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

To the Microsoft Research FATE 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.

Potential Project: Prediction Arbitrariness, Replicability, and Privacy. The recent work of Cooper et al. (2023) provides further empirical evidence that the problem of model multiplicity (Black et al., 2022) is often at odds with the overarching goal of fair classification. In particular, Cooper et al. (2023) show that many learning algorithms, when trained on subsamples of the training dataset, will have high disagreement on certain individuals—hence, the prediction on these individuals can be considered *arbitrary* since they depend on the actual realized dataset used for training. Cooper et al. (2023) propose classifier *bagging* and *abstention* to ameliorate these issues, and surprisingly, they find that their methods for reducing arbitrariness appear to improve fairness as a side-effect. Nonetheless, the way that arbitrariness is measured in Cooper et al. (2023), through dataset subsampling, is a computational proxy for the true desiderata: learning algorithms which output similar classifiers when trained on any dataset drawn i.i.d. from the underlying distribution over individuals. Such a proxy is often necessary since a practitioners tend to utilize all available training data, and do not usually collect multiple separate datasets to measure the arbitrariness of their models.

The theoretical computer science community has also recently been grappling with similar issues under the area of algorithm *replicability* (Impagliazzo et al., 2022)[1]. An algorithm is replicable if it outputs the same answer (with high probability) when trained on two independently drawn training sets from the underlying distribution. From a theory perspective, it would be desirable for various learning algorithms in optimization (Ahn et al., 2022), reinforcement learning (Karbasi et al., 2023), or clustering (Esfandiari et al., 2023) to both be replicable and perform well, since, broadly speaking, the practitioner would therefore *trust* the outputs of these algorithms more. The field of replicability has been generally successful in its goal of producing

[1] Although the original paper calls the notion *reproducibility*, the community has since renamed it *replicability*.

various learning algorithms which are guaranteed to always output the same classifier no matter what the drawn training dataset looks like.

I propose formally connecting the notions of replicability and arbitrariness in both theory and practice. It is almost immediate that a replicable algorithm will not be arbitrary, since it will be consistent across multiple realized training set draws. Furthermore, replicable algorithms typically do not use classifier abstention to potentially avoid predicting on difficult test examples. The theory community has given us ingredients for producing non-arbitrary learning algorithms, and translating this theory into practice is a fertile research direction. In particular, I am excited at the recent work of Kalavasis et al. (2023), which proposes that differentially private learning algorithms may be able to provide part of the solution to the difficult lessons posed in (Cooper et al., 2023). Testing whether this is true in practice, or if we may need to modify algorithms from the replicability literature to achieve goals important to the fairness community, is a concrete direction I am interesting in pursing. Such a connection would be interesting in its own right, since fairness, privacy, and replicability may emerge as ultimately intertwined desiderata.

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[2].

One would hope that if an algorithm gives fair recommendations, then the human receiving additional input from the AI cannot be more *unfair* 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: 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 (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. 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 interesting and timely direction given the myriad of recent applications of LLMs throughout decision making settings.

Relation to MSR FATE. 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 FATE group. In particular, my second proposed project is potentially under the umbrella of "Sociotechnical Approaches to Measuring Harms Caused by AI Systems", since there may be fundamental limitations to the *helpfulness* of AI systems to human decision makers.

Beyond the projects I have described, I am generally open to any research directions at the interface of theory and practice of fair 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 potential project directions in deep learning, LLMs, prviacy, 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 Solon Barocas, Alex Chouldechova, Hanna Wallach, Miro Dudík, Jennifer Wortman Vaughan, or Danah Boyd.

Thank you for your consideration.

Best,
Siddartha Devic

[2] 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.

K. Ahn, P. Jain, Z. Ji, S. Kale, P. Netrapalli, and G. I. Shamir. Reproducibility in optimization: Theoretical framework and limits. *Advances in Neural Information Processing Systems*, 35:18022–18033, 2022.

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

E. Black, M. Raghavan, and S. Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 850–863, 2022.

A. F. Cooper, K. Lee, S. Barocas, C. De Sa, S. Sen, and B. Zhang. Arbitrariness and prediction: The confounding role of variance in fair classification. Preprint, 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.

H. Esfandiari, A. Karbasi, V. Mirrokni, G. Velegkas, and F. Zhou. Replicable clustering. *arXiv preprint arXiv:2302.10359*, 2023.

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.

R. Impagliazzo, R. Lei, T. Pitassi, and J. Sorrell. Reproducibility in learning. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 818–831, 2022.

A. Kalavasis, A. Karbasi, S. Moran, and G. Velegkas. Statistical indistinguishability of learning algorithms. In *International Conference on Machine Learning (ICML)*, 2023.

A. Karbasi, G. Velegkas, L. F. Yang, and F. Zhou. Replicability in reinforcement learning. *arXiv preprint arXiv:2305.19562*, 2023.