1 Introduction

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 artificial intelligence (AI) will play in society (Biden, 2023). Nonetheless, many current systems utilizing some form of AI or ML have demonstrated systematic unfairness and bias, sub-optimalities, or instabilities (Islam et al., 2022). Although the potential for AI/ML systems to transform the way or society operates is large, there remains a crucial need for research into improving the trustworthiness and fairness of AI models by *investigating the way that these systems operate on and interact with the underlying population.*

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.

Investigative Research Thrusts

(1) Investigating the algorithmic fairness of systems utilizing machine learned predictors, such as recommender and ranking systems, as well as two-sided marketplaces;

(2) Understanding practical limits of black-box API access to machine learned predictors and language models. In particular, in the context of black-box API systems, what is the power and limit of algorithmic auditing and model post-processing for fairness and robustness?

Together, my research directions reflect and address both (1) the integration of AI and ML into larger algorithmic systems; and (2) the shift towards powerful and closed-source models and LLMs. I believe that these key directions are essential to the broad and successful deployment of trustworthy and fair AI.

2 Fairness in Algorithmic Rankings and Marketplaces

Individuals do not interact with modern algorithmic systems and AI in isolation: algorithms rank job candidates against each other, ride-hailers compete for drivers during surge pricing, and individuals compete for financial resources via automated credit ranking systems (Artificio, 2024). Philosophers and ethicists have long ago recognized that fairness is fundamentally a *contextual* requirement (Anderson, 1999): it relies not only on the qualifications of a single individual in a vacuum, but the context of those qualifications amongst all individuals present in an evolving, dynamic system. It is therefore important to move beyond studying fairness in purely prediction tasks like loan or income classification, and also *consider fairness in larger sociotechnical systems which may use ML predictions as subroutines.* To address this, my research focuses on two such settings: algorithmic rankings of individuals, and two-sided algorithmic marketplaces between, for example, individuals and jobs or students and universities.

Prior Work: Rankings and Recommendations. One of the most ubiquitous areas in which AI interacts with society is through ranking and recommendation systems. Ranking and recommendation systems utilizing ML predictions are already used to determine who is selected for an interview (LinkedIn, 2023), what web-page or product is ranked at the top of a list (Tsioutsiouliklis et al., 2021), or what student may be recommended to receive additional learning resources by a school district (Perdomo et al., 2023). Ranking systems utilizing ML predictions have already started to be at fault for discrimination: for example, Dastin (2018) report that Amazon's internal hiring tool was systematically discriminating against women due to biased predictions, and Wall (2021) discuss how LinkedIn's applicant rankings were biased by faulty predictors. It is therefore paramount to consider how rankings *utilizing* ML predictions can be improved in service of stability, fairness, and trustworthiness.

A standard way of integrating ML to create rankings of items (such as rankings of job candidates, web pages, or videos) is to first train a *relevance score* predictor, and then rank items according to decreasing relevance score (Robertson, 1977; Calauzènes and Usunier, 2020). Companies like LinkedIn use this approach to rank which candidates show up on a recruiter search query (Quiñonero Candela et al., 2023).

There are two problems associated with the approach of sorting by decreasing relevance score: (1) the resulting ranking ignores fairness considerations at the level of individual items or groups of items; and (2) the ranking turns out to be extremely sensitive to small variations in the predictions, and is not *stable.* The importance of fairness (1) is self-explanatory: it ensures that historical injustice not be perpetuated through algorithmic decisions. The *stability* desired in (2) is more nuanced. It articulates a need for rankings to not be based on minute and potentially *arbitrary* variations in predictions (Cooper et al., 2023). As an example, randomness in the ML training process through the weight initializations of neural networks or data selected by SGD may cause certain individuals to have slightly perturbed predictions. If a recruiter was interviewing only the top ranked candidate for each query, it is desirable for the top rank to be determined by qualifications, and not by such arbitrary variations.

In Devic et al. (2024), we propose an alternative way of transforming relevance score predictions of items into item rankings that we refer to as *Uncertainty Aware (UA) Ranking.* UA ranking has multiple desirable properties. First, it *provably retains and composes with the fairness of underlying predictors.* In particular, if the underlying predictor is individually- or group-fair in the definitions of Dwork et al. (2012); Hébert-Johnson et al. (2018), then the resulting UA ranking will also retain the respective fairness notion. Most importantly, we also prove that UA ranking is **stable** towards small variations or perturbations in the given predictions, achieving our second goal of reducing arbitrariness in rankings. UA ranking is both simple to implement and effective: preliminary experiments indicate that it may lead to drastic robustness and stability improvements in ranking systems utilizing relevance-score predictors, with only a small cost to overall system utility.

Future Work: Fairness and Stability Auditing. In Devic et al. (2024), we argued that ranking systems utilizing machine learned predictors should retain both fairness and stability guarantees. Motivated by this, we plan to conduct extensive empirical audits of existing ranking systems to test these properties in state-of-the-art deployed environments.

Proposed Research Direction

Over the next two years, we will collectively audit for fairness and stability in a number of production-grade AI/ML ranking systems, including LinkedIn, Indeed, and Amazon Shopping. In particular, we will investigate (1) whether fairness conclusions are impacted or modified when considering important and under-served *intersectional groups*; and (2) to what extent instabilities exist, and whether there exist sub-populations which *systematically* experience real-world impacts of instabilities and arbitrariness in rankings.

Systems like LinkedIn have previously been audited for disparate ranking outcomes over demographic groups (Imana et al., 2021). However, such audits are *usually restricted to large demographic groups in isolation* — such as race or gender — and typically do not examine the outcomes for complicated interleaving collections of individuals and identities. Given that existing work has shown that fair ranking methods can often provide computationally and statistically tractable guarantees for these more complex groups (Hébert-Johnson et al., 2018; Devic et al., 2024), it is extremely important to understand the extent that such intersectionality notions can change fairness conclusions in real production-grade systems. To operationalize this, we plan to adapt prior auditing techniques in order to test for ranking and outcome disparities across smaller, intersectional sub-populations on ranking systems like LinkedIn.

In addition to fairness auditing, we also plan to audit for **instabilities** in rankings, and how they impact resulting downstream opportunities for job-seekers, web pages, and products (Figure 1). As we argue in Devic et al. (2024), similar queries should instead utilize randomness in order to rank candidates more appropriately. This is because the intent of the searches is identical (e.g., "ML Engineer" vs. "Machine Learning Engineer") , yet from a fairness perspective, they result in very different opportunities for the job candidates. Such instabilities are most damaging if they *systematically discriminate against groups of individuals.* For example, all three candidates in Figure 1 may be equally qualified, but it could be that no query ever ranks candidates similar to Luisa in the top position.

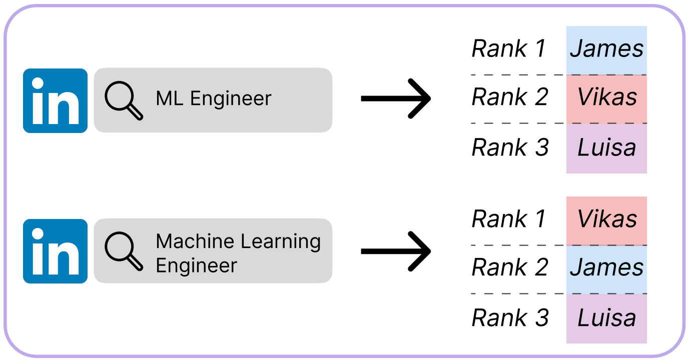

Figure 1: Instability in LinkedIn candidate search. A recruiter searches for qualified candidates using "ML Engineer" and "Machine Learning Engineer". If the two searches reveal very different candidate list orderings, the resulting ranking is both **unstable** and arbitrary.

Prior Work: Two-Sided Marketplaces. Yet another important area in which many users, consumers, and institutions interact with AI systems is through *two-sided algorithmic marketplaces.* Two-sided marketplaces already impact a large portion of our society — for example, medical residency assignments and kidney exchange marketplaces already use matching algorithms, and tens of millions Americans use dating apps or ride-sharing services. However, fairness and utility have already been seen to be at odds in this setting. For example, studies have shown that due to complicated latent dynamics, ride-hailing platforms or ad-delivery systems can routinely and systematically discriminate against minority groups (Ge et al., 2020; Imana et al., 2021). It is therefore paramount to consider *explicit* mechanisms for fairness-utility and preference-utility tradeoff considerations, which would allow for greater transparency and an increase in trustworthiness.

In a two-sided marketplace, there is a set of *individuals* (such as ride-hailers or content creators) who are to be matched with a set of *items* (drivers or users/content consumers). In large scale marketplaces or platforms, relevance score predictors are often trained to generate predictions suitable to the particular task. For example, such predictors may predict the engagement of each user with a particular piece of content, or the suitability of each driver to a ride-hailer (Sadeque and Bethard, 2019). Individuals, however, may also have heterogeneous preferences or intents which are potentially orthogonal to the simple engagement probability metric: For example, a user may have broadly selected "math and science videos" as their main interest when creating an account (Zhu et al., 2021). *How does a mechanism designer go about maximizing engagement subject to fairly respecting the preferences of the individuals?*

In Devic et al. (2023), we propose a framework which allows a mechanism designer to simultaneously account for (1) the utility of the system (e.g., overall engagement); (2) uncertainty in the predicted relevance scores; and (3) *individual level fairness* towards the preferences of the individuals. We present a way of provably *trading-off* these three desiderata in a clean fashion, and also present preliminary experiments demonstrating our method's applicability within a two-sided dating market. Our work *enhances transparency* with respect to these three desiderata by forcing the mechanism designer to *explicitly select a trade-off parameter* which quantitatively controls the relative importance of fairness, utility, and prediction uncertainty. Importantly, we also articulate an axiom based on **contextual entitlement** in the face of uncertain predictions, which we expect to be useful to future work investigating how fairness operates not only at the individual level, but at the system level with interleaving collections of individuals.

3 Fairness, Black-Box Models, and Algorithmic Auditing

Production-grade systems *utilizing* machine learning are shifting towards building on closed-source model APIs. A decade ago, a startup e-commerce site may have attempted to train a proprietary model in order to predict the relevance of items for user queries. In the present day, however, the same startup may instead build retrieval pipelines on top of highly performant closed-source model APIs such as Apple Intelligence or Anthropic's Claude. This represents a *shift* in the way that ML pipelines are designed and tested by developers and practitioners.

Unfortunately, this new paradigm of restricted "black-box" access via API rules out many existing methods for improving the fairness or robustness of these decision making pipelines. For example, model fine-tuning (Mao et al., 2023) or data re-weighting approaches (Yan et al., 2022) are popular and well-studied techniques for improving fairness. However, these techniques typically *require direct model access*. In addition, even *auditing* base models for fairness or robustness can be challenging via API and without "white-box" model access (Casper et al., 2024). Given the state of the model marketplace, and considering potential future trends in the area, the second thrust of my research is focused on clarifying: *What are the fundamental limits of fairness and robustness with only black-box access to models?*

Prior Work: Empirical Aspects of Multicalibration Post-Processing. Calibration is an important property of machine learned predictors which ensures that the probabilities they output are *meaningful*. Intuitively, calibration requires that amongst all samples given score $p \in [0, 1]$ by an ML algorithm, exactly a p-fraction of those samples have positive label. Nonetheless, calibrated predictors may still be *unfair*: In particular, a calibrated predictor may still systematically underestimate the qualifications or merits of important or underrepresented sub-populations or minorities. To ameliorate this, Hébert-Johnson et al. (2018) introduced the notion of *multicalibration*, which requires that the predictor be calibrated not only overall, but also when restricted to a collection of protected sub-populations given as input. Importantly, multicalibration is a *black-box post-processing* algorithm: it can be applied on top of any existing predictor with only *query* access to the predictor. That is, applying multicalibration algorithms does not require re-training or fine-tuning the base model, which is sometimes impossible via current LLM and Vision APIs.

Since its introduction, multicalibration has led to many theoretical results both within and outside algorithmic fairness (Gopalan et al., 2022c,b,a). However, few works have investigated how multicalibration algorithms have performed in practice, and what utility they may offer to practitioners of fair AI/ML. Working with **Preetum Nakirran at Apple MLR**, we conducted the first *comprehensive* empirical investigation of multicalibration post-processing (Hansen et al., 2024). We ran multicalibration post-processing algorithms more than 45K times on a variety of dataset modalities in order to distil important takeaways to practitioners. At a high level, we found that multicalibration algorithms still face barriers to widespread practical adoption. To address this, our work provides a toolkit of observations, best-practices, and code for practitioners of algorithmic fairness to know how and when multicalibration post-processing can be successfully applied in a variety of tabular, vision, and language settings. We envision our work as helping bridge the gap between theory and practice of black-box post-processing techniques for algorithmic fairness.

Prior Work: Post-Processing Methods for Metric Optimization in Black-Box Models. In addition to being closed source, many models are also trained on *proprietary data distributions* to maximize common metrics like accuracy or calibration. However, practitioners utilizing a closed-source model for a downstream task may in fact be interested in optimizing more complicated metrics of interest. For example, a company using a closed-source LLM via

API for a *ranking task* — e.g., ranking items on Amazon shopping — may wish to optimize performance on group fairness for item producers, recall/precision, or even more complex metrics. Since underlying black-box models are usually trained only to maximize accuracy on an unknown distribution, they may have limited test-time performance when evaluated on other metrics or data distributions of interest.

During a summer internship at Amazon AWS, we proposed a *plugin* method which can post-process arbitrary black-box classifiers in order to optimize performance on simple metrics like group fairness or F-measure/recall. The plugin method also provably helps in the face of certain kinds of distribution shift. Importantly, the plugin method — like multicalibration — *does not re-train or fine-tune the model.* Instead, using recent techniques from black-box metric optimization (Hiranandani et al., 2021), it learns a *class-reweighing* function which can adapt classifiers on a target distribution for a metric of interest with only a limited number of available labeled training samples. We have empirically demonstrated that the method can improve a variety of metrics for both vision and language models on out of distribution classification tasks. We are currently preparing the work for publication.

Future Work: Power and Limits of Black-Box Access. There is a growing literature which shows that LLMs may sometimes be biased or unfair (see, e.g., the survey of Gallegos et al. (2023)). These results are to be expected given the extremely large and diverse training corpus used to train production grade models (Longpre et al., 2023). In high stakes decision settings like hiring or clinical use, it is paramount to have a certificate of fairness, robustness, or reliability. Nonetheless, *it is still unclear whether black-box models like LLMs can be efficiently and comprehensively audited for these important properties* (Casper et al., 2024). Furthermore, given a black-box model, we do not yet understand to what extent we can efficiently *improve the fairness or robustness* of the model in a post-hoc fashion. Both directions are increasingly important as governments and policy makers begin exploring potential AI regulation: If the ability to certify a property such as unfairness is *provably impossible* with only black-box access, regulation may be unenforceable without mandating that auditors have more comprehensive model access.

Proposed Research Direction

Over the next two years, I will investigate (1) whether a black-box AI provider can efficiently and securely provide proof of fairness or robustness to an auditor concerned with checking a particular property or metric of import; and (2) to what extent AI models can be post-processed for fairness and robustness.

For (1), the AI provider would ideally — due to competitive advantage — provide such certificates without releasing the model or training data. However, it is not clear that this is fundamentally possible. Existing work proposes using *zero-knowledge proofs* to allow the provider to convince the auditor of fairness in a secure manner, however, this approach makes the strong assumption that the model provider has not tampered with the *training data* (Waiwitlikhit et al., 2024). Such an assumption is typically impossible to verify given that training data itself is often proprietary.

To ameliorate this, we propose viewing auditing as a two-person verification protocol, similar to the framework for verifying accuracy presented in Goldwasser et al. (2021). By examining the levels of access that the auditor has to the training data and model, we believe that the fundamental limits of black-box auditing may be revealed. In particular, for the setting where an auditor may have *only* black-box model access, our work may result in either impossibility results for fairness or robustness auditing, or new algorithms for achieving fairness/robustness certificates. Either outcome is highly beneficial towards achieving more trustworthy AI systems.

Finally, to address (2), I will investigate new sample and computational complexity lower bounds for post-processing arbitrary black-box predictors. Existing sample complexity results do not often take into account metrics other than calibration or accuracy, and as such, there remains ample room for further work in this important area of black-box model applications. Further, post-processing algorithms only help in certain *regimes* which can depend on (at least) the distribution and entropy of predictions, the distribution of underlying data, and model architectures (Hansen et al., 2024). Characterizing these regimes, and understanding when we expect post-hoc methods to supply improvements in practice, is an important avenue for future work. The implications are great for practitioners, since such results could inform when improvements to model fairness or robustness can be expected on top of closed-source models.

4 Conclusion

My research goals are broadly centered around creating and auditing fair and trustworthy algorithmic decision-making systems in both practice and theory. I strongly believe this goal sits at the core Apple's research mission. My research directions are especially salient given the recent explosion of large models, growing consumer access and interest, and potentially shrinking access to open-source models. I firmly believe that investigating fundamental practical and 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.

References

E. S. Anderson. What is the point of equality? *Ethics*, 109(2): 287–337, 1999.

Artificio, 2024. URL https://artificio.ai/solutions/loan-processing.

J. Biden. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence, Oct 2023.

C. Calauzènes and N. Usunier. On ranking via sorting by estimated expected utility. *Advances in Neural Information Processing Systems*, 33, 2020.

S. Casper et al. Black-box access is insufficient for rigorous ai audits. *arXiv preprint arXiv:2401.14446*, 2024.

A. F. Cooper, K. Lee, M. Choksi, S. Barocas, C. De Sa, J. Grimmelmann, J. Kleinberg, S. Sen, and B. Zhang. Arbitrariness and prediction: The confounding role of variance in fair classification. *arXiv preprint arXiv:2301.11562*, 2023.

J. Dastin. Amazon scraps secret ai recruiting tool that showed bias against women, Oct 2018.

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 multigroup fairness in ranking with uncertain predictions. In *International Conference on Machine Learning (ICML).*, 2024.

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

I. O. Gallegos et al. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*, 2023.

Y. Ge, C. R. Knittel, D. MacKenzie, and S. Zoepf. Racial discrimination in transportation network companies. *Journal of Public Economics*, 190, 2020.

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.

P. Gopalan, A. T. Kalai, O. Reingold, V. Sharan, and U. Wieder. Omnipredictors. In *Innovations in Theoretical Computer Science*, 2022a.

P. Gopalan, M. P. Kim, M. A. Singhal, and S. Zhao. Low-degree multicalibration. In *Conference on Learning Theory*, pages 3193–3234. PMLR, 2022b.

P. Gopalan, N. Narodytska, O. Reingold, V. Sharan, and U. Wieder. Kl divergence estimation with multi-group attribution. *arXiv preprint arXiv:2202.13576*, 2022c.

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

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.

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.

B. Imana, A. Korolova, and J. Heidemann. Auditing for discrimination in algorithms delivering job ads. In *Proceedings of the web conference 2021*, pages 3767–3778, 2021.

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.

B. Lira, M. Gardner, et al. Using artificial intelligence to assess personal qualities in college admissions. *Science Advances*, 9(41), 2023.

S. Longpre, G. Yauney, E. Reif, K. Lee, A. Roberts, B. Zoph, D. Zhou, J. Wei, K. Robinson, D. Mimno, et al. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. arxiv, 2023.

Y. Mao, Z. Deng, H. Yao, T. Ye, K. Kawaguchi, and J. Zou. Last-layer fairness fine-tuning is simple and effective for neural networks. *arXiv preprint arXiv:2304.03935*, 2023.

J. C. Perdomo, T. Britton, M. Hardt, and R. Abebe. Difficult lessons on social prediction from wisconsin public schools. *arXiv preprint arXiv:2304.06205*, 2023.

J. Quiñonero Candela, Y. Wu, B. Hsu, S. Jain, J. Ramos, J. Adams, R. Hallman, and K. Basu. Disentangling and operationalizing ai fairness at linkedin. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023.

S. E. Robertson. The probability ranking principle in ir. *Journal of documentation*, 33(4):294–304, 1977.

F. Sadeque and S. Bethard. Predicting engagement in online social networks: Challenges and opportunities. *arXiv preprint arXiv:1907.05442*, 2019.

S. Tsioutsiouliklis, E. Pitoura, P. Tsaparas, I. Kleftakis, and N. Mamoulis. Fairness-aware pagerank. In *Proceedings of the Web Conference 2021*, pages 3815–3826, 2021.

S. Waiwitlikhit, I. Stoica, Y. Sun, T. Hashimoto, and D. Kang. Trustless audits without revealing data or models. *arXiv preprint arXiv:2404.04500*, 2024.

S. Wall. Linkedin's job-matching ai was biased. the company's solution? more ai., Jun 2021.

B. Yan, S. Seto, and N. Apostoloff. Forml: Learning to reweight data for fairness. *arXiv preprint arXiv:2202.01719*, 2022.

Z. Zhu, J. Cao, T. Zhou, H. Min, and B. Liu. Understanding user topic preferences across multiple social networks. In *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021.