Making AI Policies Transparent to Humans through Demonstrations

Michael S. Lee

Robotics Institute, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
ml5@andrew.cmu.edu

Abstract

Demonstrations are a powerful way of increasing the transparency of AI policies to humans. Though we can approximately model human learning from demonstrations as inverse reinforcement learning, we note that human learning can differ from algorithmic learning in key ways, e.g. humans are computationally limited and may sometimes struggle to understand all of the nuances of a demonstration. Unlike related work that provide demonstrations to humans that simply maximize information gain, I leverage concepts from the human education literature, such as the zone of proximal development and scaffolding, to show demonstrations that balance informativeness and difficulty of understanding to maximize human learning.

Introduction

As complex policies learned through reinforcement learning increasingly pervade society, it is paramount that their underlying reward functions and subsequent behaviors are *transparent*, i.e. predictable and understandable to humans. A natural way that humans communicate and comprehend each others' policies is through demonstrations. Thus, one way to increase the transparency of AI policies is also through demonstrations. Furthermore, human behavior is commonly modeled as being driven by reward functions, which can be inferred by other humans through reasoning akin to inverse reinforcement learning (IRL) (Jara-Ettinger 2019). **My research thus models humans as IRL learners, and explores how AI can teach its reward function to humans using informative demonstrations.**

Though I borrow from the IRL literature to model human learning from demonstrations, I note that human learning differs from algorithmic learning in a key way: humans are limited in their computational capacity and may struggle to understand all the nuanced implications of a demonstration given their current beliefs. In contrast to related work that provide demonstrations that simply maximize information gain (Lage et al. 2019; Huang et al. 2019; Qian and Unhelkar 2022), I crucially observe that *informativeness and difficulty of understanding are often two sides of the same coin to humans and thus show demonstrations that balance the two to maximize human learning*.

Teaching Reward Functions in the ZPD

Completed work: Instructional material that is not too easy but also not too difficult for a learner is said to belong in the zone of proximal development (ZPD), also often referred to as the "Goldilocks" zone. Teaching and testing in the ZPD is common in human education to maximize learning.

Key to teaching and testing reward functions in the ZPD is modeling human beliefs and counterfactual reasoning. When considering which demonstration or test to provide next, the AI must ask "How does the human expect me to behave given their current beliefs?" Our insight is to provide a behavior that differs from the human's counterfactual expectation just enough to be meaningfully informative. Too small of a difference and the reconciliation in the human's mind is trivial, and too large of a difference and the gap is irreconcilable in one shot. **My core research contribution to date is selecting teaching demonstrations and tests that lie in the ZPD to maximize human learning (and thus the transparency) of AI reward functions and policies.**

As a case study of these ideas, imagine that a human sees a delivery robot for the first time as it takes a two-action detour around one mud patch (Fig. 1a). Because the robot does not take arbitrary actions and does not go through the mud, human may infer using IRL-like reasoning that this robot deems actions costly and that entering mud must be at least twice as costly as an action. These two relations can be represented as the two half-space constraints in Fig. 1b. Note that this information is gained by comparing the robot's behavior against a counterfactual, i.e. an alternative behavior.

In choosing what to demonstrate next, the robot examines counterfactuals likely to be considered by the human by rolling out trajectories consistent with reward functions sampled from its model of the human's beliefs (Fig. 1b) in various environments (Lee, Admoni, and Simmons 2022a). E.g. when given the environment in Fig. 1c, the human may expect the robot to also detour around two mud patches. And because it would instead go through the mud, the robot considers this an informative next demonstration that may lowerbound the mud cost in the human's mind (Fig. 1d).

And while it may be tempting to always provide demonstrations that yield the highest information gain, my first user study suggests that information gain often correlates with the effort required for the learner to process it (Lee, Admoni, and Simmons 2021). In education, teachers lever-

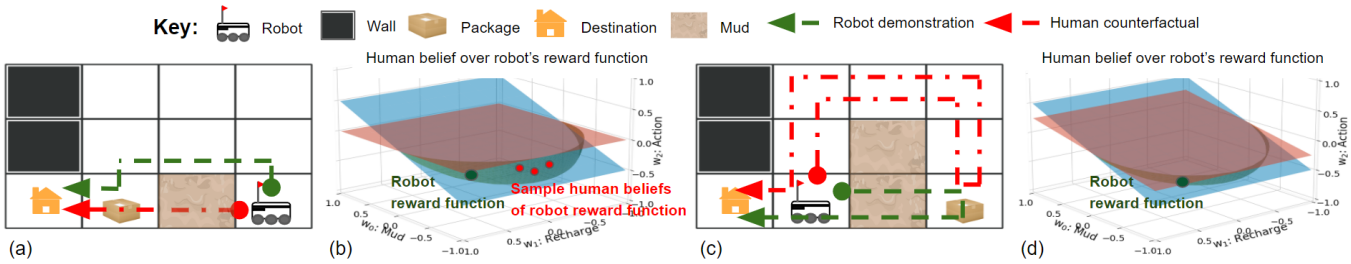

Figure 1: A sequence of demonstrations and the corresponding updated model of human beliefs over robot's reward function.

age the technique of scaffolding to teach in the ZPD, i.e. by providing structure that helps a learner accomplish a task beyond their current abilities. I thus propose an algorithm that scaffolds a curriculum of demonstrations that incrementally increases in information to ease humans into learning (Lee, Admoni, and Simmons 2021).

My second user study taught a robot's reward function via pre-selected demonstrations, then tested the participants' ability to predict robot behavior in unseen scenarios. Results showed that my algorithm for scaffolding demonstrations increased performance on tests examining understanding of later-demonstrated (reward) concepts, but also decreased participants' performance on tests on early-demonstrated concepts, suggesting that I perhaps challenged participants too early without feedback regarding their understanding (Lee, Admoni, and Simmons 2022b). I address the shortcomings of such an open-loop teaching paradigm next.

Current work: An effective teacher engages the learner in a closed-loop fashion, constantly updating their model of the learner's beliefs based on the instruction provided and test responses, then updating the next lesson accordingly.

Each half-space constraint generated by IRL can be treated as a "knowledge concept" (KC) (Koedinger, Corbett, and Perfetti 2012) that encapsulates a characteristic of the reward function (e.g. mud is at least twice as costly as an action) that the human may have internalized. However, a model of human beliefs purely comprised of half-spaces cannot handle conflicts that arise when the human incorrectly applies a KC during testing that was assumed to be learned during teaching (as you cannot reconcile two half-space constraints that point in directly opposite directions).

I thus move to a probabilistic human model in the form of a particle filter. Each particle represents a potential human belief of the robot's reward function, and particle weights are updated in a Bayesian fashion based on constraints conveyed through teaching demonstrations and test responses. By leveraging a particle filter, my algorithm not only selects demonstrations and tests in the ZPD that provide the right amount of information, but also gracefully affords iterative updates to the human model during teaching and testing.

I propose a closed-loop teaching algorithm (Lee, Admoni, and Simmons 2023) that incrementally teaches a set of related KCs (e.g. upper- and lowerbounds on the mud cost) in a series of *units*. For each unit, it provides scaffolded demonstrations, then presents the human with *diagnostic tests* that require understanding of the conveyed KCs. For each missed KC, it provides feedback and a *remedial demonstration* that teaches the KC again as simply as possible. Finally, it ends each unit by continually testing the learner on this KC using *remedial tests* and corrective feedback until they get it right. These remedial tests leverage the *testing effect*, where leveraging tests not as assessments but teaching tools leads to better learning over passively studying (e.g. seeing more demonstrations). A user study finds our proposed closed-loop algorithm reduces the regret in human test responses by 41% over a baseline and is rated as more usable by users in one of the two considered domains (under review).

Future work: My goal is for AI and humans to be able to fluently identify and reconcile gaps in their understanding of each other's reward functions in high dimensional and complex domains. Toward realizing this goal, I am next interested in exploring the following three questions.

- As humans struggle to reason beyond the interaction of three variables, how can we decompose high dimensional reward functions into lower dimensional abstractions?
- *The diversity of environments drives the diversity of demonstrations*. Moving beyond grid worlds, how may we generate sufficiently expressive environments in which an AI can demonstrate the desired information?
- How can demonstrations work in conjunction with local explainable AI techniques (e.g. conveying feature importances for an action) to increase policy transparency?

References

Huang, S.; Held, D.; Abbeel, P.; and Dragan, A. 2019. Enabling robots to communicate their objectives. *Autonomous Robots*.

Jara-Ettinger, J. 2019. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*.

Koedinger, K.; Corbett, A.; and Perfetti, C. 2012. The Knowledge-Learning-Instruction framework. *Cognitive science*.

Lage, I.; Lifschitz, D.; Doshi-Velez, F.; and Amir, O. 2019. Exploring Computational User Models for Agent Policy Summarization. In *International Joint Conference on Artificial Intelligence*.

Lee, M. S.; Admoni, H.; and Simmons, R. 2021. Machine teaching for human inverse reinforcement learning. *Frontiers in Rob. & AI*.

Lee, M. S.; Admoni, H.; and Simmons, R. 2022a. Counterfactual Examples for Human Inverse Reinforcement Learning. *Workshop on Explainable Agency in AAAI Conference*.

Lee, M. S.; Admoni, H.; and Simmons, R. 2022b. Reasoning about Counterfactuals to Improve Human Inverse Reinforcement Learning. In *International Conference on Intelligent Robots and Systems*.

Lee, M. S.; Admoni, H.; and Simmons, R. 2023. Closed-loop Reasoning about Counterfactuals to Improve Policy Transparency. *ICML Workshop on Counterfactuals in Minds and Machines*.

Qian, P.; and Unhelkar, V. 2022. Evaluating the Role of Interactivity on Improving Transparency in Autonomous Agents. In *AAMAS*.