Behavior-Informed AI: Factoring in Human Behavior for Robust Learning and Improved Decision-Making Assistance

Chien-Ju Ho, Washington University in St. Louis

Overview

The rapid advancement of artificial intelligence (AI) has revolutionized the way we approach problem-solving, innovation, and interaction with technology. Central to this evolution is AI's ability to learn from vast amounts of human-generated and annotated data, empowering machines to mimic and even surpass human capabilities across various domains. As AI continues to progress, there lies immense potential to harness its power to augment and enhance human decision-making, ultimately fostering more informed choices and improved outcomes. However, humans are known to make imperfect or even biased decisions. This imperfection impedes AI development by introducing biases into the data used to train AI. Furthermore, to optimally assist humans and prevent them from succumbing to their biases, it is crucial for AI to understand and incorporate knowledge of human behavior and biases.

In this proposal, our goal is to design behavior-informed AI that accounts for human behavior in data generation and decision-making, leading to AI systems that are robust to biased training data and are able to enhance human decisions. To achieve this goal, we will pursue the following research threads:

- Understand and model human behavior: We plan to conduct behavioral experiments to examine human behavior in the context of data annotation and decision-making. Subsequently, we will develop interpretable and accurate human models by leveraging cognitive sciences and machine learning.
- Train AI to be robust to human biases: The framework involves curating bias-mitigated datasets through designing cognitive-grounded bias-mitigation interventions during data collection, as well as developing post-hoc learning algorithms that account for human biases during training.
- Develop behavior-aware assistive AI: We will design assistive AI systems that account for human behavior and biases to enhance human decision making. The assistive AI will adaptively provide structured assistance and update the decision-making environment based on insights derived from human behavior.

The proposed research will be assessed within the application domains of crowdsourcing data collection, homelessness prevention, and pilot augmentation. We will collaborate with domain experts to refine our methodologies, ensuring alignment with the targeted domains, and develop comprehensive evaluation plans.