1 Abstract

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 more broadly:

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

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

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

2 Fairness in Algorithmic Rankings and Marketplaces

Individuals do not interact with modern algorithmic systems in isolation: algorithms rank job candidates against each other or match ride-sharers with drivers. Fairness is fundamentally a *contextual* requirement: 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 the system. It is therefore important to move beyond studying fairness in purely prediction tasks like loan or income classification, and consider fairness in larger systems which may use ML predictions as subroutines. To address this, I focus on two such settings: algorithmic rankings of individuals, and two-sided algorithmic marketplaces between individuals and items.

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. According to the Pew Research Center, 80% of Americans use YouTube, and 70% use some form of social media (Gottfried, 2024). Outside of social media, rankings and recommendation systems utilizing ML predictions are already used to determine who is selected for an interview, what web-page is ranked at the top of a list (Tsioutsiouliklis et al., 2021), or what movie is recommended.

Ranking systems utilizing ML predictions have already started to be at fault for discrimination. Dastin (2018) report that Amazon's internal hiring tool was systematically discriminating against women due to bias 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 a way of transforming relevance score predictions into rankings which *provably* (a) retains and composes with the underlying individual and group fairness guarantees of the predictors; and (b) is stable towards small variations or perturbations in the given predictions. Given the nature of negative results in the *composition* of both individual and group fairness properties (Dwork and Ilvento, 2018), we find (a) to be surprising and extremely desirable. Our work is 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.

Algorithmic Marketplaces. Another important area in which many consumers and employees interact with AI systems within is two-sided algorithmic marketplaces such as ride-hailing, Google AdSense, or large scale social media systems like dating apps. 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 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 engagement of a user with a particular piece of content, or suitability of a driver to a ride-hailer (Sadeque and Bethard, 2019; Zhao et al., 2019). Furthermore, individuals may also have heterogeneous preferences which are orthogonal to this engagement probability (see, e.g., Zhu et al. (2021)). How does a mechanism designer go about maximizing the 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) 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 showing their applicability within a two-sided dating market. Our work enhances transparency with respect to these three desiderata by forcing the mechanism designer to explicitly choose a trade-off *parameter* which quantitatively controls the relative importance of fairness, utility, and prediction uncertainty. Importantly, we 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 Robust and Optimal Learning Algorithms

To create trustworthy AI systems, the process of *learning* itself is an important phenomenon to study. Arguably, the most standard technique to train ML models is Empirical Risk Minimization (ERM). ERM consists of minimizing the empirical loss over a sufficiently large dataset. Nonetheless, in theory and practice, there seem to be settings where ERM may fail to achieve an optimal predictor. This includes, for example, clinical settings with smaller amounts of data (Volovici et al., 2022), or settings with distribution shift between train and test time (Imani et al., 2022). In such settings, a common technique to improve the training algorithm is to utilize a *regularizer*: a function which encourages the algorithm to pick *simpler* solutions. Performing ERM in tandem with a regularization function is known as *Structural* Risk Minimization (SRM). Examples of common implicit and explicit SRM in machine learning include L_2-norm regularization (Lewkowycz and Gur-Ari, 2020) or dropout in the training of neural networks (Srivastava et al., 2014).

Regularization is a key component of training modern ML systems. However, the power of regularization and when it may be *provably* useful or necessary for learning to take place is not currently well-understood. In Asilis et al. (2023), we provide a partial answer to this question: we introduce a set of increasingly powerful regularizers which capture the *learnability* of every multiclass classification problem. In particular, we prove that every multiclass classification problem which is learnable can be (optimally) learned by SRM with our proposed regularizers. Previous work has demonstrated that this is not the case for ERM (Shalev-Shwartz et al., 2010). In more recent followup work, we also show that some of our results can be extended to deal with multiclass classification with a potentially *infinite* number of labels, and prove various results regarding the *compactness* of learning problems (Asilis et al., 2024).

Our work is important in that it lays the foundation for better understanding the fundamental limits of ML systems, and in particular understanding the power and requirements of SRM and regularization. We believe that our work will eventually allow practitioners to better understand when and where regularization may be helpful for performance and robustness improvements, especially in scarce data or distribution shift settings.

4 Trustworthy Large Language Models

Organizations training and releasing state-of-the-art AI models typically allow differential levels of access: Google may allow black-box API access, while other companies may allow limited access to next token log probabilities, or even the weights of the models themselves. Many companies and governments are building algorithmic systems upon these different API levels. However, it is not well understood as to what the fundamental limits are with respect to the different levels of access (Casper et al., 2024). This is further complicated by the well-known *pre-train then fine-tune* paradigm, where companies may allow fine-tuning their pre-trained black-box models to a particular task, via API calls or otherwise (Google, 2024).

To address this black-box nature of recent AI advances, I propose studying the following: (1) is it *provably* possible to audit models for privacy, fairness, or optimality with only black-box query access? (2) How does the well known *pre-train fine-tune* paradigm interact with black-box API access? Is the fine-tune step necessary in the first place? Or is black-box pre-trained model access equivalent capability-wise to black-box fine-tuned models?

Limits of Black-Box Auditing 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). Nonetheless, in high stakes decision settings like hiring or clinical use, it is paramount to have a certificate of fairness or privacy.

We propose investigating whether a black-box AI provider can efficiently and securely provide proof of fairness or privacy to an auditor concerned with checking a particular property of import. This property may be, for example, whether a model may be providing decisions biased towards one subgroup or another (Kearns et al., 2018). The AI provider would ideally, due to competitive advantage, provide such a certificate without releasing the model or training data. However, it is not clear that this is fundamentally possible: existing work proposes *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).

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 (1) impossibility results for fairness or privacy auditing; or (2) new algorithms for achieving fairness/privacy certificates. Either outcome is highly beneficial towards achieving more trustworthy AI systems.

Black-Box Fine-Tuning Traditional fine-tuning approaches involve taking a pre-trained model, changing the *model head*, and then continuing the model training with a specific (and usually *smaller*) dataset. For example, a company may fine-tune Gemini in order to screen resumes for a particular job posting. Such an approach allows the model to achieve better results on a more narrow task. However, fine-tuning also assumes that the original weights of the model can themselves be modified (either via API or open source weights). Not all model APIs allow this function, and furthermore, the APIs that do allow it may be more costly to run than simple inference APIs.

We propose studying the following: if we are only allowed to query an LLM via black-box API calls, is it possible to fine-tune that LLM for a particular downstream classification task? We propose utilizing recent techniques from black-box metric optimization (Hiranandani et al., 2021) to allow for post-processing the outputs of LLMs by modifying their raw predicted text on specific tasks. In particular, we plan to train a *class-reweighing* function which learns the internal bias of the LLM on a particular tasks, and attempts to correct it via post-processing. Importantly, unlike current methods such as Sun et al. (2022), we are fine-tuning only the raw textual outputs of a model, and are *not* modifying the internal weights or prompt at all. Overtime, we believe that this technique may reveal fundamental limits of black-box fine-tuning. It may also lead to more interpretable models than traditional fine-tuning, since we learn a simple *vector* of probabilities to re-weigh the predictions of a black-box model.

5 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 Google's research mission. This is especially salient given the recent 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 the 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. Under submission, 2023.

J. Asilis, S. Devic, S. Dughmi, V. Sharan, and S.-H. Teng. Learnability is a compact property. Under submission, 2024.

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/.

C. Calauzènes and N. Usunier. On ranking via sorting by estimated expected utility. *Advances in Neural Information Processing Systems*, 33:3699–3710, 2020.

S. Casper, C. Ezell, C. Siegmann, N. Kolt, T. L. Curtis, B. Bucknall, A. Haupt, K. Wei, J. Scheurer, M. Hobbhahn, 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. URL https://www.reuters.com/article/idUSKCN1MK0AG/.

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. Non-archival at Symposium on Foundations of Responsible Computing (FORC 2024). Under submission, 2024.

C. Dwork and C. Ilvento. Fairness under composition. *arXiv preprint arXiv:1806.06122*, 2018.

I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed. 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*, 190:104205, 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)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2021.

Google, 2024. URL https://ai.google.dev/gemini-api/docs/model-tuning.

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.

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. PMLR, 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.

E. Imani, G. Zhang, J. Luo, P. Poupart, P. H. Torr, and Y. Pan. Label alignment regularization for distribution shift. *arXiv preprint arXiv:2211.14960*, 2022.

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.

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

A. Lewkowycz and G. Gur-Ari. On the training dynamics of deep networks with l_2 regularization. *Advances in Neural Information Processing Systems*, 33:4790–4799, 2020.

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.

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 preprint arXiv:2305.13169*, 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*, pages 1213–1228, 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. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

T. Sun, Y. Shao, H. Qian, X. Huang, and X. Qiu. Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning*, pages 20841–20855. PMLR, 2022.

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.

V. Volovici, N. L. Syn, A. Ercole, J. J. Zhao, and N. Liu. Steps to avoid overuse and misuse of machine learning in clinical research. *Nature Medicine*, 28(10):1996–1999, 2022.

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. URL https://www.technologyreview.com/2021/06/23/1026825/linkedin-ai-bias-ziprecruiter-monster-artificial-intelligence/.

Z. Zhao, R. Anand, and M. Wang. Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform. In *2019 IEEE international conference on data science and advanced analytics (DSAA)*, pages 442–452. IEEE, 2019.

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)*, pages 590–599. IEEE, 2021.