Yizhou Harry Tian tianh@uchicago.edu https://harry-tian.github.io/

Ph.D. Statement of Purpose

Building effective human-AI collaboration requires not only high-performing ML models and explanations but also an understanding of how humans and society use AI, such as retaining human agency and appropriate reliance on AI. Intrigued by this interdisciplinary direction, I started to develop a human-centered perspective and formed my research interest in **Human-AI interaction (HAI)**. My recent research focuses on designing human-compatible frameworks for AI-decision-support and machine teaching. In general, I am interested in building more effective and responsible human-AI collaboration by leveraging AI that model humans.

Finding My Research Interest in HAI via AI Decision Support

My interest in HAI stems from my undergraduate research experiences in machine learning (ML). At an REU program (Research Experiences for Undergraduates) at the University of North Texas, I learned basic ML knowledge and the experience-focused nature of ML. The following year, I attended an REU at the University of Minnesota. Working with Prof. Zhiwei Steven Wu in Differential Privacy (DP), I experimented with hyperparamter optimization methods and found random search to outperform bayesian optimization. I presented my work at a National REU Poster Symposium.

Then, I learned about the human side of ML at the predoctoral program at UChicago, working with Prof. Chenhao Tan at the CHAI (Chicago Human-AI) lab. In our first project, we designed a human-compatible model for case-based decision-support [1]. In explainable AI (XAI), example-based explanations retrieve the nearest neighbors to a test instance, but the similarity metric for retrieval is often based on some AI model embedding or ground-truth features, so the retrieved examples may not be informative to humans. Towards this end, we developed a human-compatible model that learns both image classification and human visual similarity judgment. Such a model produced representations more aligned with human perception, thus providing more effective nearest-neighbor explanations as case-based decision-support.

To test the potential of our model, I conducted synthetic experiments by building an artificial dataset with controllable features. I ran experiments with different conditions by varying simulated humans, generated by tuning feature weights, and varying decision boundaries, showing our human-compatible model resulted in better decision-support than baselines. We also showed our method's effectiveness with human studies on a natural image dataset. However, our first submission was unconvincing to reviewers as our tasks were too simple and our datasets too small. Thus, I led an additional round of human studies on a much larger chest X-ray dataset with a pneumonia diagnosis task. I recruited Prolific crowdworkers and conducted studies using our web app; results show our model also provides effective decision support for pneumonia diagnosis, highlighting its potential for more complicated, high-stake tasks. The work led to a publication to the Workshop on Human-Machine Collaboration and Teaming in ICML 2022 [1] as well as an under-review submission to a major ML conference.

Coming from an ML background, I initially treated our decision-support framework as another ML model, focusing on training better models and higher accuracies, but I soon learned that human factors are equally, if not more important. For example, in designing our decision-support policy, we encountered the tradeoff between human agency and performance: policies that achieved the best performance comes at the expense of persuading humans to follow AI.

Intrigued by this dilemma, I was drawn to the complexities of human-AI interaction. From a Human-Centered Machine Learning course and a Human+AI conference, as well as numerous related papers, I learned more about HAI. For example, I learned that in human-AI teams, performance could be suboptimal due to psychological factors on the human side, like lack of cognitive engagement [2] or the presence of assistance altering humans' decision-making process [3]. Even if humans use AI assistance in "expected ways", AI could still decrease human agency, as our work showed, or induce bias [3]. I found the questions exciting, but moreover, I learned of the misalignments between HAI systems in lab settings and real-world scenarios, including misalignment in the task [4], objective and outcomes [5], in diseradata of the assistance [6], and, a direction I am especially interested in, the experiment subjects. I gradually developed a human-centered perspective where I would take a step back and question whether the task at hand is appropriate for real-world scenarios humans use in practice.

Future Directions

I am drawn toward HAI research because of its interdisciplinary nature and human-centered approach. Specifically, I am currently interested in the following directions:
1. Machine teaching and modeling human learners. The problem of AI teaching knowledge to human learners is important, difficult and understudied. For one, I think an effective human learner model is lacking, as existing work often assume simplistic learner models like linear separators. In my past and current work, we are working on modeling human representation space from perceptual judgment annotations, an effective yet costly method. Another problem is the human learning process: while there are decades of psychological research on how humans learn, there is limited work on leveraging those findings for AI teachers. If we can more accurately simulate human learning with AI models, AI teachers can provide better teaching. For example, in my current project on designing AI-driven tutorials for radiology training, we are currently investigating whether teaching with contrastive examples can improve learning.
2. The gap between laypeople and real-world practitioners in HAI experiments. Most HAI empirical studies use crowdworkers due to their lower cost and higher accessibility compared to real-world practitioners; often they list the potential gap between laypeople and practitioners as a limitation at the end of the paper but do not expand further. I find this extremely unsatisfying and I envision two directions toward this issue. The first is to bridge the gap by bringing laypeoples' experiment environment closer to the actual task's. This could be done by improving their task knowledge through short training sessions, one direction I am currently working on, or by simulating other factors like risk, cognitive engagement, or the user interface. Another direction is to make more sense of the gap. For example, can we design a framework that infers meaningful information on how practitioners would behave (reliance towards AI, what type of explanations are useful) from experiment results on laypeople? An extension to this may be to design different forms of AI assistance for different levels of user expertise; this is motivated by AI's varying effects on people with different expertise [7,8] and is similar in spirit to disaggregated evaluations [9].

All my experiences collectively shaped my research interests and motivated me to pursue graduate studies. I believe the School of Information at University of Michigan aligns with my interest in HAI due to its focus on human-computer interaction (HCI) and connecting people and technology. Specifically, I would be excited to work with Professor Ben Green. From his works I learned of the many issues in human-AI decision making, such as explanations not aligning with human decision-making process [6] and the flaws of human overseeing algorithms [10]. I was particularly enlightened by the discussion in [10] about the upper bound of human oversight, making me realize that the human in HAI is not limited to human individuals. I would be excited to work with Professor Green on institutional oversight policies, which I think has potential beyond the government and public policy field. My research interests also overlap with Professor Etyan Adar's recent work on viewing explainable AI through Sensemaking Theory and Professor Anhong Guo's focus on hybrid human-AI intelligent interactive systems. I believe the strong HCI and computational social science research at UM-SI will foster my research development in HAI.

References

[1] Han Liu, **Yizhou Tian**, Chacha Chen, Shi Feng, Yuxin Chen, and Chenhao Tan. Towards effective case-based decision support with human-compatible representations. *Proceedings of the 1st ICML 2022 Workshop on Human-Machine Collaboration and Teaming*, 2022.

[2] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, April 2021.

[3] Ben Green and Yiling Chen. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–33, October 2021.

[4] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, New York, NY, USA, March 2020. Association for Computing Machinery.

[5] Luke Guerdan, Kenneth Holstein, and Zhiwei Steven Wu. Under-reliance or misalignment? how proxy outcomes limit measurement of appropriate reliance in ai-assisted decision-making. *CHI TRAIT '22: Workshop on Trust and Reliance in AI-Human Teams*, 2022.

[6] Yaniv Yacoby, Ben Green, Christopher L. Griffin Jr., and Finale Doshi-Velez. "if it didn't happen, why would i change my decision?": How judges respond to counterfactual explanations for the public safety assessment. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 10(1):219–230, Oct. 2022.

[7] Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, IUI '21, pages 318–328, New York, NY, USA, 2021. Association for Computing Machinery.

[8] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. Human-computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, August 2020.

[9] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W. Duncan Wadsworth, and Hanna Wallach. Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 368–378, New York, NY, USA, 2021. Association for Computing Machinery.

[10] Ben Green. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45:105681, July 2022.