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 approach and human-centered perspective, I 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.

Finding My Research Interest in HAI via AI Decision Support

My interest in HAI stems from my work with Prof. Chenhao Tan in the predoctoral program at UChicago. 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.

From this work, apart from experiencing the complexities of an entire research project lifecycle, I also began to develop my interest in human-AI interaction. In designing our decision-support policy, we encountered the tradeoff between human agency and performance: policies that achieved the best performance came at the expense of humans' overreliance on AI. I was intrigued by this dilemma as, coming from an ML background, it could not be solved just by training better models and obtaining higher accuracies. Rather, it required accounting for how humans use and interact with AI assistance. To gain a deeper understanding of HAI, I learned from a Human-Centered Machine Learning course, a Human+AI conference, and related papers. I began to notice problems like how to leverage human and AI complementary strengths, how humans use assistive technologies and how we should design AI assistance, and the many misalignments between HAI systems in lab settings and real-world scenarios. Currently, I am interested in the following specific directions:

Future Directions

(i). Machine teaching and modeling human learners. The problem of AI teaching knowledge to human learners is important, difficult, and understudied. For instance, I think an effective human learner model is lacking, as existing work often assumes 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 lack of work on leveraging psychology theories for teaching, such as simulating human learning with AI models. For example, in my current project on designing AI-driven tutorials for radiology training, we are investigating whether contrastive teaching examples can improve learning. Inspired by research in psychology and machine teaching [2, 3,4], we are simulating human learners using neural networks, contrastive loss functions, and gradient descent. Ultimately, we hope to provide more effective training for radiology residents in prostate cancer diagnosis.
(ii). The gap between laypeople and real-world practitioners in HAI experiments. Most HAI empirical studies use crowdworkers due to 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 unsatisfying and I envision two possible solutions. The first is to bridge the gap by bringing laypeoples' experiment environment closer to the actual task. This could be done by improving their task knowledge through short training sessions, by effective machine teaching, 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 [5,6] and is similar in spirit to disaggregated evaluations [7].

All my experiences collectively shaped my research interests and motivated me to enroll in a PhD program, where I can further develop my research career and pursue my goal of building more effective and responsible human-AI collaboration. University of Utah's PhD program in Computer Science is appealing to me because of the department's strong AI and HCI research and the many faculty members whose interests align with mine. I am interested in Professor Ana Marasović's work on explainable and interpretable AI. I was inspired her work *Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI* [8]. In ML and HAI research, many terms and concepts, e.g., "trust", "complementarity", are frequently used without clear definitions, so formalizing those concepts is a important and fundamental step. I like their definition of human-AI trust as it emphasizes vulnerability and warranted trust, which are prevalent factors in real-life, high risk task settings. On the other hand, this also points to flaws in HAI lab-based research on non-expert laypeople: there is little risk in the experiment setting so they would not feel vulnerable to the AI; laypeople's lack of risk and lack domain knowledge could also lead to unwarranted trust. I would be excited to leverage this human-AI trust framework to bridge the between laypeople and real-world practitioners in HAI experiments. I am also interested in Professor Ellen Riloff's direction on sentiment analysis. My interests in HAI are well represented at University of Utah, and I believe a PhD in Computer Science will allow me to build more effective and responsible human-AI collaboration.

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] Daniel N Barry and Bradley C Love. Human learning follows the dynamics of gradient descent, Dec 2021.

[3] Linsey A. Smith and Dedre Gentner. The role of difference-detection in learning contrastive categories. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society, CogSci 2014*, Proceedings of the 36th Annual Meeting of the Cognitive Science Society, CogSci 2014, pages 1473–1478. The Cognitive Science Society, 2014.

[4] Neehar Kondapaneni, Pietro Perona, and Oisin Mac Aodha. Visual knowledge tracing, 2022.

[5] 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.

[6] 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.

[7] 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.

[8] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai, 2020.