novel theoretical analysis, also conduct empirical algorithmic audits of large systems utilizing ML predictions, such as LinkedIn [7] and Amazon shopping.

Analysis. Together, my research addresses (1) the shift towards powerful and closed-source models and LLMs; and (2) integration of AI and ML into larger algorithmic systems. I firmly believe that investigating fundamental theoretical limitations of algorithms and black-box models — as well as designing new algorithms which have guaranteed performance, fairness, and robustness — is an important research direction for the current and future applications of AI and ML in high stakes decision-making settings.

Characters: 1976

References

[1] S. Devic, D. Kempe, V. Sharan, and A. Korolova. Fairness in matching under uncertainty. In *International Conference on Machine Learning (ICML)*, 2023.

[2] S. Devic, A. Korolova, D. Kempe, and V. Sharan. Stability and multigroup fairness in ranking with uncertain predictions. In *International Conference on Machine Learning (ICML).*, 2024.

[3] S. Goldwasser, G. N. Rothblum, J. Shafer, and A. Yehudayoff. Interactive proofs for verifying machine learning. In *12th Innovations in Theoretical Computer Science Conference (ITCS 2021)*, 2021.

[4] D. Hansen, S. Devic, P. Nakkiran, and V. Sharan. When is multicalibration post-processing necessary? *arXiv preprint arXiv:2406.06487*, 2024.

[5] U. Hébert-Johnson, M. Kim, O. Reingold, and G. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning (ICML)*, pages 1939–1948. PMLR, 2018.

[6] G. Hiranandani, J. Mathur, H. Narasimhan, M. M. Fard, and S. Koyejo. Optimizing black-box metrics with iterative example weighting. In *International Conference on Machine Learning*, pages 4239–4249, 2021.

[7] LinkedIn. Reimagining hiring and learning with the power of ai, Oct 2023. URL https://www.linkedin.com/business/talent/blog/talent-acquisition/reimagining-hiring-and-learning-with-power-of-ai.