Human-Aware AI – A Foundational Framework for Human-AI Interaction

Sarath Sreedharan

Colorado State University
ssreedh3@colostate.edu

Abstract

We are living through a revolutionary moment in AI history. We are seeing the development of impressive new AI systems at a rate that was unimaginable just a few years ago. However, AI's true potential to transform society remains unrealized, in no small part due to the inability of current systems to work effectively with people. A major hurdle to achieving such coordination is the inherent asymmetry between the AI system and its users. In this talk, I will discuss how the framework of Human-Aware AI (HAAI) provides us with the tools required to bridge this gap and support fluent and intuitive coordination between the AI system and its users.

Double empathy problem (Milton 2012) is a psychological theory developed in the early 2010s to accurately explain the difficulty experienced by autistic people in social interactions and communication. The theory posits that one of the core reasons for the difficulty is the underlying differences between the communication style and other social/cognitive characteristics of autistic and non-autistic people. The current AI landscape is characterized by a large number of highly capable AI systems, whose usefulness is nonetheless limited by a lack of explainability (Gunning 2017) and controllability (Hadfield-Menell et al. 2016). This points to an existence of a double empathy problem between AI systems and their users. However, in the context of AI systems, the problem is even more pronounced as the users may be forced to work with a truly alien intelligence (Kim 2022).

Human-Aware AI (HAAI) (Sreedharan, Kulkarni, and Kambhampati 2022) is a framework that attempts to address this problem by leveraging the very core mechanisms that enable human-human coordination, i.e., mental modeling. Figure 1, presents a visualization of human-AI interaction as conceptualized within HAAI. As evident from the figure, one of the core causes of confusion in this setting could be the mismatch between the human's expectations and the system-generated behavior. However, bridging this expectation mismatch in turn requires the system to be aware of three salient dimensions of asymmetry between the AI system and the user, namely (a) asymmetry in knowledge, (b) asymmetry in inferential capabilities and (c) asymmetry in vocabulary. We have already used the HAAI framework to develop a number of highly effective explanation generation methods that focus on bridging these individual asymmetries (Sreedharan 2022). Going forward I propose to leverage the framework to address challenges related to human-advising of AI systems, including achieving value alignment.

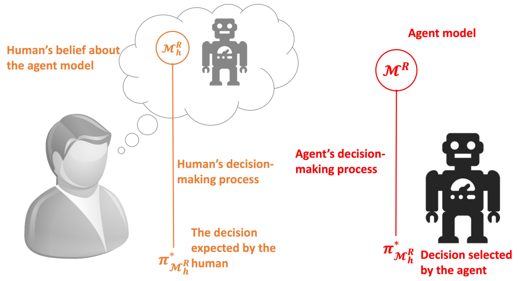
Human's belief about the agent model
$\mathcal{M}_h^R$
Human's decision-making process
The decision expected by the human
$\pi^*_{\mathcal{M}_h^R}$
Agent model
$\mathcal{M}^R$
Agent's decision-making process
$\pi^*_{\mathcal{M}_h^R}$
Decision selected by the agent

Figure 1: A diagrammatic overview of the generalized Human-aware AI (HAAI) framework. The human user their beliefs about the system ($\mathcal{M}_h^R$) and cognitive capabilities to generate an expectation over the system behavior, which could be quite different from the behavior system might choose on their own.

References

Gunning, D. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2(2).

Hadfield-Menell, D.; Russell, S. J.; Abbeel, P.; and Dragan, A. 2016. Cooperative Inverse Reinforcement Learning. In *NIPS*.

Kim, B. 2022. Beyond interpretability: developing a language to shape our relationships with AI. ICLR.

Milton, D. E. 2012. On the ontological status of autism: the 'double empathy problem'. *Disability & Society*, 27(6): 883–887.

Sreedharan, S. 2022. *Foundations of Human-Aware Explanations for Sequential Decision-Making Problems*. Ph.D. thesis, Arizona State University.

Sreedharan, S.; Kulkarni, A.; and Kambhampati, S. 2022. Explainable Human–AI Interaction: A Planning Perspective. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 16(1): 1–184.