Systems utilizing machine learning (ML) will increasingly interact with individuals from the broader population. Companies have already begun using ML models for hiring (LinkedIn, 2023), colleges and universities may have started using LLMs to sift through applications (Lira et al., 2023), and the US government has put out notices on the importance of investigating the role that AI will play in society (Biden, 2023). Nonetheless, most current systems utilizing some form of ML have demonstrated sub-optimalities, instabilities, or unfairness (Islam et al., 2022). Although the potential for ML and AI systems to transform the way or society operates is large, there remains ample need for research into improving the trustworthiness of AI models by investigating the way that these systems operate and interact with the underlying population.

The central goal of my PhD research is to make AI systems more trustworthy via investigating fairness, stability, and robustness. I will discover and implement algorithms which not only achieve good performance, but also address reliability and safety while simultaneously deepening our understanding of the underlying mechanisms of learning. To achieve my goals, I plan on researching three key areas related to trustworthy AI and ML:

1.) Algorithmic fairness and stability of systems utilizing machine learned predictors, such as recommender systems, rankings, and two-sided marketplaces;

2.) Proposing and developing fair, robust, and optimal ML algorithms; and

3.) Extending trustworthy AI auditing and post-processing algorithms to large language models.

Rankings and Recommendations. One of the most ubiquitous areas in which algorithms interact with society is through ranking and recommendation systems. This is especially salient through interaction with content platforms which are constantly performing ranking and recommendation tasks (Gottfried, 2024). My ongoing work to in this area is centered on ensuring that ranking and matching systems utilizing machine learned predictions are both *fair* and *stable* when uncertainty is present. This setting is important since many algorithmic marketplaces like LinkedIn, Amazon search, and ride-hailing or online dating platforms utilize machine learning algorithms in order to predict *relevance scores* with varying amounts of uncertainty. My work gives *formal guarantees* for the fairness and stability of converting uncertain relevance scores into rankings or matchings for downstream users (Devic et al., 2023, 2024).

Fundamental Limits. Trustworthy AI must also be developed alongside an understanding of what is and is not possible in machine learning. To address this, I have multiple works investigating the fundamental limits of machine learned predictions for classification and reinforcement learning tasks (Asilis et al., 2024a,b; Deng et al., 2022). In Hansen et al. (2024), my co-authors and I also consider the usefulness and practicality of fairness post-processing algorithms in the context of *group-wise calibration measures*. I believe that my works in this area contribute to a better understanding of when optimal and fair algorithms may or may not be possible, informing both the theory and practice of machine learning.

Trustworthy Large Language Models. At Amazon, I have just began a Summer internship with the Search team (Rufus) and am also collaborating with the AWS personalization team. Since API access is already becoming the dominant LLM paradigm, it is important to understand when we can and cannot improve the fairness of using LLMs for high-impact downstream decision tasks such as filtering job candidates or college applicants (Lira et al., 2023; LinkedIn, 2023). I am working on creating a new way to post-process the outputs of a LLM in a black-box manner in order to ensure fairness and optimality on a downstream task without requiring model fine-tuning (which may or may not be available).

Conclusion. My research goals are broadly centered around creating fair and trustworthy algorithmic decision-making systems in both practice and theory. I strongly believe this goal sits at the core Amazon's research mission. This is especially salient given the explosion of large models and growing consumer access and interest. I firmly believe that investigating fundamental theoretical limitations of algorithms — as well as designing new algorithms which have guaranteed performance, fairness, or privacy — is an important research direction for current and future applications of AI and ML in high stakes decision-making settings.

References

J. Asilis, S. Devic, S. Dughmi, V. Sharan, and S.-H. Teng. Regularization and optimal multiclass learning. In *Conference on Learning Theory (COLT)*, 2024a.

J. Asilis, S. Devic, S. Dughmi, V. Sharan, and S.-H. Teng. Transductive sample complexities are compact. Under submission, 2024b.

J. Biden. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence, Oct 2023. URL https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

Z. Deng, S. Devic, and B. Juba. Polynomial time reinforcement learning in factored state mdps with linear value functions. In *International Conference on Artificial Intelligence and Statistics*, 2022.

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

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

J. Gottfried. Americans' social media use, Jan 2024. URL https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/#:~:text=Most%20Americans%20(68%25)%20report,%25)%20say%20they%20use%20Instagram.

D. Hansen, S. Devic, P. Nakkiran, and V. Sharan. When is multicalibration post-processing necessary? Under submission, 2024.

M. T. Islam, A. Fariha, A. Meliou, and B. Salimi. Through the data management lens: Experimental analysis and evaluation of fair classification. In *Proceedings of the 2022 International Conference on Management of Data*, pages 232–246, 2022.

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.

B. Lira, M. Gardner, A. Quirk, C. Stone, A. Rao, L. Ungar, S. Hutt, L. Hickman, S. K. D'Mello, and A. L. Duckworth. Using artificial intelligence to assess personal qualities in college admissions. *Science Advances*, 9(41):eadg9405, 2023.