Ph.D. Statement of Purpose

Despite the impressive capabilties of current AI models, leveraging AI in human-AI teams and human-centered task is yet an open problem, one I am deely interested in. As a predoctoral researcher at University of Chicago, my research focuses on **human-centered machine learning (HCML)**; in particular, we are designing AI-assistance frameworks for medical training. My research and experience had inspired my interests in the following problems: I am esepcially interested in (1) designing AI assistance that inspires humans' appropriate reliance instead of blind trust (2) AI learning human perception to provide better AI assistance . I intend to continue working on these problems with the larger goal of building AI asisstance that benefit humans while retaining human agency.

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 running 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 as well as an under-review submission to a major ML conference.

Besides technical skills like designing synthetic experiments and human studies, I learned of an unexpected but useful research practice: be flexible in problem formulation. 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 we detoured to a different research problem, completed it a project on it, and now we return to the problem of AI-driven tutorial.

In this project, we encountered the tradeoff between human agency and performance: showing neutral and informative explanations would retained 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.

Coming from a machine learning and not HCI background, my perspectives towards research were also largely influenced by ML. Specifically, I treated humans as another ML model, (in our project synthetic humans were linear combinations of weights) and focused solely on improving human or team task performance. To generate our ideal assistance, we would tune hyperparameters in our model until we achieved best performance on our synthetic humans.

However, I was gradually exposed to new perspectives towards HAI through a Human-Centered Machine Learning course as well as a Human+AI conference, both at UChicago. I realized that the human side of HAI was equally, if not more, important as the AI side.

For instance, besides improving human performance on a specific form of explanations, I was exposed to problem like what types of explanations are useful to humans for what tasks? Explanations that ML researchers like, like feature importantce, may be completely useless to real humans. More important problems include misalignment between HAI research and real humans, including misalignment in the task and objective, in diserideta of the assistance, in the subjects that use our assistance.

I was used to pursuing an empirical objective like a benchmark, but disregarding the big picture elments of HAI like human and stakeholder needs, how (and should) human use AI, etc. These questions are immensly difficult and open-ended but are also what exictes me about this line of research.

future directions

My experiences collectively shaped my research interests in improving human-AI interaction and collaboration. Besides my interest, I am also very motivated in this direction of research because of its interdisciplinary nature and the visible impact that results from reearch. I hope to continue this line of work in my PhD. Currently, I am interested in the following problems:

1. How can we design AI assistance that inspires humans' appropriate reliance?

In our project, we spent a large amount of time devising a decision support system that differed from existing literature providing model explanations: our goal was to provide faithful and neutral evidence for humans to make independent decisions. Thus we eventually devised a neutral decision support policy that did not reveal AI's predicted label and provided nearest-neighbor explanations from all classes. We showed such a policy provided effective decision support, but it was less effective than a persuasive decision support policy closer to model explanations.

This alludes to a larger problem: the tradeoff between human agency and human-AI team performance. Many human-AI teams provide humans with "persuasive" AI assistance that improves humans performance but at the expense of humans' overreliance and blind trust towards AI [2]. This is undesired and unethical, especially in high-stake domains where humans should have the last say. On the other hand, many work [2], including ours, also show that neutral, non-persuasive assistance that dissuade humans from blindly following AI perform no better than persuasive assistance. Thus, I am interested in solving this dilemma.

2. AI learning human perception to provide better AI assistance

As our work showed, AI learning human perception can provide better decisions support performance; I believe in general, incorporating more human knowledge and intution to AI frameworks has potential to provide more effective AI assistance, but many issues remain unsolved in this nascent direction. For example, our work focuses on visual perception on images as it is the easiest form of perception for humans; perception for other modalities like audio and text are also important but much more difficult to collect and utilize. The limitations and bounds of AI learning human perception is also unexplored, including the amount of data needed, the format of data (we used triplet annotations), how well neural networks can learn, etc.

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. *In Proceedings of the 1st ICML 2022 Workshop on Human-Machine Collaboration and Teaming (HMCaT, ICML 2022)*, 2022.

[2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.