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 impedes AI development by introducing biases into the data used to train AI. Furthermore, to optimally assist humans and prevent them from succumbing to these biases, it is crucial for machine learning algorithms to understand and incorporate knowledge of human behavior and biases.

In this proposal, our goal is to design behavior-informed AI that understands and 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 decision making. 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.

Intellectual Merit

This proposed research will contribute to the empirical understanding of human behavior in the context of human-AI interactions. It will also provide theoretical foundations for studying the interactions between humans and AI algorithms, through incorporating human models in both learning frameworks and assistive AI design. The results of the proposal will provide insights into developing human-centered machine learning algorithms and in combining humans and machines to solve problems neither can solve alone. This research is interdisciplinary in nature, combining ideas and techniques from machine learning, algorithmic economics, and online behavioral social science.

Broader Impacts

This research has a direct impact on the design of a broad range of online platforms with active human participation. Moreover, it also contributes to improve policy making for societal issues. In particular, the PI has existing collaborations with domain experts in the department of Psychology and Brain Sciences, Brown School of Social Work, and Medical School that apply computational approaches to practical problems such as allocation of scarce resources for homeless prevention and living donor kidney transplantation. The PI plans to continue and expand the collaborations through the Center for Collaborative Human-AI Learning and Operation (HALO) and the Division for Computational and Data Sciences (DCDS) at the Washington University in Saint Louis to address societal issues. This proposal also includes a comprehensive plan for enhancing the education and broadening the research outreach, including creating a course in human-AI collaborations, engaging undergraduate research, and hosting summer workshops for high-school students.