Making Robot Policies Transparent to Humans Through Demonstrations

Michael S. Lee

I. Introduction

Much progress has been made in robots' ability to obtain complex behaviors through reinforcement learning and reward functions. Learning behaviors using rewards provides two key flexibilities: rewards can be learned through many modalities (e.g. demonstrations, preferences, labels) [6], and rewards are often more transferable than policies to new environments [21]. However, using rewards to generate behavior suffers from a big caveat: it is very difficult to ensure predictable behavior in all possible scenarios. For instance, a racing boat trained to maximize its game score ended up doing so by looping over reward targets rather than quickly finishing the race [3].

As robots enter society, it is paramount that their reward functions and subsequent behaviors are *transparent*, such that robot actions are understandable and predictable to humans [5]. Transparency is critical for not only robot developers in reviewing and ensuring proper robot function, but also for end-users in having calibrated expectations for robots, preventing undertrust and disuse, or overtrust and misuse [22].

A natural way that humans communicate and comprehend each others' policies is through demonstrations. Thus, one way to increase the understandability and predictability of robot policies is also through demonstrations [1, 2, 11, 12, 23]. Furthermore, human behavior is commonly modeled as being driven by reward functions [14], which can be inferred by other humans through reasoning akin to inverse reinforcement learning (IRL) [13]. **My research thus models humans as IRL learners, and explores how a robot can teach its reward function to humans using informative demonstrations.**

Though we borrow from the IRL literature to model human learning from demonstrations, we note that humans differ from algorithmic learning in a key way: humans are limited in their computational capacity [7] and may struggle to fully understand all the nuanced implications of a demonstration given their current understanding. Instead of providing demonstrations that simply maximize information gain, we crucially observe that *informativeness and difficulty are often two sides of the same coin to humans [16] and thus show demonstrations that balance the two to maximize human learning*.

II. Teaching Reward Functions in the ZPD

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), often also referred to as the "Goldilocks" zone [10, 26]. While teaching in the ZPD to maximize learning and engagement is both intuitive and has been empirically tested [20, 27], the same benefits can be reaped by testing in the ZPD (e.g. Duolingo defines ZPD for their questions as those that the user is 81% likely to get correctly [25]).

Key to teaching and testing reward functions in the ZPD is counterfactual reasoning. When considering which demonstration or test to provide next, the robot must ask "How would the human expect me to behave given their current beliefs?" The key 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 primary research contribution is selecting teaching demonstrations and tests that lie in the ZPD to maximize human learning (and thus the transparency) of robot reward functions and policies.**

A. Scaffolding Demonstrations of Increasing Informativeness

Teaching in the ZPD requires accurately measuring and accounting for a demonstration's informativeness at revealing the robot's reward function to a human. Our key insight is that *a demonstration's informativeness to a human is not intrinsic, but is inherently tied to that human's prior beliefs and their subsequent interpretation of the demonstration* [16–19].

Imagine that a human sees a delivery robot for the first time as it take 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 can 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.

The robot thus hallucinates counterfactuals that the human is likely to consider in new environments by rolling out reward functions sampled from its running model of the human's beliefs [19]. For instance, when faced with the environment in Fig. 1c, the human may think the robot would also detour around two mud patches. And because it would instead go through the mud in this case, the robot knows that this would be an informative demonstration to provide, which would then lowerbound the cost of the mud in the human's mind (Fig. 1d).

And while it may be tempting to provide demonstrations that yield the largest information gain, our first user study suggests that information gain often correlates to the effort required for the learner to process it [16]. In education, teachers leverage the technique of scaffolding to provide additional structure that helps a learner accomplish a task beyond their current abilities [28]. We thus propose an algorithm to scaffold

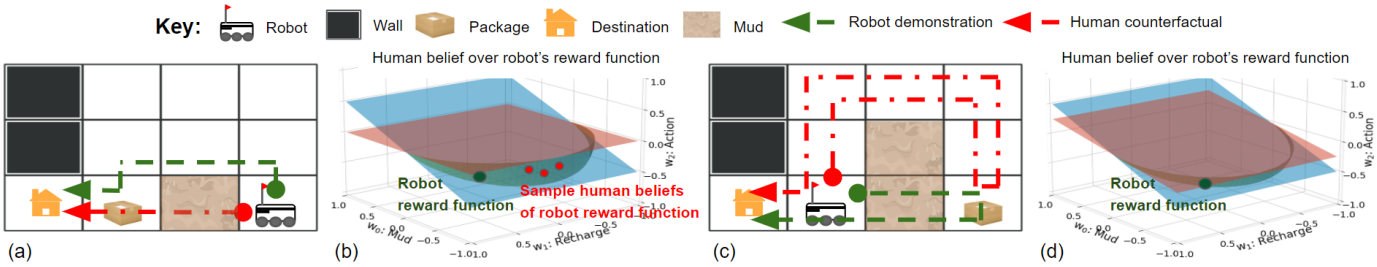

Fig. 1. **(a)** A robot's optimal demonstration (green) is shown in contrast to a suboptimal counterfactual alternative (red). **(b)** A model of the human's belief over the robot's reward function following the demonstration in (a). **(c)** A robot's optimal demonstration is shown in contrast to a counterfactual likely considered by the human (corresponding to the three red belief samples in (b)). **(d)** A model of the human's belief following the demonstration in (c).

demonstrations that incrementally increase in information gain and simultaneously ease humans into learning [17].

Our second user study in which we taught a robot's reward function via pre-selected demonstrations, then tested the participants' ability to correctly predict robot behavior in unseen scenarios showed that our algorithm for scaffolding demonstrations increased performance on difficult tests [18]. However our method also decreased participants' performance on easy tests, suggesting that we perhaps challenged participants too early without any feedback regarding their understanding. We address the shortcomings of such open-loop teaching next.

B. Closing the Teaching Loop with Targeted Tests

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 the learner's test responses, then updating the next lesson accordingly.

Each half-space constraint generated by IRL [21] can be treated as a "knowledge concept" (KC) [15] that encapsulates a characteristic of the reward function (e.g. mud is at least twice as costly as an action) that the human has hopefully 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 opposite directions).

We thus move to a probabilistic human model in the form of a particle filter [4]. Each particle represents a potential human belief regarding 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, our 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.*

We propose a closed-loop teaching algorithm 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 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 immediate corrective feedback until they get it right. These remedial tests leverage the *testing effect* [24], where leveraging tests not as assessment but teaching tools leads to better learning over passively studying (e.g. seeing more demonstrations). We are creating a user study to assess the effectiveness of the proposed closed-loop teaching method.

III. Future Work

My long-term goal is for robots 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.

Teaching high dimensional reward functions: Our work thus far has focused on conveying reward functions comprised of three features, but robots must be able to reason about more features to navigate many real-world scenarios. Interestingly, humans struggle to reason about statistical correlations beyond three variables [9] and yet operate fluently in many complex domains. Humans are adept at causal reasoning, filtering out irrelevant variables, and learning hierarchical abstractions (e.g. *force = mass·acc, work = force·dist*). How can we use these insights to factor, abstract, and decompose high dimensional reward functions into lower dimensional embeddings that can be easily and semantically communicated to a human?

Conceiving of demonstrations in a generative fashion: A key insight of our work is that the *diversity of environments drives the diversity of demonstrations*. Our work has focused on grid world environments in which we can enumerate all possible demonstrations and evaluate their informativeness. But as we move to higher dimensions and continuous domains, such enumeration and discriminative reasoning will no longer be feasible. Instead, how can the robot take a generative approach in constructing a sufficiently expressive environment in which it can demonstrate and convey the desired information?

Incorporating demonstration feedback into robot learning: Our work has explored a robot demonstrating its "optimal" reward function to a human learner. However, I hypothesize that providing demonstrations of a robot's "suboptimal" reward function *during robot learning* will provide human teachers with feedback on current robot deficiencies and how they be remedied. Given the information asymmetry in which the robot and human each understands their own reward function but has an imperfect model of the other's [8], how and when should the robot ask for a demonstration to learn or provide a demonstration to clarify its current reward function?

References

[1] Dan Amir and Ofra Amir. Highlights: Summarizing agent behavior to people. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 2018.

[2] Ofra Amir, Finale Doshi-Velez, and David Sarne. Summarizing agent strategies. *Autonomous Agents and Multi-Agent Systems*, Sep 2019. ISSN 1573-7454.

[3] Dario Amodei and Jack Clark. Faulty reward functions in the wild, 2016. URL https://openai.com/blog/faulty-reward-functions/.

[4] Arnaud Doucet, Adam M Johansen, et al. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704):3, 2009.

[5] Mica R Endsley. From here to autonomy: lessons learned from human–automation research. *Human factors*, 59(1): 5–27, 2017.

[6] Tesca Fitzgerald, Pallavi Koppol, Patrick Callaghan, Russel Q. Wong, Reid Simmons, Oliver Kroemer, and Henny Admoni. Inquire: Interactive querying for user-aware informative reasoning. *Conference on Robot Learning (CoRL)*, 2022.

[7] Thomas L Griffiths. Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*, 2020.

[8] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

[9] Graeme S Halford, Rosemary Baker, Julie E McCredden, and John D Bain. How many variables can humans process? *Psychological science*, 2005.

[10] John Hattie and Shirley Clarke. *Visible learning: feedback*. Routledge, 2018.

[11] Sandy H Huang, Kush Bhatia, Pieter Abbeel, and Anca D Dragan. Establishing appropriate trust via critical states. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[12] Sandy H Huang, David Held, Pieter Abbeel, and Anca D Dragan. Enabling robots to communicate their objectives. *Autonomous Robots*, 2019.

[13] Julian Jara-Ettinger. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 2019.

[14] Julian Jara-Ettinger, Hyowon Gweon, Laura E Schulz, and Joshua B Tenenbaum. The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, 2016.

[15] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, 36(5):757–798, 2012.

[16] Michael S Lee, Henny Admoni, and Reid Simmons. Machine teaching for human inverse reinforcement learning. *Frontiers in Robotics and AI*, 8:693050, 2021.

[17] Michael S Lee, Henny Admoni, and Reid Simmons. Robot teaching for human inverse reinforcement learning. *Workshop on Robots for Learning at ACM/IEEE International Conference on Human-Robot Interaction*, 2022.

[18] Michael S Lee, Henny Admoni, and Reid Simmons. Reasoning about counterfactuals to improve human inverse reinforcement learning. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9140–9147. IEEE, 2022.

[19] Michael S Lee, Henny Admoni, and Reid Simmons. Counterfactual examples for human inverse reinforcement learning. *Workshop on Explainable Agency in Artificial Intelligence at AAAI Conference on Artificial Intelligence*, 2022.

[20] Saul McLeod. What is the zone of proximal development?, 2019. URL https://simplypsychology.org/Zone-of-Proximal-Development.html.

[21] Andrew Y Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *International Conf. on Machine Learning*, 2000.

[22] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39 (2):230–253, 1997.

[23] Peizhu Qian and Vaibhav Unhelkar. Evaluating the role of interactivity on improving transparency in autonomous agents. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 1083–1091, 2022.

[24] Henry L Roediger III and Jeffrey D Karpicke. The power of testing memory: Basic research and implications for educational practice. *Perspectives on psychological science*, 1(3):181–210, 2006.

[25] Luis von Ahn. How duolingo uses ai to assess, engage, and teach better. Advances in Neural Information Processing Systems, 2021. URL https://nips.cc/virtual/2021/invited-talk/22280.

[26] Lev Semenovich Vygotsky. *Mind in society: The development of higher psychological processes*. Harvard university press, 1980.

[27] David Wood and David Middleton. A study of assisted problem-solving. *British journal of psychology*, 66(2): 181–191, 1975.

[28] David Wood, Jerome S Bruner, and Gail Ross. The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, 1976.