Ph.D. Statement of Purpose

Building effective human-AI collaboration requires more than mere good models and assistance as defined by some numeric metric, a valuable lesson I learned along my research process. For instance, (how) can we achieve strong human-AI team while retaining human agency? What are the mismatches between human-AI systems in lab settings and real-world practice and how can we bridge the gap? These compliated questions that require interdisciplinary knowledge and perspectives inspired my research interest in **Human-AI interaction (HAI)**. As a predoctoral researcher at the Universtiy of Chicago, my research focuses on designing human-compatible AI frameworks for decision support and machine teaching. Ultimately, my mission is to design HAI systems that help humans and society organically while reducing the harms of automated systems.

Finding My Research Interest

My research journey began with my work with Prof. Chenhao Tan at UChicago, where we designed a human-compatible model for case-based decision support [1]. In AI explanations, example-based explanations retrieve the nearest neighbors to a test instance, but the similarity metric for retrieval remains understudied. In fact, the metric is often a distance on some AI model embedding, so the retreived examples may not be informative to humans. Towards this end, we developed a human-compatible model that learns both classification and human perception judgement. Such a model produced represenations more aligned with human similarity, thus providing more effective nearest-neighbor explanations, or case-based decision-support.

My contributions include conducting synthetic experiments: using an artificial dataset with controllable features, I ran experiments with varying simulated humans, generated by tuning feature weights, and varying decision boundaries, showing our human-compatible model resulted in better decision-support than ML model baselines. With positive results from synthetic experiments, we moved on to real humans. I conducted a human study on a chest X-ray dataset with Prolific crowdworkers, showing our model also provides effective decision support for pneumonia diagnosis. 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.

Besides technical skills in experiment design, I learned of an unexpected but useful research practice: being flexible in problem formulation and the research process. Our initial goal was to design an AI-driven tutorial for radiology training, largely a machine teaching problem. However, in our experimentations with modeling human learners, we found that a neural network that learned both classification and human similarity judgement produced an interesting representation, one that encoded patterns from human perception. We did not know how to leverage our human-compatible model for teaching, but we thought it may be useful for case-based decision support. Thus, from our finding we detoured to a different research problem, showed its usefulness through empirical studies, and now we return to the problem of AI-driven tutorial.

Coming from a machine learning background, I initially treated our decision-supoprt framework as another ML model: I thought that if we could build a more generalizable model, produce better explanations and achieve higher performance on some metric, we would achieve better human-AI collaboration. This perspective was limiting as we soon encountered the tradeoff between human agency and performance: showing neutral and informative explanations would retain human agency but results in lower task performance than showing persuasive explanations that nudged humans to follow the AI. I was intrigued by this dilemma and was drawn to the complexities of human-AI interaction. I learned more about HAI through a Human-Centered Machine Learning course and a Human+AI conference, both at UChicago, as well as from numerous related papers, gradually

forming a new, more thorough perspective towards HAI.

I first learned of the many "non-ML" factors that could affect human-AI team performance. On the human side, AI asisstance could alter humans' decision-making process and induce bias [2]; lack of human cognitive engagement could cause overreliance or underreliance towards AI [3]. On the AI side, explanations may not align with humans' decision-making process [4]. More importantly, I learned of the misalignments between HAI systems in lab settings and real world scenarios, including misalignment in the task [5], objective and outcomes [6], in diserideta of the assistance [4], 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 a real-world scenarios used by humans in practice.

Future Directions

Thus, I am drawn towards HAI research because of the different perspectives and fields of knowledge it requires and its human-centered approach, differing drastically from ML. Specifically, I am currently interested in the problem of the gap between laypeople and real-world practitioners in HAI experiments. Crucially, AI assistance affects people differently depending on their task expertise. [7] suggest explanations may be more effective on tasks that people perceive themselves as more knowledable. [8] report that in AI-assisted skin-cancer recognition, clinicians with different expertise show differences in not only diagnosis accuracy but also reliance on AI. Most HAI empirical work use crowdworkers due to its their lower cost and higher accessibiilty 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 towards this issue:

1) Bridging the gap. One straightforward solution is to reduce the gap between laypeple and real-world practitioners in HAI experimental setting. There are many approaches to this, one of which I am currently working on is improving laypeoples' task expertise through machine teaching. Such teaching frameworks can be deployed as short, effective training phases before an HAI experiment procedure. We are currently investigating whether grouping teaching examples in pairs or tuples can improve teaching by inducing contrastive learning. Another way to bridge the gap is to simulate practitioners' real-world task environment in terms of risk, cognitive engagement, user-interface, etc.

2) Making sense of the gap. Another direction is to design a framework that infers meaningful information from experiments of laypeople. For example, ideally such a framework could infer patterns on practitioners 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 evidence suggesting AI has varying effects on people with different expertise [7,8] and similar in spirit to disaggregrated evaluations [9].

All my experiences collectively shaped my research interests and motivated me to pursue graduate studies. I believe HCI research at the School of Information at University of Michigan aligns with my goal of designing HAI systems that help humans in high-stake domains. 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 [4] and the flaws of human overseeing algorithms [10]. I was particularly enlightened by the discussion in [10] about the upper bound of human oversight and the need for broader frameworks that do not soley depend on human individuals. I would be exicted to work with Professor Green by combining our interests in new frameworks for. My research interests also overlap with Professor Etyan Adar's recent work on viewing explainable AI through Sensemaking Theory. I believe the strong HCI research at UM-SI will foster my research development in human-AI interaction.

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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[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.