Enhancing Human-AI Collaboration through Adaptive Interaction and Explainability

ZhaoBin Li

University of California, Irvine
zhaobin.li@uci.edu

Abstract

AI is rapidly evolving, and human-AI collaboration is becoming more prevalent. Developing robust, adaptive, and transparent models is key to improving human-AI collaboration. My research explores the intersection of AI explainability and adaptive interaction to enhance collaborative decision-making. Building on my previous work in AI explainability, adversarial robustness, and adaptive algorithms, I aim to develop adaptive interaction mechanisms that are resilient to adversarial attacks and intuitively understandable to human collaborators.

Background

The integration of AI systems into critical decision-making processes highlights the need for models that are both adaptable and explainable. Previous research (Ribeiro, Singh, and Guestrin 2016) has shown that because AI systems are opaque in their decision-making processes, human users cannot easily trust and collaborate with these systems. My earlier studies on adversarial robustness and explainable AI provide a solid foundation for addressing these challenges.

Central Question

How can we design robust AI systems that beneficially collaborate with humans through improved explainability and adaptive interaction?

Previous Research

I published these research papers during college and as a post-bac research assistant:

AI Explainability for Human Oversight

In ”Explainable AI for Natural Adversarial Images” (Folke et al. 2021), we demonstrated that saliency maps and example-based explanations significantly improve human ability to predict AI errors in adversarial scenarios. This research underscores the importance of explainable AI models in enhancing human oversight and trust.

Adaptive Algorithms in Online Learning

My research on using multi-armed bandits in educational technologies (Li et al. 2021) demonstrated that adaptive personalization is a double-edged sword. While it can improve student experiences in some cases, it can also lead to slower adaptation and increased variability in others. Educators need to be discerning when implementing personalization in the classroom.

Intent Obfuscating Attacks

My study on intent obfuscating attacks (Li and Shafto 2023), which I will present at AIES, revealed that using intent obfuscation to generate adversarial examples for object detectors is feasible. This study highlighted the need to strengthen both technical and legal protections against such attacks.

Proposed Research

Building upon these papers, my doctoral research aims to develop and evaluate adaptive AI systems to enhance human-AI collaboration.

Adaptive Interaction Systems

First, I will develop AI systems that dynamically change their behavior based on user response. This adaptive mechanism will be informed by my paper on personalization algorithms in education (Li et al. 2021), which emphasizes the appropriate use of personalized features in improving outcomes. I aim to create systems that can decide whether to bootstrap on pre-trained generic data or to spend time learning and adapting to individual users, thereby enhancing the overall collaborative experience.

Explainable AI Techniques

Explainable AI systems are crucial for building trust and improving user interaction. I will leverage the findings in my papers on using explainable AI to improve human understanding of AI decisions in image categorization (Folke et al. 2021; Bokadia et al. 2022) to develop methods for humans to understand the AI’s decision-making process. This will include the use of saliency maps and example-based explanations to help users understand and anticipate AI behavior. I will also investigate whether integrating these existing methods into a question-and-answer based explanation system will improve human interpretation.

Human-AI Complementarity

Humans and AI have complementary strengths, and I would like to explore whether combining human and AI predictions can improve adversarial robustness. Building on my paper on intent obfuscating attacks (Li and Shafto 2023) and my adviser's paper on human-AI complementarity (Steyvers et al. 2022), I will develop a Bayesian framework to integrate human and machine opinions and improve the overall accuracy of collaborative decision-making.

Empirical Studies

To evaluate these systems, I will conduct empirical studies in various domains such as healthcare, education, and autonomous driving. These studies will assess whether adaptive and explainable AI systems can improve user trust and collaboration outcomes. These findings will provide valuable insights into the practical applications and limitations of human-AI systems.

Contributions and Significance

This research will make significant contributions to human-AI interaction. Developing adaptive AI systems will improve AI's ability to personalize and respond to individual needs, thereby improving collaboration between humans and AI. Moreover, the integration of explainable AI techniques will enhance the transparency of AI systems, helping us to better understand and trust the AI's decisions. Not only will this improve user experience, but it will also enable AI to be utilized in critical domains where trust and reliability are essential.

In conclusion, I aim to build upon my existing research on AI explainability, adversarial robustness, and online adaptive interaction to improve human-AI collaboration. My research will provide new theoretical and empirical insights into developing robust and accurate AI systems that align with human values and needs.

References

Bokadia, H.; Yang, S. C.-H.; Li, Z.; Folke, T.; and Shafto, P. 2022. Evaluating perceptual and semantic interpretability of saliency methods: A case study of melanoma. *Applied AI Letters*, 3(3): e77.

Folke, T.; Li, Z.; Sojitra, R. B.; Yang, S. C.-H.; and Shafto, P. 2021. Explainable AI for Natural Adversarial Images. In *The Ninth International Conference on Learning Representations: Responsible AI Workshop*.

Li, Z.; and Shafto, P. 2023. On feasibility of intent obfuscating attacks. In *The Fortieth International Conference on Machine Learning: The Second Workshop on New Frontiers in Adversarial Machine Learning*.

Li, Z.; Yee, L.; Sauerberg, N.; Sakson, I.; Williams, J. J.; and Rafferty, A. N. 2021. Getting Too Personal (ized): The Importance of Feature Choice in Online Adaptive Algorithms. In *The 14th International Conference on Educational Data Mining: Reinforcement Learning for Education Workshop*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Steyvers, M.; Tejeda, H.; Kerrigan, G.; and Smyth, P. 2022. Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences*, 119(11): e2111547119.