Ph.D. Statement of Purpose

1 introduction

My research interest is **human-centered artificial intelligence**, lying in the intersection of artificial intelligence (AI) and human-computer interaction (HCI). My experience and research at University of Chicago has inspired my interest in the following problems: (i) generating AI assistance that inspires appropriate trust and reliance on the AI and (ii) using AI to learn and model human perception and intuition.

2 general interest

2.1 interest

My interest in HAI is largely inspired by my research with Dr. Chenhao Tan at the University of Chicago as well as his course in Human-Center Machine Learning. Through the course I learned of the human-related side of AI and the numerous complexities in human-AI interaction: When should we fully delegate a task to AI who is responsible for the decisions? When AI is used as assistance to humans, how do we design fair and objective assistance instead of merely persuading humans to follow the AI? I am deeply intrigued by the interactions between human and AIs as black-box decision-making entities and I am deeply interested in solving the problems that arise. My past research has inspired my interested in designing neutral and responsible AI assistance that aims to retain human agency instead of persuading humans to follow AI.

2.2 direct social impact

But besides my interest, human-AI interaction and human-centered AI is an important direction of research as AI has increasingly significant influence in humans' decision-making. A core field of human-centered AI lies in high-stake domains like medical or juridical fields where full automation is undesired, so we need to come up with neutral and responsible AI assistance for humans to make a better final decision. I like HAI's potential to produce direct and visible social impact: it is rewarding and motivates me.

2.3 interdisciplinary

Finally I enjoy the interdisciplinary aspects of human-centered AI research. My past research in human-compatible AI decision-support involved modeling human perception, an issue explored by experimental and cognitive psychologists. My current research aims to leverage our human-compatible AI to provide more effective teaching frameworks for radiology residents and we are exploring psychology literature in learning and categorization. More generally, human-centered AI revolves around how humans interact with a decision-making entity and thus involves many many different fields like economics, sociology, ethics, legal, etc. I greatly enjoy learning and combining knowledge from multiple fields.

3 specific interest

3.1 research experience

At the Pre-Doctoral Masters program at University of Chicago, I worked with Dr. Chenhao Tan on designing AI-driven decision support and training systems for medical teaching. Motivated by AI learning human intuition, we devised a human-compatible model that learned both a classification task and predicting human perception. Such a human-compatible representation learned some form of human similarity function and could be leveraged for case-based decision support: providing assistance as the test case's nearest-neighbor in the training set using our learned similarity function. I was responsible for conducting experiments on a synthetic datasets and a human study on a chest X-ray dataset with Prolific crowdworkers. For the synthetic experiment, we used a binary dataset with controllable features and synthetic human agents by tuning weights on the features; this allowed to exhaustively explore our design, hyperparameters, and the limitations of our model.

Our experiment results showed our human-compatible representation leads to better decision support performance than an AI that only learns classification. While the results are positive, many issues remain unsolved in this nascent direction. This research project inspired my interest in the following problems:

3.2 future directions

(i) generating AI assistance that inspires appropriate trust and reliance on the AI. Many past work provides AI's decision and explanation as assistance and show improved human-AI team performance, but a great part of the improved performance can be attributed to humans simply following AI's suggestion, suggesting overreliance on the AI and a lack of human agency. Our decision support framework instead aims to provide neutral support that retains as much human agency as possible: we provide example explanations from each class and do not reveal AI's predicted label. However, our work also shows providing neutral support leads to lower performance than providing evidence for model prediction. I believe human-AI teams, especially in high-stake domains, are inherently human-centered and such neutral supports are desired, but the tradeoff between human agency and task performance is still an significant open problem that I want to solve.

TODO: potential solutions/proposals to the problem

(ii) using AI to learn and model human perception and intuition. 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.

TODO: potential solutions/proposals to the problem

4 my fit with the school/program

I want to pursue a Ph.D. in human-centered AI because of I enjoy solving problems in human-AI interaction and aspire to do research that has direct social impact. I believe the research at CMU HCII

aligns with my interest and motivation.

TODO:

- sherry wu
- ken
- haiyi zhu