Leveraging Human Input to Enable Robust, Interactive, and Aligned AI Systems

Daniel S. Brown

Kahlert School of Computing, Robotics Center, University of Utah
daniel.s.brown@utah.edu

Abstract

Ensuring that AI systems do what we, as humans, actually want them to do, is one of the biggest open research challenges in AI alignment and safety. My research seeks to directly address this challenge by enabling AI systems to interact with humans to learn aligned and robust behaviors. The way robots and other AI systems behave is often the result of optimizing a reward function. However, manually designing good reward functions is highly challenging and error prone, even for domain experts. Although reward functions are often difficult to manually specify, human feedback in the form of demonstrations or preferences is often much easier to obtain. However, human data is often difficult to interpret due to ambiguity and noise. Thus, it is critical that AI systems take into account epistemic uncertainty over the human's true intent. My talk will give an overview of my lab's progress along the following fundamental research areas: (1) efficiently **maintaining uncertainty** over human intent, (2) directly optimizing behavior to be **robust to uncertainty** over human intent, and (3) actively querying for additional human input to **reduce uncertainty** over human intent.

Maintaining Uncertainty

Maintaining uncertainty is a critical first step towards robust AI systems. I will start by discussing my work that seeks to addresses this problem by developing the first **scalable Bayesian reward inference algorithm for visual imitation learning domains.** This research combines my work on learning reward functions from small numbers of ranked, suboptimal demonstrations with self-supervised deep learning techniques to learn a lower-dimensional latent state-representation where Bayesian reward inference becomes tractable. In comparison to prior Bayesian reward inference approaches, which would take *days to run*, my research enables Bayesian reward inference in only a *few minutes* by leveraging a small number of human pairwise preferences over trajectories. This research enables AI systems the ability to self-assess performance, when learning from small numbers of human demonstrations and demonstrated the first practical approach to **automatically detect reward hacking or gaming behaviors**, cases where the AI system's learned reward function results in misaligned behavior.

Robustness to Uncertainty

I will next discuss the challenge of how AI systems should optimize their behavior when they have uncertainty over their objective function. I will highlight my work which proposes a novel risk-aware approach to policy optimization from human feedback that explicitly accounts for uncertainty over the human's true reward function. This work resulted in the first robust reinforcement learning algorithm that optimizes an AI system's behavior over multiple, possibly competing, reward functions inferred from human input and enables AI systems to effectively deal with ambiguous human preferences by **hedging against uncertainty, rather than seeking to uniquely identify the human's reward function**. I will also discuss my labs work on developing the first Bayesian methods for learning safety constraints from demonstrations and preferences.

Actively Reducing Uncertainty

The final question I will cover is what an AI system should do if it has so much uncertainty over the human's true reward function that it does not know how to behave? I will discuss active learning methods that generate queries for additional human input in states where the AI system believes it may have high generalization error. By combining active risk-aware active learning with my work on high-confidence performance bounds, we recently showed for the first time that an **AI system can know with high confidence how many demonstrations it needs to learn a particular task**. I will also discuss how we developed a novel active learning strategy to achieve state-of-the-art interactive imitation learning while also reducing context switches by a human supervisor. This research enables a single human supervisor to **simultaneously manage an entire fleet of robots with minimal cognitive workload** and I will discuss how we recently extended this work to enable effective interactive imitaiton learning for complex surgical retraction tasks. Finally, I will highlight our recent work that allows AI systems to actively query from multiple forms of human feedback based on models of human rationality and bias and the agent's own uncertainty over the human's reward function.

References

Brown, D. S. 2020. *Safe and efficient inverse reinforcement learning*. PhD Dissertation. UT Austin.