Name: Siddartha Devic
Website: https://sid.devic.us/
Email: devic@usc.edu

To the MSR New England ML and Statistics Group,

I am writing to express my strong interest in a PhD internship position in your group at Microsoft Research. I am a PhD student in the theory group at the University of Southern California. My research goals center on bridging the gap between theory and practice in the design of algorithms that address real-world challenges while considering factors such as fairness, robustness, and optimality. My career goal is to straddle industry and academia to combine stakeholder input and practice with foundational contributions, which is why I am excited at the opportunity to learn from industry research at MSR.

My current research interests encompass the following three key areas, amongst which I highlight my relevant work.

1. Algorithmic fairness, especially in settings beyond classification such as recommender systems, rankings (Devic et al., 2023b), two-sided marketplaces (Devic et al., 2023a), and fairness in the presence of uncertainty or finite resources more broadly;
2. Theoretical machine learning, with a focus on developing robust and optimal algorithms (Asilis et al., 2023; Deng et al., 2022); and
3. The intersection of these areas, including statistical aspects of fairness in classification/resource allocation such as multicalibration.

Research Methodologies. I have primarily worked on theoretical research during my PhD. However, I am also reasonably experienced with empirical aspects of deep learning. I am currently working in Pytorch on empirically studying notions of multicalibration (Hébert-Johnson et al., 2018) in large neural networks (NNs) and LLMs. Earlier in my PhD, I worked extensively on running experiments in PyTorch to analyze the spectral bias of NNs. I also have previous empirical works utilizing Tensorflow and Pytorch on applying deep learning to reinforcement learning and federated learning tasks in computer networking, and spent one summer working on object detection with deep learning at the Johns Hopkins Applied Physics Labs.

At MSR, I would be particularly interested in working on any of the following projects, which build on recent advancements on the theory of human-AI collaboration, language models, and two-sided marketplaces.

Potential Project: Two-sided Fairness in Algorithmic Marketplaces. In previous work, I examined one-sided fairness in two-sided matching (Devic et al., 2023a). As an example, such a setting may include fairness towards job-seekers in a applicant-job matching market. In modern algorithmic recommendation systems, however, two-sided fairness is often desired. Consider a social media platform which matches content producers and content consumers. Fairness towards content producers may allow smaller producers to get recognized, instead of winner-take-all dominated marketplaces common in modern algorithmic recommendation systems. Furthermore, fairness towards content consumers is also important to ensure non-discriminatory content delivery (Imana et al., 2021).

Although fairness in two-sided marketplaces has been previously considered (e.g. in Do et al. (2021)), I propose considering the entire market as being induced by *predictions* of one or multiple machine learning algorithms. My previous work examined how the predictions of machine learning algorithms, when combined with preferences of one side of the market, induced matchings according to our fairness desiderata. An interesting further direction is to understand what matching desiderata may be important in the setting where machine learning algorithms predict relevance for both sides of the market. Examining how matching algorithms emerge as induced by one or multiple machine learning systems in this setting is extremely interesting, especially when one takes into account utility and/or fairness considerations.

Potential Project: Fair Decision Making through Human-AI Collaboration. In many high stakes decision making settings, humans will—for the forseeable future—be the final decision maker. Nonetheless, they may recieve *algorithmic advice* from a model such as a classifier or LLM. For example, in college admissions,

for better or for worse, consultants are increasingly proposing that schools use probabilistic classifiers or LLMs to streamline their admissions process by reviewing transcripts, summarizing essays, or scoring applicants[1].

One would hope that if an algorithm gives fair (or accurate) recommendations, then the human receiving additional input from the AI cannot be more *unfair* (or less accurate) than if the human had no access to the AI recommendations in the first place. Surprisingly, the recent work of Ge et al. (2024) shows that this is not always the case: they prove that recommendations by a fair algorithm can, in theory, be *selectively applied* to amplify unfairness or decrease the quality of decision making. To an extent, this has already been observed in the practice of radiology, where in one study, algorithmic recommendations provided to doctors did not improve the overall quality of decisions (Agarwal et al., 2023).

I am particularly interested in critically examining the work of Ge et al. (2024), who make fundamental assumptions on the availability of human predictions which may call into question their conclusions. Another direction is to consider how their model may differ when recommendations are given by an LLM instead of a traditional score based classifier: does greater expressiveness of the algorithmic recommendation afford a different model for human compliance? Furthermore, I believe that some conclusions may also change when considering settings with *finite resources*—such as cohort selection or ranking for admissions—as opposed to just classification. Theoretically examining the role of recommendations in Human-AI collaboration is an extremely timely direction given the myriad of recent applications of LLMs in decision making settings.

Potential Project: Voting Theory and Alignment. Modern LLMs are made suitable for release through Reinforcement learning from human feedback (RLHF). RLHF fine-tuning is essentially a preference learning algorithm, and as such, it is natural to connect RLHF with social choice theory, and in particular, voting theory. I believe that with a stylized LLM model, like the distributional model used in (Kalai and Vempala, 2023), one can potentially show alignment lower bounds against an underlying population with diverse opinions (similar to the *distortion* of voting schemes). Nonetheless, lower bounds on their own may not be very interesting, since it is well known that alignment is difficult. A naive way to deal with this is to align a model multiple times to different sub-populations, and serve the different resulting models appropriately. A natural question is therefore: what is the trade-off in compute/distortion between having multiple aligned models to satisfy different segments of an underlying population? Characterizing pareto optimality would be an interesting research direction here, since any such characterization should depend on how *polarized* the opinions of the underlying population are. A separate but related direction is to examine alignment under population distribution shift, motivated by the fact that currently deployed RLHF systems appear to have been trained on data collected in countries with potentially differing value systems or preferences (source).

Any formal connection between voting theory and RLHF should provide a fun algorithmic sandbox for exploring fundamental limits of LLMs, and could potentially uncover new algorithms for aligning the model with the preferences of population.

Relation to MSR. My research goals are broadly centered around creating fair and provably trustworthy algorithmic decision making systems, which I believe sits within the larger mission of the MSR ML and Statistics group. Beyond the projects I have described, I am generally open to any research directions at the interface of theory and practice of machine learning and decision making. I firmly believe that investigating fundamental theoretical limitations of algorithms, as well as designing new algorithms which have guaranteed/provable performance, is an important research direction for the current and future applications of machine learning in high stakes decision making settings. Furthermore, I believe that the application of these ideas to settings with *finite resources* is under-studied within our community.

Nonetheless, the technical toolset I have developed through both theoretical and applied projects gives me confidence that I can quickly learn new skills for any potential project directions in deep learning, LLMs, privacy, and beyond. I hope to place my contributions at the intersection of society and algorithms, and as such, I would be particularly interested in working with Lester Mackey, Adam Kalai, Amanda Coston, or Nicole Immorlica.

Thank you for your consideration.

Best,
Siddartha Devic

[1] As examples, see the adverts 1, 2, or 3

References

N. Agarwal, A. Moehring, P. Rajpurkar, and T. Salz. Combining human expertise with artificial intelligence: Experimental evidence from radiology. Technical report, National Bureau of Economic Research, 2023.

J. Asilis, S. Devic, S. Dughmi, V. Sharan, and S.-H. Teng. Regularization and optimal multiclass learning. *arXiv preprint arXiv:2309.13692*, 2023.

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)*, 2023a.

S. Devic, A. Korolova, D. Kempe, and V. Sharan. Stability and group fairness in ranking with uncertain predictions. Submitted, 2023b.

V. Do, S. Corbett-Davies, J. Atif, and N. Usunier. Two-sided fairness in rankings via lorenz dominance. *Advances in Neural Information Processing Systems*, 34:8596–8608, 2021.

H. Ge, H. Bastani, and O. Bastani. Rethinking fairness for human-ai collaboration. In *To appear in Innovations in Theoretical Computer Science (ITCS)*, 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.

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.

A. T. Kalai and S. S. Vempala. Calibrated language models must hallucinate. *arXiv preprint arXiv:2311.14648*, 2023.