Improving the Transparency of Agent Decision Making to Humans Using Demonstrations

Michael S. Lee

CMU-RI-TR-24-05

February 28, 2024

The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:

Reid Simmons, *Co-chair*
Henny Admoni, *Co-chair*
David Held, *Carnegie Mellon University*
Scott Niekum, *University of Massachusetts Amherst*

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Robotics.

Abstract

For intelligent agents (e.g. robots) to be seamlessly integrated into human society, humans must be able to understand their decision making. For example, the decision making of autonomous cars must be clear to the engineers certifying their safety, passengers riding them, and nearby drivers negotiating the road simultaneously. As an agent's decision making depends on its reward function to a great extent, we focus on teaching agent reward functions to humans. Through reasoning that resembles inverse reinforcement learning (IRL), humans naturally infer reward functions that underlie demonstrations of decision-making. Thus agents can teach their reward functions through demonstrations that are *informative* for IRL. However, we critically note that IRL does not consider the *difficulty* for a human to learn from each demonstration. Thus, this thesis proposes to augment teaching for IRL with principles from the education literature to provide demonstrations that belong in a human's zone of proximal development (ZPD) or their "Goldilocks" zone, i.e. demonstrations that are not too easy nor too difficult given their current beliefs. This thesis provides contributions in the following three areas.

We first consider the problem of *teaching* reward functions through select demonstrations. Based on ZPD, we use scaffolding to convey demonstrations that gradually increase in information gain and difficulty and ease the human into learning. Importantly, we argue that a demonstration's information gain is not intrinsic to the demonstration itself but must be conditioned on the human's current beliefs. An informative demonstration is accordingly one that meaningfully differs from the human's expectations (i.e. counterfactuals) of what the agent will do given their current understanding of the agent's decision making.

We secondly consider the problem of *testing* how much the human has learned from demonstrations, by asking humans to predict the agent's actions in new environments. We demonstrate two ways of measuring the difficulty of a test for a human. The first is a gross measure of difficulty that correlates test difficulty with the answer's information gain at revealing the agent's reward function. The second is a more tailored measure that conditions the difficulty of a test on the human's current beliefs of the reward function, estimating difficulty as the proportion of the human's beliefs that would yield the correct answer.

Finally, we introduce a *closed-loop teaching framework* that brings together teaching and testing. While informative teaching demonstrations may be selected *a priori*, student learning may deviate from the preselected curriculum *in situ*. Our teaching framework thus provides intermittent tests and feedback in between groups of related demonstrations to support tailored instruction in two ways. First, we are able to maintain a novel particle filter model of human beliefs and provide demonstrations targeted to the human's current understanding. And second, we are able to leverage tests not only as a tool for assessment but also for teaching, according to the *testing effect* in the education literature.

Through various user studies, we find that our demonstrations targeted for a human's ZPD increase learning outcomes (e.g. the human's ability to predict the agent's actions in new environments). However, we find that learning gains can be associated with increased mental effort for the human to update their beliefs, highlighting again the importance of selecting demonstrations that differ just enough from human expectations to be informative but not too difficult to understand. We also see that such informative demonstrations often illuminate trade-offs inherent in an agent's reward function that may be subtle and difficult to predict a priori, such as how far an agent is willing to detour around a potentially dangerous terrain like mud. And finally, we find interesting interaction effects between our various grid world domains and our results in our later user studies, and we provide further insights on how domains may be characterized in light of the observation that the best teaching method is likely domain-dependent.

Acknowledgments

As this chapter of my life comes to a close, I want to thank the many people who made these years in Pittsburgh so memorable.

I first want to thank my thesis committee. Reid and Henny, thank you for being such incredible mentors who truly care about your students both as researchers and holistically as people. The level of feedback, guidance, and care that you provide your students is unparalleled, to the point where others comment that my advising experience "sounds too good to be true." I count myself very lucky to have been part of the few who were privileged to have you as the co-advising power duo with complementary strengths: Henny's ability to see the bigger picture and shape the narrative masterfully, and Reid's ability to provide technical guidance, notice the finest of details, and ask no-nonsense questions with his very direct gaze. Dave, thank you for providing valuable guidance right from the start, as I first navigated my research qualifier and then my thesis. Noting your prolific lab and many acts of service to the department, RI is very lucky to have you. And thank you Scott for always being available for insightful feedback despite your busyness with your transition between institutions. I've always admired your work and I'm proud to say that your paper on teaching IRL learners through Set Cover Optimal Teaching is what inspired this whole thesis.

A perk of being co-advised is getting to be a part of two awesome labs, HARP lab and RASL. First, to Pallavi and Sarthak, the original MURI crew. Thanks for all of the camaraderie, support, and inspiration over the years Pallavi; I'm so thankful that we got to walk through the PhD together side by side. And borrowing Pallavi's words for you, you have 'such a zest for life' Sarthak and it's contagious! From our time together in our first year PhD office to now, I've always enjoyed learning from your many varied interests and knowledge Ada (like games, baking, etc)! Suresh, I'm so thankful that you came on as a post-doc and extended this line of work. Thanks for the many fruitful discussions, your patience, and your understanding. Michelle, you have such positive social energy – I loved hearing about your adventures and living vicariously through them! Pat, you're one of the friendliest people that I know – thanks for the sincere conversations and the laughs. Thank Roshni for always being so supportive and for being the lab's plant and baking guru. Having entered CMU at the same time, it was nice celebrating and commiserating through all of the ups and downs of grad school with you Ben. Pranay,

you hands down have the best successes + challenges updates; it's been fun sharing laughs together. Thanks for being my gym buddy Abhijat, forever showing me PPL and also helping me understand the finer sides of culture. I'll remember our final months of being office squatting buddies Daphne in rooms with and without windows! Thanks, Reuben for sharing your wisdom about what it means to be a good lab member and building up goodwill. Thanks, Tesca for being a fantastic baker and for opening up your place for a memorable post-covid Friendsgiving! And Gavin, you're a superstar and I'm excited to see the place you're headed. And finally, thanks to Shenai, Rithika, and Vignesh for being wonderful undergrads to work with and for helping make the user studies come to life!

Carnegie Mellon is a special place that is and has been home to many incredible individuals in its broader community. Thank you, Red for creating an environment in which I was able to thrive during my Master's. Your ability to not only think but also live larger-than-life is truly inspirational. Thank you Nate for taking me under your wing in 2015 and for being patient with me throughout that summer, my senior year of undergrad, and throughout my Master's. I would not be where I am today with you and I am forever grateful. A huge thank you to Rachel and John for continuing to invest in RISS, a program that provides such formative research experiences to so many young students. I count myself very lucky to be an alumnus, and I would not have gone to grad school if not for RISS. Thank you Aaron for leading the SUCCESS-MURI project that I got to be part of for my entire PhD, which also led to fruitful collaborations with colleagues at UMass Lowell, BYU, and Tufts. To the Shadyside Crisps: Cormac, I'm so grateful that we got to walk alongside one another for almost 5 years. In so many ways, you modeled how to live life more fully (going all in on celebrations and hobbies), more generously, and to think more deeply. Nadine, our conversations always end up being so deep and insightful. I'm so glad that we're always able to catch up like old friends regardless of how much time has passed. And Tess, thanks for being the fun but responsible mom of the group and always taking care of us! And a shoutout to the honorary crisps Yun and Kevin – always chill, friendly, and thoughtful. To Angela, an unexpected blessing and a friend whom I could confide in, grow with, and wrestle through some of life's biggest questions. Ravi and Thomas, I'll remember our trip to Japan together with Michelle, and very much appreciate your guys' friendship through our periodic meet-ups and conversations. Pragna, I'll miss being (nearby) officemates and your cheerful and empathetic presence! Matt, you'll always be my DOE Trainee buddy from our time during our Master's.

Thanks for rooting for me throughout my PhD, and always treating me to food with that good industry money. To friends in TBD lab: Sam and Zhi, I'll remember our trip to Korea together fondly, and with Kate and Allan, our good conversations in the hallways and during HRI socials. To my very first lab, RISLab. Thank you Shaurya, John, Vishnu, and Micah for being some of my first mentors at CMU and guiding me patiently. From pulling an all-nighter for our RISS 2015 posters to now, we're both finally on the other side Arjav! I enjoyed transitioning to PhD with you Tab, first talking about potential advisers, then also talking about many deeper topics as we experienced various milestones. To the Quatros Amigos, Xuning, Aditya, and (honorary member) Jerry, for a wonderful and memorable beginning to the PhD in Spain! And finally, a thank you to many others from RISLab, including Wennie, Alex, Vibhav, Logan, Lauren, Kshitij, Mosam, Shobhit, and others. As we all finally graduate soon, I'll miss our meals and conversations Ashwin and Shivam. Finally, thank you to the admins who made all of the all-important logistics a breeze – Suzanne, Nora, Keyla, Dayna, Stephanie, and Devin.

And to friends near and far. Clement, you've walked with me through all of the highs and lows of the past twelve years and I'm so thankful to have someone that I can completely be myself around every week. Nate, I'm so glad that we got to be housemates for the last couple of years, always supporting one another and also taking much joy in the simple things of life like Costco runs. Richard, thanks for believing in me and investing in me. I've learned so much from you on how to live in the Spirit, think like a research scientist, to expect great things while remaining humbly reliant on grace. Ben, it's been an honor turning 30 together and growing through our joint trips to conferences last year. And finally, to Jin, my oldest friend since 4th grade – I'm glad that we kept in touch all these years and still get to celebrate each others' big moments.

To friends at Central Church, my spiritual home during my time in Pittsburgh. Chang, JHo, Joseph, Sunny, and Nick, thanks for being brothers that I could confide in during good weeks or bad weeks. To the Veld crew – Keziah, Anna, Eric, and Lisa – you guys helped make the pandemic one of the most meaningful periods of my life and I am so grateful. To everyone in the Brenda and Karen's Knitting Crew – April, Brian, Eric, Izzy, Laura, Matt, Sarah, SLam, Timbox, thanks for the community and the many hangouts. And a special thanks to Timbox for always opening up his home and also being a mentor to me. And finally, thanks to Tobin for being one of the first welcoming faces I met when I

first arrived in Pittsburgh.

Thank you to my wonderful family for their unconditional love, ceaseless prayers, and unwavering support. I honestly could not have asked for better and love you all dearly. Mom and Dad, thank you for always loving me and David so selflessly. And David, I am so proud of the man that you've become and I'm excited to see how you'll continue to grow. Finally, thank you God for always being with me and leading me through this extended season. Soli Deo Gloria.

Funding

This work was supported by the Office of Naval Research under award N00014-181-2503.

Contents

When this dissertation is viewed as a PDF, the page header is a link to this Table of Contents.

List of Figures

List of Tables

1 Introduction

As intelligent agents (e.g. robots) become ubiquitous in our world, our capacity to deploy, collaborate, and co-exist fluently with them as humans is contingent on our ability to understand and predict their decision making. For example, an engineer certifying the navigation policy of a ground delivery robot may ask, "Does the robot understand all the terrains it might encounter well enough to successfully balance efficiency and safety?" A human driver may wonder, "Will this autonomous car slightly ahead of me try and merge into my heavily crowded lane?" And finally, new owners of an autonomous vacuum may wonder, ""How much clutter will the robot tolerate in an area before it steers clear to ensure it does not get stuck?" An incorrect model of agent decision making may lead to premature deployments, unsafe interactions, and inefficient use of such agents. It is thus imperative that agent decision making is *predictable and understandable*, and thus *transparent* [23], to developers, collaborators, and end-users of intelligent agents.

A critical way in which people communicate and comprehend each others' decision-making is through demonstrations. Humans will naturally observe agent behavior over time and continuously refine their belief of the agent's decision making in a process called familiarization [19]. But as Huang et al. [35] note, this *passive* process can be inefficient. Instead, we explicitly model both how humans learn and their current belief over the agent's decision making to *actively* select informative demonstrations that help the human converge quickly to the agent's true model. Predictably, the effectiveness of such a method hinges in part on the accuracy of our models of human learning and human beliefs. Cognitive science suggests that humans often model

one another's behavior as exactly or approximately maximizing a reward function [38, 39, 62], which they can infer through reasoning resembling inverse reinforcement learning (IRL) [10, 11, 37, 66].

Though standard IRL [66] (henceforth referred to simply as IRL for the sake of brevity) is a good foundation for modeling human learning of potential reward functions underlying demonstrations, it does not fully capture the multi-faceted nature of human learning. One key aspect of human learning that it fails to consider is the difficulty for a human to learn from a demonstration. Humans are limited in their computational capacity [28] and may struggle to fully understand all of the nuanced implications of a demonstration given their current knowledge. This relates to the influential idea of Lev Vygotsky's zone of proximal development (ZPD) [92], which suggests that learning best occurs in the region between what a student can accomplish on their own and what they can accomplish when they are supported with the right level of assistance. ZPD is a general principle that has been leveraged in analyzing and informing instruction across various knowledge-based learning subjects, such as mathematics [25, 46, 86] and language [3, 42, 50], e.g. the popular language learning app Duolingo defines ZPD for their questions as those that the user is 81% likely to get correctly [24, 89]. ZPD can inform subject-agnostic learning environments such as game-based learning [73] as well. This thesis builds on the **key insight** that information gain and difficulty are often correlated with one another for human learners. Thus, in contrast to the standard paradigm of providing humans with demonstrations that simply maximize information gain [35, 47, 76], we **hypothesize** that teaching in the ZPD (i.e. engaging at the right level of information gain and difficulty conditioned on the human beliefs) will be informative to the human, and will significantly improve human learning of agent decision making. In our hypothesis, we take care to differentiate between *information gain*, which corresponds to the *expected* reduction in human uncertainty over agent reward function, and *informativeness*, which corresponds to the *actual* reduction in human uncertainty over agent reward function.

Consider how an autonomous car may convey its reward function to a human. For example, the car could ease in someone with no prior knowledge by first demonstrating successful driving in nominal conditions that have moderate information gain but also easy to comprehend (e.g. in open roads with minimal required turns or lane changes).

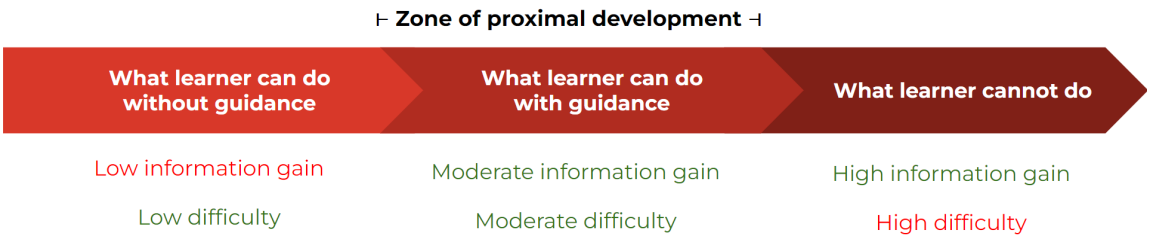

Figure 1.1: This thesis aims to provide instruction at the right level of information gain and difficulty for learners, i.e. in their zone of proximal development.

Then it could proceed to demonstrate more nuanced behavior, such as how it trades off efficiency and caution in more challenging scenarios that have high information gain but perhaps more difficult to comprehend (e.g. needing to make a quick lane change to make an upcoming exit). This intuitive sequence of demonstrations carefully balances the goal of communicating information with the difficulty of comprehension, namely by gradually increasing in both information gain and difficulty.

In sequencing informative yet comprehensible demonstrations, it is important to model the human's prior knowledge and tailor the demonstrations accordingly. For example, a "novice" would benefit from starter demonstrations that convey introductory information that are comprehensible with no prior knowledge. However, these same demonstrations would elicit boredom or frustration when shown to an "expert", who should instead directly be shown more nuanced behavior that further refines their current knowledge.

To test the human's understanding of the agent's reward function after seeing demonstrations, we can ask how they believe the agent would behave in unseen environments. For example, we could test a human's understanding of an autonomous car's reward function by asking how they believe the car would attempt to reach a quickly approaching exit if they were one lane away, two lanes away, in light traffic, or in heavy traffic, etc.

Finally, though a curriculum of informative demonstrations may be selected *a priori*, student learning may deviate from the preselected curriculum *in situ*. Thus testing is not only critical for assessing understanding after the full curriculum has been shown, but also intermittently throughout the curriculum to ensure an up-to-date model of human understanding and to provide tailored instruction. And the *testing effect* [79] from the education literature predicts an increase in learning

outcomes when a portion of the teaching budget is devoted to testing the student, i.e. leveraging testing not only as a tool for assessment but also for teaching.

Thesis contributions

The following chapters provide algorithms for both teaching and testing with various contributions (C1 - C9), with corresponding user studies to test our hypotheses.

Chapter 4 primarily concerns how to measure and account for the information gain of a demonstration (at revealing the agent's underlying reward function) for teaching and testing.

C1. We posit that the information gain of a demonstration during teaching also corresponds to the effort required for a human to extract that information. Thus we propose an algorithm for *scaffolding* when conveying demonstrations to a human, such that they gradually increase in information gain and difficulty and ease the human into learning.

C2. We show that the information gain of a demonstration during teaching correlates directly to the *difficulty* for a human to predict it in an unseen environment during testing.

C3. We also note that human learners are likely also influenced by the *visual features and sequence* in which demonstrations are conveyed. We show how promoting the simplicity of visuals and affordance of a discernible pattern during teaching can improve learning outcomes.

Chapter 5 primarily concerns how a human's beliefs over the agent's reward function impact the measure of a demonstration's information gain during teaching and testing.

C4. We update our measure of a demonstration's information gain by leveraging the idea that a teaching demonstration that provides information gain is one that differs meaningfully from the learner's expectations (i.e. *counterfactuals*) of what the agent will do given their current understanding of the agent's decision making. We provide an algorithm that scaffolds demonstrations accordingly

which yields mixed results.

C5. We update our measure of test difficulty by *conditioning* it on a human's current beliefs and measuring how many of that individual's beliefs would yield the correct agent behavior, and show that it correlates with test performance.

C6. With the hypothesis that encouraging simplicity in not only the visuals but also in reward features will assist in teaching, we propose a method for gradually scaling up of the number of features conveyed to further ease information transfer.

Chapter 6 explores augmenting a curriculum of informative demonstrations (selected using the ideas from the prior chapters) with a closed feedback loop to provide tailored instruction in real-time. And while this thesis focuses primarily on increasing transparency via increasing predictability, this chapter also briefly discusses understandability.

C7. We propose a closed-loop teaching framework based on insights from the education literature that complements demonstrations with intermittent tests and feedback and show that the framework reduces regret of human test responses by 43% over an open-loop baseline.

C8. We develop a novel particle filter model of human beliefs that simultaneously supports iterative updates and calibrated predictions of the counterfactuals likely considered by the human for each subsequent demonstration that could be provided.

C9. We finally contextualize our results in light of another common reward teaching technique (direct numerical reward explanation) with an exploratory user study, finding a strong interaction effect of domain on learning outcome.

2 Related Work

2.1 Principles from Education and Cognitive Science for Teaching Humans

The science of teaching and explaining concepts to humans is a multifaceted process that has been studied extensively. Thus, we take inspiration from the education and cognitive science literature on how humans provide explanations in informing how an agent may teach a skill to a human learner so that the learner may correctly reproduce that skill in new situations. We highlight below the major principles that guide the teaching and testing frameworks developed in this thesis. For a list of just a few of the many other educational principles and factors that influence student learning, we refer the reader to [44] and [30] respectively.

Teaching

First, we focus on teaching in the zone of proximal development (ZPD) [92], which suggests that learning best occurs in the region between what a student can accomplish on their own and what they can accomplish when they are supported with the right level of assistance. To do so, we combine the following principles (from the education literature and cognitive science) to inform our approach.

Scaffolding is a well-established pedagogical technique in which a more knowledgeable teacher assists a learner in accomplishing a task currently beyond the learner's abilities, e.g. by reducing the degrees of freedom of the problem and/or

by demonstrating partial solutions to the task [96]. Noting the benefits seen by automated scaffolding to date (e.g. [80]), we implement the first recommendation made by [78] for software-based scaffolding, which is to reduce the complexity of the learning problem through additional structure. Specifically, we incorporate this technique when teaching a skill by providing demonstrations that sequentially increase in information gain and difficulty.

In our work, we note that information gain is not an intrinsic quantity but is dependent on the human's current beliefs in two ways. First, the literature on how humans explain to one another notes that "explanations are contrastive—they are sought in response to particular counterfactual cases," and that it is critical that the learner's counterfactuals matches the ones intended by the teacher [65]. Thus, we select demonstrations that provide information with respect to the learner's current beliefs and the **counterfactuals** that they will likely consider. Second, Reiser [78] suggests that scaffolding should sometimes challenge the learner by inducing cognitive conflict. Thus, we not only wish to provide demonstrations that are supported by the learner's beliefs but demonstrations differ just enough from the human's current expectations 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.

Finally, though the information gain of a demonstration is a critical factor in scaffolding, human learning is multi-faceted and is also influenced by other factors. For example, studies on explanations preferred by humans indicate a bias toward those that are simpler and have fewer causes [61]. Furthermore, [95] found that explanations can be detrimental if they do not help the learner notice useful patterns or even mislead them with false patterns. Together, these two works support the idea that explanations should minimize distractions that potentially inspire false correlations and instead highlight and reinforce the minimal set of causes. We thus also optimize for **simplicity and pattern discovery** when selecting demonstrations that naturally "explain" the agent's underlying reward function.

Testing

Effective scaffolding requires an accurate diagnosis of the learner's current abilities to provide the appropriate level of assistance throughout the teaching process [15]. A common diagnostic method is presenting the learner with tests of **varying difficulties** and assessing their understanding of a skill. Toward this, we propose a way to quantify the difficulty of a test that specifically assesses the student's ability to predict the right behavior in a new situation.

Finally, the **testing effect** from the education literature [79] predicts an increase in learning outcomes when a portion of the teaching budget is devoted to testing the student. We thus leverage tests not only for assessment but also intermittently during teaching. And when testing, immediate **feedback** on errors made in the test responses can be leveraged to yield better learning outcomes [44].

2.2 Explainable Reinforcement Learning

The field of explainable reinforcement learning (RL) focuses on assisting humans in understanding the decision making of RL agents. Recent surveys [64, 75, 94] highlight a variety of approaches, such as approximating a black box RL policy via an interpretable model (e.g. a decision tree [85]), using saliency maps to highlight features of a state used for decision making [27], visualizing minimally different counterfactual states that would have yielded a different action [67], and identification of critical training points (e.g. for estimating Q-values [26]). The most recent survey by Milani et al. [64] divides the work in this field into three categories of methods: *feature importance* methods that highlight the features that influenced the agent's decision making, *learning process and MDP* methods that highlight relevant past experiences or MDP components that lead to the agent's current action, and *policy-level* methods that convey the agent's general long-term behavior. In this work, we contribute a *policy-level* method that conveys an understanding of an agent's overall behavior to a human through representative demonstrations.

2.2.1 Policy Summarization

The problem of policy summarization considers which states and actions (i.e. demonstrations) should be conveyed to help a user obtain a global understanding of an agent's policy [6]. There are two primary approaches to this problem.

Heuristic-driven

The first relies on heuristics to evaluate the value of communicating certain agent states and actions to humans. Huang et al. [34] considers conveying "critical states" in which it is much worse to act randomly than optimally, measuring the entropy of the agent's action distribution for a policy trained using maximum-entropy reinforcement learning. Amir and Amir [5] similarly measures a state's "importance" for being conveyed to a human as the difference between its best and worst Q-values. Finally, this class of methods can be extended to comparing and contrasting the policies of two agents [7].

Machine Teaching

We build on the second approach, which follows the machine teaching paradigm [104]. Given an assumed learning model of the student (e.g. IRL to learn a reward function), the machine teaching objective is to select the minimal set of teaching examples (i.e. demonstrations) that will help the learner arrive at a specific target model (e.g. a policy). Though machine teaching was first applied to classification and regression [60, 103], it has also recently been employed to convey reward functions from which the corresponding policy can be reconstructed. Various methods for conveying a reward function to humans are surveyed by Sanneman et al. [81] and we summarize a couple of relevant works below.

Huang et al. [35] selected informative demonstrations for humans modeled to employ approximate Bayesian IRL for recovering the reward. This technique requires the true reward function to be within a candidate set of reward functions over which to perform Bayesian inference, and computation scales linearly with the size of the set. And as the candidate set of reward functions is not updated with additional demonstrations (e.g. to remove unlikely original candidate reward functions and

resample new candidates in a more likely region), this method is sensitive to the initial sampling and perhaps leads to slower convergence to the robot's reward function. Cakmak and Lopes [13] instead focused on IRL learners and selected demonstrations that maximally reduced uncertainty over all viable reward parameters, posed as a volume removal problem. Brown and Niekum [12] improved this method (particularly for high dimensions) by solving an equivalent set cover problem instead with their Set Cover Optimal Teaching (SCOT) algorithm. However, SCOT is not explicitly designed for human learners and this thesis builds on SCOT to address that gap.

And finally, though this field has traditionally focused on methods where the agent directs human learning, recent methods explore giving control to the human. Qian and Unhelkar [76] explore interactive policy summarization in which they explore how allowing the student to request specific demonstrations impacts their learning via a GUI. They find that a hybrid strategy of AI-selected and human-selected demonstrations yields the best objective and subjective results; our proposed methods could supply the former demonstrations in their framework. Finally, Amitai et al. [8] develop a GUI-based interactive tool that allows a human to request video clips of the agent acting in ways that satisfy the requested temporal properties of interest. Future work could continue to look at leveraging other forms of communication between the human and the agent, e.g. the agent supplying language-based summaries of its behaviors to the human [20, 21] and the human being able to request demonstrations via language as well.

3 Approach

We seek to teach the reward functions underlying intelligent agents to humans by accurately modeling human beliefs and leveraging teaching strategies to provide instruction at the right level of informativeness and difficulty.

This teaching process can be characterized as a forward pass and a backward pass that inform one another. The full forward pass involves first selecting good teaching demonstrations (by considering how they will influence the learner's beliefs), conveying the demonstrations and updating a model of their beliefs accordingly, then selecting tests that assess their true beliefs. The full backward pass involves assessing the learner's test responses, which provides insight into their current beliefs, and can in turn help select the next teaching demonstration to provide. This teaching process can be summarized by the block diagram in Fig. 3.1 and the rest of this chapter will enumerate each of the key components of this diagram. First, the *assumptions* that we make regarding the agent's world model, reward function & policy, and how the human learns from demonstrations. And second, the three key components of teaching, modeling of human beliefs, and testing, with the educational and algorithmic principles that inform each component as well as how they each influence one another.

As a running example in this section, we will consider the delivery domain in which a robot is rewarded for efficiently delivering the package to the destination while avoiding the mud if the detour is not too costly. Two sample teaching demonstrations and a sample test in this domain are shown for reference (Fig. 3.2).

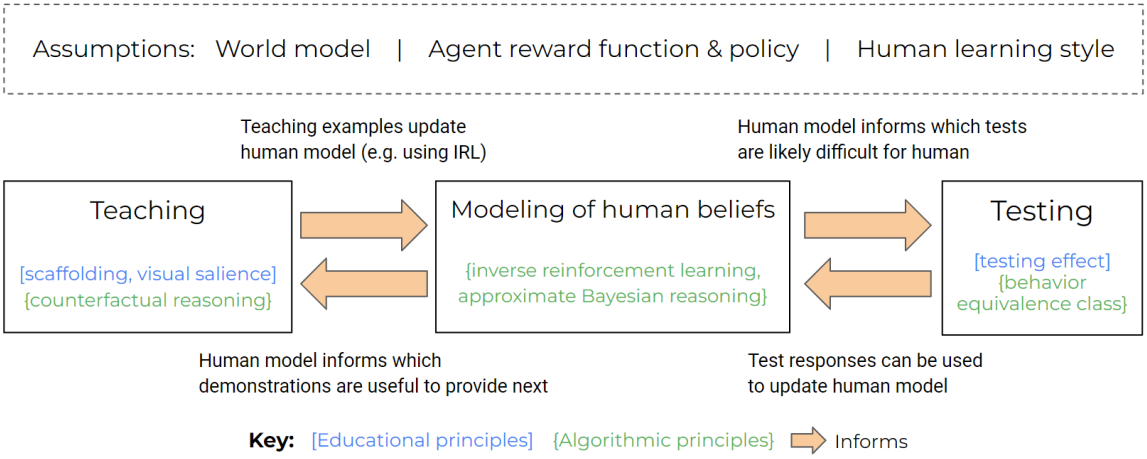

Figure 3.1: An overview of the teaching process explored in this thesis.

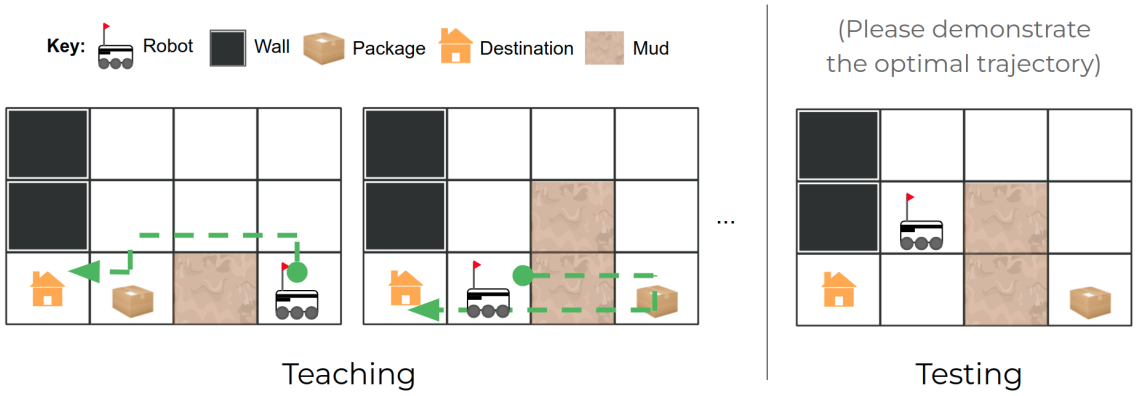

Figure 3.2: Sample teaching demonstrations and a sample test in the delivery domain. The green dotted line demonstrates the robot's chosen path for delivering the package to the destination, while avoiding mud if the detour is not too long. After seeing a series of teaching demonstrations, the human is asked to demonstrate the robot's path for delivering a package in a new environment to test the human's understanding of the robot's reward function.

3.1 Assumptions

World model: The agent's environment is represented as an instance (indexed by i) of a deterministic Markov decision process $MDP_i := (\mathcal{S}_i, \mathcal{A}, T_i, \gamma, S_i^0, R)$. As the methods described here naturally generalize to MDPs with stochastic transition functions and policies through the use of an expectation, we assume a deterministic

MDP for simplicity. $\mathcal{S}_i$ and $\mathcal{A}$ denote the state and action sets, $T_i : \mathcal{S}_i \times \mathcal{A} \times \mathcal{S}_i$ the transition function, $\gamma \in [0, 1]$ the discount factor, and S_i^0 the initial state distribution, $\mathcal{S} : \bigcup_i \mathcal{S}_i$ the union over the states of all related instances of MDPs, which we call a domain (described below), and $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the reward function that operates over domains.

Let a domain refer to a collection of related MDPs that share $\mathcal{A}, R, \gamma$ but differ in $\mathcal{S}_i$, T_i and S_i^0. Take for example the delivery domain. Though MDPs in this domain may vary in the number and locations of mud patches and subsequently offer a diverse set of demonstrations (e.g. see the two distinct MDPs that underly the teaching demonstrations in Fig. 3.2), they importantly share the same reward function R, which we describe next.

Reward function & policy: A reward function R is a function that prescribes how an agent ought to behave, rewarding certain states and actions and punishing others. Ng and Russell [66] famously note that "The reward function, rather than the policy, is the most succinct, robust, and transferable definition of a task," and we focus in this thesis on conveying an agent's reward function to a human. Following prior work [1], reward R is represented as a weighted linear combination of reward features ϕ: $R = \mathbf{w}^{*\top}\phi(s, a, s')$. Though the reward features can theoretically be nonlinear with respect to states and actions and capture arbitrarily complex reward functions, the methods proposed in this thesis were designed for domains with reward functions that cleanly decompose into a set of disentangled, semantic features. Finally, also assume that the human is aware of all aspects of an MDP (including the reward features) but not the weights $\mathbf{w}^*$.

The agent has an optimal policy $\pi_i^* : \mathcal{S}_i \to \mathcal{A}$ that maps each state in an MDP to the action that will optimize the reward in an infinite horizon. A sequence of (s_i, a, s_i') tuples obtained by following π^* gives rise to an optimal trajectory (i.e. a demonstration) ξ^*, where $s_i, s_i' \in \mathcal{S}_i, a \in \mathcal{A}$.

Because instances of a domain share R, the various demonstrations all support inference over the same $\mathbf{w}^*$ through IRL. Thus, we overload the notation π^* to refer to any policy of a domain instance that optimizes a reward with $\mathbf{w}^*$. Furthermore, while a demonstration strictly consists of both an optimal trajectory ξ^* (obtained by following π^*) and the corresponding MDP (minus $\mathbf{w}^*$), we will refer to a demonstration

only by ξ^* in this thesis for notational simplicity.

We emphasize that the agent has a separate policy for each environment (represented as an MDP) and our problem formulation focuses on teaching the agent's *decision making*, i.e. how the agent would generate its policy for any environment given $\mathbf{w}^*$, rather than a single policy.

Human learning style: Cognitive science suggests that humans also often model one another's behavior as exactly or approximately maximizing a reward function [38, 39, 62]. And given demonstrations of behavior, they can infer the underlying reward function through reasoning resembling inverse reinforcement learning (IRL) [10, 11, 37, 66]. Thus, we assume in this work that humans will employ standard IRL [66] to infer an agent's reward function from demonstrations of its policy.

And while an understanding of the reward function is arguably necessary for an understanding of the agent's decision making that generalizes across environments, it may not be sufficient. For example, an agent could simply convey $\mathbf{w}^*$ explicitly to a human instead of having the human infer $\mathbf{w}^*$ given demonstrations. However, it can be nontrivial for humans to map precise numerical reward weights to the corresponding optimal behavior through reasoning resembling planning [84, 98], especially if there is a large number of reward features or the reward features interact in complex ways. Thus, providing demonstrations that inherently carry information regarding $\mathbf{w}^*$ and directly conveying the optimal behavior can be more a effective teaching method for human learners. We in fact verify this through a user study in which people struggle to utilize explicitly provided $\mathbf{w}^*$ to predict agent behavior in our domains of interest (see Section 6.5).

Finally, we briefly note that another common learning style is imitation learning[1] (IL) [47]. This models humans as learning the optimal behavior directly from demonstrations (as opposed to through an intermediate reward function like IRL) and also has a basis in neuroscience [17]. In this work, we focus only on IRL-based learning and leave incorporating IL (e.g. determining which style the human is currently employing and catering to it accordingly) for future work.

[1]Note that the term 'behavior cloning' is sometimes used instead to refer to the process of directly learning the optimal behavior. Accordingly, 'imitation learning' is sometimes used to refer to the broad class of techniques that learn optimal behavior from demonstrations, encompassing both behavior cloning and IRL [68].

3.2 Teaching Components

Here, we introduce the teaching components and concepts that are relevant to our system (shown in Fig. 3.1 as a block diagram).

Teaching: This component considers which demonstrations of the agent's optimal behavior to provide to best teach the agent's reward function to a human. In order to tailor the demonstrations to human learners, we utilize techniques from social constructivist learning theory such as scaffolding [96] (which serves as the foundational method, Section 4.2), insights from cognitive science on how humans provide explanations that optimize for simplicity and pattern matching [61, 95] (Section 4.2) and accounting for the counterfactuals likely considered by the human [65] (Section 5.2). We finally incorporate demonstrations in the form of feedback [44] to tests (Section 6.1).

Teaching importantly informs the model of the human's beliefs (i.e. what information each demonstration that is conveyed to the human provides regarding the agent's reward function). By modeling expected change in human beliefs due to each demonstration, and scaffolding demonstrations such that they do not change the human's beliefs too drastically at each step, we aim to teach in an incrementally informative yet comprehensible manner.

Modeling of human beliefs: We maintain an up-to-date model of the human's expected beliefs of the agent's reward function given a series of demonstrations and tests. We constrain the human's expected beliefs on $\mathbf{w}^*$ to lie on the surface of the $N-1$ sphere where $N = |\mathbf{w}^*|$, the dimensionality of the weight vector. As a reward function will yield the same policy even when multiplied by an arbitrary scale factor, we require $||\mathbf{w}^*||_2 = 1$ to bypass both the subsequent scale invariance of IRL and the degenerate all-zero reward function without loss of generality. See Fig. 4.1 for an example human belief model when $|\mathbf{w}^*| = 3$.

To translate a set of demonstrations and tests into a corresponding model of human belief, we use inverse reinforcement learning (IRL). In Chapters 4 and 5, we assume an exact IRL model [66] that makes precise, and efficient inference on the underlying reward function through constraints that remove sets of candidate reward

functions.

However, research has shown that humans often employ more approximate reasoning in inferring reward functions from demonstrations [35, 90]. Though it may be less computationally efficient, this more gracefully accounts for the inherent uncertainty in the useful but imperfect model of humans as IRL learners, and better supports iterative updates through continual teaching and testing. We thus develop a particle filter-based model of human beliefs that can be updated using approximate IRL (Section 6.1).

A human's current beliefs will inform how they will interpret a demonstration, and thus also inform how much information a demonstration may convey to them (Section 5.2). A human's currently beliefs will also inform how difficult a test will likely be to a human, which is addressed next.

Testing: After providing a number of instructive examples selected to teach a concept, a natural way to assess student learning is to test them. In this work, we do not ask the human learner for the agent's exact reward function in terms of reward weights, as humans are likely approximate in their reasoning as noted previously. Instead, we ask the human to predict the agent's behavior in unseen environments, querying the human's understanding of the agent's policy as a proxy. Defining a test as a prediction of the agent's behavior, we provide two novel measures of how difficult a test will likely be for a human to answer correctly in Sections 4.2 and 5.2. The assigned difficulty of a test and the correctness of the human's corresponding answer provide an approximate measure of the accuracy of their beliefs regarding the agent's reward function.

Finally, the testing effect [79] is a well-established idea in the education literature that suggests that tests are not only useful for assessing but also in learning (e.g. by strengthening retrieval from memory). Thus, we integrate intermittent testing in our closed-loop teaching framework developed in Section 6.1, and leverage tests not only for assessment but also for teaching.

A human's response to a test can importantly be used to further inform and update our model of the human's beliefs, which we explore in Section 6.1.

4 Incorporating Scaffolding and Visual Saliency in Demonstration Selection

This chapter primarily explores how to select a sequence of demonstrations that will effectively teach a robot's reward function to a human. First, we leverage the educational principle of *scaffolding* to select demonstrations that gradually increase in information gain and difficulty and ease the human into learning. Second, we note cognitive science literature that suggests humans favor simple explanations that follow a discernible pattern [61, 95] and also optimize for visual *simplicity and pattern discovery* when selecting demonstrations, which we refer to jointly as visual saliency. And toward effective *testing* of the learner's understanding, we finally show that the expected difficulty for a human to predict an agent's particular behavior as a test inversely correlates to that behavior's expected information gain as a potential teaching demonstration[1].

We assess our methods with user studies and find that our measure of test difficulty corresponds well with human performance and confidence, and also find that favoring visual simplicity and pattern discovery increases human performance on difficult tests. However, we did not find a strong effect for our method of scaffolding, revealing likely shortcomings in our measure of demonstration information gain to a human, which we address in Chapter 5.

[1]The contents of this chapter were published in [51].

4.1 Problem formulation:

The problem of selecting informative demonstrations for teaching the robot's reward function and subsequent policy can be formulated as an instance of machine teaching.

Machine teaching for policies: As formalized by Lage et al. [47], machine teaching for policies seeks to convey a set of demonstrations $\mathcal{D}$ of size n (i.e. the allotted budget for teaching set) that will maximize the similarity ρ between π^* and the policy $\hat{\pi}$ recovered using a model $\mathcal{M}$ on $\mathcal{D}$

$$\underset{\mathcal{D}\subseteq\Xi}{\operatorname{argmax}}\, \rho(\hat{\pi}(\mathcal{D},\mathcal{M}),\pi^*) \;\text{ s.t. }\; |\mathcal{D}| = n \tag{4.1}$$

where Ξ is the set of all optimal demonstrations of π^* in a domain. We assume that the $\mathcal{M}$ employed by humans to approximate the underlying $\mathbf{w}^*$ is IRL. Once $\mathbf{w}^*$ (and the subsequent reward function) is approximated, we assume that human learners are able to arrive at π^* through planning on the underlying MDP.

Thus, the teaching objective reduces to effectively conveying $\mathbf{w}^*$ through well-selected demonstrations. In order to quantify the information a demonstration provides on $\mathbf{w}^*$, we leverage the idea of behavior equivalence classes.

Behavior equivalence class: The *behavior equivalence class* (BEC) of π is the set of (viable) reward weights under which π is still optimal. The larger the BEC(π) is, the greater the potential uncertainty over $\mathbf{w}^*$ that is underlying the robot's optimal policy.

$$\text{BEC}(\pi) = \left\{\mathbf{w}\in\mathbb{R}^l \mid \pi \text{ optimal w.r.t. } R = \mathbf{w}^\top\phi(s,a,s')\right\} \tag{4.2}$$

The BEC(π) can be calculated as the intersection of the following half-space constraints generated by the standard IRL equation [66]

$$\begin{gathered}\mathbf{w}^\top\left(\mu_\pi^{(s,a^*)} - \mu_\pi^{(s,b)}\right) \geq 0 \\ \forall a^* \in \arg\max_{a\in\mathcal{A}} Q^*(s,a), b\in\mathcal{A}, s\in\mathcal{S}\end{gathered} \tag{4.3}$$

where $\mu_\pi^{(s,c)} = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\phi(s_t) \mid \pi, s_0 = s, a_0 = c\right]$ is the vector of expected reward feature counts accrued from taking action c in s, then following π after, and $Q^*(s,a)$

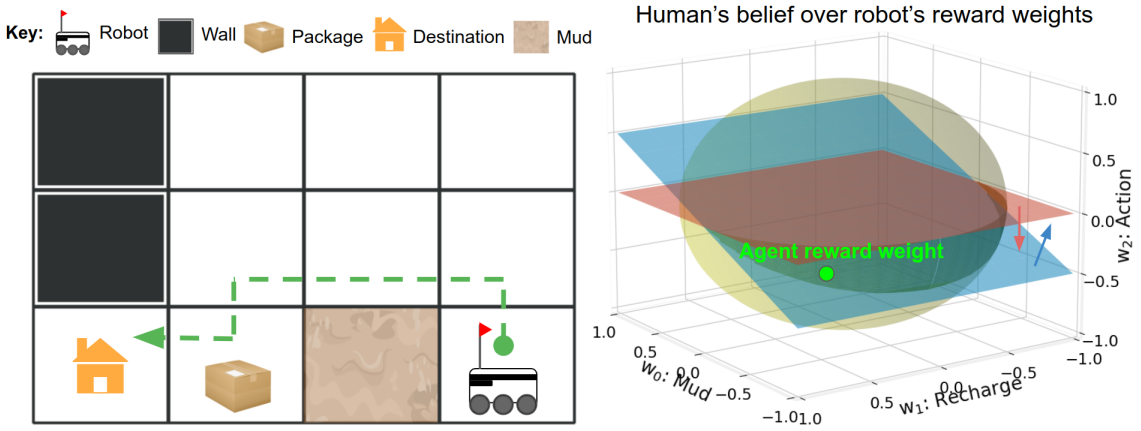

Figure 4.1: **(a)**: A demonstration $\mathcal{D}$ of an optimal policy in the delivery domain. Agent aims to deliver the package to the destination while avoiding walls and mud if the detour is not too costly.

(b): The demonstration can be translated into a set of half-space constraints (the red and blue half-spaces) on the possible underlying reward function using standard IRL (Eq. 4.4). The set of reward functions that obey the constraints (which includes the agent's true reward function) corresponds to BEC($\mathcal{D}|\pi$), and can be used to model the human's subsequent belief over the agent's reward function.

refers to the optimal Q-value in a state and a possible action [93].

Brown et al. [12] proved that the BEC($\mathcal{D}|\pi$) of a set of demonstrations $\mathcal{D}$ of a policy π can be formulated similarly as the intersection of the following half-spaces

$$\mathbf{w}^\top \left(\mu_\pi^{(s,a^*)} - \mu_\pi^{(s,b)}\right) \geq 0, \forall (s,a) \in \mathcal{D}, b \in \mathcal{A}. \tag{4.4}$$

Using the Eq. 4.4, every demonstration can be translated into a set of constraints on the viable reward weights. Whereas Eq. 4.3 generates constraints from rollouts from all states that comprise the state space of a domain, Eq. 4.4 generates constraints from only rollouts that start from states that comprise the demonstration of interest.

To gain an intuition for how IRL translates a demonstration into a set of half-space constraints on the possible underlying reward function, consider again the delivery domain where the agent is tasked with delivering the package to the destination while generally taking the fewest number of actions, avoiding mud, and recharging

its battery at a recharge station (not shown in Fig. 4.1) if possible[2]. The domain accordingly has three binary reward features $\phi = [$*traversed mud*, *battery recharged*, *action taken*$]$ and $\mathbf{w}^* \propto [-3, 3.5, -1]$, where $\mathbf{w}^*$ is shown as proportional to the vector because of the requirement for $||\mathbf{w}^*||_2 = 1$ to bypass both the scale invariance of IRL and the degenerate all-zero reward function (as previously noted in Section 3.2). Thus, with no prior information, the expected space of reward weights in the human's mind for the delivery domain is the surface of the 2-sphere. Imagine that the robot provides the demonstration in Fig. 4.1a, which yields the constraints in Fig. 4.1b that leaves only the sliver of the 2-sphere that contains the robot's true reward weights. The red constraint plane in Fig. 4.1b intuitively indicates that $w_2^* \leq 0$, since no arbitrary additional actions were taken in delivering the package, and jointly with the blue constraint plane indicates that $w_0^* \leq 0$ and $w_0^* \leq 2w_2^*$, since two actions were taken to detour around the mud. Note that because this demonstration environment does not have a recharge station that could be visited or not (e.g. if it is too far away), the constraints do not convey any information on the 'battery recharged' weight in Fig. 4.1b. Additional demonstrations that contain a recharge station will be needed accordingly to convey the recharge weight.

Importantly, the surface area of the $N-1$ sphere that remains after incorporating a demonstration's constraints can be used as a measure of its information gain. The smaller the area, the fewer viable reward weights remain, and the higher the information gain. However, we note that we must also consider the difficulty of human learners to extract the information from demonstrations, which we address next.

4.2 Methods

Scaffolding

The SCOT algorithm [12] efficiently selects the minimum number of demonstrations that results in the smallest BEC area for a pure IRL learner. Such a learner is assumed to fully grasp these few highly nuanced examples that delicately straddle

[2]Note that this version of the delivery domain is slightly different from the one used in the original paper [51] and in the experiments in this chapter. A change in one of the reward features and the use of the L^2 norm on the weights rather than the L^1 norm to accommodate IRL's scale invariance allows for consistency with subsequent chapters in this thesis.

decision-making boundaries and find any other demonstrations redundant. However, *we posit that the BEC area of a demonstration not only inversely corresponds to the amount of information it contains about the possible values of* $\mathbf{w}^*$*, but also inversely corresponds to the effort required for a human to extract that information.* Thus humans will likely benefit from additional scaffolded examples that ease them in and incrementally relax the degrees of freedom of the learning problem.

We develop a scaffolding method for a learner without any prior knowledge, outlined as follows. First, obtain the SCOT demonstrations that contain the maximum information on $\mathbf{w}^*$. To do so, the robot first translates all possible demonstrations of its policy in a domain into a corresponding set of BEC constraints. After taking a union of these constraints, redundant constraints are removed using linear programming [71]. These non-redundant constraints together form the minimal representation of BEC(π^*). SCOT now iteratively runs through all possible demonstrations again and greedily adds to the teaching set $\mathcal{D}$ the demonstration that covers as many of the remaining constraints in BEC(π^*), until all constraints are covered. These steps for obtaining SCOT demonstrations correspond to lines 2-14, as part of our overarching scaffolding method detailed in Algorithm 1

If space remains in the teaching budget n for additional demonstrations after selecting j SCOT demonstrations, begin scaffolding by sorting all possible demonstrations in a domain according to their BEC areas. Then cluster them using k-means into $2(n-j)$ clusters to ensure that no two consecutive demonstrations are nearly identical in BEC area (see Fig. 4.2). Randomly draw m candidate demonstrations from every other cluster, where stochasticity in demonstration selection increases as m decreases; in the limit, candidate demonstration selection will be deterministic when m is set to the size of the cluster. We randomly drew one-sixth of the possible demonstrations from each cluster for some stochasticity (e.g. to avoid showing a special type demonstration early on that would 'exit' rather than complete the task, which will be further explained in the discussion section). Finally from these pools of candidate demonstrations, select the ones that best optimize visuals for the teaching set $\mathcal{D}$ (as described in the next section) to show in addition to the SCOT demonstrations. See lines 17-22 in Algorithm 1. In our experiments, the algorithm divided the BEC areas into 6 clusters, considering every other cluster (starting with the second cluster) to correspond to "low", "medium", and "high" information respectively and

m was set as a quarter of the number of demonstrations in the cluster.

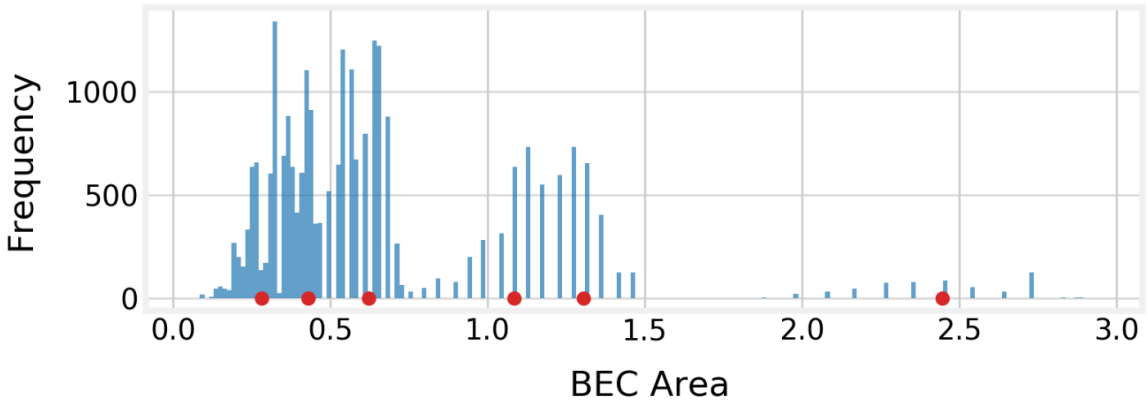

Figure 4.2: Histogram of BEC areas of the 25,600 possible demonstrations in the delivery domain, where the agent, passenger, mud, and recharge station locations are allowed to vary. Cluster centers returned by k-means (for k = 6) are shown as red circles along the x-axis. Demonstrations from every other cluster are selected and shown in order of largest to smallest BEC area for scaffolded machine teaching.

Visual Saliency

Studies on explanations preferred by humans indicate a bias toward those that are simpler and have fewer causes [61]. Furthermore, [95] found that explanations can be detrimental if they do not help the learner notice useful patterns or even mislead them with false patterns. Together, these two works support the idea that explanations should minimize distractions that potentially inspire false correlations and instead highlight and reinforce the minimal set of causes. We thus also optimize for simplicity and pattern discovery when selecting demonstrations that naturally explain the underlying reward function.

Though the BEC area of a demonstration provides an unbiased, quantitative measure of the information transferred to a pure IRL learner, *human learners are likely also influenced by the medium of the demonstration, e.g. visuals, and the simplicity and patterns it affords.* For example, visible differences between sequential demonstrations can highlight relevant aspects, while visual clutter that does not actually influence the robot's behavior (e.g. extraneous mud not in the path of the delivery robot) may distract or even mislead the human.

We perform a greedy sequential optimization for visual pattern discovery and then for visual simplicity, which we collectively term visual saliency. We first encourage pattern matching by considering candidates from different BEC clusters (which often exhibit qualitatively different behaviors) that are most visually similar to the previous demonstration. We measure the visual similarity of two states by defining a binary hash function over a domain's state space (such that each non-zero value in the state hash corresponds to the presence of a consistent visual feature, like a mud patch at a particular location) and calculating the edit distance between the two corresponding binary state hashes. The aim is to highlight a change in environment (e.g. a new mud patch) that caused the change in behavior (e.g. robot takes a detour) while keeping all other elements constant. We then optimize for simplicity. A measure of visual simplicity is manually defined for each domain (e.g. the number of mud patches in the delivery domain), and out of the scaffolding candidates, the visually simplest demonstration is selected.

The proposed methods for scaffolding and visual saliency come together in Algorithm 1[3]. Since the highest information SCOT demonstrations are selected first then demonstrations are selected via k-means clustering from high to low information, the algorithm concludes by reversing the demonstration list to order the demonstrations from easiest to hardest (line 29). An example of a sequence of demonstrations that exhibits scaffolding, simplicity, and pattern discovery are shown in Fig. 4.3.

Testing

An optimal trajectory's BEC area theoretically correlates to its information gain as a teaching demonstration. The smaller the area, the less uncertainty there is regarding the value of $\mathbf{w}^*$.

We propose a complementary and novel idea: *that the BEC area can be inverted as a measure of a trajectory's difficulty as a question during testing*, i.e. when a human is asked to predict the robot's trajectory in a new situation. Intuitively, a large BEC area indicates that there are many viable reward weights for a demonstration, and thus the human does not need to precisely understand $\mathbf{w}^*$ to correctly predict the

[3]An implementation is available at https://github.com/SUCCESS-MURI/machine-teaching-human-IRL.

Algorithm 1 Machine Teaching for Human Learners

Require: π^*: optimal policy, $\mathbb{D}$: set of all MDPs belonging to a domain, Ξ: all possible demonstrations of π^* in a domain, n: teaching budget, m: cluster pool size

// Obtain SCOT demos
$U = \varnothing$
for $MDP \in \mathbb{D}$ **do**
    // Obtain BEC(π^*) using Eq. 4.3 on each MDP comprising a domain. $\hat{\mathbf{N}}[\cdot]$ // extracts unit normal vectors corresponding to a set of half-space constraints.
    $U = U \cup \hat{\mathbf{N}}[\text{BEC}(\pi^*)]$
end for
$U = \text{removeRedundantConstraints}(U)$ ▷ See [71]
$\mathcal{D} = [\,], C = \varnothing$
while $|U \setminus C| \neq 0$ **do** ▷ $\setminus$ denotes set subtraction
    $\xi^* = \underset{\xi \in \Xi}{\text{argmax}} |\hat{\mathbf{N}}[\text{BEC}(\xi|\pi^*)] \cap (U \setminus C)|$ ▷ Eq. 4.4
    $\mathcal{D}.\text{append}(\xi^*)$
    $C = C \cup \hat{\mathbf{N}}[\text{BEC}(\xi|\pi^*)]$
    $\Xi = \Xi \setminus \xi^*$
end while
// Select candidates to fill the teaching budget via scaffolding
if $|\mathcal{D}| < n$ **then**
    $\mathbf{D}_{cand} = \varnothing$ ▷ Set of sets
    $\Xi_{sorted} = \text{sortByIncreasingBECArea}(\Xi)$
    $\Xi_{cluster} = \text{kMeans}\Big(\Xi_{sorted}, 2(n - |\mathcal{D}|)\Big)$
    for $\Big(i = 1, i = 2(n - |\mathcal{D}|), i \mathrel{+}= 2\Big)$ **do**
        $\mathbf{D}_{cand} = \mathbf{D}_{cand} \cup \{\text{sampleTraj}(m, \Xi_{cluster}[i])\}$
    end for
    // Downselect from candidates based on visuals
    for $\mathcal{D}_{cand} \in \mathbf{D}_{cand}$ **do**
        $\mathcal{D}_{prelim} = \text{maximizeVisualSimilarity}(\mathcal{D}_{cand}, \mathcal{D})$
        $\xi^* = \text{maximizeVisualSimplicity}(\mathcal{D}_{prelim})$
        $\mathcal{D}.\text{append}(\xi^*)$
    end for
    $\mathcal{D} = \text{reverse}(\mathcal{D})$ ▷ Order demonstrations from easiest to hardest
end if
return $\mathcal{D}$ ▷ Final demonstration set to show human

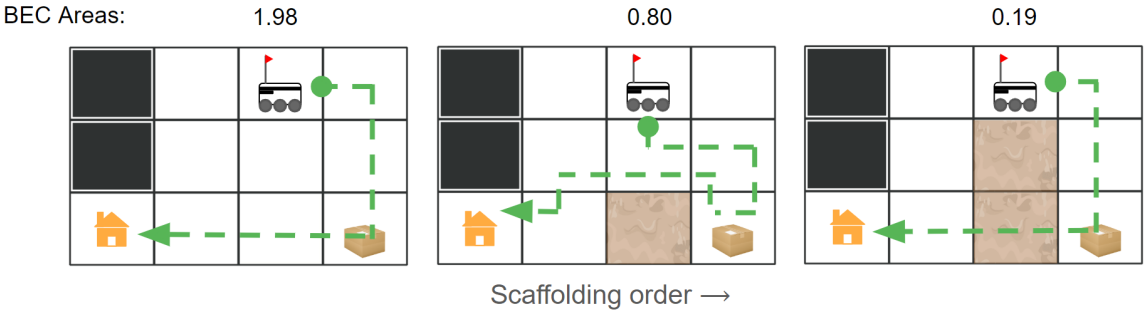

Figure 4.3: Demonstrations hand-picked to illustrate ideal scaffolding, simplicity, and pattern discovery. We scaffold by showing demonstrations that incrementally decrease in BEC area (which appears to correlate inversely with information gain and difficulty). Simplicity is encouraged by minimizing visual clutter (e.g. unnecessary mud patches). Pattern discovery is encouraged by holding the agent and passenger locations constant while highlighting the single additional mud patch between demonstrations that changes the optimal behavior.

robot's trajectory. We can also use this measure to select tests of varying difficulties to assess the human's final understanding of $\mathbf{w}^*$ and subsequently π^* after having seen a set of teaching demonstrations.

4.3 User Studies

We ran two online user studies that involved participants watching demonstrations of a 2D agent's policy and predicting the optimal trajectory in new test environments[4]. The first study explored how BEC area of provided demonstrations correlates with a human's subsequent understanding of the underlying policy, while the second study explored how incorporating human learning strategies into demonstration selection impacts a human's understanding of the underlying policy.

Domains

Three simple grid world domains were designed for the two studies (see Fig. 4.4). The available actions were {*up*, *down*, *left*, *right*, *pick up*, *drop*, *exit*}. Each domain

[4]Code for the user studies, videos of teaching and testing demonstrations, and the collected data are available at https://github.com/SUCCESS-MURI/psiturk-machine-teaching.

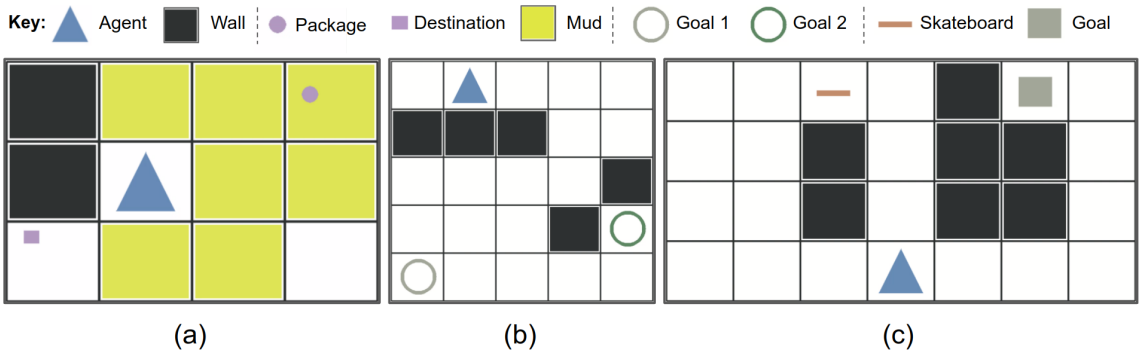

Figure 4.4: Three domains were presented in the user study, each with a different set of reward weights to infer from demonstrations using inverse reinforcement learning. **(a)**: delivery, **(b)**: two-goal, **(c)**: skateboard

consisted of one shared reward feature of unit action cost, and two unique reward features as follows.

Delivery domain. The agent is rewarded for bringing a package to the destination and penalized for moving into mud.

Two-goal domain. The agent is rewarded for reaching one of two goals, with each goal having a different reward.

Skateboard domain. The agent is rewarded for reaching the goal. It is penalized less per action if it has picked up a skateboard (i.e. riding a skateboard is less costly than walking).

To convey an upper bound on the positive reward weight, the agent exited from the game immediately if it encountered an environment where working toward the positive reward would yield a lower overall reward (e.g. too much mud along its path). The semantics of each domain were masked with basic geometric shapes and colors to prevent biasing human learners with priors. All domains were implemented using the simple_rl framework [2].

Study Design

The first and second user studies (US1 and US2, respectively) used the same domains, procedures, and measures, though they differed in which variable was manipulated.

US1 explored how BEC area of demonstrations correlates with a human's understanding of the underlying policy. Thus, the between-subjects variable was

information class, with three levels: low, medium, and maximum (i.e. SCOT). The low and medium information demonstrations were selected from the 5th and 3rd BEC clusters respectively (see Fig. 4.2). When selecting multiple demonstrations from a *single* cluster, we optimized for visual simplicity and *dissimilarity* as coverage of demonstrations has been shown to improve human learning [5, 35][5]. Thus, each person only saw low, medium, or maximum (i.e. SCOT) information demonstrations across each domain, where the number of demonstrations shown in each domain was set to equal the number of SCOT demonstrations for fair comparison (2 for delivery and skateboard, 3 for two-goal).

US2 explored how incorporating human learning strategies impacts a human's understanding of the underlying policy. Specifically, it examined how the presence and direction of scaffolding, and optimization of visuals, would impact the human's test performance. The between-subjects variables were *scaffolding class* (none, forward, and backward), and *visual saliency* (positive and negative). For scaffolding class, forward scaffolding showed demonstrations according to Algorithm 1, backward scaffolding showed forward scaffolding's demonstrations in reverse, and no scaffolding showed all high information gain examples from the 1st BEC cluster (Fig. 4.2). Five demonstrations were shown for each domain.

Both US1 and US2 had two additional within-subject variables: *domain* (delivery, two-goal, and skateboard, described above and *test difficulty* (low, medium, and high, determined by the BEC area of the test). As such, each participant was shown teaching demonstrations in all three domains and was tested with tests spanning all three difficulty levels (in randomized order).

For both user studies, participants first completed a series of tutorials that introduced them to the mechanics of the domains they would encounter. In the tutorials, participants learned that the agent would be rewarded or penalized according to key events (i.e. reward features) specific to each domain. They were then asked to generate a few predetermined trajectories in a practice domain with a live reward counter to familiarize themselves with the keyboard controls and a practice reward function. Finally, participants entered the main user study and completed a single

[5]Note that Algorithm 1 already achieves coverage by scaffolding demonstrations across *different* BEC clusters and thus benefits instead from optimization of visual similarity amongst consecutive demonstrations that highlights changes in environments that lead to different behaviors.

trial in each of the delivery, two-goal, and skateboard domains. Each trial involved a teaching portion and a test portion. In the teaching portion, participants watched videos of optimal trajectories that maximized reward in that domain, then answered subjective questions about the demonstrations (M2-M4, see below). In the subsequent test portion, participants were given six new test environments and asked to provide the optimal trajectory. The tests always included two low, two medium, and two high difficulty environments shown in random order. For each of the tests, participants also provided their confidence in their response (M5). The teaching videos for each condition were pulled from a filtered pool of 3 exemplary sets of demonstrations proposed by Algorithm 1 to control for bias in the results (e.g. to remove the confound of showing demonstrations that simply 'exit' early on in the teaching set, which will be explained in the discussion section). The tests were likewise pulled from a filtered pool of 3 exemplary sets of demonstrations for each of the low, medium, and high difficulty test conditions. Please refer to the scaffolding subsection in Section 4.5 for additional discussion on how filtering was employed.

Finally, though the methods described in this chapter are designed for a human with no prior knowledge regarding any of the weights, the agent in our user studies assumed that the human was aware of the step cost and only needed to learn the relationship between the remaining two weights in each domain. This simplified the problem at the expense of a less accurate human model and measure of a demonstration's information gain via BEC area. However, the effect was likely mitigated in part by the clustering and sampling in Algorithm 1, which only makes use of coarse BEC areas.

Hypotheses

H1: The BEC area of a demonstration correlates 1) inversely to the expected difficulty for a human to correctly predict that exact demonstration *during testing*, and 2) directly to their confidence in that prediction.

H2: The BEC area of a demonstration also correlates 1) inversely to the information transferred to a human *during teaching* and thus inversely to the subsequent human test performance as measured by a suite of test, and 2) leads to better qualitative assessments on informativeness, mental effort, or puzzlement.

H3: Forward scaffolding (demonstrations shown in increasing difficulty) will result in better qualitative assessments of the teaching set and better participant test performance over no scaffolding (only high difficulty demonstrations shown) and backward scaffolding (demonstrations shown in decreasing difficulty), in that order.
H4: Positive visual saliency will result in better qualitative assessments of the teaching set and better test performance over negative visual saliency (with positive and negative visual saliency corresponding to the maximization and minimization, respectively, of both simplicity and pattern discovery).

The two user studies jointly tested H1. The first study tested H2 and the second study tested H3 and H4.

Measures

The following objective and subjective measures were recorded to evaluate the aforementioned hypotheses. The Likert scales corresponding to M2-M4 were provided after all of the demonstrations but before the tests. The Likert scale corresponding to M5 was provided after each test.
M1. Optimal response: For each test, whether the participant's trajectory received the optimal reward was recorded.
M2. Informativeness rating: 5-point Likert scale with prompt "How informative were these demonstrations in understanding how to score well in this game?"
M3. Mental effort rating: 5-point Likert scale with prompt "How much mental effort was required to process these demonstrations?"
M4. Puzzlement rating: 5-point Likert scale with prompt "How puzzled were you by these demonstrations?"
M5. Confidence rating: 5-point Likert scale with prompt "How confident are you that you obtained the optimal score?"

4.4 Results

162 participants were recruited using Prolific [70] for the two user studies. Each of the nine possible between-subjects conditions across the two user studies was randomly assigned 18 participants (such that US1 and US2 contained 54 and 108

participants respectively), and the order of the domains presented to each participant was counterbalanced. Participants' ages for US1 ranged from 18 to 57 (M = 25.07, SD = 8.37). Participants reported gender to be roughly 74% male, 22% female, 2% non-binary, and 2% preferred to not disclose. Participants' ages for US2 ranged from 18 to 52 (M = 26.57, SD = 8.33). Participants reported gender to be roughly 64% male, 34% female, 2% non-binary, and 0% preferred to not disclose. The recruitment process and studies was approved by Carnegie Mellon University's Institutional Review Board.

The three domains were designed to vary in the difficulty of their respective optimal trajectories. We calculated an intraclass coefficient (ICC) based on a mean-rating (k = 3), consistency-based, 2-way mixed effects model [45] to evaluate the consistency of each participant's performance across domains. A low ICC value of 0.37 ($p < .001$) indicated that performance in fact varied considerably across domains for each participant. We subsequently average each participant's scores across the domains in all following analyses, potentially yielding results that are representative of domains with a range of difficulties.

H1: We combine the test responses from both user studies as they share the same pool of tests. A one-way repeated measures ANOVA revealed a statistically significant difference in the percentage of optimal responses (M1) across test difficulty ($F(2, 322) = 275.35, p < .001$). Post-hoc pairwise Tukey analyses further revealed significant differences between each of the three groups, with the percentage of optimal responses dropping from low ($M = 0.89$), to medium ($M = 0.68$), to high ($M = 0.36$) test difficulties ($p < .001$ in all cases).

Spearman's rank-order correlation further showed a significant inverse correlation between test difficulty and confidence (M5, $r_s = -.40, p < .001, N = 486$). See Fig. 4.5 for the raw confidence data.

Objective and subjective results both support H1, that BEC area can indeed be used as a measure of difficulty for testing. We thus proceed with the rest of the analyses with "test difficulty" as a validated independent variable.

H2: A two-way mixed ANOVA on percentage of optimal responses (M1) did not reveal a significant effect of information class of the teaching set ($F(2, 51) = 1.23, p = .30$), though test difficulty had a significant effect consistent with the H1 analysis ($F(2, 102) = 118.58, p < .001$). There was no interaction between information class

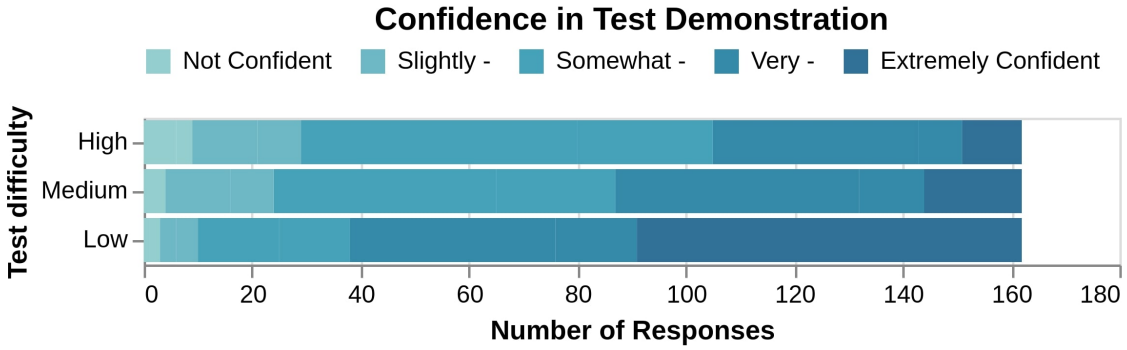

Figure 4.5: Participants were significantly more confident of their responses as test difficulty decreased.

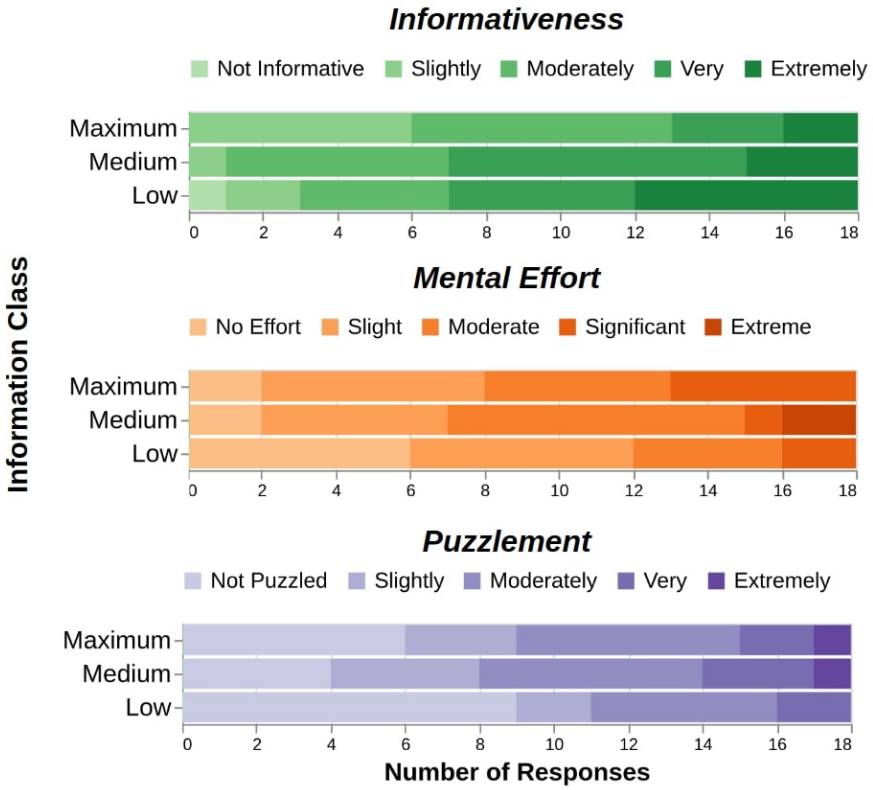

Figure 4.6: The information class of demonstrations only significantly influences their perceived informativeness, ironically decreasing from low to maximum information class. This suggests that a demonstration's intrinsic information content (as measured by its BEC area) does not always correlate with the information transferred to human learners. No significant effects were found between information class and mental effort or puzzlement.

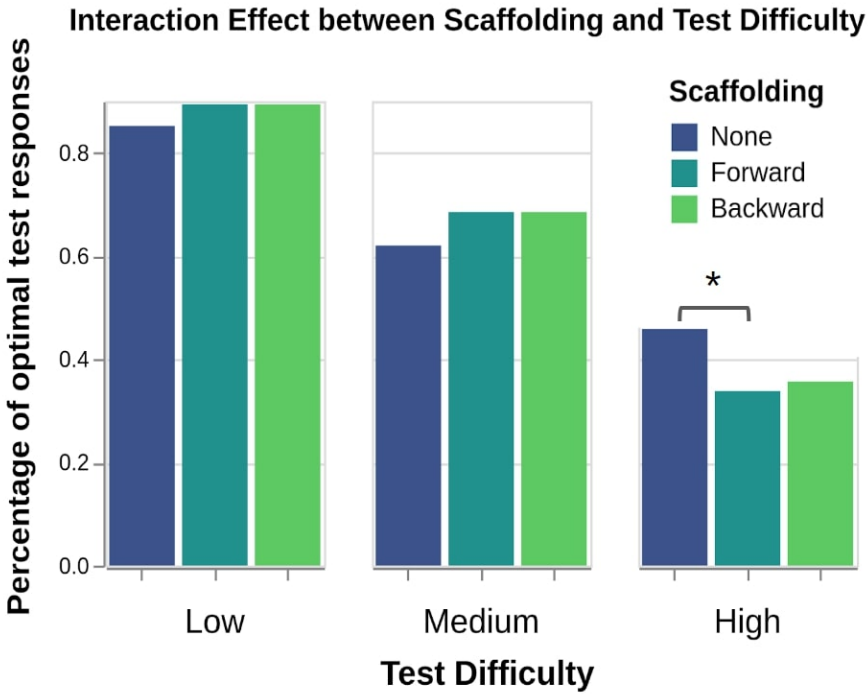

Figure 4.7: Though the three scaffolding conditions perform similarly in aggregate across all tests, 'no scaffolding' significantly increases performance for high difficulty tests.

and test difficulty ($F(4, 102) = 0.67, p = .61$).

Spearman's correlation test only found a significant negative correlation between information class and perceived informativeness (M2, $r_s = -0.28, p = .04, N = 54$). Neither mental effort (M3, $p = .08$) nor puzzlement (M4, $p = .36$) were found to have significant correlations with information class. See Fig. 4.6 for the raw subjective ratings.

The data failed to support H2. The data suggest that IRL alone is indeed an imperfect model of human learning, motivating the use of human teaching techniques to better accommodate human learners.

H3: A two-way mixed ANOVA on percentage of optimal responses (M1) revealed a significant interaction effect between scaffolding and test difficulty ($F(4, 210) = 2.79, p = .03$). Tukey analyses showed that no scaffolding ($M = 0.46$) yielded

significantly better test performance than forward scaffolding ($M = 0.34$) for high difficulty tests ($p = .05$). Though not statistically significant, a trend of forward and backward scaffolding outperforming no scaffolding on low ($M = 0.89, 0.89, 0.85$ respectively) and medium difficulty tests ($M = 0.69, 0.69, 0.62$ respectively) can be observed as well (see Fig. 4.7).

Seeing the strong effect of domain on the results of the user studies in Chapter 6, we explore whether the significant effect noted above was also driven by domain. Indeed, a t-test revealed that only in the skateboard domain did no scaffolding ($M = 0.61$) yield significantly higher learning outcomes over forward scaffolding ($M = 0.39$) for high difficulty tests at Bonferroni adjusted $p = 0.04$.

A two-way mixed ANOVA did not reveal a significant effect from scaffolding ($F(2, 105) = 0.02, p = .98$) but did find a significant effect for test difficulty ($F(2, 210) = 167.63, p < .001$) on percentage of optimal responses (M1) as hypothesized.

A Kruskal-Wallis test did not find differences between the informativeness ($H(2) = 5.18, p = .07$), mental effort ($H(2) = 1.16, p = .56$), or puzzlement ($H(2) = 0.59, p = .74$) ratings (M2–M4) of differently scaffolded teaching sets.

The data largely failed to support H3. Forward and backward scaffolding surprisingly led to nearly identical test performance. Though no scaffolding performed similarly overall, it yielded a significant increase in performance specifically for high difficulty tests. The subjective measures did not indicate any clear relationships.

H4: A two-way mixed ANOVA on percentage of optimal responses (M1) revealed significant effects of test difficulty ($F(2, 212) = 169.21, p < .001$) and an interaction effect between optimized visuals and test difficulty ($F(2, 212) = 5.61, p = .004$). Exploring the interaction effect with Tukey analyses revealed that visual saliency had no effect on test performance on low ($p = .24$) and medium ($p = .90$) difficulty tests, but led to a significant improvement in performance in high ($p < .001$) difficulty tests for positive visual saliency ($M = 0.45$) over negative ($M = 0.31$), see Fig. 4.8. The two-way mixed ANOVA did not reveal a significant effect from optimized visuals alone ($F(1, 106) = 2.27, p = .13$).

Again, seeing the strong effect of domain on the results of the user studies in Chapter 6, we explore whether the significant effect noted above was also driven by domain. Indeed, a t-test revealed that only in the skateboard domain did positive

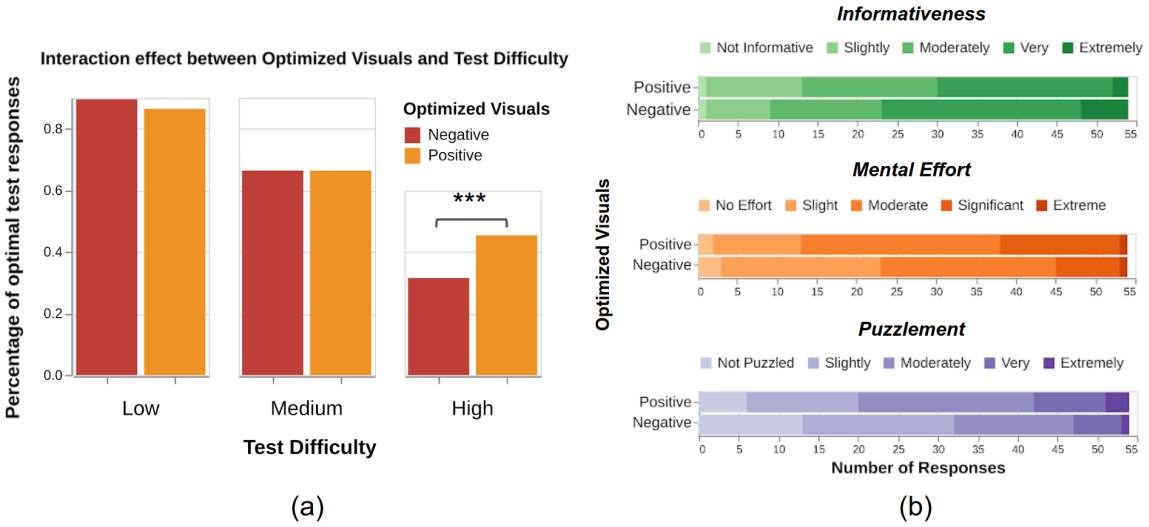

Figure 4.8: **(a)**: Optimizing teaching demonstration visuals does not significantly affect performance on low and medium difficulty tests, but leads to a significant improvement on high difficulty tests.

(b): Ratings on mental effort and puzzlement surprisingly increased for positive visual saliency, likely an artifact of unforeseen study design effects. No significant effects were found for ratings on informativeness.

visual saliency ($M = 0.62$) yield significantly higher learning outcomes over negative visual saliency ($M = 0.33$) for high difficulty tests at Bonferroni adjusted $p < .001$.

A Mann-Whitney U test found that ratings for mental effort ($U(N_{neg} = 54, N_{pos} = 54) = 1131.5, p = .03$) and puzzlement ($U(N_{neg} = 54, N_{pos} = 54) = 1082.5, p = .02$) (M3 and M4) increased for positive visual saliency. Informativeness ratings were not found to differ significantly between the two visual saliency conditions ($p = .11$).

The data partially supports H4. Optimizing visuals improved test performance for high difficulty tests. However, optimizing visuals also yielded counterintuitive results for the subjective measures on mental effort and puzzlement, which we address in the following section.

4.5 Discussion

Learning styles: The data refuting H2 suggests that IRL alone is indeed an imperfect model of human learning, motivating the use of human teaching techniques to better accommodate human learners.

There was no correlation between information class and test performance, likely a result of two factors. First, the number of demonstrations provided (two or three) across the conditions in US1 was likely too few for human learners, who are not pure IRL learners and can sometimes benefit from 'redundant' examples that reinforce a concept. Second, as will be discussed under the scaffolding subsection in Section 4.5, BEC area is likely an insufficient model of a demonstration's information gain to a human and warrants further iteration.

Accordingly, maximum information demonstrations provided by SCOT ($M = 0.61$) failed to significantly improve the percentage of optimal responses compared to medium ($M = 0.65$) and low ($M = 0.67$) information demonstrations, as IRL would have predicted. The subjective results further indicate that people ironically found the demonstrations with the highest information gain the least informative. We hypothesize that participants struggled to digest the information contained within the SCOT demonstrations all at once, motivating the use of scaffolding to stage learning into manageable segments.

Furthermore, analyzing the free-form comments provided by participants throughout the user studies revealed insights about their learning styles. Though this chapter

assumed that participant learning would resemble IRL, we discovered it sometimes resembled imitation learning, which models humans as learning the optimal behavior directly from demonstrations (as opposed to through an intermediate reward function like IRL) [17, 47]. For example, one participant expounded upon their mental effort Likert rating (M3) with the following description of IRL-style learning: "You need to make a moderate amount of mental effort to understand all the rules and outweight [sic] everything and see what is worth it or not in the game." In contrast, another expounded upon their used mental effort rating with the following description of IL-style learning: "The primary 'mental effort' was in memorising the patterns of each level/stage and matching the optimal movements for them."

To better understand the types of learning employed by our participants, we analyzed their optional responses to the following questions: "Feel free to explain any of your selections above if you wish:" (asked in conjunction with prompts for ratings of informativeness, mental effort, and puzzlement of demonstrations in each domain, i.e. three times) and "Do you have any comments or feedback on the study?" (asked once, after the completion of the full study). Similar to Lage et al. [47], we coded relevant responses from participants regarding their thought process as resembling IRL (e.g. "So, the yellow squares should be avoided if possible and they possibly remove 2 points when crossed but I'm not sure") or as resembling IL (e.g. "I did not understand the rule regarding yellow tiles. It seems they should be avoided, but not always. Interesting..."), or as 'unclear' (e.g. "After some examples I feel like I'm understanding way better these puzzles."). A second coder uninvolved in the study independently labeled the same set of responses, assigning the same label to 79% of the responses. A Cohen's kappa of 0.64 between the two sets of codings further indicates moderate to substantial agreement [4, 49, 63]. Please refer to Section 9.1 of the appendix for the responses, labels, and further details on the coding process.

As Table 4.1 conveys, both coders agreed that more responses resembled IRL than IL and 'unclear' combined, suggesting that people perhaps employed IRL more often than not. However, we note that the way the tutorials introduced the domains may have influenced this result. For example, explicitly conveying each domain's unique reward features and clarifying that a trajectory's reward is determined by a weighting over those features may have encouraged participants to first infer the reward weights from optimal demonstrations (e.g. through IRL) and then infer the optimal policy

Table 4.1: Coding of qualitative participant responses as resembling inverse reinforcement learning (IRL) or imitation learning (IL), or 'unclear'.

Learning style	Raw counts (across user studies)		Percentages (across coders)	
	Coder 1	Coder 2	User study 1	User study 2
IRL	25	27	32%	68%
IL	7	9	27%	12%
Unclear	15	11	41%	20%

(as opposed to directly inferring the optimal policy e.g. through IL).

Examining the percentage of each response across the two user studies reveals another interesting trend. Responses were far more likely to be coded as IRL in US2, where participants got to see five demonstrations as opposed to US1, where participants only got to see two or three demonstrations. This echoes the observation of [47] that people may be more inclined to use IL over IRL in less familiar situations, which may be moderated in future studies through more extensive pre-study practice and/or additional informative demonstrations that better familiarize the participant to the domains.

Finally, out of 11 participants who provided more than one response, coders agreed that 8 appeared to employ the same learning style throughout the user study (e.g. participants 129 and 142 in US2 only provided responses resembling IRL), 4 appeared to have changed styles through the user study (e.g. participants 59 in US1 and 20 in US2 provided various responses that resembled IL, IRL, or were unclear), and 3 were ambiguous (i.e. one coder coded a consistent learning style while the other did not). Though we controlled for learning effects by counterbalancing the order of the domains, participants likely found the domains to vary in the difficulty of their respective optimal trajectories (as suggested by the ICC score). Furthermore, certain conditions led to significant differences in subjective and objective outcomes (e.g. demonstrations with the highest information gain were ironically perceived to be least informative (H2) and positive visual saliency improved performance for high difficulty tests (H4)). We thus hypothesize that the varying difficulties in domains and conditions non-trivially influenced the learning styles at different times (e.g. by moderating familiarity [47]).

Scaffolding: Though BEC area is a well-motivated preliminary model of a demonstration's information gain to a human, backward scaffolding's unexpected on-par performance with forward scaffolding suggests that it is insufficient and our scaffolding order likely was not clear cut in either direction. In considering possible explanations, we note that Eq. 4.4 presents a computationally elegant method of generating BEC constraints via sub-optimal, one-action deviations from the optimal trajectory. However, *these suboptimal trajectories do not always correspond to the suboptimal trajectories in the human's mind (e.g. which may allow more than one-action deviations).* This sometimes leads to a disconnect between a demonstration's information gain as measured by BEC area and its informativeness from the point of view of the human.

Furthermore, forward and backward scaffolding (each comprised of low, medium, and high information demonstrations) yielded higher performance for low and medium difficulty tests, and no scaffolding (comprised of only high information demonstrations) yielded significantly higher performance for high difficulty tests. Improved performance when matching the information gain and difficulty of teaching and testing demonstrations respectively (which yields similar demonstrations) further suggests that *IL-style learning may have also been at play.*

Additionally, participants across each condition never achieved a mean score of greater than 0.5 for high difficulty tests, indicating that they were largely unable to grasp the more subtle aspects of the agent's optimal behavior. While the five demonstrations shown in US2 should have conveyed the maximum possible information (in an IRL-sense), they were not as effective in reality. One reason may be that human cognition is constrained by limited time and computation [28], and at times may opt for approximate, rather than exact, inference [35, 90]. Approximate inference indeed would have struggled with high difficulty tests whose optimal behavior could often only be discerned through exact computation of rewards. We move to an approximate inference model of human belief later in Chapter 6.

Finally, the current method of scaffolded teaching assumes that the learner has no prior knowledge when calculating a demonstration's information gain (e.g. Algorithm 1 considers a repeat showing of a demonstration to a learner as providing equal information gain as the first showing). But when filtering for teaching and testing sets for the user studies, we sometimes observed and accounted for the fact that

demonstrations with the same BEC area could further vary in informativeness or difficulty to different learners in two primary ways.

First, demonstrations with the same BEC area could differ in informativeness depending on whether it presented an expected behavior, given the human's expected prior knowledge. While we always placed the SCOT demonstrations at the end of the scaffolded sequence, a SCOT demonstration could have a large BEC area (e.g. if it only contributes a single constraint) and could be shown earlier. However, a SCOT demonstration always contributes a (highest information gain) constraint of BEC(π^*) that is guaranteed to reduce BEC area of the running model of human knowledge, and showing this SCOT demonstration too soon could render a later non-SCOT demonstration's constraints to be redundant. Instead, showing non-SCOT demonstrations that iteratively decrease in BEC area first, then showing SCOT demonstrations ensures that the learner always receives non-redundant constraints on $\mathbf{w}^*$ at each step. *These observations highlight that information gain cannot be calculated based solely based on the demonstration itself (and its BEC area), but must be calculated with respect to its effect on an explicit model of human prior knowledge.* We believe that providing demonstrations that incrementally deviate from the human's current model will be more informative to a human and would be better suited to scaffolding, which is addressed in the next chapter.

one could include SCOT demonstrations in between the other demonstrations in theory in order of increasing BEC area However, a SCOT demonstration that contributes a (highest information gain) constraint of BEC(π^*) may in fact have a large BEC area. Thus, showing this SCOT demonstration early on may actually render a later k-means demonstration as uninformative (i.e. the SCOT demonstration's BEC(π^*) constraint may cause a later k-means demonstration's constraints to be redundant). Instead, showing k-means demonstrations that iteratively decrease in BEC area, then showing SCOT demonstrations ensures that the learner receives non-redundant constraints on $\mathbf{w}^*$ at each step. *These two observations highlight the critical importance of maintaining an explicit model of human prior knowledge when calculating a demonstration's potential information gain.* We believe that providing demonstrations that incrementally deviate from the human's current model will be more informative to a human and would be better suited to scaffolding, which is addressed in the next chapter.

Second, we observed that demonstrations where the agent would simply 'exit' rather than complete the task (as the latter would yield a lower overall reward than exiting) were always associated with a relatively large BEC area as considering one-step deviations (Eq. 4.4) would often only yield a single constraint that would correspond to a BEC area of 2π, or half of the surface of a sphere in our setting. But despite the large BEC area, we observed during piloting that 'exits' were quite difficult to understand and predict during teaching and testing, respectively. We hypothesize that people naturally had a bias toward figuring out *how* to complete the task presented to them (e.g. what path should they take) rather than *if* they should complete the path. The latter appears to require a more difficult multi-step process of first determining the best trajectory for completing the task, then comparing that trajectory's reward to that of exiting, which is arguably more challenging than simply determining the best trajectory. The 'exit' action was originally incorporated into the domains in order to support direct inference over the upper and lower bounds of a reward feature weight given the action weight. The domains in subsequent chapters remove this action to focus simply on how the agent will complete a task given its reward function, rather than if it will at all.

Visual Saliency: Optimizing visuals improved test performance, but only for high difficulty tests. This suggests that simplicity and pattern discovery could produce a meaningful reduction in complexity for only high information demonstrations (which contain the insights necessary to do well on the high difficulty tests), while those of low and medium information were already comprehensible.

We found counterintuitive results on mental effort or puzzlement ratings (M3–M4) for H4, where ratings for mental effort and puzzlement increased from negative to positive visual saliency. One factor may have been the open-ended phrasing of the corresponding Likert prompts that failed to always elicit the intended measure. For example, one participant expounded upon their mental effort rating by saying "it takes a bit of efford [sic] remembering that you can quit at any time," referencing the difficulty of remembering all available actions rather than the intended difficulty of performing inference over the optimal behavior.

Similarly, the open-ended prompt for puzzlement failed to always query specifically for potential puzzlement arising from (a potentially counterintuitive) ordering of the demonstrations. Instead, it sometimes invited comments on puzzlement arising from

other factors, e.g. "The main puzzling thought is why did the triangle exited in a configuration and in the next one it decided to do it even though it was the same", and interestingly informed us of unforeseen confounders such as limited memory. As participants could not rewatch previous demonstrations (to enforce scaffolding order), consecutive demonstrations selected to be as similar as possible (in the positive visual saliency condition) sometimes led to greater confusion as participants believed they saw different behaviors in the same environment. To address this issue of limited memory, in subsequent studies participants were not only allowed to rewatch current demonstrations but also allowed to freely view any previous demonstrations.

Finally, we note that our selected demonstrations often revealed information about multiple reward weights at once, which could be difficult to process. Instead, we can further scaffold by teaching about one weight at a time, when possible, which we explore in the next chapter.

Testing: Objective and subjective results strongly support BEC area as a measure of test difficulty, and following studies thus used tests of varying BEC areas to evaluate and track the learner's understanding throughout the learning process. One limitation of this measure of test difficulty is that it is agnostic to the human's current belief and is a gross measure (for perhaps an average learner). For example, a test classified as medium difficulty test may in fact be of medium difficulty to a novice but may be easy for an expert. In the next chapter, we provide a way to condition on the current beliefs of a human to provide a more personalized measure of difficulty.

Domain: Exploratory analyses show that two of our key experimental results are domain-dependent. That is, our findings that no scaffolding increases test performance over forward scaffolding for high difficulty tests, and positive visual saliency increases test performance over negative visual saliency for high difficulty tests are both largely driven by the skateboard domain. As noted previously, the three domains were designed to vary in the difficulty of their respective trajectories and participant performance in fact varied across the domains. Seeing how even seemingly similar grid world domains can impact the efficacy of the proposed methods, we explicitly consider domain as an independent variable in the user studies in Chapter 6 and provide further discussion in Chapter 7.

5 Demonstration Selection by Reasoning over Human Counterfactual Beliefs and Feature Spaces

This chapter builds on the previous by modifying how the information gain of demonstrations are calculated during scaffolding, which was originally done without consideration for the human's current beliefs. Instead, we now explicitly model the human's beliefs over the robot's decision making and calculate a demonstration's information gain based on how it differs from the human's expectations (i.e. *counterfactuals*) of what the robot will do. A calibrated measure of demonstration information gain aids in the selection of demonstrations that fall in the zone of proximal development. Second, we aim to further improve demonstration selection by incrementally increasing (and *scaffolding*) the number of *unique reward features* that are conveyed. And finally, we also update our measure for estimating the *difficulty* for a human to predict instances of a robot's behavior in unseen environments as *tests* by conditioning it on a human's current beliefs of the reward function, measuring how many of that individual's beliefs would yield the correct behavior[1].

A user study finds that our test difficulty measure correlates well with human performance and confidence but finds no effect for feature scaffolding. Considering human beliefs on robot decision-making in selecting informative demonstrations decreases human performance on easy tests, but increases performance for difficult tests, providing insight on how to best utilize such human models.

[1]The contents of this chapter were presented at IROS and is available at [52].

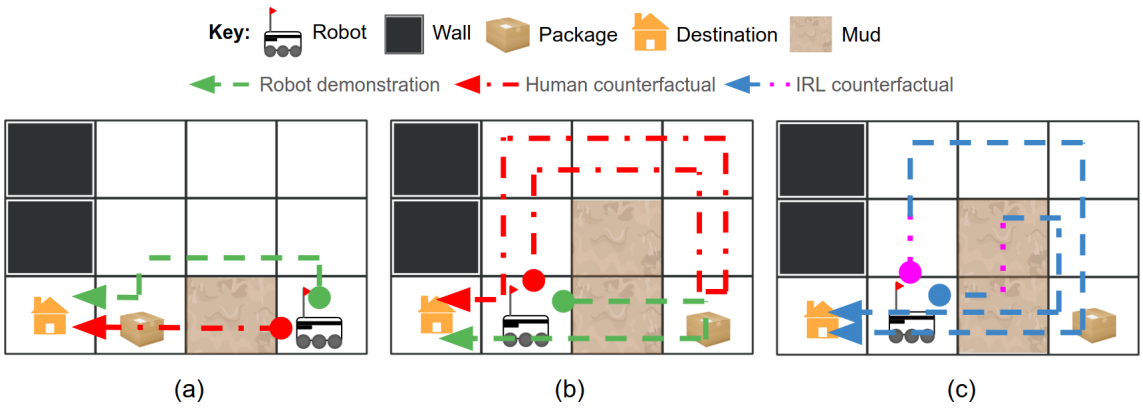

Figure 5.1: **(a)** A robot's optimal demonstration (green) is shown in contrast to a suboptimal counterfactual alternative (red). **(b)** A robot's optimal demonstration is shown in contrast to a counterfactual likely considered by a human who has seen the demonstration in (a). **(c)** Sample counterfactual alternatives to the robot's trajectory in (b) that are considered by standard IRL, generated by deviating from the robot's path by one action (pink), then following the robot's optimal policy afterward (blue). Note that neither matches the human's counterfactual.

5.1 Motivation

For IRL, the information a demonstration reveals regarding the underlying reward function inherently depends on the *counterfactuals* (i.e. alternative, suboptimal demonstrations) that are considered. Imagine a human who encounters a robot in the delivery domain for the first time. To convey its reward function, imagine the robot providing a human with the demonstration in Fig. 5.1a. Because the robot takes a two-action detour to avoid the mud instead of going through it (a natural counterfactual), the human may infer that the robot associates a negative reward with going through mud.

After providing this first demonstration, the robot considers what to demonstrate next to convey more information regarding its reward function. Importantly it knows that the human likely knows that mud is costly from the first demonstration, but does not know *how* costly. For instance, the human may reasonably believe that the robot would take a four-action detour when faced with two mud patches (Fig. 5.1b). However, the robot knows that its ratio of mud to action reward is -3 to -1 and that consequently, it would simply go through the mud in Fig. 5.1b to maximize

its reward. Seeing how its direct path meaningfully differs from the human's likely counterfactual that detours heavily, the robot considers this to be a very informative demonstration to provide next to upper-bound the human's belief of the cost of mud.

Standard IRL [66], however, does not model the learner's beliefs and would fail to consider this detouring human counterfactual. Instead, standard IRL enumerates all trajectories that differ by a single initial action as counterfactuals. Two sample IRL counterfactuals are shown in Fig. 5.1c, but neither matches the intuitive human counterfactual of completely detouring around the mud. As a result, IRL has the potential to both under- and over-estimate the information gain of a demonstration to a human by considering the wrong counterfactuals or considering too many, respectively.

Looking to the related literature on how humans explain to one another, Miller notes that "explanations are contrastive—they are sought in response to particular counterfactual cases," and that it is critical that the learner's counterfactuals matches the ones intended by the teacher [65]. In our work, we tailor demonstrations to the learner given the agent's estimate of the human's current belief and the counterfactuals that the human will likely consider.

Furthermore, Reiser [78] suggests that scaffolding should sometimes challenge the learner by inducing cognitive conflict whose reconciliation results in learning (e.g. by providing examples that challenge and refine the learner's current understanding). Thus, this chapter aims to ensure that the robot correctly anticipates the human's likely counterfactuals and provides demonstrations that *differ* from those counterfactuals to provide information, which we explore next.

5.2 Methods

This section develops methods that leverage a model of the learner's current beliefs and likely counterfactuals in 1) selecting informative teaching demonstrations and 2) rating the difficulty of potential tests. We also consider providing further scaffolding on teaching demonstrations by incrementally increasing the number of reward weights communicated by each demonstration.

Counterfactual Scaffolding

We make two observations regarding the BEC. First, Eq. 4.4 captures the key idea that IRL depends not only on the robot's optimal trajectory but also on the suboptimal counterfactual trajectories that are considered, represented by $\mu_\pi^{(s,a)}$ and $\mu_\pi^{(s,b)}$ respectively. Second, BEC($\mathcal{D}|\pi$) could be used to model the human's belief over the robot's possible reward weights after having seen $\mathcal{D}$. We now build on these concepts to leverage a human model and select demonstrations that account for human counterfactuals[2].

As previously mentioned, a demonstration's ability to reveal the underlying reward function via IRL hinges on the counterfactuals considered. However, many counterfactuals proposed by IRL can seem nonsensical to humans as they fail to consider the human's beliefs. Instead, IRL generates counterfactuals in the following way—at each state s along the robot's optimal trajectory, it first takes a potentially suboptimal action b before following the optimal policy afterward (4.4). This process generates the two sample counterfactuals in Fig. 5.1c, which do not correspond to the human counterfactual in Fig. 5.1b. While such one-action deviations from the optimal trajectory are computationally sensible and efficient (i.e. multi-action deviations often yield only redundant constraints), these are unlikely to be the counterfactuals on the human's mind for a number of reasons.

First, humans are unlikely to methodically go through each state of the robot's trajectory and consider all alternative actions. Instead, humans naturally incline toward a few causes and a few counterfactuals for explanation [65]. This can lead IRL to overvalue the information gain of a demonstration if counterfactuals beyond those on the human's mind are considered. Second, IRL counterfactuals are generated by "perturbing" the demonstration directly (by taking a suboptimal action) and may not be consistent with any reward function (e.g. no reward function in the delivery domain would first avoid mud, then later go through mud like one of the counterfactuals in Fig. 5.1c). While humans may consider a reward function that differs from the robot's, their counterfactuals are likely to be consistent with that differing reward function (e.g. avoiding the mud both ways in Fig. 5.1b). This

[2]Code for the methods, domains, and relevant hyper-parameters (e.g. reward weights, sample rate) used in this study can be found at https://github.com/SUCCESS-MURI/counterfactual_human_IRL.

can lead IRL to also undervalue a demonstration's information gain if the human's counterfactuals are not considered.

In selecting effective explanations, we posit that you must consider not only the learner's learning model (i.e. IRL) but also their beliefs and subsequently what counterfactuals they would consider. We thus extend our work in the previous chapter to evaluate a demonstration's information gain based on counterfactuals generated via potential reward functions estimated to be on the human's mind as opposed to counterfactuals generated via one-action deviations, and scaffold by showing demonstrations of increasing information gain.

To account for human beliefs and counterfactuals when evaluating the information gain of potential demonstrations, we do the following. First, we instantiate a prior model of the human's beliefs over the reward feature weights $\mathbf{w}^*$, $B(\mathbf{w}^*)$. This model could be the full $N-1$ sphere if the human has no prior knowledge, or it may be a partial sphere due to prior knowledge (e.g. that action reward is negative). Then we sample m weights from $B(\mathbf{w}^*)$. Each weight represents a particular belief that the human could have over the robot's reward function. For every robot possible demonstration in a domain, and for each of the m weights, we simulate what the human counterfactual to each demonstration would be if the human had this reward weight in mind and generate the corresponding constraints using (4.4). For each possible demonstration by the robot, we consolidate the corresponding m human counterfactuals by taking a union of all corresponding constraints. Finally, we select the demonstration that maximizes information gain, i.e. select the demonstration that maximizes the ratio between $B(\mathbf{w}^*)$ before and after the human sees this demonstration. We take the ratio rather than the difference as we empirically observed that the latter does not faithfully capture information gain in instances where $B(\mathbf{w}^*)$ has unequal uncertainty across multiple feature weights (e.g. the ratio between a narrow, long $B(\mathbf{w}^*)$ and a narrow, short $B(\mathbf{w}^*)$ is large whereas the difference is small even though much information was gained on the uncertain feature). Once we have shown the selected demonstration and updated $B(\mathbf{w}^*)$, we select the next demonstration to show by sampling m weights from the updated $B(\mathbf{w}^*)$ and repeating the steps above. This method is summarized in Alg. 2.

Algorithm 2 Counterfactual Machine Teaching for Humans

Require: π^*: robot policy, Ξ: all possible demonstrations of π^* in a domain, m: number of beliefs to sample, $B(\mathbf{w}^*)$: human prior over robot reward weights

infoGain $= \infty$
$\mathcal{D} = [\,]$
while infoGain $\neq 0$ **do**
    $B_{dict} = \varnothing$
    // Sample human beliefs on $\mathbf{w}^*$
    $\mathbf{W} = \text{sample}(m, \text{B}(\mathbf{w}^*))$
    // Obtain constraints yielded by each possible demonstration, conditioned on
    // the sampled human beliefs
    for $\xi \in \Xi$ **do**
        $\mathcal{C} = \varnothing$
        for $\mathbf{w} \in \mathbf{W}$ **do**
            // Constraints given "human" counterfactual. $\hat{\mathbf{N}}[\cdot]$ extracts unit normal
            // vectors corresponding to a set of half-space constraints
            $C = C \cup \hat{\mathbf{N}}[\text{BEC}(\xi|\pi_{\mathbf{w}})]$
        end for
        // Store updated belief given this demonstration
        $B_{dict}[\xi] = C \cup \hat{\mathbf{N}}[B(\mathbf{w}^*))]$
    end for
    // Select the trajectory that maximizes information gain
    $\xi^* = \underset{\xi \in \Xi}{\text{argmax}}\, \text{BECArea}(B(\mathbf{w}^*)) \,/\, \text{BECArea}(B_{dict}[\xi])$
    infoGain $= \text{BECArea}(B(\mathbf{w}^*)) \,/\, \text{BECArea}(B_{dict}[\xi^*])$
    if infoGain $\neq 1$ **then**
        $\mathcal{D}$.append(ξ^*)
        $\Xi = \Xi \setminus \xi^*$ ▷ $\setminus$ denotes set subtraction
        $B(\mathbf{w}^*) = B_{dict}[\xi]$ ▷ Update human belief
    end if
end while
return $\mathcal{D}$ ▷ Final demonstration set to show human

Feature Scaffolding

A standard scaffolding technique suggested by Wood et al. [96] is to reduce the degrees of freedom of the problem. Accordingly, one could initially show demonstrations that limit the number of reward features over which information is conveyed. In the delivery domain, one could show demonstrations that convey information on the mud and action weights first, then on battery recharge and action weights, then on mud, battery, and action weights to show potentially nuanced tradeoffs. Put another way, this sequence of demonstrations first "masks" the battery recharge weight, then the mud weight, then no weights, filtering demonstrations such that they do not convey information on a masked weight. In general, we can scaffold k features by showing demonstrations that iteratively mask $k-2$, $k-3$, ..., 0 features (because solutions of IRL are scale-invariant, we must show at least two features at a time relative to one another, e.g. how many actions the robot is willing to take to avoid mud). At every iteration in which we wish to mask n features, there are $\binom{k}{n}$ possible masks that can be applied. We now discuss how to order the possible masks at every iteration.

The key idea behind ordering the possible masks is to hide features that appear infrequently, as infrequently appearing features are less likely to be able to initially support fine-grain comparisons. To do so, we first obtain all possible constraints that could be generated using Eq. 4.4 for all possible agent demonstrations in a domain. Then we tally the number of nonzero entries across all of the constraints (e.g. the sample constraint $\mathbf{w}^\top[2, 0, -5] \geq 0$ has nonzero entries for the first and third features) for each feature. For each of the $\binom{k}{n}$ masks, we simply sum the frequency of each of the masked features and order them from the lowest sum to the highest sum (which will allow the features with the highest frequency to be conveyed first and the features with the lowest frequency to be conveyed last).

Once the order of masks has been decided, we apply each of the masks in this iteration in turn (where each iteration corresponds to masking n features). For each mask that is applied, we remove any demonstrations that convey information about a masked feature from consideration (i.e. any demonstrations that convey constraints in which the entry for a masked feature is nonzero). From this reduced set of demonstrations, we run counterfactual scaffolding as described in the previous subsection until there are no more demonstrations that can provide additional information gain.

We then move on to the next mask in this iteration, removing any demonstrations that convey information about a masked feature from consideration and running counterfactual scaffolding until there are no more demonstrations that can provide additional information gain. And so on and so forth until all masks in this iteration have been applied. We then move on to the set of masks for the next iteration (which are also ordered by the frequency of masked features), until we have gone through every possible mask. In the delivery domain, the first iteration would first mask recharging and only show demonstrations that trade off mud and action weights, then would mask mud and only show demonstrations that trade off recharge and action weights. The second iteration would not mask any feature but would show trade-offs that involve mud, recharge, and action weights. We note that the approach as conveyed here is exponential in the number of reward features, and we leave the formulation of a more efficient, non-exhaustive method for future work.

Testing

The area of a demonstration's BEC intuitively correlates with its information gain during teaching as smaller areas indicate less uncertainty regarding $\mathbf{w}^*$. The previous chapter showed that a demonstration's BEC area may also be inverted to measure the difficulty of correctly predicting the demonstration as a test if the human has not seen it before (e.g. so that a smaller BEC area indicates a more difficult test). But as previously mentioned, this measure is perhaps a gross measure as it is agnostic to a human's current beliefs.

We hypothesize that the overlap between BEC($\xi|\pi^*$) and $B(\mathbf{w}^*)$, a model of a human's beliefs over the robot's reward weight, better captures the difficulty of a demonstration ξ as a test for this particular individual. This overlap intuitively represents the number of candidate reward functions in the human's mind that would generate the correct behavior. As seen in Fig. 5.2, a demonstration may have an intrinsically large BEC area but may not overlap much with the human's belief and may therefore be a difficult test for this individual.

To estimate the expected difficulty of each ξ that could be shown in a domain, we first obtain the BEC($\xi|\pi^*$) using (4.4). Noting that one-action deviation does not always consider all reasonable counterfactual trajectories, we take a union over

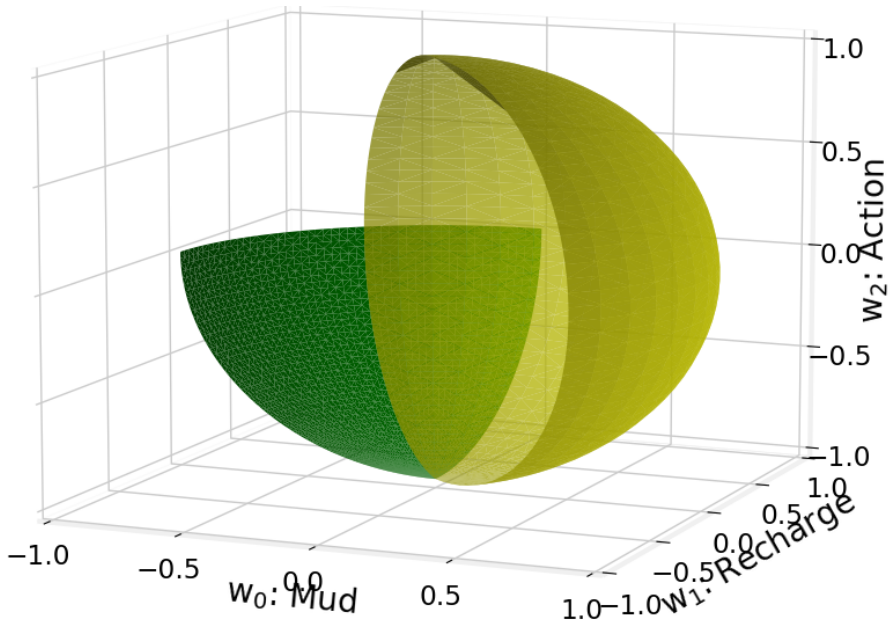

Figure 5.2: There are many reward weights BEC($\xi|\pi^*$) (yellow) that will generate the demonstration ξ. However, only a portion overlaps with the weights currently on the human's mind $B(\mathbf{w}^*)$ (green), making it difficult for the human to correctly predict ξ during testing.

Algorithm 3 Measuring Test Difficulty of a Demonstration

Require: π^*: robot policy, ξ: demo, m: number of beliefs to sample, $B(\mathbf{w}^*)$: human prior over robot reward weights

1: $\text{BEC}'(\xi|\pi^*) = \varnothing$
2: // Sample possible beliefs on $\mathbf{w}^*$
3: $\mathbf{W} = \text{sample}(m, \text{B}(\mathbf{w}^*))$
4: // Obtain constraints yielded by each possible demonstration, using both standard IRL and human counterfactuals
5: $\text{BEC}'(\xi|\pi^*) = \text{BEC}'(\xi|\pi^*) \cup \hat{\mathbf{N}}[\text{BEC}(\xi|\pi^*)]$
6: **for** $\mathbf{w} \in \mathbf{W}$ **do**
7: $\quad \text{BEC}'(\xi|\pi^*) = \text{BEC}'(\xi|\pi^*) \cup \hat{\mathbf{N}}[\text{BEC}(\xi|\pi_{\mathbf{w}})]$
8: **end for**
9: // The overlap is inversely correlated to difficulty
10: $\text{difficulty} = 1/\text{measureOverlap}(B(\mathbf{w}^*), \text{BEC}'(\xi|\pi^*))$
11: **return** difficulty

constraints that define $\text{BEC}(\xi|\pi^*)$ and constraints obtained from counterfactual scaffolding using m models sampled from the human's belief over the agent's reward weights $B(\mathbf{w}^*)$. These combined constraints for each demonstration will give a better estimate of the set of all weights that yield the correct demonstration, denoted by $\text{BEC}'(\xi|\pi^*)$. Finally, to measure the difficulty of a demonstration ξ as a test for this human, we simply take the overlap between $B(\mathbf{w}^*)$ and $\text{BEC}'(\xi|\pi^*)$. The smaller the overlap, the fewer of the reward weights in the human's mind will generate the correct demonstration and the harder the test. This method is summarized in Alg. 3.

5.3 User Study

We ran an online user study[3] that explored whether demonstrations selected using our proposed methods of counterfactual and feature scaffolding improves a human's understanding of a robot's policy. Similar to the user studies of Chapter 4, this study involved participants watching robot demonstrations in three domains and predicting the robot's behavior in new test environments. Each domain consisted of one shared action reward feature (that penalized each action with a reward weight of

[3]Code for the user study, data, and analyses can be found at https://github.com/SUCCESS-MURI/counterfactual_human_IRL_study.

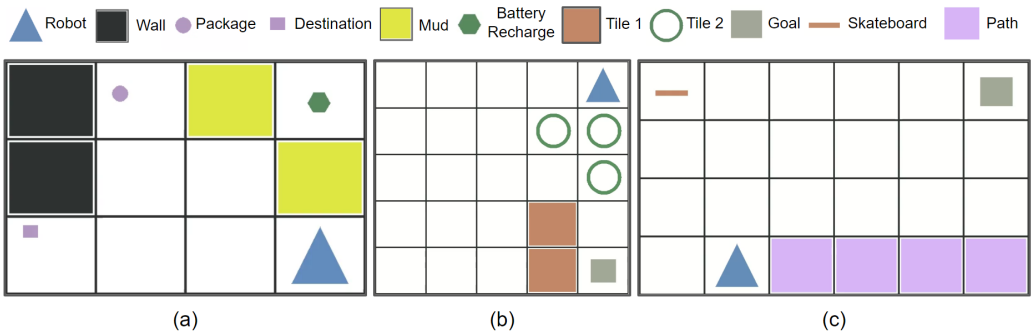

Figure 5.3: Three domains were designed for the user study, each with a different set of reward weights to infer from demonstrations: **(a)** delivery **(b)** tiles **(c)** skateboard. The semantics of the various objects were hidden using abstract geometric shapes and colors.

-1), and two unique reward features as follows with the corresponding reward weight in parentheses (see Fig. 5.3).

Domains

Delivery domain. The robot is penalized for moving out of mud (-3) and rewarded for recharging (+3.5). Five demonstrations were shown in this domain.

Tiles domain[4]**.** The robot is penalized differently for traversing the two differently shaped tiles (-6.5 and -5.25 respectively). Five demonstrations were shown in this domain.

Skateboard domain. The robot is penalized less per action if it has either picked up a skateboard (i.e. riding a skateboard is less costly than walking, +0.825) or is traversing through a designated path (-5.25). Seven demonstrations were shown in this domain.

The number of demonstrations shown in each domain was determined by the number needed by counterfactual and feature scaffolding to arrive at a resultant BEC area that matched that of SCOT demonstrations. More demonstrations were selected in the skateboard domain than others steadily work up to the nuanced trade-offs

[4]The tiles domain replaces the two-goal domain in the previous chapter, as the reward weights in the latter reduce to a single trade-off regarding which goal to go to without an 'exit' action (i.e. the weights would lie on the perimeter of a unit circle rather than the surface of the unit sphere like the other domains).

likely arising from the more fine-grained reward weights.

Study design

The participants were explicitly informed of each domain's reward features, but had to infer the respective reward weights by watching demonstrations. The demonstrations they were provided were determined by the one of four between-subjects conditions they were assigned.

The between-subjects variables were *BEC scaffolding* (counterfactual and baseline), and *feature scaffolding* (yes and no). Baseline scaffolding followed the method proposed by our prior work 4, using one-action deviations to generate counterfactuals and selecting demonstrations that iteratively decreased in BEC area. As a brief refresher, baseline scaffolding ordered demonstrations by their BEC area, then clustered in (n - j) * 2 - 1 clusters (where n is the teaching budget and j is the number of SCOT demonstrations), such that demonstrations could be selected from every other cluster. Baseline scaffolding demonstrations always ended with SCOT demonstrations. And no feature scaffolding did not iteratively mask features or hold out corresponding demonstrations as feature scaffolding did. When counterfactual scaffolding was paired with no feature scaffolding, a demonstration that provided 70% of the maximal information gain was always selected for the delivery and skateboard domains to allow all between-subjects conditions to provide the same number of demonstrations (otherwise this condition would have shown fewer demonstrations than the other conditions). Importantly, we note that demonstrations from all conditions resulted in the same final BEC area for each domain, theoretically providing the same amount of information in the end. Finally, we conservatively modeled $B(\mathbf{w}^*)$ prior to any demonstrations having been shown as knowing that the action reward is negative (assuming a human bias for efficiency).

All four between-subjects conditions optimized visual similarity amongst consecutive demonstrations and visual simplicity within demonstrations, as suggested by our prior work 4. Thus, for any given set of demonstrations with equal information gain, the one that looked most similar to the previously shown demonstration (e.g. location of mud patches) and also had the fewest visual clutter (e.g. number of mud patches) was selected to be shown next.

There were also two within-subject variables: *domain* (delivery, tiles, skateboard) and *test difficulty* (high, medium, and low). The tests were pulled from three representative sets of demonstrations that had high, medium, and low overlap between $B(\mathbf{w}^*)$ and $\text{BEC}'(\xi|\pi^*))$ for the low, medium, and high difficulty test conditions. Specifically, the overlaps for every possible demonstration in a domain were ordered from high to low, grouped into five clusters using K-means, and high, medium, and low overlap demonstrations were taken from the 1st, 3rd, and 5th cluster respectively. We conservatively modeled the $B(\mathbf{w}^*)$ of a person who watched all of the teaching demonstrations with constraints as knowing the correct sign of each of the reward weights (e.g. knowing that mud is negative and battery is positive in the delivery domain).

The user study itself consisted of three trials, with each trial comprising a teaching portion and a testing portion in a unique domain. During teaching, participants were explicitly informed of the reward features of the domain, then they inferred the corresponding reward weights by watching demonstrations. To mitigate the effects of limited memory, participants were allowed to watch a demonstration as many times as they wished, and were also allowed to rewatch previous demonstrations during the teaching portion. Before moving on to the testing portion, the participants provided subjective observations regarding the demonstrations. For testing, participants were tasked with predicting the optimal trajectory in six unseen test environments (a random order of two high, medium, and low difficulty environments each) and rating their confidence in their responses.

Hypotheses

H1: The overlap between a human's belief over the weights $B(\mathbf{w}^*)$ and the BEC of a demonstration $\text{BEC}'(\xi|\pi^*)$ during teaching correlates inversely to the difficulty of predicting it *during testing* and correlates directly to their prediction confidence.

H2: Demonstrations selected with counterfactual scaffolding will result in higher perceived informativeness during teaching and better participant test performance over those selected with baseline scaffolding [51].

H3: Demonstrations selected with feature scaffolding will result in lower mental effort during teaching and better participant test performance over those selected

without.

H4: Demonstrations selected with counterfactual scaffolding and feature scaffolding will result in the highest perceived informativeness of teaching demonstrations, lowest mental effort, and best participant test performance compared to the other possible conditions.

Measures

The following objective and subjective measures were recorded to evaluate the aforementioned hypotheses. The Likert scales corresponding to M2-M3 were provided after all of the demonstrations but before the tests. The Likert scale corresponding to M4 was provided after each test.

M1. Optimal response: Participants were assigned a binary score depending on the optimality of their test trajectory.

M2. Informativeness rating: "How informative were these demonstrations in understanding the best strategy [robot's policy] in this game?", answered with a 5-point Likert scale

M3. Mental effort rating: "How much mental effort was required to understand the best strategy [robot's policy] in this game?", answered with a 5-point Likert scale

M4. Confidence rating: "How confident are you that you minimized [the robot's] energy loss while completing the task [i.e. performed the task optimally in this unseen test environment]?", answered with a 5-point Likert scale

These measures correspond to those used in the user studies in the previous chapter, with the exception of 'puzzlement rating', which was originally included to potentially highlight the expected counterintuitive ordering of backward-scaffolded demonstrations.

5.4 Results

We collected data from 216 participants using Prolific. Participants were roughly 67% male, 32% female, 1% non-binary, and ages varied from 18 to 69 (M = 28.39, SD = 9.48). The recruitment process and study was approved by Carnegie Mellon University's Institutional Review Board. 54 participants were randomly assigned to

each of the four between-subjects conditions and the order in which the domains were shown was fully counterbalanced. We removed data from 4 participants whose aggregate test performance (compared to the optimal answer)[5] or individual test performances (compared to other participants)[6] were 3 standard deviations below their respective means as outliers.

The three domains varied in the difficulties of their optimal policies. We calculated a mean-rating (k = 3), 2-way mixed effects, consistency-based intraclass coefficient (ICC) to see how the performance of each participant varied across domains [45]. Given an ICC of 0.32 that indicates significant variance ($p < .001$), we average the performance of every participant across domains and provide findings that may represent a range of domains and difficulties.

H1: A one-way repeated measures ANOVA on percentage of optimal responses revealed a statistically significant difference across test difficulty ($F(2, 422) = 289.78, p < .001$). Post-hoc pairwise Tukey analyses confirmed significant differences between high ($M = 0.40$), medium ($M = 0.71$), and low ($M = 0.86$) test difficulties ($p < .001$ in all cases).

Spearman's rank-order correlation revealed that confidence inversely correlated significantly with test difficulty ($r_s = -.36, p < .001, N = 636$).

The data strongly support H1 that the overlap between $B(\mathbf{w}^*)$ and $\text{BEC}'(\xi|\pi^*)$ captures a demonstration's difficulty for testing. We also confirm *test difficulty* as a valid within-subjects variable.

H2: A two-way mixed ANOVA revealed a significant interaction between counterfactual scaffolding and test difficulty for percentage of optimal responses ($F(2, 420) = 6.56, p = .002$). Tukey analyses revealed that for low difficulty tests ($p = .002$), no counterfactual scaffolding ($M = 0.90$) significantly improved performance over counterfactual scaffolding ($M = 0.83$). However, the relationship was reversed for high difficulty tests ($p = .048$) with counterfactual scaffolding ($M = 0.44$) outperforming no counterfactual scaffolding ($M = 0.37$), as Fig. 5.4 shows. A significant effect was not revealed for counterfactual scaffolding by a two-way mixed ANOVA

[5]Calculated by averaging each participant's 18 test responses (i.e. six tests in three domains) into a percentage of tests that the participant got correct.

[6]Calculated by comparing an individual's test performances against other participants. The number of times a participant's reward for a test trajectory was 3 standard deviations below the mean reward was compared.

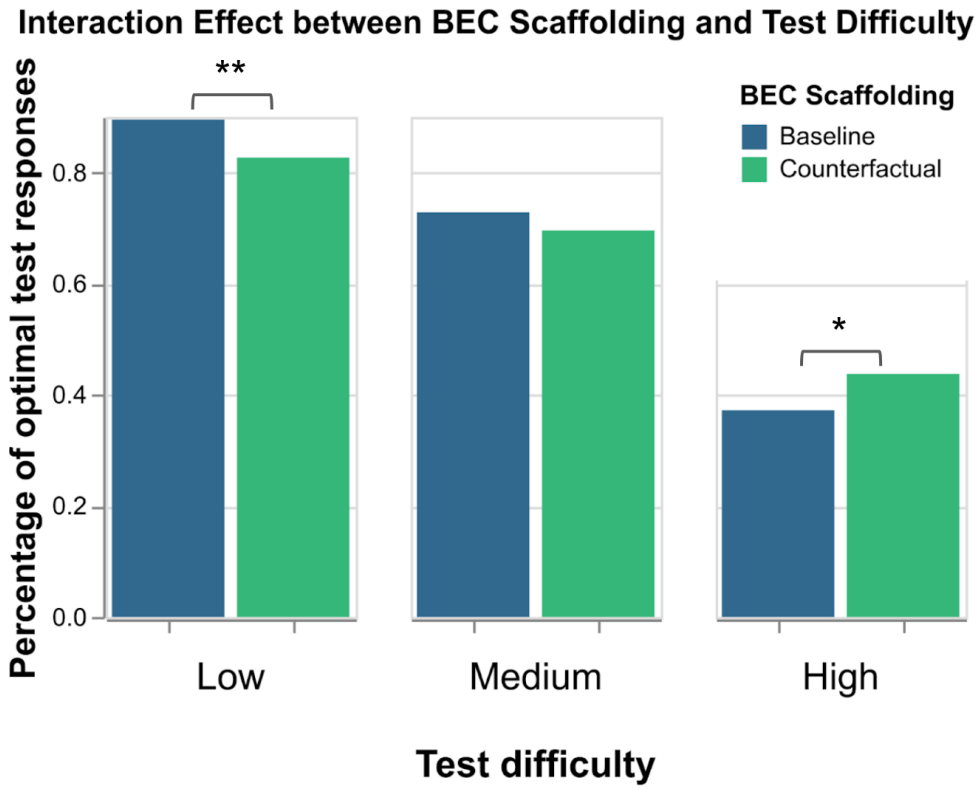

Figure 5.4: While baseline scaffolding significantly increases performance on low difficulty tests over counterfactual scaffolding, the effect is reversed for high difficulty tests.

$(F(1, 210) = 0.47, p = .49)$.

Ratings for mental effort was found by a Mann-Whitney U test to be significantly higher for counterfactual scaffolding ($U(N_{baseline} = 108, N_{counterfactual} = 104) = 4690.0, p = .03$). The two counterfactual scaffolding conditions did not differ significantly in informativeness ratings ($p = .08$).

As exploratory measures, we also recorded the average number of times a participant watched each demonstration and the time taken for a participant to provide a test demonstration and rate their confidence. Interestingly, Tukey analyses revealed that counterfactual scaffolding significantly increased the average number of times a teaching demonstration was watched ($M = 1.23$) over baseline scaffolding ($M = 1.15, p = .02$) and also significantly increased the time taken to complete a test ($M = 2.95$ sec) over baseline scaffolding ($M = 2.49$ sec, $p = .01$).

Finally, seeing the strong effect of domain on the results of the user studies in Chapter 6, we explore whether the significant effects above were also driven by domain. Indeed, t-tests revealed that only in the skateboard domain did counterfactual scaffolding yield significantly higher learning outcomes over baseline scaffolding for high difficulty tests ($M = 0.49$ vs $M = 0.27$, at Bonferroni adjusted $p < .001$) and the relationship significantly reverse for low difficulty tests ($M = 0.59$ vs $M = 0.77$, at Bonferroni adjusted $p = .006$). Interestingly, Mann-Whitney U tests showed that the only significant difference in ratings of mental effort for counterfactual scaffolding was in colored tiles, where counterfactual scaffolding was rated to require more mental effort ($M = 2.32$) over no counterfactual scaffolding ($M = 2.02$) at Bonferroni adjusted $p = .042$.

The data partially support H2. Counterfactual scaffolding fails to outperform baseline scaffolding in test performance. However, counterfactual scaffolding appears to improve test performance on high difficulty tests at the cost of increased mental effort (as indicated by both objective and subjective measures).

H3: A two-way mixed ANOVA revealed that feature scaffolding had no significant effect on percentage of optimal responses ($F(1, 210) = 1.79, p = .18$), and no interaction between feature scaffolding and test difficulty ($F(2, 420) = 1.72, p = .18$). Mann-Whitney U tests found that feature scaffolding did not impact ratings on informativeness ($p = .81$) nor mental effort ($p = 0.14$). Again, we did an exploratory analysis to see if feature scaffolding would have an interaction effect with domain on percentage of optimal responses but did not find any.

The data does not support H3. We did not observe any effect for feature scaffolding.

H4: A two-way mixed ANOVA revealed a significant interaction effect between the four possible between-subjects conditions and test difficulty on percentage of optimal responses ($F(6, 416) = 4.40, p < .001$). Tukey analyses showed that counterfactual scaffolding with no feature scaffolding ($M = 0.75$) was significantly outperformed by baseline scaffolding with ($M = 0.88$) and without ($M = 0.91$) feature scaffolding, and also by counterfactual scaffolding with feature scaffolding ($M = 0.91$) for low test difficulty. The conditions did not significantly affect test performance ($F(3, 208) = 1.98, p = .12$).

Mann-Whitney U tests revealed that counterfactual scaffolding with feature scaffolding (the proposed method in this work) required the most mental effort (Mdn

= 3) over baseline scaffolding with (Mdn = 2, $p < .001$) and without (Mdn = 2, $p = .007$) feature scaffolding, and also by counterfactual scaffolding without feature scaffolding (Mdn = 2, $p = .005$).

The data does not support H4. The observations that counterfactual scaffolding can decrease performance on low difficulty tests and requires more mental effort corroborates the findings for H2.

5.5 Discussion

The overlap between a human's belief over the weights $B(\mathbf{w}^*)$ and the BEC of a demonstration $\text{BEC}'(\xi|\pi^*)$ correlated inversely to the difficulty of predicting it during testing and correlated directly to their prediction confidence, such that H1 was strongly supported. Whereas the prior chapter's test difficulty measure solely relied on demonstration (i.e. test answer) BEC and was intrinsic to the test, the overlap is a more personalized test difficulty measure.

Whereas the measure of test difficulty from Chapter 4 solely based on demonstration (i.e. test answer) BEC is intrinsic to the test, this new measure of test difficulty based on the overlap with the human's belief over the weights and allows for a more personalized measure of difficulty.

Contrary to expectation, feature scaffolding did not yield any objective or subjective results and H3 was not supported. The domains each only had three reward features, which perhaps were already too few to significantly benefit from scaffolding. We hypothesize that domains with a higher number of reward features may stand more to gain from feature scaffolding.

The effect of counterfactual scaffolding was more nuanced than H2 and H4 initially expected. First, counterfactual scaffolding failed to outperform baseline scaffolding in test performance as a main effect. However, counterfactual scaffolding improved test performance on high difficulty tests at the cost of increased mental effort (as indicated by both objective and subjective measures) as an interaction effect. Along the same vein, counterfactual scaffolding with feature scaffolding (the experimental condition) required the most mental effort over any other condition but also yielded the highest *overall* and *high test difficulty performance* of counterfactual scaffolding with feature scaffolding ($M = 0.68, M = 0.45$ respectively), over baseline scaffolding

with ($M = 0.66, M = 0.35$) and without ($M = 0.67, M = 0.39$) feature scaffolding and counterfactual scaffolding without feature scaffolding ($M = 0.63, M = 0.42$).

The aforementioned results highlighted a tension that we did not initially expect, but makes sense in hindsight. As previously noted, Reiser [78] suggests that scaffolding should sometimes challenge and engage the learner by inducing cognitive conflict. Indeed counterfactual scaffolding explicitly selects demonstrations that would not be anticipated by the learner and requires the learner to reconcile the gap by updating their belief. It is unsurprising in retrospect that mental effort is often required to learn new material; the key is ensuring that the material belongs to the student's zone of proximal development and that the mental effort required is just right. This is arguably reminiscent of the famous Yerkes-Dodson law that has shown that performance increases with physiological and mental arousal up to a point, then performance decreases with arousal [99].

Counterfactual scaffolding also surprisingly performed worse than baseline scaffolding for low difficulty tests (despite performing better than baseline scaffolding for high difficulty tests as previously mentioned). One possible explanation may lie in the fact that demonstrations conveying information necessary for answering low and high difficulty tests appeared in earlier and later demonstrations, respectively. Given that higher performance on high difficulty tests (which in theory requires a more focused understanding of the agent's reward function – i.e. a smaller BEC) did not translate to higher performance on low difficulty tests suggests that participants may have again learned from demonstrations using a more imitation learning-style of reasoning (where they were able to simply recall and reproduce the later demonstration better). Additionally, our counterfactual scaffolding method always presented demonstrations with high information gain given the user's current belief. However, a person's learning ability is likely more context-dependent (e.g. on their prior knowledge, current stage of learning, etc) and the pace of learning should be more personalized, which we address through a closed-loop teaching framework in the next chapter. And as we again see key significant results for counterfactual scaffolding and mental effort only holding for a subset of the domains, we consider domain as an independent variable in the user studies of Chapter 6 and provide further discussion on the impact of domain in Chapter 7.

6 Closing the Teaching Loop with in situ Demonstration Selection

This thesis has thus far explored how an agent can select informative demonstrations that reveal its reward function to a human. Each characteristic of the reward function can be thought of as a *knowledge component* (KC), which is broadly defined in the education literature as "a concept, principle, fact, or skill inferred from performance on a set of related tasks" [43]. In this thesis, KCs are operationalized as discrete constraints on the reward function, like mud being at least twice as costly as an action, or mud being less costly than four actions.

Though machine teaching can assist in selecting a principled curriculum of demonstrations that teach a set of *a priori*, student learning may deviate from the modeled learning trajectory *in situ*. In the previous chapter, machine teaching-selected demonstrations improved human performance on tests examining understanding of early-demonstrated concepts but decreased performance on tests examining understanding of later-demonstrated concepts, suggesting perhaps that the curriculum moved too quickly past the early concepts without testing and providing additional instruction as necessary.

Thus, *our key idea is to complement a curriculum of machine teaching-selected demonstrations with a closed-loop teaching framework inspired by the education literature to provide tailored instruction in real-time* (Fig 6.1). A guiding educational concept is teaching in the *zone of proximal development* (ZPD) or "Goldilocks zone" [31, 91], which suggests that the examples provided to the learner should not be too easy nor too difficult, given their current understanding. However, the ZPD often changes at different rates for different students based on their personal learning rate,

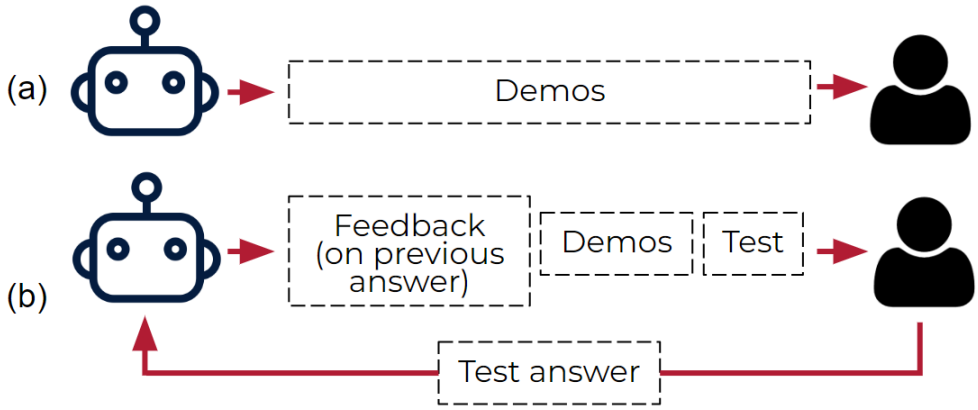

Figure 6.1: (a) Previous work aimed to improve policy transparency via a set of demonstrations selected *a priori*, but student learning may deviate from the expected trajectory. (b) We propose a closed-loop teaching framework using tests, feedback, etc., to detect and correct for such deviations *in situ*.

which must be assessed periodically through testing. We inform the testing cadence with the educational concept of the *testing effect* [79], which predicts an increase in learning outcomes when a portion of the teaching budget is devoted to testing the student (leveraging testing not only as a tool for assessment but also for teaching). By incorporating tests and feedback in a closed teaching loop, we maintain an up-to-date model of human beliefs and promote subsequent demonstrations that are provided at the right level of difficulty.

Our contributions are as follows: First, a closed-loop teaching framework based on insights from the education literature that provides demonstrations, tests, and feedback as necessary. Second, a particle filter model of human beliefs that supports iterative updates and a calibrated prediction of the counterfactuals likely considered by the human for each demonstration that could be provided. Third, a user study finds that our proposed framework reduces the regret of human test responses by 43% over a baseline and is rated as more usable by users in one of the two considered domains.

6.1 Methods

The example of the delivery robot in Section 5.1 highlights the importance of maintaining an up-to-date model of human beliefs and likely counterfactuals when selecting a

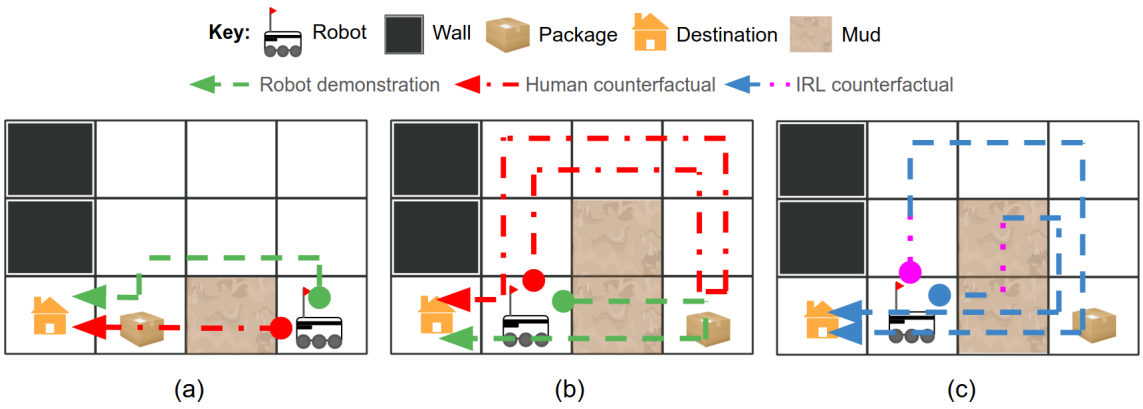

Figure 6.2: **(a)** A robot's optimal demonstration (green) is shown in contrast to a suboptimal counterfactual alternative (red). **(b)** A robot's optimal demonstration is shown in contrast to a counterfactual likely considered by a human who has seen the demonstration in (a). **(c)** Sample counterfactual alternatives to the robot's trajectory in (b) that are considered by standard IRL, generated by deviating from the robot's path by one action (pink), then following the robot's optimal policy afterward (blue). Note that neither matches the human's counterfactual.

demonstration; we wish to provide an informative demonstration that *differs* from the human's expectations (see Fig. 6.2, which is copied from Section 5.1 for convenience). In this section, we propose a particle filter-based model of human beliefs that is amenable to iterative Bayesian updates and sampling for counterfactual reasoning. We then leverage this model in a closed-loop teaching framework that leverages insights from the education literature to select demonstrations that target gaps identified through testing.

Particle Filter Human Model

Though our prior work previously modeled the human as an exact IRL learner 5, this choice falls short for two reasons. First, people are more likely to perform approximate, rather than exact, inference [35]. Second, a model of human beliefs solely comprised of half-spaces cannot handle conflicts that arise when the human incorrectly applies a knowledge component (KC) during testing that was assumed learned during teaching (as you cannot reconcile two identical half-space constraints that point in opposite directions).

We thus move to a probabilistic human model in the form of a particle filter. Each

Algorithm 4 Particle Filter for Modeling Human Beliefs

Initialize particles $x_0^{(i)} \sim p(x_0)$ for $i = 1, \ldots, N$
for $t = 1, \ldots, T$ **do**
  // Update filter given new demonstration or test at t
  for $i = 1, \ldots, N$ **do**
    Compute weight $\check{w}_t^{(i)} = \check{w}_{t-1}^{(i)} \cdot p(x_t^{(i)}|y_t)$ ▷ Update Particle Positions and Weights
  end for
  if $\sum_{j=1}^{N} \check{w}_t^{(j)} < \check{w}_{threshold}$ **then**
    Perform a particle filter reset
  end if
  Normalize weights $\tilde{w}_t^{(i)} = \frac{\check{w}_t^{(i)}}{\sum_{j=1}^{N} \check{w}_t^{(j)}}$
  Compute effective sample size $n_{\text{eff}} = \frac{1}{\sum_{i=1}^{N} (\tilde{w}_t^{(i)})^2}$
  if $n_{\text{eff}} < N_{\text{threshold}}$ **then**
    Resample $x_t^{(i)}$ with probabilities $\tilde{w}_t^{(i)}$ using KLD resampling
  end if
end for

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. *Leveraging both constraints and Bayesian updates gracefully affords both reasoning over KCs (e.g. bounds on the cost of mud) and probabilistic modeling of human understanding that is amenable to iterative updates during teaching and testing.* The particle filter routines outlined in the following paragraphs come together in Alg. 4.

Updating Particle Positions and Weights

Assume a set of particles, defined by their positions and associated weights $\{\mathbf{x}_t, \check{\mathbf{w}}_t\}$. Without loss of generality, assume that a demonstration or test response is provided at each time step t. Each demonstration generates multiple constraints by comparing the demonstration against possible counterfactuals and each incorrectly answered test will generate a single constraint by comparing the true test answer against the incorrect answer, both through Eq. 4.4. Each constraint generated via a demonstration or a test response is a half-space constraint, with one side being *consistent* with the demonstration or test response and the other side being *inconsistent*.

Each constraint y_t can then be translated into a probability distribution $p(x_t|y_t)$ that can be used to update the weights of each particle (Fig. 6.3). We propose

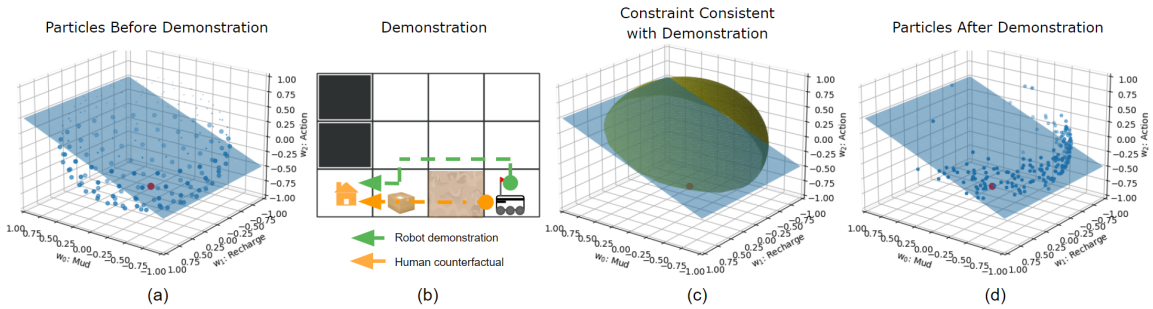

Figure 6.3: Example sequence on how a demonstration updates a particle filter model of human beliefs. The robot reward function is shown as a red dot, and the constraint consistent with the demonstration is shown in all plots for reference. **(a)** Particles before demonstration (prior). **(b)** Demonstration shown to human. **(c)** The constraint (Eq. 4.4) consistent with the demonstration that conveys that mud must be at least twice as costly as an action, visualized with the uniform distribution portion of the custom distribution (Fig. 6.4) used to update particle weights. **(d)** Particles after demonstration (posterior).

a custom probability distribution $p(x_t|y_t)$ that translates each constraint into a combination of a uniform distribution that aligns with the consistent half-space of the constraint and a von Mises-Fisher distribution (a generalization of the Gaussian distribution on a sphere) [18] whose mean direction aligns with the inconsistent half-space (Fig. 6.4). The uniform distribution asserts that any particle lying on the consistent half-space is equally valid for that demonstration, whereas the Von-Mises Fisher distribution asserts that a particle is exponentially less likely to have generated that demonstration as you move away from the constraint. The resulting probability density function (pdf) of the custom distribution is given in Eq. 6.1, with the normalizing constant c_1 that ensures that the pdf sums to 1 (Eq. 6.2), and the scaling constant c_2 that matches the probability of the Von-Mises Fisher distribution to that of the uniform distribution at meeting point of the two distributions (Eq. 6.3). Though the custom distribution naturally generalizes to high dimensions, the pdf in Eqs. 6.1 – 6.3 is specified for the 2-sphere for simplicity. In our experiments, we set the concentration parameter κ of the Von-Mises Fisher distribution to be 2, which we empirically observed as providing the right signal-to-noise ratio during the particle weight updates ($\kappa = 0$ corresponds to the uniform distribution and the distribution becomes more peaked around the mean, and less noisy, as κ increases).

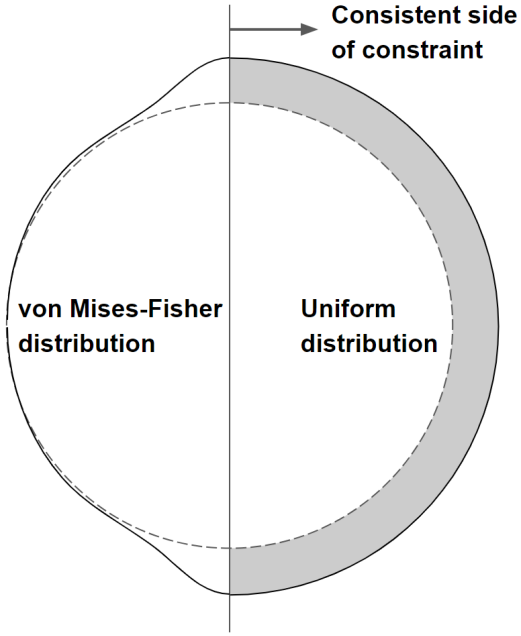

Figure 6.4: Cross-section of the spherical probability density function used to update particle weights given a constraint generated from a demonstration.

$$f(x,\mu,\kappa) = \begin{cases} \frac{1}{2\pi c_1}, & \mu^\top x \geq 0 \\ \frac{c_2 \kappa e^{\kappa \mu^\top x}}{2c_1\pi(e^\kappa - e^{-\kappa})}, & \mu^\top x < 0 \end{cases} \tag{6.1}$$

$$c_1 = \frac{1}{c_2 \int_0^\pi \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} \frac{\kappa e^{\kappa cos(\theta)\cdot sin(\phi)} sin(\phi)}{2\pi(e^\kappa - e^{-\kappa})} d\theta d\phi + 0.5} \tag{6.2}$$

$$c_2 = \frac{1}{4\pi f(y,\mu,\kappa)}, \forall y \ s.t. \ \mu^\top y = 0 \tag{6.3}$$

Sampling Human Beliefs

Given a running particle filter model, we may sample human beliefs in order to do counterfactual reasoning. We first run systematic resampling on a copy of the

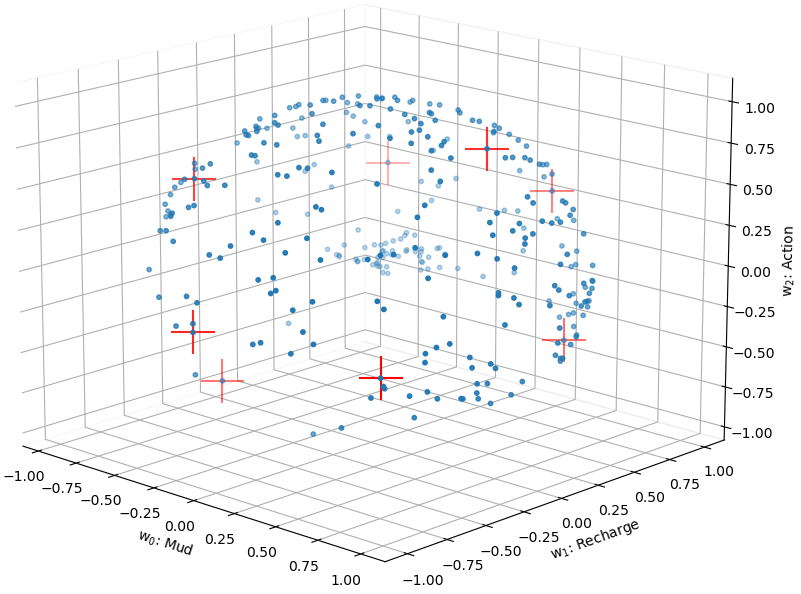

Figure 6.5: Human counterfactuals are generated by sampling beliefs from the particle filter model. As nearby particles are likely to generate similar counterfactuals, we rely on the 2-approximation algorithm for the k-center problem to sample k beliefs (marked by red crosses) that are spread out.

particles to downselect to a candidate set (not affecting the original particle filter), accounting for the differences in the weights of the particles and favoring those that are higher weighted. We then rely on the 2-approximation algorithm [33] to greedily select k distributed samples such that the maximum distance from any particle in the candidate set to one of the k samples is minimized. The algorithm iteratively picks the particle with the largest distance to the already selected samples as the next sample; this heuristic ensures that the maximum distance from any particle to any of the selected samples is never worse than twice the optimal. As nearby particles are likely to generate similar counterfactuals, we wish to sample beliefs that are approximately spread out. And as we do not require an optimal coverage of the belief space, this algorithm provides an efficient sampling method. For our experiments, we set k to be 25. To support real-time counterfactual reasoning, we also sampled 2500 beliefs from the surface of the 2-sphere (the space of possible human beliefs regarding the agent's reward function) for which we pre-computed the optimal policy. Each particle in the particle filter was then mapped to the closest pre-computed belief during experiments toward efficient selection of additional demonstrations and tests.

Resampling and Resetting the Particle Filter

We address common challenges to using particle filters in practice. Sample degeneracy occurs when successive updates to the weights of the particles cause only a few particles to have high weight and the particle filter fails to model regions of interest in the posterior with sufficient detail [58]. Furthermore, the number of particles (i.e. sample size) should adapt to the complexity of the distribution being modeled. To address both concerns, we rely on KLD-resampling [57] to obtain the sample size that bounds the Kullback-Leibler (KL) divergence between the sample-based maximum likelihood estimate and the true posterior distribution, and simultaneously rely on systematic resampling to concentrate the sampling near regions of high probability. Finally, measures to combat sample degeneracy can actually cause sample impoverishment, where the particle filter is too concentrated and not amenable to future shifts in the posterior. Thus we resample only when the effective sample size (a measure of sample degeneracy) drops below a predefined threshold and also add Gaussian noise when resampling the particles [58]. This limited resampling balances the risk of running into sample degeneracy or sample impoverishment, which are at opposite extremes.

The particle filter may converge, then suddenly obtain new information that is heavily inconsistent with the current distribution (Fig. 6.6). In this case, the filter will struggle to update, as none or very few of the particle weights would be increased to shift the distribution in a meaningful way. We thus implement particle filter resetting, taking inspiration from sensor resetting localization [54] that combats the kidnapped robot problem, where the robot has been moved without being told and must reinitialize its localization. Our particle filter resetting triggers when the weights of the particles, after accounting for $p(x_t|y_t)$ and before weight normalization (line 10 of Alg. 4), drop below a threshold. We uniformly distribute a set number of particles into the consistent half-space (Fig. 6.6b) and again rely on KLD-resampling [57] to obtain the number of particles that will bound the KL divergence between the posterior distribution following the reset and its sample-based maximum likelihood estimate. We then sample that number of particles directly from the custom distribution corresponding to $p(x_t|y_t)$ and add it to the particle filter.

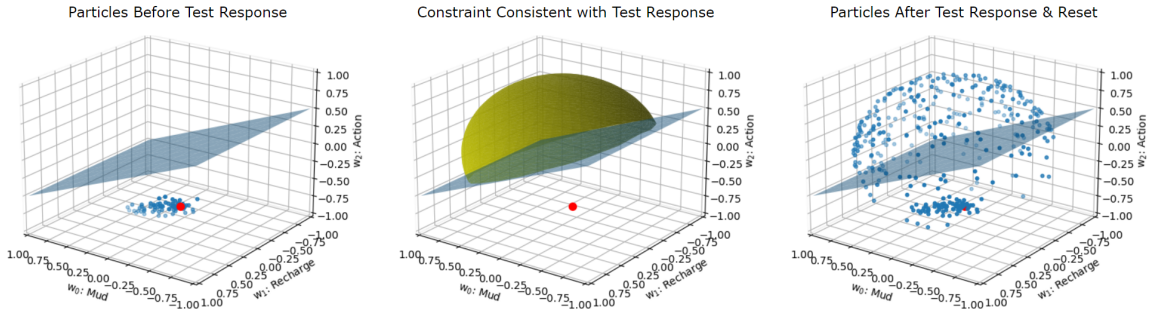

Figure 6.6: When a test response is heavily inconsistent with the current model of human beliefs, we perform a reset (Section 6.1). The constraint consistent with the test response is shown in all panels, with the consistent side shown with the uniform distribution as a yellow dome in the center panel. The robot reward function is shown as a red dot.

Algorithm 5 Closed-loop Teaching Framework

```
Group related knowledge components (KC) into batches using counterfactual scaffolding
for each batch of KCs (i.e. lesson) do
    Provide initial demonstrations and diagnostic tests
    Evaluate diagnostic test responses
    if diagnostic test responses are incorrect then
        Provide corrective feedback, remedial demo, and a remedial test
        Evaluate remedial test response
        while remedial test response is incorrect do
            Provide corrective feedback and provide new remedial test
            Evaluate remedial test response
        end while
    end if
end for
```

Closed-loop Teaching

With a particle filter model of human beliefs that is amenable to iterative updates via demonstrations and tests, we now formulate a closed-loop teaching framework for conveying a robot's reward function to a human. As we walk through the framework that is visualized in Fig. 6.7, we highlight the principles from the education literature that guide the design. A sample rollout of a teaching sequence is shown in Fig. 6.8, which may serve as a visual correspondence to the overview of the framework that is provided in Alg. 5.

We first leverage feature and counterfactual scaffolding from our prior work 5 to

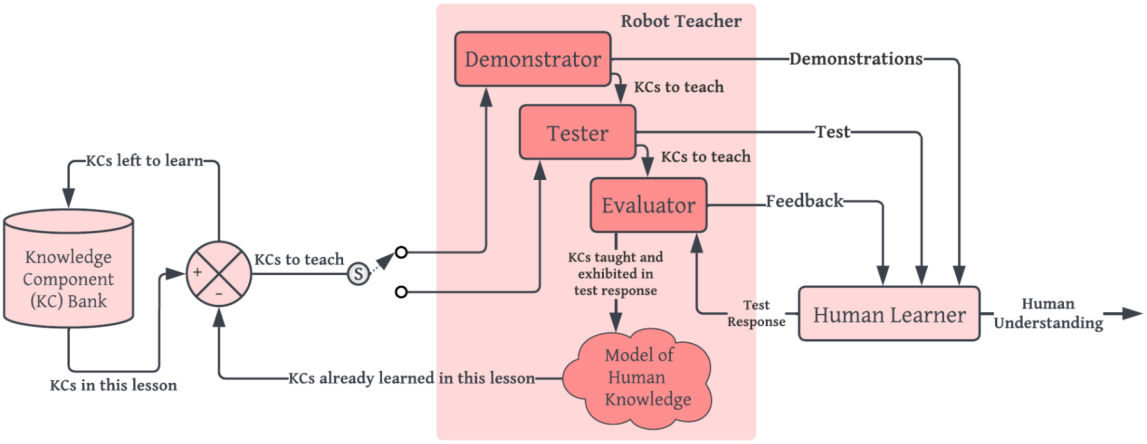

Figure 6.7: Proposed closed-loop teaching framework. A group of related knowledge components (KCs) are passed to the robot teacher as a lesson. The `demonstrator` generates demonstrations that convey the KCs, the `tester` provides test(s), and the `evaluator` analyzes the test response(s), provides feedback on its correctness, and updates the model of human knowledge. If the human fails to learn a KC through two rounds of demonstrations and tests, the switch (labeled 'S') flips such that only tests and feedback are provided until understanding of the remaining KCs is demonstrated through correct responses.

select KCs that incrementally increase in information across an increasing subset of features (e.g. mud vs action cost, recharging vs action cost, then tradeoffs between all three). This set of KCs guides the machine teaching-selection of the *curriculum* of demonstrations that can be used to teach the robot reward function to a human, where each demonstration is selected to convey a single KC whenever possible.

We begin the loop by taking a single batch of related KCs that define a *lesson* (e.g. the upper and lower bound on mud cost) and providing it to the `demonstrator` (Fig. 6.7) to select demonstrations from the curriculum that convey these KCs that belong in this lesson. Specifically, we utilize counterfactual reasoning [53] to select demonstrations that are informative with respect to the counterfactuals likely considered by the human. We simultaneously leverage the educational principles of the *ZPD* [91] to provide a sequence of demonstrations that provide information incrementally, i.e. demonstrations that convey one new KC (i.e. constraint) at a time (such as a lower-bound then an upper-bound on mud cost).

After all of the demonstrations in this lesson have been provided, the `tester`

selects *diagnostic tests* that will verify whether the human has learned all of the KCs in the lesson. These diagnostic tests optimize for visual *dissimilarity* from the teaching demonstrations and visual *complexity* (i.e. increasing distracting visual clutter) [53] to challenge the learner.

For each diagnostic test that is answered incorrectly, the `evaluator` will provide immediate *feedback* to the human on how their answer differed from the correct one, inspired by findings that immediate feedback on errors leads to better learning outcomes [44]. And for each diagnostic test that is answered incorrectly, a remedial demonstration that most closely conveys the missed KC with visual simplicity [51] will be provided to focus on the concept being taught, along with a remedial test with visual complexity to challenge the learner in demonstrating the missed KC. We note that this missed KC is determined by comparing the human's test answer with the optimal test answer; while it may or may not be the same as one of the KCs originally contained in the lesson, it addresses the human's current misunderstanding. If the human also gets the remedial test wrong, the switch in Fig. 6.7 (labeled 'S') flips and the `tester` and `evaluator` will continue to provide only visually dissimilar and complex remedial tests with corresponding feedback (but no additional demonstrations) until the human shows understanding of each iteration's missed KC. This is motivated by the testing effect [79], which supports using tests not only for assessment but also for teaching and increasing learning outcomes. Note that for each demonstration provided or test response received throughout this learning process, we update the particle filter model of the human's beliefs. And we utilize the particle filter model to consider the counterfactuals the human is likely to consider for each potential remedial demonstration or remedial test in order to select the one that will best convey or test the missed KC for the human. Once all of the missed KCs for this lesson have been demonstrated via correct remedial test responses, a fresh batch of KCs (i.e. a new lesson) is pulled from the KC bank and the switch flips upward to provide demonstrations again.

Alternatively, if all diagnostic tests in this lesson had been correctly answered initially, a fresh batch of KCs would have been pulled from the KC bank to begin the next lesson directly without remedial instruction.

To illustrate the utility of our closed-loop teaching framework, consider a robot that makes its reward function and subsequent policy more transparent to a human using

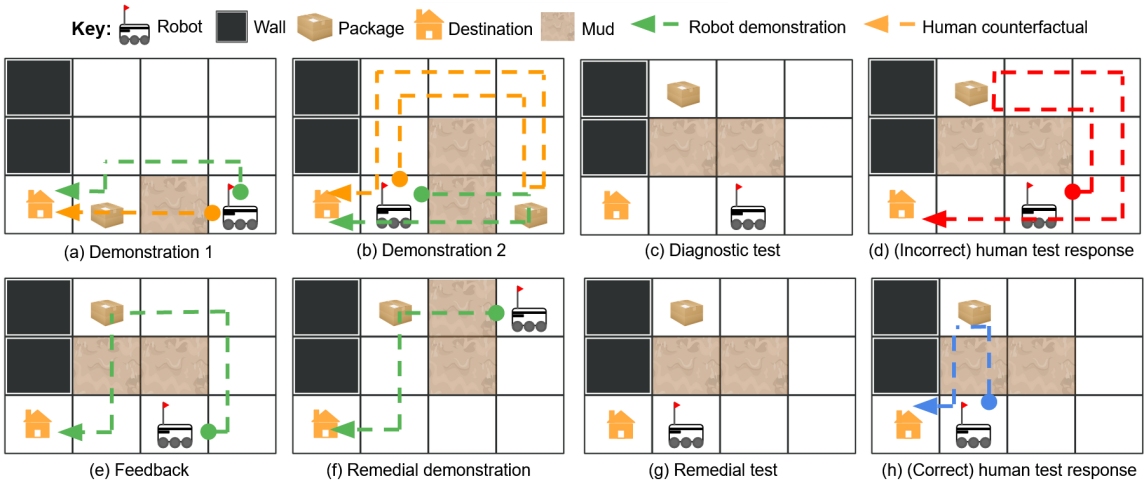

Figure 6.8: Sample teaching sequence for a batch of KCs on mud cost. **(a)** First demonstration (green) contrasts with a counterfactual alternative likely considered by a human (orange), which conveys that mud is costly. **(b)** Second demonstration lowerbounds mud cost. **(c)** Human is asked to predict the robot's behavior in a test. **(d)** Incorrect response suggests that the demonstration was not understood. **(e)** Human is given the correct response as feedback. **(f)** Remedial demonstration is provided to target the misunderstanding. **(g)** Human is given a remedial test. **(h)** Correct answer suggests understanding.

demonstrations, tests, and feedback accordingly (Fig. 6.8). The robot's objective is to deliver a package to the destination, whose reward function balances traveling through difficult terrain, like mud, and reducing the overall number of actions it takes (i.e. steps). To convey its reward function, the robot first provides a human with the demonstration in Fig. 6.8a. Because the robot takes a two-action detour to avoid the mud instead of going through it, the human may infer that the robot associates mud with a negative reward.

The robot considers what to demonstrate next to convey more information regarding its reward function. Importantly, it knows that the human likely knows that mud is costly from the first demonstration, but does not know *how* costly. For instance, the human may counterfactually believe that the robot would take a four-action detour when faced with two mud patches (Fig. 6.8b). However, the robot knows that its ratio of mud to action reward is -3 to -1 and that consequently, it would simply go through the mud in Fig. 6.8b to maximize its reward. Seeing how its direct path meaningfully differs from the human's likely detouring counterfactual (i.e.

an alternative, potentially suboptimal behavior), the robot considers this to be an informative next demonstration to provide – it aims for the ZPD as it provides a meaningful yet incremental update to the human belief through an additional KC that upper-bounds the cost of mud.

The robot then follows the two demonstrations with a diagnostic test that simultaneously challenges the human to apply their learned knowledge and reveals whether the robot's current model of the human's beliefs must be corrected (Fig. 6.8c). If the human answers incorrectly, the robot may provide feedback, a remedial demonstration, and then a sequence of remedial tests and feedback until the human demonstrates concept mastery, inspired by the testing effect (Fig. 6.8e-h). Importantly, the robot continues to update its model of the human's beliefs according to the test answer and throughout the remedial interactions to consider the right counterfactuals when estimating the information gain of demonstrations that could be provided next.

When all lessons have been taught, the human's knowledge can be evaluated via their performance on a held-out set of tests in which they predict the policy in previously unseen scenarios.

6.2 User Study

We ran an online user study[1] exploring whether our proposed closed-loop teaching method improves the transparency of a robot's policy to a human. The study involved participants learning about the robot policy in two domains through a combination of demonstrations, tests, and feedback and predicting the robot's behavior in new test environments.

Study Design

The within-subject variable was *domain*, which consisted of the following two conditions. In the *delivery* domain, the robot is penalized for moving out of mud and rewarded for recharging. In the *skateboard* domain, the robot is rewarded each time it moves with the skateboard (e.g. riding is efficient) or traverses through a

[1]Code for the methods, domains, and relevant hyper-parameters used in this study can be found at https://github.com/SUCCESS-MURI/closed_loop_teaching_study.

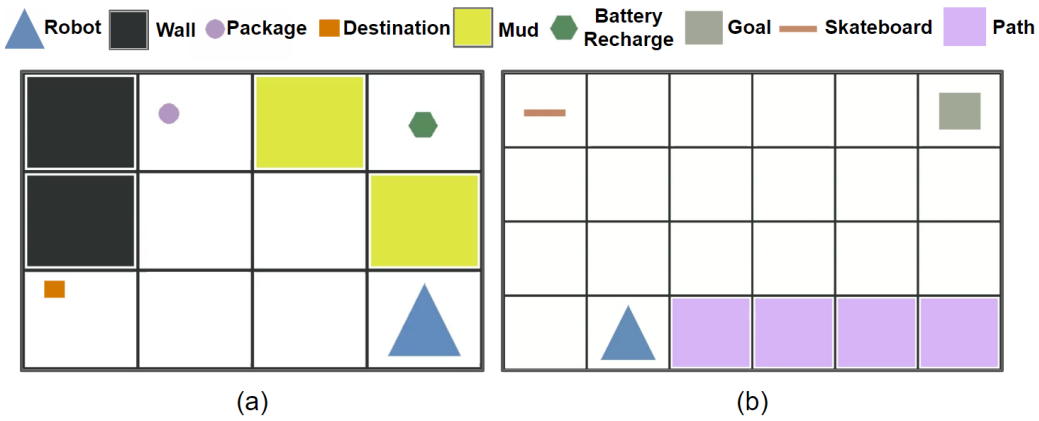

Figure 6.9: Two domains designed for user study, **(a)** delivery, **(b)** skateboard. The semantics of the objects were hidden using arbitrary shapes and colors.

designated path (see Fig. 6.9). Thus each domain consists of two unique reward features and one shared feature that penalizes each action. The *skateboard* domain was explicitly designed to be more challenging than the *delivery* domain (confirmed through pilot studies), as the value of the skateboard depends both on the distance to the skateboard and subsequent distance to the goal. The order of the domains in the study was counterbalanced.

The between-subjects variable was *feedback loop* with the following three conditions. *Open* feedback loop followed our prior work [53] in selecting a set of informative demonstrations a priori using counterfactual reasoning that incrementally decreased in BEC area, one KC at a time. *Partial* feedback loop additionally provided a diagnostic test after each lesson and provided feedback as necessary, while the *full* feedback loop additionally provided a remedial demonstration and remedial tests until the KC in question was correctly applied in a remedial test. For a fair comparison, each condition showed the same median number of demonstrations and tests (11 for delivery and 22 for skateboard)[2].

The user study consisted of two trials, with each trial comprising a teaching portion and a testing portion in one domain. During teaching, participants were first explicitly informed of the reward features of the domain. Then they inferred the corresponding

[2]Participants in the *full* feedback loop condition could receive a variable number of remedial demonstrations and tests, so we ran this condition first to determine the median number of demonstrations and tests for the other two conditions.

reward weights by watching demonstrations and perhaps diagnostic tests, corrective feedback, and further remedial instruction depending on their assigned feedback loop condition. For every interaction, participants responded to whether it improved their understanding of the policy. At the end of the teaching session, participants were asked to rate their level of attention, the usability of their assigned teaching condition, and their understanding of the policy. For testing, participants were tasked with predicting the robot's optimal trajectory in six unseen test environments in random order, which were selected according to prior work [53] to comprise two low, medium, and high difficulty environments each.

Hypotheses

H1: (a) The test responses will be best for *full* feedback loop, then *partial*, then *open*. (b) *Delivery* will result in better test responses over *skateboard*.

H2: (a) Focused attention and perceived usability will be highest for *full* feedback loop, then *partial*, then *open*. (b) *Delivery* will result in higher focused attention and perceived usability over *skateboard*.

H3: (a) Improvement ratings will be highest for *full* feedback loop, then *partial*, then *open*. (b) *Delivery* will result in higher improvement ratings over *skateboard*.

H4: (a) Understanding ratings will be highest for *full* feedback loop, then *partial*, then *open*. (b) *Delivery* will result in higher understanding ratings over *skateboard*.

Measures

The following objective and subjective measures were recorded to evaluate the aforementioned hypotheses. The Likert scales corresponding to M2 and M4 were provided after the teaching portion but before the testing portion, and Likert scales corresponding to M3 were provided after each demonstration and test in the teaching portion.

M1. Test response: The reward of the human's test response, measuring the human's ability to predict the policy.

M2. Focused attention and perceived usability: We adapted the User Engagement Scale short form [69] to ask three questions targeting focused attention:

- "I was fully engaged with learning the game strategy."
- "The time I spent learning the game strategy passed by quickly."
- "I was absorbed in this experience."

and three questions targeting perceived usability:

- "I felt frustrated while learning the game strategy."
- "I found learning the game strategy confusing."
- "Learning the game strategy was taxing."

each answered with a 5-point Likert scale.

M3. Improvement: "Did this interaction improve your understanding of the game strategy [robot policy]?", answered with a 5-point Likert scale.

M4. Understanding: "Do you feel that you now understand the game strategy?", answered with a 5-point Likert scale.

Test response (M1) relates to the 'optimal response' measure in the previous user studies, which measures the human's ability to predict the policy. Improvement (M3) relates to the (demonstration) 'informativeness rating' measure in previous user studies, but the former is asked after each interaction (as opposed to once at the end of the teaching session) and further focuses on increases with respect to the participant's current understanding. The perceived usability (M2) mirrors the 'mental effort' rating in previous user studies but uses a validated scale, and attention (M2) aims to measure the ability of the closed-loop teaching framework to engage the learner. Finally, understanding (M4) aims to measure the participant's perceived subjective level of understanding and contrast it with their objective ability to predict agent behavior (M1).

6.3 Results

We collected data from 206 participants using Prolific. Participants were roughly 70% male, 28% female, 1% non-binary, and 1% preferred not to disclose, and ages varied from 18 to 67 (M = 32.49, SD = 11.15). The recruitment process and study was approved by Carnegie Mellon University's Institutional Review Board. In the *full* feedback loop condition, we removed data from one participant who did not miss any

diagnostic tests during teaching (thus did not see any remedial instruction in either domain), and one outlier participant whose total number of interactions exceeded 3 standard deviations of the mean number of interactions in this condition (as repeated failures of similar remedial tests suggested lack of attention). This left 68 participants in each between-subjects condition.

H1: We considered analyzing test responses in two ways: binary scores measuring the optimality of human test responses, and regret measuring the degree of suboptimality of human test responses (i.e. the difference between rewards of human and optimal test responses). The former analysis was coarse and did not yield any significant results, so we opted for the latter which provides a finer resolution. We also considered normalizing the regret by the optimal test response but decided against it to prevent identical mistakes from being penalized differently based on different trajectory lengths and optimal rewards (please find further elaboration in Section 6.4). A two-way mixed ANOVA indicated a significant effect of feedback loop on regret ($F(2, 201) = 3.65, p = .028$)[3]. Tukey analyses revealed that *full* ($M = 0.24$) had 43% lower regret over *open* ($M = 0.42$, $p = .027$), with *partial* sitting in between with no significant difference to either ($M = 0.29$, Fig. 6.10a). The ANOVA also indicated a significant effect of domain on regret ($F(1, 201) = 50.75, p \leq .001$), where a t-test revealed a significant difference between the regret between *delivery* ($M = 0.18$) and *skateboard* ($M = 0.45$), $t(406) = -5.792, p < .001$.

The ANOVA also indicated an interaction effect ($F(2, 201) = 3.45, p = .03$) between feedback loop and domain. In the *skateboard* domain, Tukey analyses revealed that *full* ($M = 0.33$) had significantly lower regret over *open* ($M = 0.62$, $p = .014$),

H1a is partially supported. Though the regret for *partial* sat in between *full* and *open* as expected (being an intermediary between those two levels), it was not significantly different from either. However, *full* did indeed significantly outperform *open*. The interaction effect reveals that the difference between *full* and *open* on regret is driven by the *skateboard* domain, suggesting perhaps that the benefit of the proposed fully closed-loop teaching scheme is greater for more challenging domains.

[3]Though one participant had only 11/12 test responses recorded, we note that this does not significantly impact the reported results as responses were averaged for each participant and 2447 total test responses were recorded.

H1b is supported. Delivery resulted in a significantly lower regret over *skateboard*, as expected.

H2: We ran a Cronbach's alpha to verify the reliability of the corresponding Likert scales for measuring focused attention and perceived usability. For focused attention, we observed that the value rose from $\alpha = 0.58$ to $\alpha = 0.65$ without the second item (which asked for a response to the question "The time I spent learning the game strategy passed by quickly." on a 5-point scale) and we remove this item from the analysis accordingly. For perceived usability, we keep all items for the analysis below as removing any of them did not increase the $\alpha = 0.86$ that was obtained using all items.

A two-way mixed ANOVA did not find a significant effect of feedback loop ($F(2, 201) = 1.56, p = 0.21$), nor domain ($F(1, 201) = 0.38, p = .54$) on focused attention, nor an interaction effect between feedback loop and domain on focused attention ($F(2, 201) = 1.90, p = .15$). A two-way mixed ANOVA found a significant effect of domain on perceived usability ($F(1, 201) = 85.77, p < .001$). A t-test revealed a significant difference in the perceived usability ratings of *delivery* ($M = 3.57$) and *skateboard* ($M = 2.89$), $t(406) = 6.562, p < .001$. Finally, a two-way mixed ANOVA also found an interaction effect between feedback loop and domain on perceived usability ($F(2, 201) = 6.17, p = .003$), where Tukey revealed a significant difference between *partial* ($M = 2.64$) and *open* ($M = 3.21$) for *skateboard* ($p = .006$, Fig. 6.10b). A main effect of feedback loop on perceived usability was not found ($F(2, 201) = 2.06, p = .13$).

H2a is not supported. Though no main effects were found for feedback loop on focused attention or perceived usability, the interactions effects on the *skateboard* domain reveal that *partial* feedback loop is less usable than *open* loop. *H2b is partially supported.* The trend of the domain differences continues with *delivery* yielding significantly higher ratings of perceived usability over *skateboard*, though no difference was found between the domains for focused attention.

H3: As participants gave an *improvement* rating for each interaction (e.g. demonstration, feedback, etc), a mean is more descriptive than a median for each participant and for each domain and we use parametric analyses accordingly. A two-way mixed ANOVA indicated a significant effect of domain on improvement ($F(1, 201) = 32.17, p < .001$). A t-test revealed that the teaching in *delivery*

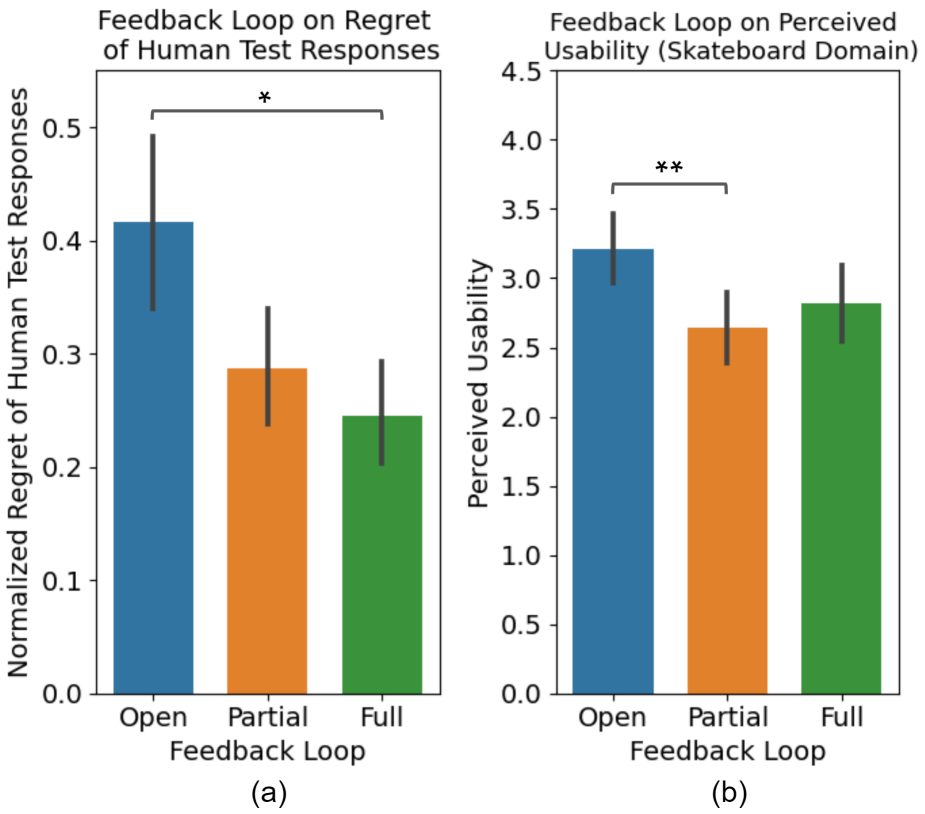

Figure 6.10: (a) *Full* closed-loop teaching yields lower regret for human tests responses than *open* across domains (lower is better). (b) *Partial* yields lower ratings on perceived usability (higher is better) than *open* in the skateboard domain. Error bars indicate 95% confidence intervals.

($M = 3.38$) was rated to yield higher improvement than in *skateboard* ($M = 3.12$), $t(406) = 3.001, p = .003$). The ANOVA did not indicate a significant effect of feedback loop ($F(2, 201) = 1.54, p = .22$) nor a significant interaction effect ($F(2, 201) = 1.23, p = .29$) between feedback loop and domain.

H3a is not supported. Feedback loop did not impact ratings of improvement. *H3b is supported.* The ratings suggest that participants learned more overall about the *delivery* domain than the *skateboard* domain.

H4: A Kruskal-Wallis H test did not reveal a statistically significant effect of feedback loop on ratings of understanding ($p = .41$). However, a Wilcoxon signed-rank test showed a statistically significant change in ratings of understanding between *delivery* and *skateboard* domains ($Z = -6.474, p < .001$). Though the median ratings

on understanding of both domains were 4, the mean for *delivery* was 3.90 and the mean for *skateboard* was 3.34.

H4a is not supported. Feedback loop did not impact ratings of understanding. *H4b is supported.* The ratings support a difference in the difficulty of the two domains.

Finally, as an exploratory measure, we asked participants at the end of each domain in the user study (having gone through the respective teaching and testing portions) to provide their best guess as to the weights of the domain's reward features. We evaluated whether the signs of each of the weights were correct as a coarse, first-pass analysis, which can be found in Table 6.1. Note that not every participant had reward weight estimates recorded due to technical difficulties in collecting this exploratory measure.

Table 6.1: Correctness of the signs of reward weight estimates from participants

	Taxi Domain		Skateboard Domain	
	Correct	Incorrect	Correct	Incorrect
Open loop	34	32	36	30
Partial closed loop	37	31	34	32
Full closed loop	40	28	37	31

6.4 Discussion

The primary hypothesis of the user study, H1a, was partially supported with *full* closed-loop teaching leading to a significantly lower regret in human test responses over *open* loop teaching. As *partial* closed-loop was explicitly designed to incorporate only a subset of *full*'s framework (i.e. diagnostic tests and feedback, but not additional remedial demonstrations or tests), it predictably led to regret that sat in between *full* and *open* without significant difference to either. Importantly, the three aforementioned conditions each provided the same median number of interactions (where each demonstration or test counts as one interaction), highlighting that the content and the interaction type matter in instruction. *Full* closed-loop teaching was designed to detect misunderstandings in human's beliefs using diagnostic tests, then address the misunderstanding with tailored remedial demonstrations and tests

until the human exhibits understanding through a correct test response. *Open* loop teaching does not provide real-time tailoring of instruction, and *partial* only provides a diagnosis of potential misunderstanding and shallow remediation through quick feedback.

Not too surprisingly, results indicated a clear difference between the two domains across all measures except focused attention (as they were designed to vary in difficulty – see Chapter 7 for a more in-depth discussion). Interestingly, there were interaction effects driven by domain. The results show that the significant improvement in objective learning outcomes from *full* closed-loop teaching over *open* comes primarily from the *skateboard* domain. However, *full* but isn't simultaneously able to significantly improve usability over *open*. Again, we see hints of the dual nature of effective learning that requires mental effort to continuously update one's knowledge (note that the usability questions in this study address a similar construct to mental effort). Indeed, one person in the *full* condition said the following in response to the open-ended question at the conclusion of the study, "Do you have any general comments or feedback on the study? Is there anything you wish [the agent] would've done to help you understand the game strategies better?"

> "I found it a little confusing. Each time I thought I understood the best strategy I was proved wrong. Nothing more [the agent] could have done except give more examples. More examples and more practice might have helped."

Full closed-loop teaching employs the counterfactual scaffolding technique from the previous chapter to explicitly select demonstrations that the human does not expect to provide maximum information. While we detect when the human has failed to successfully incorporate knowledge from counterfactual scaffolding demonstrations and remedy with remedial demonstrations and tests, these initial demonstrations can understandably be challenging to grasp. A closed-loop teaching scheme is thus critical for keeping the human learner in the zone of proximal development with intermittent testing, feedback, and target instruction.

Interestingly, we also saw another interaction effect where *partial* loop teaching is rated significantly less usable than *open* in the skateboard domain. A number of people in *partial* noted that they wanted more demonstrations to clear up confusion, e.g. saying "the strategy on the first game somewhat confused me. Maybe if

there were more demonstrations it would be easier to understand its strategy." We hypothesize that perhaps it can be frustrating to have diagnostic tests highlight gaps in understanding without providing further instruction (as in the case of *full*) or not highlight potential gaps in understanding at all and provide additional instruction instead (as in the case of *open*).

We also considered analyzing H1 using normalized regret as previously mentioned in Section 6.3. In debating whether to analyze participant test responses using regret or normalized regret, we observed a key tradeoff between the two metrics that is highlighted in Fig. 6.11. While normalizing regret by the reward of the optimal trajectory allows for a fairer comparison between tests of different domains (each with its own unique reward function), it also necessarily scales the reward of each individual error according to the reward of the optimal trajectory. For example, while one may argue that the suboptimal test responses that go through mud in Fig 6.11a and Fig 6.11b are qualitatively the same and should be penalized the same (indeed the regret for both trajectories is 0.64), the normalized regrets are different. The normalized regret for Fig 6.11a is 0.60 while the normalized regret for Fig 6.11b is only 0.43, as mistakenly going through mud comprises a smaller portion of the longer overall trajectory in Fig 6.11b. We note that there are merits to each measure and advise selecting one over the other depending on context. For instance, a 10-minute detour in a five-hour trip to a conference is negligible, but the same detour for a daily commute from the hotel to the conference that should only take 10 minutes is arguably worse and better captured by normalized regret (regret would be the same). In this thesis, we are instead interested in measuring *how much someone has learned* and thus each mistake should arguably be penalized the same, regardless of whether it is made once in a shorter trajectory or once in a longer. We thus opt for regret. Interestingly, none of the significant findings change for H1 when moving from one form of regret to the other – no new results are added nor taken away. This may be because the sizes of our domains were similar (the delivery and skateboard domains consisted of 10 and 24 grid squares respectively) and resulted in reward feature counts of the same magnitude. Furthermore, the reward feature weights were l_2-normalized such that each weight lay between 0 and 1. For domains of vastly different reward feature counts and reward weights may subsequently lead to vastly different regret and we suggest normalization for fairer comparison across domains.

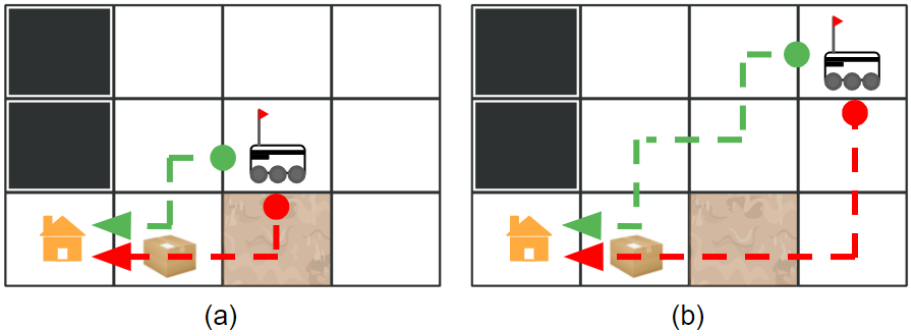

Figure 6.11: Two scenarios exemplifying the difference between regret and normalized regret, where optimal and suboptimal trajectories are shown in green and red respectively. The regret in both scenarios is 0.64, but normalized regret is 0.60 in (a) and 0.43 in (b).

Finally, the results of asking participants to guess the weights of the reward features in each domain surprised us (see Table 6.1). While there were always more, or at least as many, correct answers as incorrect answers, the number of incorrect answers was higher than expected even as we only considered the sign of the weights and not the magnitude. As we'll discuss further in Chapter 7, this further points to humans not likely performing inverse reinforcement learning (IRL) algorithmically as we model in this thesis. Furthermore, the proportion of correct answers increases from *open*, to *partial*, to *full* in order of decreasing regret for *delivery*, but not so for *skateboard*. As we alluded to earlier in Chapter 4, the more difficult and complex domain may have encouraged participants to utilize a more imitation-based learning style than IRL-based learning style.

6.5 Comparing Demonstrations with Direct Reward Explanations

While this thesis has focused so far on reward explanations in the form of demonstrations, they can take other forms. e.g. conveying direct weights of reward features, saliency maps highlighting where the agent is attending to, and reward decomposition bars that group future rewards into semantically meaningful categories [9, 82].

Interestingly, Sanneman and Shah [82] found that communicating weights of reward

features directly performed the best objectively and subjectively in their two domains (waypoints and grid world), compared with Highlights, a policy summarization technique that communicates the reward function via demonstrations from states with maximal difference between the Q-values for the best and worst actions [5].

We wondered whether direct reward explanations would also outperform our closed-loop teaching method in our domains, and also whether there would be synergy in conveying both. We thus ran an online user study exploring whether direct reward explanations improve the transparency of agent policies in the grid world domains considered in this thesis.

Study Design

Most of the details of this user study carried over from the previous user study in 6.2. The within-subject variable was again *domain*, which consisted of the same two conditions as the user study on feedback loop: delivery and skateboard.

The between-subjects variable was *explanation type* with the following three conditions.

- *Direct reward* followed the methodology of [82] and directly provided the numerical reward weights to the participant in a bar graph (we also provided the numerical values).
- *Full* implemented the full closed-loop teaching framework as described earlier in this chapter as a baseline.
- *Joint* provided both direct reward information via bar graphs and numerical values, as well as the full closed-loop teaching framework.

The user study consisted of two trials, with each trial comprising a teaching portion and a testing portion in one domain. During teaching, participants were first explicitly informed of the reward features of the domain through an informational page. In the *direct reward* or *joint* conditions, the participants were also provided the corresponding reward weights in bar graph form as well as explicit numerical values on this informational page. For these two conditions, the numerical values of the reward weights were provided on every subsequent page (e.g. alongside demonstrations and tests) to remove the confound of memory. Participants in the *direct reward* condition then moved straight from the informational page on reward weights and features

(which comprised the teaching portion) to a page of Likert items that queried their level of attention, the usability of their assigned teaching condition, and their subsequent understanding of the policy to close out their teaching portion. Participants in the *full* and *joint* conditions were provided demonstrations and perhaps diagnostic tests, corrective feedback, and further remedial instruction as necessary. For every interaction, participants also responded to whether it improved their understanding of the policy. Participants in the *full* and *joint* conditions also followed up their teaching portion by responding to Likert items that queried their level of attention, the usability of their assigned teaching condition, and their subsequent understanding of the policy to close out their teaching portion.

Following the teaching portion, participants in all conditions proceeded to the testing portion where they predicted the robot's optimal trajectory in six unseen test environments in random order, which were selected according to prior work [53] to comprise two low, medium, and high difficulty environments each.

Hypotheses

H1: (a) The test responses will be best for *joint*, then *full*, then *direct reward*. (b) *Delivery* will result in better test responses over *skateboard*.

H2: (a) Focused attention will be highest for *joint*, then *direct reward*, then *full*. Perceived usability will be highest for *direct reward*, then *joint*, then *full*. (b) *Delivery* will result in higher focused attention and perceived usability over *skateboard*.

H3: (a) Improvement ratings will be highest for *joint*, then *full* (no improvement ratings were queried for *direct reward*). (b) *Delivery* will result in higher improvement ratings over *skateboard*.

H4: (a) Understanding ratings will be highest for *joint*, then *full*, then *direct reward*. (b) *Delivery* will result in higher understanding ratings over *skateboard*.

Measures

The following objective and subjective measures were recorded to evaluate the aforementioned hypotheses (all measures are shared with the previous user study in 6.2 but are repeated here for convenience). The Likert scales corresponding to M2 and M4 were provided after the teaching portion but before the testing portion, and Likert

scales corresponding to M3 were provided after each demonstration and test in the teaching portion.

M1. Test response: The reward of the human's test response, measuring the human's ability to predict the policy.

M2. Focused attention and perceived usability: We adapted the User Engagement Scale short form [69] to ask three questions targeting focused attention:

- "I was fully engaged with learning the game strategy."
- "The time I spent learning the game strategy passed by quickly."
- "I was absorbed in this experience."

and three questions targeting perceived usability:

- "I felt frustrated while learning the game strategy."
- "I found learning the game strategy confusing."
- "Learning the game strategy was taxing."

each answered with a 5-point Likert scale.

M3. Improvement: "Did this interaction improve your understanding of the game strategy [robot policy]?", answered with a 5-point Likert scale.

M4. Understanding: "Do you feel that you now understand the game strategy?", answered with a 5-point Likert scale.

Results

We collected data from 204 participants using Prolific. Participants were roughly 72% male, 26% female, 1% non-binary, and 1% preferred not to disclose, and ages varied from 18 to 67 (M = 31.54, SD = 9.68). The recruitment process and study were approved by Carnegie Mellon University's Institutional Review Board. 68 participants were randomly assigned to each of the three between-subjects conditions and the order of the domains in the study was counterbalanced.

H1: Consistent with the previous user study, we analyze participant test responses using regret (i.e. the difference between rewards of human and optimal test response). A two-way mixed ANOVA indicated a significant effect of feedback loop on regret ($F(2, 201) = 23.72, p < .001$). Tukey analyses revealed that both *joint* ($M = 0.22$) and *full* ($M = 0.24$) had significantly lower regret compared to *direct reward* ($M =$

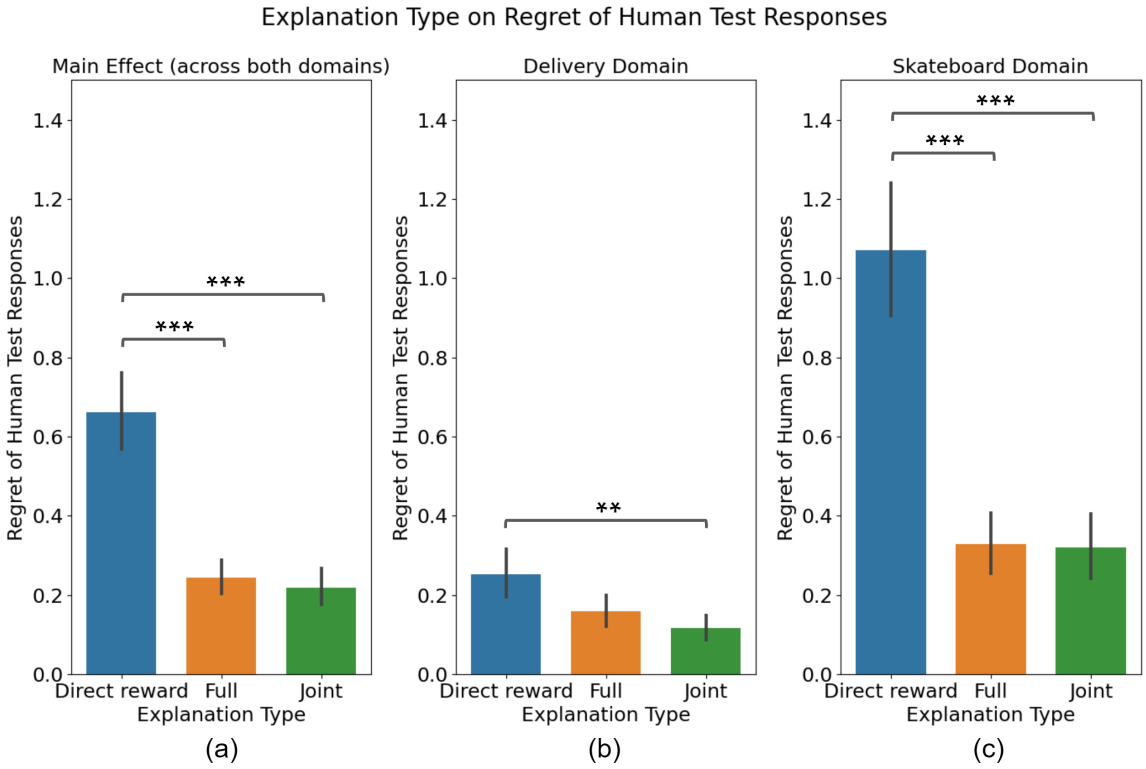

Figure 6.12: (a) *Direct reward* leads to significantly higher regret in human test responses compared to *full* and *joint*. (b-c) The gap between the regret from direct reward and the other explanation types is notably bigger in the skateboard domain than the delivery domain, where the skateboard was objectively and subjective deemed by participants to be more challenging.

0.66), with both at $p < .001$. The ANOVA also indicated a significant effect of domain on regret ($F(1, 201) = 51.62, p < .001$), where a t-test revealed a significant difference between the regret between *delivery* ($M = 0.18$) and *skateboard* ($M = 0.57$), $t(406) = -6.378, p < .001$.

Finally, the ANOVA also indicated an interaction effect ($F(2, 201) = 14.65, p < .001$) between explanation type and domain. In the *delivery* domain, Tukey revealed that *joint* ($M = 0.12$) led to significantly lower regret compared to *direct reward* ($M = 0.25$), at $p = .005$, while *full* ($M = 0.16$) trended toward significantly lower regret than *direct reward* at $p = .08$. In the *skateboard* domain, Tukey analyses revealed that both *joint* ($M = 0.32$) and *full* ($M = 0.33$) had significantly lower regret compared to *direct reward* ($M = 1.07$), with both at $p < .001$ (Fig. 6.12).

Table 6.2: Mean regret of human test responses across the five conditions of the two user studies (lower is better).

	Open loop	*Partial*	*Full* closed loop	*Joint*	*Direct reward*
Delivery	0.210	0.162	0.160	**0.118**	0.254
Skateboard	0.624	0.412	0.328	**0.320**	1.070

H1a is partially supported. While *joint* and *full* each led to significantly lower regret compared to *direct reward*, *joint* did not lead to significantly lower regret with respect to *full* as expected. An exploration of interaction effect revealed that the differences between *direct reward* and either *joint* or *full* are larger in the *skateboard* domain, again suggesting an interesting influence of domain that will be further discussed in the next section. *H1b is supported. Delivery* resulted in a significantly lower regret over *skateboard*, as expected.

A table comparing the mean regret of human test responses across the five conditions across the two user studies conducted in this chapter is found in Table 6.2. Of note are *direct reward* leading to the worst performance in both domains, and full and joint performing the best (a statistically significant difference was not found between these two conditions). We provide similar tables comparing the results of other measures across the two user studies in subsequent analyses.

H2: We ran a Cronbach's alpha to verify the reliability of the corresponding Likert scales for measuring focused attention and perceived usability. For focused attention, we observed that the value again rose from $\alpha = 0.61$ to $\alpha = 0.67$ without the second item (which asked for a response to the question "The time I spent learning the game strategy passed by quickly." on a 5-point scale) and we remove this item from the analysis accordingly. For perceived usability, we keep all items for the analysis below as removing any of them did not significantly increase the $\alpha = 0.85$ that was obtained using all items.

A two-way mixed ANOVA found a significant effect of feedback loop ($F(2, 201) = 5.63, p = 0.004$) on focused attention. Tukey analyses revealed that *joint* ($M = 4.46$) led to significantly higher ratings over *full* ($M = 4.22$) and *direct reward* ($M = 4.16$), at $p = .033$ and $p = .005$ respectively. While the ANOVA reported a significant effect of domain on focused attention ($F(1, 201) = 5.11, p = .02$), a post hoc t-test revealed

Table 6.3: Mean focused attention rating across the five conditions of the two user studies (higher is better).

	Open loop	*Partial*	*Full* closed loop	*Joint*	*Direct reward*
Delivery	4.279	4.412	4.272	**4.522**	4.169
Skateboard	4.309	4.360	4.169	**4.397**	4.147

that the difference between focused attention ratings in *delivery* ($M = 4.32$) and *skateboard* ($M = 4.24$) was not significant, $t(406) = 1.349, p = .18$. The ANOVA did not find an interaction effect between explanation type and domain on focused attention ($F(2, 201) = 0.72, p = 0.49$).

A two-way mixed ANOVA also found a significant main effect of explanation type on perceived usability ($F(2, 201) = 8.30, p < .001$), where Tukey revealed that *direct reward* ($M = 3.76$) led to significantly higher ratings over *joint* ($M = 3.22$) and *full* ($M = 3.25$), at $p = .001$ and $p = .002$ respectively. The ANOVA also revealed a significant effect of domain on perceived usability ($F(1, 201) = 78.51, p < .001$), and a post hoc t-test revealed that a significant difference between ratings in *delivery* ($M = 3.70$) and *skateboard* ($M = 3.13$), $t(406) = 5.641, p < .001$. Finally, the ANOVA also found an interaction effect between explanation type and domain on perceived usability ($F(2, 201) = 12.36, p < .001$), where Tukey revealed that *direct reward* ($M = 3.70$) led to significantly higher ratings over *joint* ($M = 2.87$) and *full* ($M = 2.82$), at $p < .001$ for both, only for *skateboard* (no significant differences were found for the *delivery* domain – see Fig. 6.13).

H2a is partially supported. Joint resulted in significantly higher focused attention ratings over *full* and *direct reward* as expected. There was no difference in focused attention ratings between *full* and *direct reward*, however. *Direct reward* resulted in significantly higher perceived usability ratings over *joint* and *full* as expected, but there was no difference in perceived usability ratings between *joint* and *full*. Interestingly, post hoc analyses of the interaction effect between domain and usability find that the significant main effects are entirely driven by *skateboard*. *H2b is partially supported.* The trend of the domain differences continues with *delivery* yielding significantly higher ratings of perceived usability over *skateboard*, though no difference was found between the domains for focused attention.

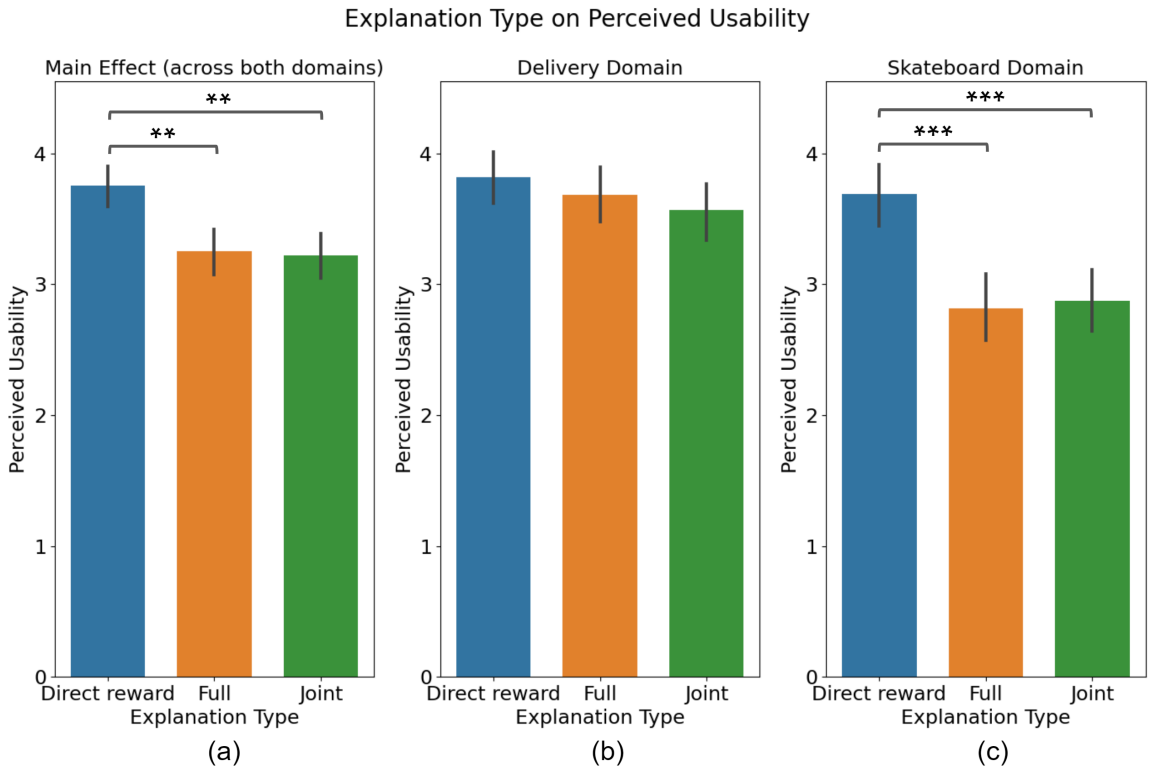

Figure 6.13: (a) *Direct reward* leads to significantly higher ratings of perceived usability compared to *full* and *joint*. (b-c) The main effect is mostly driven by the skateboard domain.

H3: As participants gave an *improvement* rating for each interaction in *joint* and *full* (e.g. demonstration, feedback, etc), a mean is more descriptive than a median for each participant and for each domain and we again use parametric analyses accordingly[4]. A two-way mixed ANOVA indicated a significant effect of explanation type on improvement ($F(1, 132) = 11.85, p = .001$). A t-test revealed that *joint* ($M = 3.77$) yielded significantly higher ratings on improvement over full ($M = 3.23$), $t(134) = 3.613, p = 0.001$. The ANOVA also indicated a significant effect of domain on improvement ($F(1, 132) = 18.23, p < .001$). A t-test revealed that the teaching in *delivery* ($M = 3.64$) was rated to yield higher improvement than in *skateboard* ($M = 3.43$), $t(270) = 1.900, p = .058$). The ANOVA did not indicate a significant interaction effect between explanation type and domain ($F(1, 134) = 3.36, p = .06$).

[4]Due to technical challenges, the improvement ratings of two out of 68 participants in the *joint* condition were not recorded.

Table 6.4: Mean perceived usability rating across the five conditions of the two user studies (higher is better).

	Open loop	*Partial*	*Full* closed loop	*Joint*	*Direct reward*
Delivery	3.525	3.485	3.686	3.569	**3.819**
Skateboard	3.211	2.637	2.819	2.873	**3.691**

Table 6.5: Mean improvement rating across the five conditions of the two user studies (higher is better).

	Open loop	*Partial*	*Full* closed loop	*Joint*	*Direct reward*
Delivery	3.430	3.269	3.440	**3.848**	N/A
Skateboard	3.270	2.953	3.125	**3.729**	N/A

H3a is supported. As expected, *joint* lead to higher ratings on improvement over *full*. *H3b is supported.* The ratings also suggest that participants learned more overall about the *delivery* domain than the *skateboard* domain.

H4: A Kruskal-Wallis H test revealed that *full* ($M = 3.65$) yielded significantly lower ratings of understanding compared to *joint* ($M = 4.19$) as well as *direct reward* ($M = 4.21$), at $p < .001$ for both. A Wilcoxon signed-rank test also showed a statistically significant change in ratings of understanding between *delivery* and *skateboard* domains ($Z = -4.83, p < .001$). Though the median ratings on understanding of both domains were 4, the mean for *delivery* was 4.17 and the mean for *skateboard* was 3.87.

H4a partially supported. While ratings on understanding were higher for *joint* over *full* as expected, ratings on understanding were also higher for *direct reward* over *full*. *H4b is supported.* The ratings on understanding were higher in *delivery* than *skateboard* as expected.

Table 6.6: Mean understanding rating across the five conditions of the two user studies (higher is better).

	Open loop	*Partial*	*Full* closed loop	*Joint*	*Direct reward*
Delivery	3.809	3.882	4.015	**4.353**	4.147
Skateboard	3.589	3.147	3.294	4.029	**4.279**

Discussion

We first observe that the best reward explanation method is likely domain-dependent, and we specifically hypothesize that conveying numerical reward weights alone is increasingly insufficient as an explanation methodology as domain complexity increases. Not only does *direct reward* lead to significantly higher regret than *joint* and *full*, the gap is larger in *skateboard* over *delivery* – where we consider former domain more complex than the latter. Domain complexity is difficult to define, but some factors to consider may be the number of reward features [82], degree of familiarity for the human [47, 76], the size of the domain, and the degree of interaction between reward features and the subtleties of the trade-offs amongst reward features afforded by the reward weights and the (resolution of) the domain. We provide further commentary on these factors in 7.1.

Second, we observe a potential synergy between different explanation types where they can help reinforce each other's information. While we found no difference in regret, *joint* has significantly higher ratings on improvement and attention than *full*. Interestingly, a few qualitative quotes from the *direct reward* or *full* conditions suggest that participants wanted the information that was not purview to them. In response to the open-ended question at the conclusion of the study, "Do you have any general comments or feedback on the study? Is there anything you wish [the agent] would've done to help you understand the game strategies better?", two participants in the *direct reward* condition replied:

> "a demonstration instead of written rules might have helped a bit more"

> "Maybe an example puzzle with optimal moves demonstrated"

indicating a desire for demonstrations as well. And one participant in the *full* condition replied:

> "If [the agent] told me the implication of moving into yellow or purple boxes, it would have helped me a lot."

indicating a desire for direct information regarding the effect of various reward features (e.g. perhaps in the form of numerical weights). And people who received both numerical weights and demonstrations in the *joint* condition, replied:

> "This demonstration reinforced to me the importance of obtaining the

> orange rectangle as moving with it results in a + 0.825% energy change."
>
> "I already knew to avoid the yellow square, and would have moved the same way as demonstrated "

which reveal the dual possibility that different explanation types may be helpfully reinforcing or unhelpfully redundant. To the latter point, one must be mindful of cognitive overload when providing too much information at once, which can lead to a worse understanding of model decision making [74].

And finally, we observe that high self-reports of understanding do not always translate to corresponding performance. While *direct reward* led to significantly higher levels of reported understanding[5] over *full*, as well as significantly high ratings on usability over *full*, *direct reward* also led to significantly worse objective performance. Our results raise the possibility that people may believe that their knowledge is sufficient and may terminate learning early (especially since effective learning often requires significant mental effort as prior results in this thesis have shown), when the subsequent testing portion will likely reveal significant gaps in their knowledge. All in all, our results point to a need for a closed-loop, AI-driven teaching that provides tests and additional instruction as needed to verify the human's true understanding. And though our results support AI-driven teaching, Qian and Unhelkar [76] found that a hybrid strategy where participants could choose between AI-selected and user-requested examples outperformed only AI-selected examples and was also subjectively preferred. However, we note that they fixed the teaching budget, and an interesting direction for future work may be in exploring how to balance AI-driven and user-driven learning when given a flexible teaching budget setting (e.g. the human may be feeling unmotivated and wish to terminate learning after a few insufficient examples).

Understandability is a multi-faceted concept that can be difficult to measure in practice. While the accuracy of a person's prediction of an agent's behavior is arguably the most common sub-measure of understandability (e.g. [9, 35, 48, 82]), other measures include coding responses to an open-ended question regarding agent decision making (e.g. [9, 82]), agent preference elicitation and feature sub-selection [82], and verification of agent response and counterfactual reasoning [48]. Our user

[5]We note that we queried participants for their perceived understanding right after the conclusion of the teaching portion of the user study, and before the testing portion. We hypothesize that perceptions on understanding may have changed when queried after the testing portion.

studies that queried for participants' ability to predict agent behavior and our single Likert-scale item concerning understanding are incomplete measures, and we also leave how one may query and measure a human's understanding of agent decision making more fully for future work.

7 Discussion

As AI systems increasingly abound in society, it is important that their decision making is *transparent*, e.g. such that the actions taken by AI are predictable and understandable to humans. Transparency is critical for not only developers in reviewing and ensuring proper AI function, but also for end-users in having calibrated expectations – preventing undertrust and disuse, or overtrust and misuse. And as humans naturally communicate and comprehend each others' policies through demonstrations, this thesis explored increasing the transparency of AI policies through demonstrations. Furthermore, as 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) [37], this collection of work modeled humans as IRL learners.

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 the standard paradigm of providing humans with demonstrations that simply maximize information gain [35, 47, 76], we crucially observe that information gain and difficulty of understanding are often correlated for humans and thus show demonstrations that balance the two in an attempt to maximize human learning. We subsequently leverage ideas and techniques from the education literature, such as teaching in the zone of proximal development (i.e. engaging at the right level of information gain and difficulty conditioned on human beliefs) by scaffolding and

providing targeted instruction in a closed-loop fashion to significantly improve human learning of agent decision making.

In Chapter 4, we confirmed through a user study that the Set Cover Optimal Teaching (SCOT) [12] demonstrations with the highest information gain for an IRL learner ironically led to the poorest learning outcomes in humans (in comparison to demonstrations with lower information gain) and also were rated least informative subjectively. Given that humans are not pure IRL learners, we aimed to scaffold demonstrations such that 1) they incrementally increased in information gain to ease humans into learning and 2) optimized for visual simplicity and similarity amongst consecutive demonstrations to highlight meaningful environmental changes that prompted qualitatively different behavior. We also hypothesized that a behavior's information gain as a demonstration during teaching could be inverted into a measure of the difficulty of predicting that (unseen) behavior in a new situation as a test. While a user study confirmed the test difficulty measure and that optimizing for visuals led to increased performance on high difficulty tests, our initial attempt at scaffolding did not have an effect. In analyzing our results, we noticed that the method for calculating a demonstration's information gain failed to account for the human's current beliefs and was likely insufficient (e.g. the same demonstration provided to the human twice in succession would be rated to provide the same information gain).

Thus in Chapter 5, we began to explicitly consider the human's beliefs during teaching and testing. We first conditioned the information gain of a demonstration on whether it would differ from the human's expectation of agent behavior (i.e. counterfactual behavior), given the human's beliefs over the agent's reward function. We similarly updated our test difficulty measure as the overlap between the beliefs of the agent's potential reward function in the human's mind and the set of reward functions consistent with agent behavior, such that smaller overlap correlated with higher difficulty. A user study once again confirmed our test difficulty measure, and also provided a more nuanced result for scaffolding teaching demonstrations that incrementally increased in information gain as measured by counterfactual reasoning. This scaffolding increased performance for tests targeting later demonstrated concepts but decreased for early demonstrated concepts, suggesting that we had perhaps moved too quickly without obtaining real-time feedback.

We addressed the shortcomings of such an open-loop teaching paradigm in Chapter

6. Though one may generate a curriculum of informative demonstrations *a priori*, student learning may deviate *in situ*. Thus, we leverage tests within a teaching loop to not only assess human understanding in real time and provide targeted instruction via feedback and additional demonstrations, but also to teach according to the *testing effect* in the education literature. A user study showed that our closed-loop teaching yielded better learning outcomes over baseline open-loop teaching, yielding lower regret in human test responses by 43% and also yielding higher usability scores for one of the two considered domains.

Demonstrations are only one means of improving the transparency of agent policies and, inspired by results by Sanneman and Shah [82], we also considered how directly conveying the agent's underlying reward weights would fare in our domains, both as a standalone method as well as in conjunction with our closed-loop teaching via demonstrations. In contrast to their findings, we found that directly conveying the agent's reward weights yielded significantly worse human test responses, although it led to reports of high understanding and usability as an explanation type. However, providing both reward weights and demonstrations provided synergy that allowed for high objective and subjective outcomes nearly across the board, highlighting that different explanation types can provide complementary information that can augment one another.

Finally, we have defined transparency as understandability and predictability, borrowing from work by Endsley [23], a leading expert in human situational awareness involving intelligent agents. One can easily imagine how understandability and predictability can be correlated to one another – high understanding could improve predictability through forward simulation, and high predictability could improve understanding through the generation of data that could support model building. However, we see that even participants who were given the reward weights explicitly in the *direct reward* condition, and they theoretically had all of the information needed to understand the agent's decision making (and subsequently rated their understanding very highly), they struggled to predict the agent's actions in the tests. Another study similarly found that providing full details behind an 8-feature linear regression model of apartment selling prices led participants to be "worst at simulating the model's predictions, followed the model's predictions less, and made less accurate predictions of the apartments' selling prices compared to participants

assigned to the other primary experimental conditions," including a simplified 2-feature linear regression model [74]. The authors of the study hypothesize that this unexpected result (arguably capturing high understanding but low predictability) likely stemmed from participants experiencing information overload, and we emphasize that understanding the relationship between understandability and predictability as a function of domain and task is an important direction for future work.

7.1 Considering Interaction Effects and Confounds by Domain

The thesis explored the central research question of how to teach through demonstrations in four total grid world domains. Each domain contained a reward function that was composed of two unique reward features (e.g. moving out of mud and recharging) with their respective reward weights and one shared action reward feature that penalized each action. Though semantics were always hidden through abstract shapes and the size of each domain was also comparable (ranging from 10 to 25 grid squares), the intraclass coefficients (ICC) values we computed in Chapters 4 and 5 to evaluate the consistency of each participant's performance across domains were low, which indicated that performance in fact varied considerably across domains for each participant. However, we averaged over the domains for simplicity of analysis in these chapters as there were already more critical between and within-subjects variables.

When we analyzed domain in Chapter 6 as a within-subjects variable, we confirmed that it had a very strong impact on the results. The skateboard and delivery domains significantly differed on almost all measures, with the skateboard domain rated as more challenging, both objectively and subjectively. In both user studies in this chapter, differences in learning outcomes between conditions were largely driven by the skateboard domain and differences in reported usability among conditions were significant in one of these two domains. These results first suggest that different domains call for different teaching methods as we see our contributions having a greater effect in the more complex skateboard domain. However, we note that learning often comes at the cost of more mental effort and simpler domains like delivery may not require as challenging of a teaching regimen. These results secondly suggest that

we must take care to analyze results from user studies based on domain. Significant findings may not transfer across domains. Furthermore, exploratory post-hoc analyses of some of the key findings in Chapters 4 and 5 are also domain-dependent, only holding for one of the three possible domains and providing further evidence for a domain-specific analysis.

Finally, characterizing a domain toward inferring when results will transfer is an open but important question. Domain complexity is difficult to define and Sanneman and Shah [82] offer a definition for the related concept of reward complexity as the number of features that comprise the reward function. Though the number of reward features is a reasonable starting point for domain complexity, our observations suggest that one must also consider the degree of interaction between features and also the subtleties of the tradeoffs that result from the reward features. Though we do not test this in our user study, another consideration when considering domain complexity could be the degree of familiarity. As Qian and Unhelkar [76] note, their navigation domain had a much smaller state space of 400 over other domains that had a state space of 3,200 and 80,000 but it was the most challenging for their participants due to some of the navigation robot's less intuitive movements.

In our study, *delivery* and *skateboard* each had three reward features but both objective and subjective results strongly indicated that the latter domain was more challenging for participants. First, *delivery* supports more "local" planning around mud patches and batteries whereas *skateboard* requires more "global" planning that considers the distance to the skateboard and the subsequent distance to the goal to determine whether it is worth detouring to pick up the skateboard on the way to the goal. In this, we'd argue that the *skateboard* domain has an implicit dependence between the action and skateboard reward features that must be carefully considered in advance before selecting between two paths. Furthermore, the grid size of the *delivery* domain was smaller than *skateboard* and the reward weights were more coarse (the reward weights for *delivery* were -3, 3.5, and 1 for moving out of mud, picking up the battery, and for each action respectively whereas the reward weights for *skateboard* were 0.825, 0.4875, -1 for moving with the skateboard, moving on the path, and for each action respectively). This allowed for more subtle trade-offs to be made in the skateboard domain such that the difference in reward between a trajectory that detoured to pick up the skateboard first, a trajectory that detoured to go on the

path instead, and a trajectory that went straight toward the goal could differ by only fractional amounts.

In summary, for complex domains with a high number of reward features, a high degree of interaction amongst the features, and subtle trade-offs, we see a proportionally higher benefit of teaching via demonstrations. Future studies may also consider how to explicitly highlight such trade-offs not only in demonstrations but explicitly in natural language using domain-level concepts, which was found to lead to higher user understanding and confidence in understanding of agent decision making [87].

Finally, we note that the domains underwent a number of iterations as the research progressed as we sought to remove potential confounds that would influence the effect that we wanted to measure, which was ultimately the effect of various teaching conditions on human learning outcomes. While we intentionally used abstract geometric shapes as opposed to semantically meaningful images that would likely bias learning with a prior (e.g. a battery would be good, and mud would be bad for a ground delivery robot), we found two potential additional confounds as we conducted pilots and user studies. First, we removed the ability for the agent to exit the domain if it deemed completing the task to be more costly than exiting (a single action). While exiting is algorithmically similar to any other action from the point of view of the agent, we found that people especially struggled to reason over this action. As explained in more detail in Chapter 4, we hypothesize that people naturally had a bias toward figuring out how to complete the task presented to them rather than if they should complete the task, as the latter seemingly requires reasoning over both how and if they should complete the task. We also changed the skateboard reward feature and weights between Chapters 5 and 6 to perhaps be more intuitive such that action weight was always present and skateboard and path weights were positive. Originally when the skateboard, path, and actions weights were mutually exclusive and each negative, we anecdotally found that people struggled even more to infer the correct signs of the reward weights (see Table 6.1). Rane et al. [77] show that a learner's ability to correctly infer the demonstrator's reward function from behaviors critically depends on the learner's ability to correctly model the demonstrator's transition function. We similarly posit that reward inference depends on the learner's ability to model the demonstrator's reward features (which is an assumption that we've

made throughout this thesis). To conclude, we posit that we ought to be mindful of cognitive biases that influence human inference over agent behaviors, such as priors over the unlikelihood of special actions like exiting. Second, future work may explore the extent to which the transparency of a policy depends on the learner's knowledge of the agent's reward features, especially for deep neural network-based representations.

7.2 Limitations & Future work

We now discuss the limitations of this thesis' approach and findings, as well as corresponding avenues for future work.

Dimensionality and form of reward function and domain

In this work, we focused on teaching a low-dimensional reward that specifically took the form of a weighted linear combination of reward features. And though the reward features can theoretically be nonlinear with respect to states and actions and capture arbitrarily complex reward functions, the methods proposed in this thesis were designed for domains with reward functions that cleanly decompose into a set of disentangled, semantic features.

Furthermore, we constrained ourselves to grid worlds of limited size and diversity (e.g. the number and locations of possible mud and path patches in the delivery and skateboard domains were decided a priori) that could support exhaustive enumeration.

Future work: One obvious extension is to consider teaching reward functions that are a weighted linear combination of many more than three reward features. In this scenario, we posit that learning abstractions that group related reward features into lower-dimensional reward features whose corresponding weights can be communicated will be key, e.g. as humans struggle to reason about statistical relationships beyond three variables at once [29]. Sanneman and Shah [82] found such abstractions to be a good compromise between conveying information while limiting the workload required of the human to understand (see [83] for additional exploration of these results). Recent work has also begun leveraging such abstractions, or often referred to concepts, to increase the interpretability of policies learned through RL [16, 101]. However, these methods require the human to hand-specify the concepts.

Automatically distilling high-dimensional reward features into low-dimensional and semantically meaningful concepts and selecting demonstrations that convey both the concepts as well as the weighting will be an important direction moving forward.

Furthermore, there are many other forms that the agent's reward function can take, such as logical conjunctions of atomic features that allow limited non-linearity (e.g. interaction effects) [14, 55], Gaussian Processes (GP) reward functions that allow for more expressive non-linearity [56], and deep representations that allow for both expressive non-linearity and faster query times than GPs [97]. Future work may explore how to convey reward functions of various forms using demonstrations. For instance, for a GP-based reward function that maps feature values to rewards, demonstrations could perhaps be chosen heuristically as those whose feature values correspond to those of inducing points that approximate the GP.

Finally, continuous domains or real-world domains may not afford an exhaustive enumeration of all possible domain instances from which to select possible demonstrations. In such cases, candidate domain instances that may support the desired knowledge component to be shown through a demonstration may need to be generated on the fly. Similar in spirit to goal recognition design [41], which aims to find a domain instance that forces an agent to reveal its objective as early as possible, this real-time enumeration of domain instances may potentially be formulated as a search problem.

Models of Human Learning and Abilities

In this thesis, we assumed that humans learn from demonstrations by inferring the underlying reward function using reasoning akin to inverse reinforcement learning (IRL). However, we found evidence in the objective results of Chapters 4 and 5, as well as in the coding of participant quotes in Chapter 4 that people may sometimes be learning from demonstrations using reasoning akin to imitation learning (IL). Furthermore, we also assumed that humans will be able to reconstruct the optimal policy given a reward function using reasoning akin to planning. However, we observed in the follow-up user study in Chapter 6 that people failed to perfectly predict the agent's behavior in tests even when they were explicitly provided the agent's reward features and weights.

Finally, Chapter 6 began to introduce some personalized instruction via teaching and testing in a closed-loop manner. Though the framework was designed with principles from the education literature in mind, it still largely adopts a one-size-fits-all approach without any real-time adaptation on key parameters (e.g. the concentration parameter κ of the Von Mises-Fisher distribution) nor based on the human's current affective state (i.e. their mood or emotions).

Future work: Though IRL and IL are both accepted styles of human learning from demonstrations [32], there are a number of possible algorithms that support both styles [68] and it is not always obvious which style or algorithm would best model human learning in a given situation. It is possible that people switch between IRL and IL-style reasoning (e.g. depending on the familiarity of the domain [47], which can even change as a function of the number of demonstrations seen – see Section 4.5), or perhaps there is yet another style of learning from demonstrations that humans employ. Findings by Lage et al. [47] additionally suggest that human learning of the agent's policy can increase if the agent correctly models the human learning style (e.g. IRL vs IL) when generating demonstrations. Determining when humans employ IRL or IL, and identifying other styles of human learning from demonstrations will be interesting future endeavors.

The closed-loop teaching framework depends on key parameters such as κ of the Von Mises-Fisher distribution which models how much the human learns from each demonstration (the higher the κ, the less likely any belief over the agent's reward function that is inconsistent with the demonstration will survive in the particle filter, such that $\kappa = 0$ corresponds to learning nothing), and k, the number of beliefs that are sampled from the running human model to forward simulate human expectations on the agent's behavior when estimating the information gain of a demonstration. These values were hard-coded, and future work may optimize these values on a population or individual level based on historical data.

Finally, while we have focused primarily on the cognitive domain of learning (e.g. selecting demonstrations that will convey a desired knowledge component while belonging to the ZPD), there is increased recognition that educational technologies must also target the affective domain of learning as well [72]. Positive affect has been shown to increase flexibility in thinking, integration of ideas, intrinsic motivation, etc [36]. Kaushik and Simmons [40] have shown that the affective behavior of a robot

teacher can influence both subjective ratings (e.g. perceived difficulty of a task) and objective learning outcomes by a human leaner, and future work should similarly consider monitoring and influencing the affective state of the human learner toward a better learning experience.

Modality of Explanations

We largely restricted ourselves to increasing the transparency of agent policies through demonstrations in this thesis. However, this is just one form that policy and reward explanations can take.

Future work: We saw in the follow-up study in Chapter 6 that direct reward explanations integrated nicely with demonstrations to yield high objective and subjective outcomes nearly across the board. This highlights the potential synergies that can arise from employing complementary explanation techniques; e.g. global *policy-level* techniques that convey an understanding of an agent's overall behavior through representative examples can be combined with with local *feature importance* techniques that highlight the contextual factors the influence an agent's single decision [64].

Finally, language is another common modality for teaching that shares strengths and weaknesses that are complementary to that of demonstrations (e.g. see [20, 21] for work on explaining agent decision making to humans using language). While language has the ability to convey complex, generalizable concepts more effectively than demonstrations, the efficacy of language is heavily dependent on shared abstraction between the teacher and the student (e.g. what a rook is in the statement "In chess, rooks move along rows and columns.") [88]. Furthermore, language can suffer from ambiguity (e.g. which can arise from stylistic differences in speech or uncertain phrases such as 'cold weather' – how cold is cold?) and may struggle to convey certain physical concepts such as spatial movement, color, etc. While demonstrations are more grounded, they require the learner to infer the underlying rules or concepts, some of which may be difficult to demonstrate exhaustively (e.g. it would be inefficient to demonstrate all the possible ways that the rook can move on a chess board). Recent work has begun exploring leveraging the complementary strengths of both language and demonstrations for humans to teach agents [59, 100], which we posit will also

be effective for conversely for agents to teach their policies and reward functions to humans.

8 Conclusion

We conclude with three insights from our thesis on what makes for informative demonstrations that increase the transparency of agent decision making.

First, informative demonstrations differ just enough from the human's current expectations 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. One way that we subsequently operationalized the zone of proximal development in our work was constraining each new demonstration to add one new knowledge component/constraint (e.g. going through mud is twice as costly as an action for a delivery robot) at a time. Interestingly, Miller [65] highlights that human explanations more broadly are inherently contrastive with respect to a specific counterfactual case, "presented relative to the explainer's beliefs about the explainee's beliefs." And Ehsan et al. [22] similarly notes that an "explanation is only explanatory if it can be consumed by the recipient." In the same vein, an effective demonstration must be grounded in the learner's beliefs (and their subsequent expectations) so that it is informative.

Second, informative demonstrations illuminate trade-offs (e.g. how many actions are you willing to take to detour around mud?) that are inherent in the agent's reward function. It is not always obvious how an agent's numerical reward function (even if given explicitly to a human) will translate into behavior, and we see that demonstrations that highlight the bounds of the trade-off (e.g. detour with two actions around a single patch of mud but not detouring with four actions around two patches) are effective. We also observe that such bounds can be very subtle; for

instance, changing the starting position of the agent by one square, all else being equal, can radically change behavior (e.g. riding the skateboard rather than going along the path). Thus we often opted for visual similarity amongst consecutive demonstrations to highlight the difference, though we also observed in our first study that visual similarity can lead to confusion if people do not recall a difference between the consecutive demonstrations and do not have access to both. Finally, we posit that explicitly highlighting trade-offs via demonstrations will become increasingly important for complex agents and systems that have to negotiate competing interests from different stakeholders in a fair manner (e.g. see work by Zhang et al. [102] that utilizes visualizations and preference elicitation to help model designers navigate the trade-offs in the objectives of different stakeholders).

Third, we see that methods for generating informative demonstrations can have different impacts in different domains. In this thesis, we see that even grid world domains comprised of state spaces of similar size and reward functions that are each comprised of three reward features can significantly differ in their complexity. People's performance across our various domains always differed significantly, and we see a proportionally higher benefit of teaching via demonstrations in more complex domains with more subtle tradeoffs and a higher degree of interaction amongst reward features. However, we also observe that informative demonstrations that increase learning gains often go hand in hand with increased mental effort for humans, which may not be necessary for less complex domains. Echoing the broader consensus in the explainable AI literature that there is no one-size-fits-all explainability method, the best teaching method is likely domain-dependent.

To close, fluent co-existence and interaction between humans and intelligent agents is contingent on the transparency of agent decision making to humans. A powerful way to communicate decision making is through through demonstrations, and we operationalize key ideas and principles from the extensive literature on human education to design agents that can convey their decision making to humans through informative demonstrations. And as effective pedagogy is often more multi-modal and multi-faceted beyond just demonstrations, e.g. using not only visuals but also speech to engage both the cognitive and affective aspects of the learner, this thesis is one contribution toward a more holistic explainability framework that will ensure that agent decision making is transparent to humans.

9 Appendix

9.1 Qualitative Responses Regarding Learning Style

Participants optionally responded to the following two questions throughout the two user studies from Chapter 4: "Feel free to explain any of your selections above if you wish:" (asked in conjunction with prompts for ratings of informativeness, mental effort, and puzzlement of demonstrations in each domain, i.e. up to three times) and "Do you have any comments or feedback on the study?" (asked after the completion of the full study, i.e. once). Thus, each participant could provide up to four responses.

While both questions were open-ended, many participants provided responses that provided insight into how they performed inference over the optimal behavior in new situations. Thus, the lead author pulled a subset of the responses to be coded that either demonstrated an attempt at understanding a specific aspect of a domain's reward structure (e.g. mud/yellow squares yielding negative reward), deducing the corresponding optimal behavior (e.g. avoid mud/yellow squares if possible), or meta-level comments on the inference performed through out the user study (e.g. seeing the user study as a "guessing game trying to figure out reward values and such..."). Other comments such as rote recalls of particular training demonstrations (which reveals data used to perform inference, but not the inference mechanism itself), imprecise remarks of confusion (perhaps over an aspect of a domain with no allusion to how it may affect the optimal behavior), and overall impressions on user study were not included in the coding set.

Comments in the coding set were independently coded by the lead author and a

second coder uninvolved in the study as resembling inverse reinforcement learning (IRL), imitation-learning (IL), or as 'unclear'. All responses to the two questions (including those unrelated to inference) can be found in the study data that is available in the following repository: https://github.com/SUCCESS-MURI/psiturk-machine-teaching. Please note that references to 'Chip' in the responses below are to the agent that behaved optimally in each domain.

Table 9.1: Coding qualitative participant responses with learning styles (User study 1)

Participant ID	Coder 1	Coder 2	Response
1	Unclear	IRL	I'm not sure at this stage why the robot would choose to go to one ring over another.
1	Unclear	IRL	The study required a certain amount of inference from me rather than following explicit instructions.
34	Unclear	IRL	I don't believe this video is as informative as the other ones because I think it should clarify the following situation: if Chip has to move the same amount of 'houses' to go to one of the circles, which one would he go to? Because we've only seen Chip going into the green circle but is that because the green is 'better' than the gray or did it choose the green circle because it was the nearer circle?
35	IRL	IRL	I'm still not sure how any of these affect point values or such

continues on next page

35	IRL	IRL	This was a really interesting survey, I like the aspect of it as some sort of a guessing game trying to find out reward values and such, thank you for this opportunity!
37	Unclear	Unclear	I'm honestly really puzzled by these games, I hope the next page will explain the scoring.
56	IL	IL	Goal seems to be to get to the grey square and avoid everything else
56	IL	IRL	Not sure if the goal is the nearest ring or the green ring
59	IL	IL	The primary 'mental effort' was in memorising the patterns of each level/stage and matching the optimal movements for them.
59	IRL	IRL	The role of the yellow squares in affecting my score was somewhat confusing in these demonstrations.
81	IL	Unclear	I did it mostly by intuition after analyzing the puzzles for a brief moment.
81	Unclear	Unclear	After completing some of the puzzles I realized it was better to probably exit some of them.
81	Unclear	Unclear	After some examples I feel like I'm understanding way better these puzzles.
98	Unclear	Unclear	Slowly i [sic] understand the game more and more

continues on next page

98	Unclear	Unclear	It took me a bit to understand how it works, but as soon i [sic] got it, it was a great game
151	IL	IL	I think there is no reason to pick up the bar if its [sic] not on the way
151	IRL	IL	Deliver the circle is priority i [sic] guess

Table 9.2: Coding qualitative participant responses with learning styles (User study 2)

Participant ID	Coder 1	Coder 2	Response
7	IRL	IRL	I think going to the square earns & moving with the rectangle. I *think* moving without the rectangle loses points...
7	IRL	IRL	I think the both rings are rewarding (green>grey) but moving is negative.
7	Unclear	IL	Deliver circle good, yellow squares bad
8	Unclear	IL	I couldn't understand in which case it was better to pick up the rectangle.
18	Unclear	Unclear	Confused a bit about which is the best ring to go to in some of these examples
20	Unclear	Unclear	The videos were moderately informative but did not explain rules at all, so I have to depend on my own interpretation which may, or may not, be correct. But that's the stated purpose I think.

continues on next page

20	IRL	IRL	I did not truly understood why green circle is preferred (worth more points?) but gray one is acceptable as well sometimes (getting to green would be too costly and getting to gray would make less profit but still better than quitting?)
20	IL	IL	I did not understand the rule regarding yellow tiles. It seems they should be avoided, but not always. Interesting...
21	IRL	IRL	not sure whether i [sic] get taxed going into the yellow squares
29	IRL	IRL	I wonder for one of the instances where the orange rectangle was very few moves away (i.e., 3 or fewer) and would be with Chip for all the remaining moves until reaching the gray square how the game points would play out. Actions taken with the orange rectangle and actions without taking the orange rectangle both affect the score, but I am not sure how (which is positive or negative).
29	IRL	IRL	Sometimes Chip grabbed the circle in more moves than necessary to retrieve it and bring it back to purple square. For another, when choosing to grab the circle, Chip moved onto white squares instead of yellow squares. Moving into a yellow square would be an action that affects the score, but in what way? This demonstration would imply negatively.

continues on next page

47	IRL	IRL	I was unsure why chip decided to exit or decided to choose x or y path. It was pretty confusing. In one instance though with the board covered in yellow, I assumed chip would end up with a pretty negative score so chip exited. Overall though, it is confusing.
64	IRL	IRL	Well the only thing really missing is the amount of points each action does
103	Unclear	Unclear	I think it's not as informative as the first one, there are a lot more of movements, sometimes it picked the red rectangule [sic] sometimes it didn't, so I'm still trying to think when to pick and when don't pick it.
103	IL	IL	I'm kind of puzzled, do I have to take one ring to another if possible?
105	IRL	IRL	I'm not sure what the best strategy is, because I don't don't know the value of the circles
105	IRL	IRL	I think moving without the orange square take more points but i'm [sic] not sure
129	IRL	IRL	You need to make a moderate amount of mental effort to understand all the rules and outweight [sic] everything and see what is worth it or not in the game.
129	IRL	IRL	I think this left me very puzzled because it wasn't easy to differentiate the value of the rings.

continues on next page

129	IRL	IRL	This required a significant effort to understand the value of the square.
136	Unclear	IL	JUst [sic] will be hard to understand when to quit or when to pick or not the orange line but i will do my best
142	IRL	IRL	If I would be able to see the demonstrations back and forth I would eventually get there, not a specific scoreline but within limits. For instance, the last demonstration tells me that if moving is -1, then scoring must be higher than 8 since chip went for it
142	IRL	IRL	Ok, so this time around I got way better because I didn't get it the first time. Also, this puzzle is easier since there are basically only two variables. Since this is a comparison between green and gray, I did a mathematical system on paint and got the information I needed. I'm still unsure about the exact values but my calculations tell me that if moving is -1, then gray is around +6 and green is around +10. This is all based on the system I've come up with. For example, if Chip would move to green in 8 moves, that tells me that moveValue * 8 + greenValue is positive, and since I'm assuming moving is -1, then this means greenValue - 8 > 0 which means greenValue > 8.
142	IRL	IRL	a = yellow value, b = white value, c = objective value.

continues on next page

			1) a + 8b + c > 0 2) 2a + 4b + c > 0 3) 2a + 6b + c > 0 4) 6a + 2b + c < 0 5) 5a + 5b + c > 0 if b = - 1 then 1) a + c - 8 > 0 2) 2a + c - 4 > 0 3) 2a + c - 6 > 0 4) 6a + c - 2 < 0 5) 5a + c - 5 > 0 with this equations overlapped this tells me that a <= -3, c >= 20 so moving to yellow is -3, getting objective is 20, assuming white is -1
145	IRL	IRL	I was trying to attribute values to the rings but weren't able, just saw that green > grey
147	Unclear	Unclear	I'm not sure about which ring I should prioritize.
147	IRL	IRL	So, the yellow squares should be avoided if possible and they possibly remove 2 points when crossed but I'm not sure
154	IRL	Unclear	I think the green ring is better than the gray ring?

continues on next page

156	IRL	IRL	I chosen Moderately [sic] informative in first question because I am not sure if there were enough differnent [sic] possibilities shown in demonstrations to assess how many points we get for specifiifc [sic] action.
157	IRL	IRL	I think that score system should be explained right away with new “mechanic”. Yellow squares made me wonder if they’re -2 but I could only guess

Bibliography

[1] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, 2004. 3.1

[2] David Abel. simple rl: Reproducible reinforcement learning in python. In *ICLR Workshop on Reproducibility in Machine Learning*, 2019. 4.3

[3] Ali Aljaafreh and James P Lantolf. Negative feedback as regulation and second language learning in the zone of proximal development. *The modern language journal*, 78(4):465–483, 1994. 1

[4] Douglas G Altman. *Practical statistics for medical research*. CRC press, 1990. 4.5

[5] 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. 2.2.1, 4.3, 6.5

[6] Ofra Amir, Finale Doshi-Velez, and David Sarne. Summarizing agent strategies. *Autonomous Agents and Multi-Agent Systems*, Sep 2019. ISSN 1573-7454. 2.2.1

[7] Yotam Amitai and Ofra Amir. "I don't think so": Disagreement-based policy summaries for comparing agents. *arXiv preprint arXiv:2102.03064*, 2021. 2.2.1

[8] Yotam Amitai, Guy Avni, and Ofra Amir. Asq-it: Interactive explanations for reinforcement-learning agents. *arXiv preprint arXiv:2301.09941*, 2023. 2.2.1

[9] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Alan Fern, and Margaret Burnett. Explaining reinforcement learning to mere mortals: An empirical study. *arXiv preprint arXiv:1903.09708*, 2019. 6.5, 6.5

[10] Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011. 1, 3.1

[11] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009. 1, 3.1

[12] Daniel S Brown and Scott Niekum. Machine teaching for inverse reinforcement learning: Algorithms and applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 2.2.1, 4.1, 4.2, 7

[13] Maya Cakmak and Manuel Lopes. Algorithmic and human teaching of sequential decision tasks. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 1536–1542, 2012. 2.2.1

[14] Jaedeug Choi and Kee-Eung Kim. Bayesian nonparametric feature construction for inverse reinforcement learning. In *Twenty-Third International Joint Conference on Artificial Intelligence*. Citeseer, 2013. 7.2

[15] Allan Collins, John Seely Brown, and Susan E Newman. Cognitive apprenticeship: Teaching the craft of reading, writing and mathematics. *Thinking: The Journal of Philosophy for Children*, 8(1):2–10, 1988. 2.1

[16] Devleena Das, Sonia Chernova, and Been Kim. State2explanation: Concept-based explanations to benefit agent learning and user understanding. *arXiv preprint arXiv:2309.12482*, 2023. 7.2

[17] Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12):1704–1711, 2005. 3.1, 4.5

[18] Inderjit S Dhillon and Suvrit Sra. Modeling data using directional distributions. Technical report, Citeseer, 2003. 6.1

[19] Anca Dragan and Siddhartha Srinivasa. Familiarization to robot motion. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 366–373, 2014. 1

[20] Upol Ehsan, Brent Harrison, Larry Chan, and Mark O Riedl. Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 81–87, 2018. 2.2.1, 7.2

[21] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. Automated rationale generation: a technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 263–274, 2019. 2.2.1, 7.2

[22] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2021. 8

[23] Mica R Endsley. Supporting human-ai teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior*, 140:107574, 2023. 1, 7

[24] Cassie Freeman, Audrey Kittredge, Hope Wilson, and Bozena Pajak. The duolingo method for app-based teaching and learning. 2023. 1

[25] Merrilyn Goos, Peter Galbraith, and Peter Renshaw. Socially mediated metacognition: Creating collaborative zones of proximal development in small group problem solving. *Educational studies in Mathematics*, 49:193–223, 2002. 1

[26] Omer Gottesman, Joseph Futoma, Yao Liu, Sonali Parbhoo, Leo Celi, Emma Brunskill, and Finale Doshi-Velez. Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions. In *ICML*, 2020. 2.2

[27] Samuel Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. Visualizing and understanding atari agents. In *International conference on machine learning*. PMLR, 2018. 2.2

[28] Thomas L Griffiths. Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*, 2020. 1, 4.5

[29] Graeme S Halford, Rosemary Baker, Julie E McCredden, and John D Bain. How many variables can humans process? *Psychological science*, 16(1):70–76, 2005. 7.2

[30] John Hattie. *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. routledge, 2008. 2.1

[31] John Hattie and Shirley Clarke. *Visible learning: feedback*. Routledge, 2018. 6

[32] Mark K Ho and Thomas L Griffiths. Cognitive science as a source of forward and inverse models of human decisions for robotics and control. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:33–53, 2022. 7.2

[33] Dorit S Hochbaum and David B Shmoys. A best possible heuristic for the k-center problem. *Mathematics of operations research*, 1985. 6.1

[34] 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. 2.2.1

[35] Sandy H Huang, David Held, Pieter Abbeel, and Anca D Dragan. Enabling robots to communicate their objectives. *Autonomous Robots*, 43(2):309–326, 2019. 1, 2.2.1, 3.2, 4.3, 4.5, 6.1, 6.5, 7

[36] Alice M Isen. Some ways in which positive affect influences decision making and problem solving. *Handbook of emotions*, 3:548–573, 2008. 7.2

[37] Julian Jara-Ettinger. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29:105–110, 2019. 1, 3.1, 7

[38] 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. 1, 3.1

[39] Alan Jern, Christopher G Lucas, and Charles Kemp. People learn other people's preferences through inverse decision-making. *Cognition*, 168:46–64, 2017. 1, 3.1

[40] Roshni Kaushik and Reid Simmons. Affective robot behavior improves learning in a sorting game. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 436–441. IEEE, 2022. 7.2

[41] Sarah Keren, Avigdor Gal, and Erez Karpas. Goal recognition design. In *Twenty-Fourth International Conference on Automated Planning and Scheduling*, 2014. 7.2

[42] Celeste Kinginger. Defining the zone of proximal development in us foreign language education. *Applied linguistics*, 23(2):240–261, 2002. 1

[43] 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*, 2012. 6

[44] Kenneth R Koedinger, Julie L Booth, and David Klahr. Instructional complexity and the science to constrain it. *Science*, 2013. 2.1, 3.2, 6.1

[45] Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016. 4.4, 5.4

[46] Imam Kusmaryono, Widya Kusumaningsih, et al. Construction of students' mathematical knowledge in the zone of proximal development and zone of potential construction. *European Journal of Educational Research*, 10(1):341–351, 2021. 1

[47] Isaac Lage, Daphna Lifschitz, Finale Doshi-Velez, and Ofra Amir. Exploring computational user models for agent policy summarization. In *International Joint Conference on Artificial Intelligence*. 1, 3.1, 4.1, 4.5, 4.5, 6.5, 7, 7.2

[48] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Samuel J Gershman, Been Kim, and Finale Doshi-Velez. An evaluation of the human-interpretability of explanation. In *Conference on Neural Information Processing Systems (NeurIPS) Workshop on Correcting and Critiquing Trends in Machine Learning*, 2018. 6.5

[49] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977. 4.5

[50] James P Lantolf and Ali Aljaafreh. Second language learning in the zone of proximal development: A revolutionary experience. *International Journal of Educational Research*, 23(7):619–632, 1995. 1

[51] Michael S Lee, Henny Admoni, and Reid Simmons. Machine teaching for human inverse reinforcement learning. *Frontiers in Robotics and AI*, 2021. 1, 2, 5.3, 6.1

[52] Michael S. Lee, Henny Admoni, and Reid Simmons. Reasoning about counterfactuals to improve human inverse reinforcement learning, 2022. URL https://arxiv.org/abs/2203.01855. 1

[53] Michael S Lee, Henny Admoni, and Reid Simmons. Reasoning about counterfactuals to improve human inverse reinforcement learning. In *IROS*. IEEE, 2022. 6.1, 6.2, 6.5

[54] Scott Lenser and Manuela Veloso. Sensor resetting localization for poorly modelled mobile robots. In *International Conference on Robotics and Automation*, 2000. 6.1

[55] Sergey Levine, Zoran Popovic, and Vladlen Koltun. Feature construction for inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 23, 2010. 7.2

[56] Sergey Levine, Zoran Popovic, and Vladlen Koltun. Nonlinear inverse reinforcement learning with gaussian processes. *Advances in neural information processing systems*, 24, 2011. 7.2

[57] Tiancheng Li, Shudong Sun, and Tariq Pervez Sattar. Adapting sample size in particle filters through kld-resampling. *Electronics Letters*, 2013. 6.1

[58] Tiancheng Li, Shudong Sun, Tariq Pervez Sattar, and Juan Manuel Corchado. Fight sample degeneracy and impoverishment in particle filters: A review of intelligent approaches. *Expert Systems with applications*, 2014. 6.1

[59] Toby Jia-Jun Li, Marissa Radensky, Justin Jia, Kirielle Singarajah, Tom M Mitchell, and Brad A Myers. Pumice: A multi-modal agent that learns concepts and conditionals from natural language and demonstrations. In *Proceedings of the 32nd annual ACM symposium on user interface software and technology*, pages 577–589, 2019. 7.2

[60] Ji Liu and Xiaojin Zhu. The teaching dimension of linear learners. *Journal of Machine Learning Research*, 17:1–25, 2016. 2.2.1

[61] Tania Lombrozo. Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, 20(10):748–759, 2016. 2.1, 3.2, 4, 4.2

[62] Christopher G Lucas, Thomas L Griffiths, Fei Xu, Christine Fawcett, Alison Gopnik, Tamar Kushnir, Lori Markson, and Jane Hu. The child as econometrician: A rational model of preference understanding in children. *PloS one*, 9(3): e92160, 2014. 1, 3.1

[63] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012. 4.5

[64] Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. Explainable reinforcement learning: A survey and comparative review. *ACM Comput. Surv.*, aug 2023. ISSN 0360-0300. 2.2, 7.2

[65] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. 2.1, 3.2, 5.1, 5.2, 8

[66] Andrew Y Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *International Conf. on Machine Learning*, 2000. 1, 3.1, 3.1, 3.2, 4.1, 5.1

[67] Matthew L Olson, Roli Khanna, Lawrence Neal, Fuxin Li, and Weng-Keen Wong. Counterfactual state explanations for reinforcement learning agents via generative deep learning. *Artificial Intelligence*, 2021. 2.2

[68] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 7(1-2):1–179, 2018. 1, 7.2

[69] Heather L O'Brien, Paul Cairns, and Mark Hall. A practical approach to measuring user engagement with the refined user engagement scale (ues) and new ues short form. *International Journal of Human-Computer Studies*, 2018. 6.2, 6.5

[70] Stefan Palan and Christian Schitter. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018. 4.4

[71] S Paulraj and P Sumathi. A comparative study of redundant constraints identification methods in linear programming problems. *Mathematical Problems in Engineering*, 2010, 2010. 4.2, 7

[72] Rosalind W Picard, Seymour Papert, Walter Bender, Bruce Blumberg, Cynthia Breazeal, David Cavallo, Tod Machover, Mitchel Resnick, Deb Roy, and Carol Strohecker. Affective learning—a manifesto. *BT technology journal*, 22(4): 253–269, 2004. 7.2

[73] Jan L Plass, Bruce D Homer, and Charles K Kinzer. Foundations of game-based learning. *Educational psychologist*, 50(4):258–283, 2015. 1

[74] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52, 2021. 6.5, 7

[75] Erika Puiutta and Eric MSP Veith. Explainable reinforcement learning: A survey. In *International cross-domain conference for machine learning and knowledge extraction*. Springer, 2020. 2.2

[76] 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. 1, 2.2.1, 6.5, 7, 7.1

[77] Sunayana Rane, Mark Ho, Ilia Sucholutsky, and Thomas L Griffiths. Concept alignment as a prerequisite for value alignment. *arXiv preprint arXiv:2310.20059*, 2023. 7.1

[78] Brian J Reiser. Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *The Journal of the Learning sciences*, 13(3): 273–304, 2004. 2.1, 5.1, 5.5

[79] Henry L Roediger III and Jeffrey D Karpicke. The power of testing memory: Basic research and implications for educational practice. *Perspectives on psychological science*, 2006. 1, 2.1, 3.2, 6, 6.1

[80] Sandra Sampayo-Vargas, Chris J Cope, Zhen He, and Graeme J Byrne. The effectiveness of adaptive difficulty adjustments on students' motivation and learning in an educational computer game. *Computers & Education*, 69:452–462, 2013. 2.1

[81] Lindsay Sanneman and Julie Shah. Explaining reward functions to humans for better human-robot collaboration. In *AAAI Fall Symposium AI-HRI Workshop*, 2021. 2.2.1

[82] Lindsay Sanneman and Julie A Shah. An empirical study of reward explanations with human-robot interaction applications. *IEEE Robotics and Automation Letters*, 7(4):8956–8963, 2022. 6.5, 6.5, 7, 7.1, 7.2

[83] Lindsay Sanneman, Mycal Tucker, and Julie Shah. An information bottleneck characterization of the understanding-workload tradeoff. *arXiv preprint arXiv:2310.07802*, 2023. 7.2

[84] Hanan Shteingart and Yonatan Loewenstein. Reinforcement learning and human behavior. *Current Opinion in Neurobiology*, 25:93–98, 2014. 3.1

[85] Andrew Silva, Matthew Gombolay, Taylor Killian, Ivan Jimenez, and Sung-Hyun Son. Optimization methods for interpretable differentiable decision trees applied to reinforcement learning. In *Intern. conference on artificial intelligence and statistics*, 2020. 2.2

[86] Sibawu Siyepu. The zone of proximal development in the learning of mathematics. *South African Journal of Education*, 33(2):1–13, 2013. 1

[87] Roykrong Sukkerd, Reid Simmons, and David Garlan. Tradeoff-focused contrastive explanation for mdp planning. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication*, 2020. 7.1

[88] Theodore R Sumers, Mark K Ho, Robert D Hawkins, and Thomas L Griffiths.

Show or tell? exploring when (and why) teaching with language outperforms demonstration. *Cognition*, 232:105326, 2023. 7.2

[89] 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. 1

[90] Edward Vul, Noah Goodman, Thomas L Griffiths, and Joshua B Tenenbaum. One and done? optimal decisions from very few samples. *Cognitive science*, 38 (4):599–637, 2014. 3.2, 4.5

[91] Lev Semenovich Vygotsky. *Mind in society: The development of higher psychological processes*. Harvard university press, 1980. 6, 6.1

[92] Lev Semenovich Vygotsky and Michael Cole. *Mind in society: Development of higher psychological processes*. Harvard university press, 1978. 1, 2.1

[93] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8 (3-4):279–292, 1992. 4.1

[94] Lindsay Wells and Tomasz Bednarz. Explainable ai and reinforcement learning—a systematic review of current approaches and trends. *Frontiers in AI*, 2021. 2.2

[95] Joseph Jay Williams, Tania Lombrozo, and Bob Rehder. Why does explaining help learning? insight from an explanation impairment effect. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32, 2010. 2.1, 3.2, 4, 4.2

[96] David Wood, Jerome S Bruner, and Gail Ross. The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, 1976. 2.1, 3.2, 5.2

[97] Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*, 2015. 7.2

[98] Klaus Wunderlich, Peter Dayan, and Raymond J Dolan. Mapping value based planning and extensively trained choice in the human brain. *Nature neuroscience*, 15(5):786–791, 2012. 3.1

[99] Robert Mearns Yerkes, John D Dodson, et al. The relation of strength of stimulus to rapidity of habit-formation. 1908. 5.5

[100] Albert Yu and Raymond J Mooney. Using both demonstrations and language instructions to efficiently learn robotic tasks. *arXiv preprint arXiv:2210.04476*, 2022. 7.2

[101] Renos Zabounidis, Joseph Campbell, Simon Stepputtis, Dana Hughes, and Katia P Sycara. Concept learning for interpretable multi-agent reinforcement learning. In *Conference on Robot Learning*, pages 1828–1837. PMLR, 2023. 7.2

[102] Yunfeng Zhang, Rachel Bellamy, and Kush Varshney. Joint optimization of

ai fairness and utility: a human-centered approach. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 400–406, 2020. 8

[103] Xiaojin Zhu. Machine teaching: an inverse problem to machine learning and an approach toward optimal education. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 4083–4087, 2015. 2.2.1

[104] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N Rafferty. An overview of machine teaching. *arXiv preprint arXiv:1801.05927*, 2018. 2.2.1