Counterfactual Examples for Human Inverse Reinforcement Learning

Michael S. Lee, Henny Admoni, Reid Simmons

The Robotics Institute
Carnegie Mellon University
ml5, hadmoni, rsimmons@andrew.cmu.edu

Abstract

Humans naturally communicate their decision-making through demonstrations, and infer others' decision-making through reasoning that resembles inverse reinforcement learning (IRL). Though robots may also convey their decision-making to humans through demonstrations, standard IRL misrepresents human inference, often yielding ineffective demonstrations. Critically, the counterfactuals considered by standard IRL often do not match those considered by humans, and can misjudge a demonstration's informativeness to humans. This paper thus aims to more effectively convey robot policies to humans by maintaining a more accurate model of human knowledge throughout the demonstration process. We present a novel approach for evaluating a demonstration's informativeness via counterfactuals likely to be considered by humans based on their current expectations of the robot's policy. In addition, we investigate how scaffolding the number of features conveyed via demonstrations can improve informativeness. Finally, we show how to evaluate the expected difficulty for a human to predict an instance of a robot's behavior based on their belief of the robot's decision-making. We conclude by proposing an experiment that will evaluate the aforementioned methods.

Introduction

Our capacity to deploy and co-exist fluently with robots is contingent in part on our ability to understand their decision-making. An engineer certifying the navigation policy of a ground delivery robot may ask, "Does it have a calibrated understanding of the terrain types it should risk traversing as it balances efficiency and safety?" Moreover, new owners of an autonomous vacuum gauging how much of their floor to keep clear may wonder, "How much clutter will the robot tolerate in an area before it steers clear to ensure it does not get stuck?"

One important way that people communicate, comprehend, and evaluate each others' decision-making is through demonstrations. Cognitive science suggests that humans often model one another's behavior as exactly or approximately maximizing a reward function (Jern, Lucas, and Kemp 2017; Jara-Ettinger et al. 2016; Lucas et al. 2014), which they can infer through reasoning resembling inverse reinforcement learning (IRL) (Ng and Russell 2000; Jara-Ettinger 2019; Baker, Saxe, and Tenenbaum 2009, 2011). Once they know a reward function, humans are often able to deduce a behavior that (approximately) maximizes it through planning (Shteingart and Loewenstein 2014; Wunderlich, Dayan, and Dolan 2012). Thus we can often expect humans to be able to understand one another's decision-making through IRL and behavior through planning, linked by the reward function[1] underlying demonstrations. And though a robot could convey its reward function directly, a study by Sukkerd, Simmons, and Garlan suggests that humans better understand the corresponding policy if demonstrations accompany the communicated reward functions, further motivating our use of demonstrations.

Importantly, the informativeness of a demonstration can be quantified by how much information it reveals regarding the reward function. For IRL, the information inherently depends on the *counterfactuals* (i.e. alternative, suboptimal demonstrations) that are considered. Picture an agent in a delivery domain whose objective is to bring the package to the destination, and its reward is determined by how much mud it traverses and its total number of actions (i.e. steps). To convey its reward function, imagine the agent provides a human with the demonstration in Fig. 1a. Intuitively, because the agent takes a two-action detour to avoid the mud instead of going through it (a natural counterfactual), the human knows that the agent associates a negative reward with going through mud.

After providing this first demonstration, the agent considers what to demonstrate next to convey more information regarding its reward function. Given the first demonstration, a human knows mud is costly, but doesn't know *how* costly. For instance, what should the agent do if detouring around the mud takes four actions, as in Fig. 1c? In our example, suppose the ratio of mud to action reward is -3 to -1 for the agent; consequently, the agent should simply go through the mud in Fig. 1c to maximize its reward. Intuitively, this would be a very informative demonstration for the human to see, as it upper-bounds the cost of the mud by contrasting the agent's direct path against a suboptimal human counterfac-

[1] Ng and Russell (2000) suggest that "the reward function, rather than the policy, is the most succinct, robust, and transferable definition of the task."

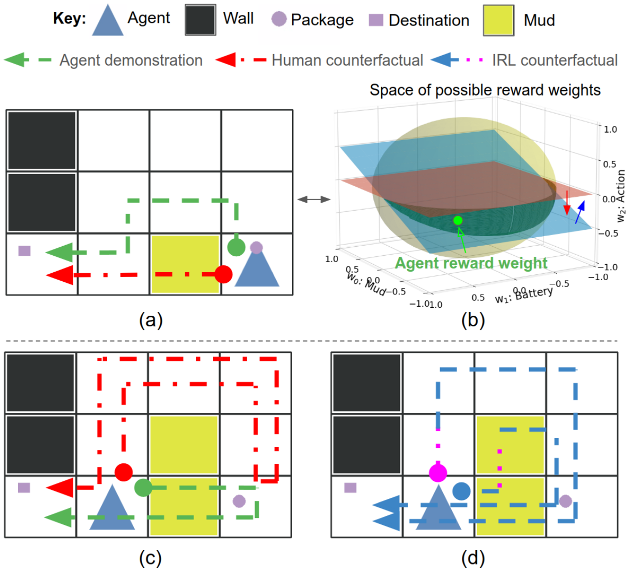

Figure 1: **(a)** An agent's optimal demonstration is shown in contrast to a suboptimal counterfactual alternative. **(b)** Inverse reinforcement learning constrains the possible reward functions underlying the agent's demonstration by comparing against the counterfactuals in (a). **(c)** An agent's optimal demonstration is shown in contrast to a counterfactual likely considered by a human who has seen the demonstration in (a). **(d)** Sample counterfactual alternatives considered by standard IRL, which are generated by incrementally deviating by one action (pink) along the agent's path, then following the agent's optimal policy afterward (blue). Note that neither matches the human's counterfactual.

tual that detours heavily.

Standard IRL (Ng and Russell 2000), however, does not model the learner's beliefs and fails to consider this detouring human counterfactual. Instead, standard IRL (we will henceforth simply refer to this method as IRL for the sake of brevity) enumerates possible counterfactual trajectories by considering all others actions that could have been taken at each step in the optimal trajectory (described in greater detail in the Counterfactual Scaffolding section). Two sample IRL counterfactuals to the agent's trajectory in Fig. 1c are shown in Fig. 1d, but neither matches the intuitive human counterfactual of completely detouring around the mud. Not only is IRL unable to consider the human counterfactual, a human is unlikely to follow IRL's step-wise method of enumerating many counterfactuals. As a result, IRL has the potential to both undershoot or overshoot the informativeness of a demonstration to a human by incorrectly considering the wrong counterfactuals or considering too many counterfactuals, respectively. Though IRL is a principled formalism for extracting reward information from demonstrations, it must be adapted before being used to model human learning.

This work thus aims to more effectively convey robot policies to humans by maintaining a more accurate model of human knowledge throughout the demonstration process. First, we evaluate the informativeness of demonstrations based on counterfactual trajectories likely to be considered by the human rather than those generated via one-action deviations as in standard IRL. Second, we improve scaffolding demonstrations by incrementally increasing the number of unique features (e.g. the delivery domain not only has mud, but also a spare battery) that are conveyed. And finally, we condition the expected difficulty of a human to predict an agent's behavior during testing on the human's current beliefs of the agent's reward.

Related Work

Policy Summarization & Machine Teaching

The problem of policy summarization considers which states and actions should be conveyed to help a user obtain a global understanding of a robot's policy (Amir, Doshi-Velez, and Sarne 2019). There are two primary approaches to this problem. The first relies on heuristics to evaluate the value of communicating certain states and actions, such as entropy (Huang et al. 2018), differences in Q-values (Amir and Amir 2018), and differences between the policies of two agents (Amitai and Amir 2021).

We build on the second approach, which follows the machine teaching paradigm (Zhu et al. 2018). Given an assumed learning model of the student (e.g. IRL), 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 (Zhu 2015; Liu and Zhu 2016), it has also recently been employed to convey reward functions from which the corresponding policy can be reconstructed. Sanneman and Shah (2021) provide a survey of such methods for explaining agent reward function to humans. The related works in this section, including our own, fall under their categorization of policy space techniques that convey information regarding the reward function using demonstrations of the agent's policy. We summarize a few relevant works below.

Huang et al. (2019) selected informative demonstrations for humans modeled to employ approximate Bayesian IRL for recovering the reward function. 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. Cakmak and Lopes (2012) 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 (2019) improved this method (particularly for high dimensions) by solving an equivalent set cover problem instead with their Set Cover Optimal Teaching (SCOT) algorithm. Also assuming that humans sometimes employ IRL-like reasoning to understand others' policies, Lage et al. (2019) used SCOT to select demonstrations to show to a human learner. However, SCOT is not explicitly designed for human learners and so our prior work (Lee, Admoni, and Simmons 2021) built on SCOT by incorporating human learning techniques and concepts such as scaffolding, simplicity, and similarity. Noting that humans are not pure IRL learners that can fully grasp a few highly informative but nuanced examples, we sought to scaffold demonstrations of increasing informativeness and difficulty to ease the learning. However, our initial method of scaffolding via IRL did not yield significant learning gains, which we aim to improve in this work by incorporating counterfactuals that are based on the human's expected knowledge.

Techniques for Human Teaching

We take inspiration from cognitive science in informing how a robot may teach and convey their decision-making to a human learner so that the learner may correctly predict the robot's behavior in new situations.

Scaffolding: Scaffolding is a well-established pedagogical technique in which a 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 (Wood, Bruner, and Ross 1976). We implement this by showing demonstrations that convey information on an increasing number of unique reward features. Following Reiser (2004)'s recommendation for software-based scaffolding to reduce the complexity of the learning problem through additional structure, we also provide demonstrations that sequentially increase in both informativeness and difficulty (as determined by counterfactual trajectories that likely mirror the human's beliefs).

Counterfactuals: In surveying the literature on how humans explain to each other, Miller (2019) notes that "explanations are constrastive – they are sought in response to particular counterfactual cases." Miller also notes that explanations are contextual and that it is important that the explainee is cognizant of the counterfactual intended by the explainer. Demonstrations likewise must be tailored to the learner given their current knowledge and the counterfactuals that the learner will probably consider.

Furthermore, Reiser (2004) suggests that scaffolding should not only provide structure that reduces problem complexity but at times induce cognitive conflict to challenge and engage the learner. As noted previously, the information provided by a demonstration is contingent on the counterfactual alternatives that are considered. It is subsequently important to ensure that the agent's demonstration *differs* from the counterfactuals likely to be considered by the human to provide information.

Testing: Effective scaffolding requires an accurate diagnosis of the learner's current abilities to provide the appropriate level of assistance throughout the teaching process (Collins, Brown, and Newman 1988). A common diagnostic method is presenting the learner with tests of varying difficulties and assessing their understanding. Our tests consist of presenting humans with unseen instances of a domain, then asking them to demonstrate the agent's optimal behavior (akin to the "best demonstration" or "simulation" reward understanding assessment (Sanneman and Shah 2021; Lage et al. 2018)). In our prior work (Lee, Admoni, and Simmons 2021), we showed a demonstration's expected informativeness (determined by IRL) could simply be inverted into a measure of the expected difficulty of a human to predict that exact demonstration during testing. In this work, we propose to update the difficulty measure by explicitly accounting for and conditioning on the learner's current knowledge.

Technical background

Much of the technical background is shared with our prior work (Lee, Admoni, and Simmons 2021) and is repeated below for completeness.

Markov decision process: The agent models its world as an instance (indexed by i) of a deterministic[2] Markov decision process, $MDP_i := (\mathcal{S}_i, \mathcal{A}, T_i, R, \gamma, S_i^0)$, where $\mathcal{S}_i$ and $\mathcal{A}$ denote the state and action sets, $T_i : \mathcal{S}_i \times \mathcal{A} \to \mathcal{S}_i$ the transition function, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the reward function, $\gamma \in [0, 1]$ the discount factor, and S_i^0 the initial state distribution, and $\mathcal{S} : \bigcup_i \mathcal{S}_i$ the union over the states of all related instances of MDPs, which we call a domain (to be described in the following paragraphs).

The robot 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 $s_i, s_i' \in \mathcal{S}_i, a \in \mathcal{A}$. We assume that R can be expressed as a weighted linear combination of l reward features[3] $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^l$, i.e. $R = \mathbf{w}^{*\top} \phi(s, a, s')$ (Abbeel and Ng 2004). We also assume that the human is aware of all aspects of an MDP apart from the weights $\mathbf{w}^*$.

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. Fig. 1a and 1c), they importantly share the same reward function R.

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 work for notational simplicity.

Having represented the agent's world and policy, we now define the problem of generating demonstrations for teaching that policy through the lens of machine teaching.

Machine teaching for policies: As formalized by Lage et al. (2019), machine teaching for policies seeks to convey a set of demonstrations $\mathcal{D}$ of size n (i.e. the allotted budget for the teaching set) that will maximize the similarity ρ between π^* and the policy $\hat{\pi}$ recovered using a model $\mathcal{M}$ on $\mathcal{D}$

$$\arg\max_{\mathcal{D} \subset \Xi} \rho(\hat{\pi}(\mathcal{D}, \mathcal{M}), \pi^*) \ \text{ s.t. } \ |\mathcal{D}| = n \qquad (1)$$

[2] Though we assume a deterministic MDP, methods described here naturally generalize to MDPs with stochastic transition functions and policies.

[3] This assumption can be made without loss of generality as the reward features can be nonlinear with respect to states and actions and be arbitrarily complex.

where Ξ is the set of all optimal demonstrations of π^* in a domain. We assume that the human is aware of the reward features and employs IRL (Ng and Russell 2000) as their model $\mathcal{M}$ for approximating the $\mathbf{w}^*$ underlying demonstrations. 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[4]. 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 region of (viable) reward weights under which π is still optimal. The larger the area of BEC(π) is, the greater the potential uncertainty over $\mathbf{w}^*$ that is underlying the robot's optimal policy.

$$\mathrm{BEC}(\pi) = \left\{\mathbf{w} \in \mathbb{R}^l \mid \pi \text{ optimal w.r.t. } R = \mathbf{w}^\top \phi(s, a, s')\right\} \tag{2}$$

The BEC(π) can be approximated as the intersection of the following half-space constraints generated by the central IRL equation (Ng and Russell 2000; Abbeel and Ng 2004)

$$\begin{gathered} \mathbf{w}^\top \left(\mu_\pi^{(s,a)} - \mu_\pi^{(s,b)}\right) \geq 0 \\ \forall a \in \underset{a' \in \mathcal{A}}{\arg\max}\, Q^*\left(s, a'\right), b \in \mathcal{A}, s \in \mathcal{S} \end{gathered} \tag{3}$$

where $\mu_\pi^{(s,a)} = \left[\sum_{t=0}^{\infty} \gamma^t \phi\left(s_t\right) \mid \pi, s_0 = s, a_0 = a\right]$ is the vector of reward feature counts accrued from taking the optimal action a in s, then following π after, and $Q^*(s, a)$ refers to the optimal Q-value in a state and a possible action (Watkins and Dayan 1992).

Brown and Niekum (2019) 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

$$\mathbf{w}^\top \left(\mu_\pi^{(s,a)} - \mu_\pi^{(s,b)}\right) \geq 0, \forall (s, a) \in \mathcal{D}, b \in \mathcal{A}. \tag{4}$$

Using Eq. 4, every demonstration can be translated into a set of constraints on the viable reward weights.

We make two observations. First, Eq. 3 and Eq. 4 capture the key idea that IRL depends not only on the agent'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, assuming that the human has been shown the set of demonstrations $\mathcal{D}$, BEC($\mathcal{D}|\pi$) could subsequently be used to model the human's belief over the agent's possible reward weights. We will later build on this concept when incorporating a human model to select demonstrations that account for human counterfactuals.

Consider an example in the delivery domain with $A = \{$*up*, *down*, *left*, *right*, *pick up*, *drop*$\}$, binary reward features $\phi = [$*traversed mud*, *picked up battery*, *action taken*$]$, $\mathbf{w}^* \propto [-3, 3.5, -1]$. In practice, we require that $||\mathbf{w}||_2 = 1$ to circumvent the scaling invariance of IRL solutions and to eliminate the degenerate all-zero reward function (Brown and Niekum 2018). The space of models in the human's mind without information is the $n-1$ sphere due to the L^2 norm constraint on $\mathbf{w}$, where n is the number of domain features. The demonstration in Fig. 1a corresponds to the constraints in Fig. 1b, which has cut away regions of the 2-sphere and has left only the sliver which contains the agent's true reward weights. The constraints on viable reward weights intuitively indicate that $w_2^* < 0$ (conveyed by the red halfspace constraint in Fig. 1b) since no unnecessary actions were taken in delivering the package, and $w_0^* < 0$ and $w_0^* < 2w_2^*$ since two actions were taken to detour around the mud (conveyed by the red and blue constraints jointly). Importantly, the size of the $n-1$ sphere that remains after considering the constraints generated by a demonstration can be used as a measure of its informativeness. The smaller the area, the fewer viable reward weights remain, and the more informative the demonstration.

Proposed Techniques for Teaching Humans

Counterfactual Scaffolding

As conveyed in the introduction and Fig. 1, a demonstration's ability to reveal the underlying reward function via IRL critically hinges on the counterfactuals considered.

Many counterfactual alternatives proposed by IRL can seem nonsensical to humans due to the rote process used to generate them. As expressed by Eq. 4, IRL generates counterfactuals by taking the agent's optimal trajectory and at each state s, taking a potentially suboptimal action b then following the optimal policy afterward. This process generates the two sample counterfactuals that are seen in Fig. 1d, which importantly do not correspond to human counterfactual in Fig. 1c. While such one-action deviations from the optimal trajectory are computationally sensible and often efficient[5], these are unlikely to be the counterfactuals on the human's mind for a number of reasons.

First, humans are likely unlikely to methodically go through each state of the agent's trajectory and consider all possible alternative actions. Instead, humans naturally incline toward a few causes and a few counterfactuals out of many potential ones for explanation (Miller 2019). This can lead IRL to oversell the informativeness if more counterfactuals than ones in the human's mind are considered. Second, the counterfactuals that IRL considers are generated by "perturbing" the demonstration directly (by taking a single suboptimal action, then following the optimal policy henceforth) and may not be consistent with any reward function (e.g. no reward function considered in the delivery domain would first avoid the mud, then later go through the mud

[4] In principle, a robot could simply convey $\mathbf{w}^*$ explicitly to a human. However, it can be nontrivial for humans to map precise numerical reward weights to the corresponding optimal behavior through planning, especially if there is large number of reward features. 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.

[5] One could consider n-action deviations for trajectory ξ where n can be an integer greater than one and the length of the trajectory is $|\xi|$. However, this results in an exponential branching factor where the number of counterfactual trajectories grows by $|\xi| \times \mathcal{A}^n$, many of which yield redundant constraints that are looser than those generated by one-action deviation.

as one of the counterfactuals in Fig. 1d does). Instead, humans may consider a reward function that is different than the agent's, but their counterfactuals are likely to be consistent with that "perturbed" reward function (e.g. avoiding the mud both ways as in Fig. 1c). This can cause IRL to also undersell the informativeness of a demonstration if the human's counterfactual alternatives are not considered.

In selecting effective explanations, we posit that you must not only consider the learner's learning model (i.e. IRL) but also their prior knowledge and subsequently what counterfactuals they would consider. We thus extend our prior work (Lee, Admoni, and Simmons 2021) to evaluate a demonstration's informativeness based on counterfactuals generated via potential reward functions on the human's mind as opposed to counterfactuals generated via one-action deviations, and scaffold by showing demonstrations of increasing informativeness.

To incorporate a human model and account for human counterfactuals when evaluating the informativeness of potential demonstrations, do the following. First instantiate a prior model of the human's beliefs over the agent's reward 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 and corresponding constraints (e.g. that action reward is negative). Then sample m weights from $B(\mathbf{w}^*)$, e.g. using the Gon algorithm (Gonzalez 1985) to ensure that weights are evenly distributed. Each weight represents a particular belief that the human could have over the agent's reward function. For every possible demonstration in a domain[6] by the agent, and for each of the m weights, simulate what the "human" counterfactual to each demonstration would be if the human had this weight (and subsequent reward function) in mind and generate the corresponding constraints using Eq. 4[7]. For each possible demonstration by the agent, consolidate the corresponding m possible human counterfactuals by taking a union of all corresponding constraints. Finally, select the demonstration that maximizes information gain, i.e. select the demonstration that maximizes the difference between $B(\mathbf{w}^*)$ before and after the human sees this demonstration. Once you have shown the selected demo and updated $B(\mathbf{w}^*)$, you may select the next demonstration to show by sampling m weights from the updated $B(\mathbf{w}^*)$ and repeating the steps above.

Feature Scaffolding

In one of the first papers on scaffolding, Wood, Bruner, and Ross (1976) note that one can scaffold by reducing the degrees of freedom of the problem. One natural way to implement this is to selectively show demonstrations in which the number of reward features ϕ over which information is conveyed is limited. In the delivery domain for example, one could show demonstrations that convey information on the mud and action weights first, then on the battery and action weights, then on the mud, battery, and action weights to show potentially nuanced tradeoffs. Note that because solutions of IRL are scale invariant (i.e. weights $\mathbf{w}$ and $2\mathbf{w}$ will yield the same behavior) we must show at least two features at a time such that information on reward weights are conveyed relative to one another (e.g. how many actions are you willing to take to avoid mud?). Any feature over which information should not be conveyed can be considered as being "masked". Though we only propose a method for scaffolding three features in this work, you could extend this method to an arbitrary number of k features by showing demonstrations that iteratively mask $k-2$, $k-3$, ..., 0 features and showing combinations of two, three, ..., all features respectively. We leave a more principled scheme for scaffolding a higher number of features for future work.

To employ feature scaffolding for three features, first determine the order in which you will mask the features. For all of the demonstrations that the agent could show in a domain, obtain all possible constraints that could be generated using Eq. 4. The order of the masking will be from the feature that has the least number of nonzero entries across all of the constraints to the feature that has the largest number of nonzero entries (a sample constraint generated by Eq. 4 could be [2, 0, -5] in which the first and third features have nonzero entries). A feature with a high number of nonzero entries (e.g. action reward) often serves as a good reference feature that also allows for fine-grained comparisons, and are thus masked last. Once the masking order has been decided, remove any demonstrations that convey information about the first masked feature from consideration (i.e. any demonstrations that conveys constraints in which the feature count for a masked feature is nonzero). From this reduced set of demonstration, run counterfactual scaffolding as described in the previous subsection until there are no more demonstrations that can provide additional information gain. Then remove any demonstrations that convey information about the second masked feature and run counterfactual scaffolding until there are no more demonstrations that can provide additional information gain. Repeat for the third masked feature. Finally, consider all possible demonstrations in the domain and run counterfactual scaffolding until there are no more demonstrations that can provide additional information gain.

Testing

The size of a demonstration's BEC area intuitively captures its informativeness during teaching; the smaller the area, the less uncertainty there is regarding the value of $\mathbf{w}^*$. In our prior work (Lee, Admoni, and Simmons 2021), we showed that the BEC area can also 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.

However, the learner's current knowledge also likely plays a role. We hypothesize that the overlap in area between BEC($\xi|\pi^*$) and $B(\mathbf{w}^*)$ better captures the difficulty of a demonstration ξ as test for a human, for this overlap in-

[6] As we work in relatively small gridworld domains with hundreds of states, the possible demonstration set can simply be collected by rolling out the agent's policy from each possible state. For larger or continuous state spaces, a more sophisticated method for obtaining the set of possible demonstrations may be needed.

[7] Amongst equally rewarding counterfactuals, we give the "human" the benefit of the doubt by selecting the counterfactuals the maximizes the agent's true reward.

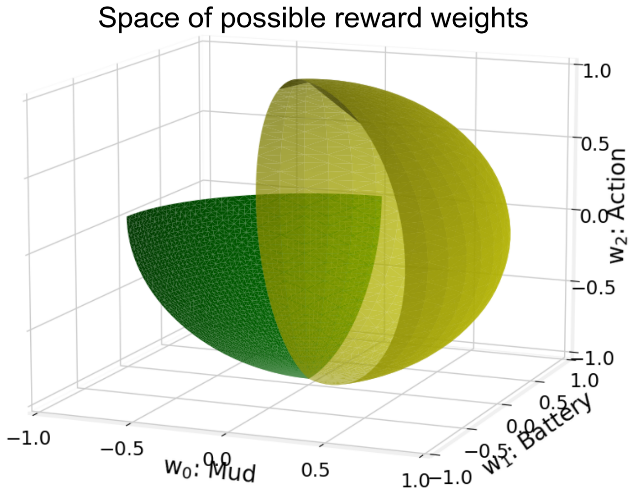

Figure 2: There are many reward weights BEC($\xi|\pi^*$) (yellow) that will generate the optimal demonstration ξ, indicated by the half-sphere. However, only a portion of it overlaps with the reward weights currently on the human's mind $B(\mathbf{w}^*)$ (green), making demonstration ξ likely difficult for the human to correctly predict it during testing.

tuitively represents the percentage of current models in the human's mind that would generate the correct behavior. As seen in Fig. 2, a demonstration may have an intrinsically large BEC area but the human's current knowledge may not overlap much.

To estimate the expected difficulty of each demonstration ξ that could be shown in a domain, first obtain the BEC($\xi|\pi^*$) using Eq. 4. Noting that one-action deviation does not always consider all reasonable counterfactual trajectories, we combine constraints that define BEC($\xi|\pi^*$) and constraints obtained from human counterfactual trajectories generated using m models sampled from $B(\mathbf{w}^*)$. These combined constraints for each demonstration will give a better estimate of the set of all weights that yield the correct demonstration, denoted by BEC$'$($\xi|\pi^*$)[8]. Finally, to measure the difficulty of a demonstration ξ as a test for this human, simply take the overlap between $B(\mathbf{w}^*)$ and BEC$'$($\xi|\pi^*$) as an estimate of how many weights in $B(\mathbf{w}^*)$ would produce the correct demonstration. The smaller the overlap, the fewer of the reward functions in the human's mind will generate the correct demonstration and harder the test.

Proposed experiment

We plan on running an online user study that involves participants watching demonstrations of a 2D agent's policy and predicting the optimal trajectory in new test environments. The study will evaluate the following hypotheses.

H1: The overlap between the human's belief over the agent's weights $B(\mathbf{w}^*)$ and the BEC area of a demonstration BEC$'$($\xi|\pi^*$) correlates 1) inversely to the expected difficulty for a human to correctly predict it during testing, and 2) directly to their confidence in that prediction.

H2: Using counterfactual scaffolding in selecting training demonstration will result in higher perceived informativeness of training demonstrations and better participant test performance over the baseline scaffolding proposed by Lee, Admoni, and Simmons (2021).

H3: Using feature scaffolding in selecting training demonstration will result in lower mental effort during training and better participant test performance over no feature scaffolding.

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

Three simple gridworld domains will be used for this study, with each domain consisting of one shared reward feature of action, and two unique reward features as follows. The humans are explicitly told the features, but must infer the respective reward weights by watching demonstrations.

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

Colored tiles domain. The agent penalized differently for traversing the two differently colored tiles throughout the environment.

Skateboard domain. The agent is penalized less per action if it has either picked up a skateboard (i.e. riding a skateboard is less costly than walking) or is traversing through a designated path.

The user study will explore whether demonstrations selected using counterfactual and feature scaffolding improves a human's understanding of agent's reward function and policy. The between-subjects variables will be *counterfactual scaffolding* (yes and no), and *feature scaffolding* (yes and no). There will be two within-subject variables: *domain* (delivery, colored tiles, and skateboard) and *test difficulty* (low, medium, and high, determined by the aforementioned overlap between $B(\mathbf{w}^*)$ and BEC$'$($\xi|\pi^*$)).

Finally, the following objective and subjective measures will be recorded to evaluate the aforementioned hypotheses.

M1. Optimal response: For each test, whether the participant's trajectory received the optimal reward/score or not.

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 why these demonstrations were optimal?"

M4. Confidence rating: 5-point Likert scale with prompt "How confident are you that you obtained the optimal score?"

[8] Our intuition is that in theory, the best estimate of BEC$'$($\xi|\pi^*$) would naively require considering every single possible suboptimal demonstration (e.g. obtain via one-, two-, three-action deviations and so on) but the branching factor quickly becomes unmanageable.

References

Abbeel, P.; and Ng, A. Y. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*.

Amir, D.; and Amir, O. 2018. Highlights: Summarizing agent behavior to people. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 1168–1176. International Foundation for Autonomous Agents and Multiagent Systems.

Amir, O.; Doshi-Velez, F.; and Sarne, D. 2019. Summarizing agent strategies. *Autonomous Agents and Multi-Agent Systems*, 33(5): 628–644.

Amitai, Y.; and Amir, O. 2021. "I Don't Think So": Disagreement-Based Policy Summaries for Comparing Agents. *arXiv preprint arXiv:2102.03064*.

Baker, C.; Saxe, R.; and Tenenbaum, J. 2011. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Baker, C. L.; Saxe, R.; and Tenenbaum, J. B. 2009. Action understanding as inverse planning. *Cognition*, 113(3): 329–349.

Brown, D. S.; and Niekum, S. 2018. Efficient probabilistic performance bounds for inverse reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Brown, D. S.; and Niekum, S. 2019. Machine teaching for inverse reinforcement learning: Algorithms and applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7749–7758.

Cakmak, M.; and Lopes, M. 2012. Algorithmic and human teaching of sequential decision tasks. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 1536–1542.

Collins, A.; Brown, J. S.; and Newman, S. E. 1988. Cognitive apprenticeship: Teaching the craft of reading, writing and mathematics. *Thinking: The Journal of Philosophy for Children*, 8(1): 2–10.

Gonzalez, T. F. 1985. Clustering to minimize the maximum intercluster distance. *Theoretical computer science*, 38: 293–306.

Huang, S. H.; Bhatia, K.; Abbeel, P.; and Dragan, A. D. 2018. Establishing appropriate trust via critical states. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3929–3936. IEEE.

Huang, S. H.; Held, D.; Abbeel, P.; and Dragan, A. D. 2019. Enabling robots to communicate their objectives. *Autonomous Robots*, 43(2): 309–326.

Jara-Ettinger, J. 2019. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29: 105–110.

Jara-Ettinger, J.; Gweon, H.; Schulz, L. E.; and Tenenbaum, J. B. 2016. The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, 20(8): 589–604.

Jern, A.; Lucas, C. G.; and Kemp, C. 2017. People learn other people's preferences through inverse decision-making. *Cognition*, 168: 46–64.

Lage, I.; Chen, E.; He, J.; Narayanan, M.; Gershman, S. J.; Kim, B.; and Doshi-Velez, F. 2018. An Evaluation of the Human-Interpretability of Explanation.

Lage, I.; Lifschitz, D.; Doshi-Velez, F.; and Amir, O. 2019. Exploring Computational User Models for Agent Policy Summarization. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 1401–1407. International Joint Conferences on Artificial Intelligence Organization.

Lee, M. S.; Admoni, H.; and Simmons, R. 2021. Machine Teaching for Human Inverse Reinforcement Learning. *Frontiers in Robotics and AI*, 8: 188.

Liu, J.; and Zhu, X. 2016. The Teaching Dimension of Linear Learners. *Journal of Machine Learning Research*, 17: 1–25.

Lucas, C. G.; Griffiths, T. L.; Xu, F.; Fawcett, C.; Gopnik, A.; Kushnir, T.; Markson, L.; and Hu, J. 2014. The child as econometrician: A rational model of preference understanding in children. *PloS one*, 9(3): e92160.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267: 1–38.

Ng, A. Y.; and Russell, S. 2000. Algorithms for Inverse Reinforcement Learning. In *in Proc. 17th International Conf. on Machine Learning*.

Reiser, B. J. 2004. Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *The Journal of the Learning sciences*, 13(3): 273–304.

Sanneman, L.; and Shah, J. 2021. Explaining Reward Functions to Humans for Better Human-Robot Collaboration. In *AAAI Fall Symposium AI-HRI Workshop*.

Shteingart, H.; and Loewenstein, Y. 2014. Reinforcement learning and human behavior. *Current Opinion in Neurobiology*, 25: 93–98.

Sukkerd, R.; Simmons, R.; and Garlan, D. ???? Tradeoff-focused contrastive explanation for MDP planning. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 1041–1048. IEEE.

Watkins, C. J.; and Dayan, P. 1992. Q-learning. *Machine learning*, 8(3-4): 279–292.

Wood, D.; Bruner, J. S.; and Ross, G. 1976. The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, 17(2): 89–100.

Wunderlich, K.; Dayan, P.; and Dolan, R. J. 2012. Mapping value based planning and extensively trained choice in the human brain. *Nature neuroscience*, 15(5): 786–791.

Zhu, X. 2015. 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*, 4083–4087.

Zhu, X.; Singla, A.; Zilles, S.; and Rafferty, A. N. 2018. An overview of machine teaching. *arXiv preprint arXiv:1801.05927*.