

I Know What You Meant: Learning Human Objectives by (Under)estimating Their Choice Set

Ananth Jonnavittula and Dylan P. Losey

Abstract—Assistive robots have the potential to help people perform everyday tasks. However, these robots first need to learn what it is their user wants them to do. Teaching assistive robots is hard for inexperienced users, elderly users, and users living with physical disabilities, since often these individuals are unable to show the robot their desired behavior. We know that *inclusive* learners should give human teachers credit for what they cannot demonstrate. But today’s robots do the opposite: they assume *every* user is capable of providing *any* demonstration. As a result, these robots learn to mimic the demonstrated behavior, even when that behavior is not what the human really meant! Here we propose a different approach to reward learning: robots that reason about the user’s demonstrations in the context of similar or simpler alternatives. Unlike prior works — which err towards overestimating the human’s capabilities — here we err towards *underestimating* what the human can input (i.e., their choice set). Our theoretical analysis proves that underestimating the human’s choice set is risk-averse, with better worst-case performance than overestimating. We formalize three properties to generate similar and simpler alternatives. Across simulations and a user study, our resulting algorithm better extrapolates the human’s objective. See the user study here: <https://youtu.be/RgbH2YULVRo>.

I. INTRODUCTION

Imagine that you’re using an assistive robot arm, and you want to teach this arm to carry a cup of coffee upright. You control the robot’s motion using a joystick [1]–[3]. The robot has many degrees-of-freedom, and — as you teleoperate the arm to provide your demonstration — you must carefully orchestrate each joint to keep the coffee upright. It’s possible for a dexterous, experienced human to teleoperate this task correctly. But you’re not as skilled with the control interface: *because of your limitations* the best you can do is to avoid turning over the cup and spilling all the coffee (see Fig. 1).

We refer to the set of possible demonstrations that a given human can show the robot as the human’s *choice set* [4], [5]. Robots learn what the human wants by reasoning about the human’s actual demonstrations in the context of this choice set. For instance, if the human’s demonstrations keep the coffee more vertical than the alternatives within their choice set, it’s likely that the human values keeping the coffee upright. Importantly, when we interact with robots, our choice sets are *inherently constrained* by our own physical and cognitive capabilities — and these limitations vary from one person to another [6]–[8]. Take our example: although an expert is able to show a trajectory with the coffee roughly upright, the best we can do is to keep the cup from flipping

over. Uncertainty over what this choice set is makes it hard for the robot to learn: did we tilt the coffee because we meant to, or because we were unable to keep it upright?

Existing work on learning human objectives from demonstrations often ignores user limitations, and assumes that the human’s choice set includes *every* demonstration that is consistent with the robot’s dynamics [9]–[13]. This results in robots that learn to *match* your demonstrated behavior, so that if all your demonstrations spill some coffee, the robot also learns to spill. We explore the opposite perspective:

Inclusive robots should assume that the human may have limitations, and can only show alternative behaviors that are similar to or simpler than their actual demonstrations.

Returning to our coffee example, now the robot narrows down its estimate of the human’s choice set to only include the observed human trajectories and sparse and noisy alternatives. Compared to these alternatives, our actual demonstrations better keep the cup closer to vertical. Here the robot *extrapolates* what we really meant, despite the fact that we spilled coffee in all our demonstrations (see Fig. 1).

Overall, we make the following contributions:

Underestimating vs. Overestimating Choice Sets. Robots will inevitably get the human’s choice set wrong. We theoretically compare overestimating the human’s choice set (i.e., assuming users can provide any demonstration) to underestimating the human’s choice set (i.e., users can only provide a few demonstrations). We prove that underestimating is a risk-averse approach with better worst-case learning.

Generating Choice Sets from Demonstrations. How does the robot get its estimate of a human’s choice set? Our analysis finds that waiting for human teachers to demonstrate their choice set is intractable. To address this, we formalize three properties that robots can leverage to generate trajectories that are similar to or simpler than the human’s actual choices.

Conducting a User Study. We compare our inclusive approach to state-of-the-art baselines. Our results suggest that robots which err towards underestimating the human’s choice set better extrapolate the human’s underlying objective.

II. RELATED WORK

Application – Assistive Robots. Assistive robots, such as wheelchair-mounted robot arms, promise to improve the autonomy and independence of users living with physical disabilities [14]. However, people’s ability to control assistive robots is often restricted, both by their own physical impairments and the teleoperation interfaces they leverage [1]–[3], [15]–[17]. Expecting all users to produce the same demonstrations is unfair [18]. We therefore develop a *personalized*

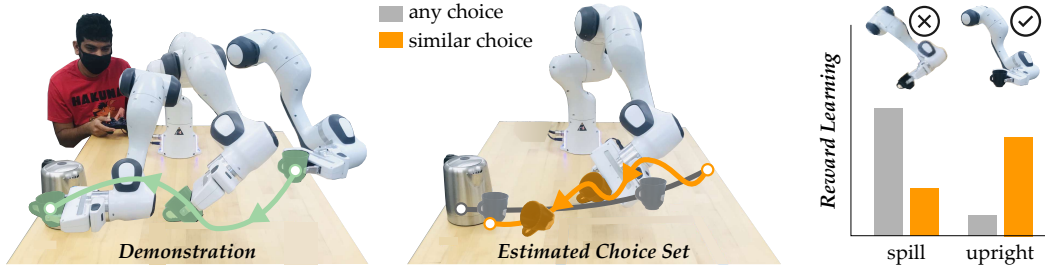


Fig. 1. (Left) User teaches an assistive robot arm to carry coffee upright. Because of their limitations, this user’s best demonstration unintentionally tilts the cup. (Middle) When the robot ignores these limitations — and compares the human’s input to **any trajectory** — it learns to match the human’s demonstration and also spill coffee. To make learning more inclusive, we instead compare the human’s demonstration to **simpler and similar trajectories**. (Right) Because these alternatives spill more coffee than the human, the robot now extrapolates what the user really wanted: keeping the coffee upright.

approach, where the robot learns from each user by placing their demonstrations in the context of similar behaviors.

Learning Rewards from Demonstrations. Prior work studies how assistive arms and other robots can learn from human demonstrations. Inverse reinforcement learning (IRL) infers what the user wants — i.e., their underlying reward function — from their demonstrated trajectories. To do this, the robot searches for the reward function that makes the human’s trajectories appear roughly optimal [9]. Crucially, today’s methods assume that the human is optimizing over all possible trajectories; accordingly, when human demonstrates suboptimal behavior, the robot thinks this behavior is better than every other option, and learns the reward function that best reproduces what the human demonstrated [10]–[13].

Recent research explores how this learning model *misses out* on the way humans actually behave [19]–[21]. Most relevant are [4] — where the authors formalize choice sets in reward learning — and [5] — where the authors experimentally compare different classes of choice set misspecification. We also build upon inverse reinforcement learning, but develop a formalism for generating smaller, constrained choice sets from the human’s demonstrations.

Outperforming Imperfect Demonstrations. Our goal in generating these choice sets is for robots to learn rewards which better align with what the human wants than the human’s demonstrations. Other works outperform imperfect and suboptimal demonstrations by eliciting additional types of human feedback [22]–[26]. We do not gather additional feedback; instead, our approach is most similar to [27]–[29]. Here the robot considers noisy perturbations of the human’s original demonstrations, and extrapolates a reward function which maximizes the difference between the human’s demonstrations and these noisy alternatives. Viewed within our formalism, the alternative demonstrations form the human’s choice set, and leveraging noisy perturbations becomes *one instance* of generating this set.

III. PROBLEM STATEMENT

Let’s return to our motivating example, where we are teleoperating an assistive robot arm, and want to teach this arm to hold coffee upright. Each time we guide the robot through the process of picking up, carrying, and putting down our coffee cup, we show the robot a trajectory $\xi \in \Xi_{\mathcal{H}}$. Here ξ is a sequence of robot state-action pairs, and $\Xi_{\mathcal{H}}$ is our

choice set: i.e., the set of all trajectories *we are capable* of showing to the robot. Of course, the robot does not know what our choice set is — we could be an expert user, capable of keeping the cup perfectly vertical, or an inexperienced user, who struggles to keep the cup from flipping over. The robot therefore works with $\Xi_{\mathcal{R}}$, an *estimate* of $\Xi_{\mathcal{H}}$.

Over repeated interactions we show the robot N trajectories from our choice set. Given all of these user demonstrations $\mathcal{D} = \{\xi_1, \xi_2, \dots, \xi_N\}$, the robot recovers our reward function $r_{\theta}(\xi) \rightarrow \mathbb{R}$, i.e., what we want the robot to optimize for across a trajectory¹. More formally, the robot tries to infer r_{θ} (what we value) given \mathcal{D} (the trajectories we have shown) and $\Xi_{\mathcal{R}}$ (the trajectories the robot thinks we can show).

Bayesian Inference. Let belief b denote the robot’s learned probability distribution over r_{θ} . Applying Bayes’ Theorem:

$$b(r_{\theta}) = P(r_{\theta} \mid \mathcal{D}, \Xi_{\mathcal{R}}) \propto P(\mathcal{D} \mid r_{\theta}, \Xi_{\mathcal{R}}) \cdot P(r_{\theta}) \quad (1)$$

where $P(r_{\theta})$ is the robot’s prior over the human’s reward. Since we assume the human’s demonstrations are conditionally independent [9], [24], Equation (1) simplifies to:

$$b(r_{\theta}) \propto P(r_{\theta}) \prod_{\xi \in \mathcal{D}} P(\xi \mid r_{\theta}, \Xi_{\mathcal{R}}) \quad (2)$$

The crucial term is $P(\xi \mid r_{\theta}, \Xi_{\mathcal{R}})$, the likelihood of observing ξ given that the human has reward r_{θ} and choice set $\Xi_{\mathcal{R}}$. Within robot learning [11] and cognitive science [30], this is commonly written using a Boltzmann-rational model:

$$P(\xi \mid r_{\theta}, \Xi_{\mathcal{R}}) = \frac{\exp(\beta \cdot r_{\theta}(\xi))}{\sum_{\xi' \in \Xi_{\mathcal{R}}} \exp(\beta \cdot r_{\theta}(\xi'))} \quad (3)$$

Intuitively, this model asserts that a rational human picks the trajectory from their choice set that noisily maximizes their reward, and the hyperparameter $\beta \geq 0$ captures how close-to-rational the human is. Substituting this likelihood function into Equation (2), we arrive at the *robot’s learning rule*:

$$b(r_{\theta}) \propto \frac{\exp(\beta \cdot \sum_{\xi \in \mathcal{D}} r_{\theta}(\xi))}{(\sum_{\xi' \in \Xi_{\mathcal{R}}} \exp(\beta \cdot r_{\theta}(\xi')))^N} \cdot P(r_{\theta}) \quad (4)$$

Importance of Choice Sets. When the robot misunderstands the human’s choice set (i.e., $\Xi_{\mathcal{R}} \neq \Xi_{\mathcal{H}}$) this learning rule may

¹In our experiments we are consistent with previous work on inverse reinforcement learning and use $r_{\theta}(\xi) = \theta \cdot \Phi(\xi)$, where $\Phi(\xi) \in \mathbb{R}^k$ are known features and $\theta \in \mathbb{R}^k$ are unknown reward weights [4], [9]–[11].

spectacularly fail. Imagine that the robot from Fig. 1 has a uniform prior over two opposite rewards that optimize for staying upright ($r_{upright}$) or spilling coffee (r_{spill}). As an inexperienced user the best we can do is to spill at least 75% of the coffee, and we demonstrate this ξ to the robot — hoping the robot will recognize our limitations and learn $r_{upright}$. However, the robot overestimates our capabilities. Let $\Xi_{\mathcal{R}}$ include a full spectrum of trajectories, from spilling the entire cup to spilling none at all: importantly, our ξ pours more coffee than most of these alternatives in $\Xi_{\mathcal{R}}$. Hence, if we apply Equation (4) with this choice set, for any $\beta > 0$ the robot believes r_{spill} is actually what we wanted to optimize. The robot gets it wrong here because it *incorrectly estimates* what trajectories the human can input.

IV. THEORETICAL ANALYSIS

The problem is that robots will *inevitably* get the human’s choice set wrong. This is especially true in assistive robotics, where physical and control limitations vary greatly from one user to another [3]. If we are doomed to mistake $\Xi_{\mathcal{H}}$, on which side should we err? One option is to *overestimate* the choices that the human can make, such that $\Xi_{\mathcal{R}} \supseteq \Xi_{\mathcal{H}}$. This is traditionally done in inverse reinforcement learning, where $\Xi_{\mathcal{R}}$ includes all feasible trajectories [10]–[13]. By contrast, our insight is to err towards *underestimating* the human’s capabilities, such that $\Xi_{\mathcal{R}} \subseteq \Xi_{\mathcal{H}}$. In this section we analyze the pros and cons of each approach.

Assumptions. To better understand how the system behaves at the limit, we assume that the human is perfectly rational, and always inputs the $\xi \in \Xi_{\mathcal{H}}$ that maximizes their reward (i.e., $\beta \rightarrow \infty$). We further assume that the human’s choice ξ is an element of $\Xi_{\mathcal{R}}$: because the robot observes ξ prior to learning, it is trivial to add ξ to $\Xi_{\mathcal{R}}$ if it is not already included. Finally, for the sake of clarity, we assume that the mapping from trajectories ξ to rewards $r_{\theta}(\xi)$ is injective.

In practice, putting these assumptions together with Equation (3) implies that either $P(\xi | r_{\theta}, \Xi_{\mathcal{R}}) = 1$ (if $\xi \in \Xi_{\mathcal{R}}$ maximizes r_{θ}) or $P(\xi | r_{\theta}, \Xi_{\mathcal{R}}) = 0$ (if $\xi \in \Xi_{\mathcal{R}}$ does not maximize r_{θ}). These assumptions are purely for analysis, and are removed in our simulations and user study.

A. Risk-Sensitivity and Learning

In the context of learning with Equation (4), one definition of risk relates to *confidence*. A risk-seeking robot becomes confident in its estimate of r_{θ} after just a few demonstrations, while a risk-averse robot remains uncertain even when many demonstrations are available. We quantify this confidence as the Shannon entropy over the robot’s belief b [31].

But how confident should the robot actually be? Our point of reference is an ideal learner that knows $\Xi_{\mathcal{H}}$. This is the gold standard, because if we knew $\Xi_{\mathcal{H}}$, we’d always have the right context for learning from the human’s demonstrations. We therefore define risk as the difference between the robot’s *actual* entropy and the *gold standard* entropy. A risk-seeking robot has less entropy than the gold standard (indicating it is overly confident in its estimate) while a risk-averse robot has

more entropy than the gold standard (indicating it is being unnecessarily conservative).

We find that how we estimate the choice set affects risk:

Proposition 1. *Robots which overestimate the human’s capabilities are risk-seeking, while robots which underestimate the human’s choice set are risk-averse.*

Proof. Let $\xi \in \mathcal{D}$ be an observed human demonstration. When the robot overestimates the human’s capabilities, $\Xi_{\mathcal{R}} \supseteq \Xi_{\mathcal{H}}$. This increases the denominator of Equation (3), so that $P(\xi | r_{\theta}, \Xi_{\mathcal{R}}) \leq P(\xi | r_{\theta}, \Xi_{\mathcal{H}})$ for all r . But from our assumptions we know that $P(\xi | r_{\theta}, \Xi_{\mathcal{R}})$ is either 0 or 1. Thus, $\Xi_{\mathcal{R}} \supseteq \Xi_{\mathcal{H}}$ concentrates the robot’s belief b around the same or fewer rewards than if $\Xi_{\mathcal{R}} = \Xi_{\mathcal{H}}$.

Conversely, if the robot underestimates the human’s capabilities, $\Xi_{\mathcal{R}} \subseteq \Xi_{\mathcal{H}}$. Removing choices decreases the denominator of Equation (3), so that $P(\xi | r_{\theta}, \Xi_{\mathcal{R}}) \geq P(\xi | r_{\theta}, \Xi_{\mathcal{H}})$ for all r_{θ} . Hence, $\Xi_{\mathcal{R}} \subseteq \Xi_{\mathcal{H}}$ distributes the robot’s belief b around the same or more rewards than if $\Xi_{\mathcal{R}} = \Xi_{\mathcal{H}}$. \square

To see this in action let’s return to our motivating example, where we input a trajectory which spills some coffee. If the robot overestimates our capabilities, the robot becomes confident that we are optimizing for spilling, since there are better trajectories to choose if we preferred to keep the cup upright. But a robot that underestimates our capabilities is not as quick to eliminate other explanations. This robot realizes that the trajectory we have shown is our best choice if we intend to spill some coffee, as well as our best choice for spilling no coffee at all. Hence, the risk-averse robot thinks both of these rewards are still likely.

B. Worst-Case Learning

Next let’s look at the worst that could happen when we get the human’s choice set wrong. We’ll think about this error in terms of Equation (4), where the robot learns the likelihood of each reward. In the best case the robot learns the human’s true reward r_{θ} , so that $b(r_{\theta}) = 1$.

The worst-case depends on how we estimate $\Xi_{\mathcal{H}}$:

Proposition 2. *In the worst case, robots that overestimate the human’s capabilities learn the wrong reward.*

Proof. The rational human chooses $\xi \in \Xi_{\mathcal{H}}$ to maximize their reward r_{θ} . But $\Xi_{\mathcal{R}} \supset \Xi_{\mathcal{H}}$, and in $\Xi_{\mathcal{R}}$ there might be another trajectory ξ' where $r_{\theta}(\xi') > r_{\theta}(\xi)$. So the human’s choice ξ does not maximize reward for r_{θ} ; instead, $\xi \in \Xi_{\mathcal{R}}$ is the best choice for reward r'_{θ} . Here $b(r'_{\theta}) = 1$ and $b(r_{\theta}) = 0$. \square

Proposition 3. *In the worst case, robots that underestimate the human’s capabilities learn nothing from demonstrations.*

Proof. The rational human chooses $\xi \in \Xi_{\mathcal{H}}$ to maximize r_{θ} . But $\Xi_{\mathcal{R}} \subset \Xi_{\mathcal{H}}$ only contains this single choice, $\Xi_{\mathcal{R}} = \{\xi\}$. So while $\xi \in \Xi_{\mathcal{R}}$ is the best choice for reward r_{θ} , it is also the best choice for all other rewards. Now $P(\xi | r_{\theta}, \Xi_{\mathcal{R}}) = 1$ for every reward, and $b(r_{\theta}) = P(r_{\theta})$. \square

We can intuitively connect this worst-case performance to our earlier analysis of risk-sensitivity. Robots that overestimate the human’s capabilities are risk-seeking, and quickly become overconfident in the reward they have learned. When

these risky robots get it wrong — for instance, thinking we want to spill the coffee — they *commit* to their mistakes, resulting in complete confidence in the wrong reward. Robots that underestimate the human’s choice set err in the opposite direction. These risk-averse robots play it safe, and maintain several possible explanations for the human’s behavior. When these conservative robots are overly *cautious* — e.g., severely underestimating our capabilities — they ignore the information our demonstrations actually contain.

C. Can We Rely on Human Teachers?

To avoid this worst-case performance, one option is to rely on the human *to show the robot* their choice set. Every time the human teleoperates the robot along a trajectory ξ , they are showing the robot another element of $\Xi_{\mathcal{H}}$. A naïve robot may assume that — given enough time — the human will demonstrate *all* the trajectories they are capable of inputting. Using this wait-and-see approach, the robot sets $\Xi_{\mathcal{R}} = \mathcal{D}$.

The problem here is that the user isn’t picking ξ to convey their choice set; instead, the human is inputting trajectories to teach the robot. Referring back to Equation (3), the human chooses $\xi \in \Xi_{\mathcal{H}}$ to noisily maximize their reward r . When users follow this Boltzmann-rational model, passively waiting for the user to show a diverse set of choices from $\Xi_{\mathcal{H}}$ becomes prohibitively time consuming:

Proposition 4. *If we normalize the rewards over $\Xi_{\mathcal{H}}$ between 0 and 1, the probability that the human will show a minimal reward trajectory ξ in N demonstrations is bounded by:*

$$P(\xi \in \mathcal{D}) \leq 1 - \left[\frac{|\Xi_{\mathcal{H}}| - 2 + \exp \beta}{|\Xi_{\mathcal{H}}| - 1 + \exp \beta} \right]^N \quad (5)$$

Proof. We seek to maximize Equation (3) when $r_{\theta}(\xi) = 0$ and there exists at least one ξ' where $r_{\theta}(\xi') = 1$. To do this we minimize the denominator, which occurs when $r_{\theta}(\xi') = 1$ only once and 0 otherwise. Now $P(\xi \mid r_{\theta}, \Xi_{\mathcal{H}}) = 1/(\exp \beta + |\Xi_{\mathcal{H}}| - 1)$, and the probability of not picking ξ during the current interaction is $1 - P(\xi \mid r_{\theta}, \Xi_{\mathcal{H}})$. \square

Jumping back to the motivating example, let’s say there are $|\Xi_{\mathcal{H}}| = 2$ different types of trajectories we can demonstrate: either tilting the mug or completely flipping it over. For a Boltzmann-rational human teacher with $\beta = 5$, *even if the robot waits for 50 demonstrations*, there is at most a 29% likelihood that the human demonstrates flipping the cup over. Hence, when humans are optimizing for teaching, we *cannot* rely on them to convey their choice set to the robot.

V. GENERATING THE CHOICE SET

The last section explored how underestimating or overestimating the human’s choice set affects reward learning — but how does the robot estimate this choice set in the first place? Although we can’t rely on the human to show us the entire $\Xi_{\mathcal{H}}$ (see Proposition 4), we can still leverage the choices the human makes. To estimate $\Xi_{\mathcal{H}}$ from an individual user’s demonstrations, we apply our insight: $\Xi_{\mathcal{H}}$ only includes trajectories that are *similar to* or *simpler than* \mathcal{D} . Unlike prior work, this approach errs towards *underestimating* (i.e.,

Algorithm 1 Inclusive reward learning from demonstrations

- 1: Collect human demonstrations: $\mathcal{D} = \{\xi_1, \xi_2, \dots, \xi_N\}$
- 2: Initialize estimate of human’s choice set: $\Xi_{\mathcal{R}} = \mathcal{D}$
- 3: **for** $\xi \in \mathcal{D}$ **do**
- 4: Generate counterfactual ξ' from ξ
- 5: Add the resulting trajectory ξ' to $\Xi_{\mathcal{R}}$
- 6: **end for**
- 7: Update robot’s belief given \mathcal{D} and $\Xi_{\mathcal{R}}$:

$$P(r_{\theta} \mid \mathcal{D}, \Xi_{\mathcal{R}}) \propto \frac{\exp(\beta \cdot \sum_{\xi \in \mathcal{D}} r_{\theta}(\xi))}{(\sum_{\xi' \in \Xi_{\mathcal{R}}} \exp(\beta \cdot r_{\theta}(\xi')))^N} \cdot P(r_{\theta})$$

$\Xi_{\mathcal{R}} \subseteq \Xi_{\mathcal{H}}$), since we are implicitly assuming that the user’s demonstrations are the limits of their capabilities.

In what follows we formalize three properties that result in similar and simpler trajectories. These properties are not meant to be an exhaustive list, but rather a starting point for generating intuitive and inclusive choice sets.

Noisy Deformations. When humans teleoperate robots they make minor mistakes: e.g., unintentionally pressing the joystick up instead of right. We capture these *local perturbations* as noisy trajectory deformations. Let $\sigma \sim p(\cdot)$ be a noise parameter. After sampling σ , we deform the human trajectory $\xi \in \mathcal{D}$ to get a similar alternative: $\xi' = f(\xi, \sigma)$. One concrete way to deform the human’s trajectory is $\xi' = \xi + A\sigma$, where A defines the deformation shape [32]. In practice, this produces alternate trajectories the user *could have input* if they made small, stochastic changes to their inputs.

Sparse Inputs. To guide the robot throughout the process of picking up, carrying, and placing a coffee cup, the user must input a complex sequence of teleoperation commands. Each of these individual inputs requires user *effort* and *intention*. Hence, we hypothesize that some simpler trajectories are the result of *less user oversight*. Let u be a human teleoperation input, let $U = [u^1, u^2, \dots]$ be the sequence of human inputs, and let $g(U) \rightarrow \xi'$ be the trajectory that results from input sequence U . Formally, this property searches for trajectories that are similar to $\xi \in \mathcal{D}$ but produced by sparser inputs:

$$\xi' = g(U^*), \quad U^* = \arg \min \| \xi - g(U) \|^2 + \lambda \cdot \|U\|_1^2 \quad (6)$$

where scalar $\lambda > 0$ determines the relative trade-off between trajectory similarity and input sparsity.

Consistent Inputs. Precisely controlling the robot arm often requires rapidly *changing* inputs: e.g., adjusting the position of the cup, then rotating the cup upright, and then fixing its position again [19]. Instead of expecting the user to make all of these different commands, we recognize that it is easier to *maintain consistent inputs*. Using the same notation from the last property, we now search for similar trajectories where the user minimizes their changes in teleoperation input:

$$U^* = \arg \min \| \xi - g(U) \|^2 + \lambda \cdot \sum \|u^t - u^{t-1}\|^2 \quad (7)$$

Solving Equation (7) outputs trajectories $\xi' = g(U^*)$ where the human inputs are held roughly constant.

Algorithm. The listed properties produce *counterfactuals*: i.e., what *would have happened* if the human was more noisy, more sparse, or more consistent. In Algorithm 1 we add these trajectories to $\Xi_{\mathcal{R}}$ before leveraging our learning rule from Equation (4), where $\Xi_{\mathcal{R}}$ determines the partition function. Overall, Algorithm 1 learns from human demonstrations by comparing \mathcal{D} only to a personalized choice set that we are confident the human is capable of inputting.

VI. SIMULATIONS

In this section we apply our learning algorithm across a spectrum of simulated humans with different levels of rationality (β). To better relate these findings to our theoretical results, for now we assume a *discrete* set of possible reward functions (we will move to a continuous hypothesis space in the subsequent user study).

Baselines. We compare Algorithm 1 (**Ours**) to three baselines. First we consider a hypothetical world (**Ideal**) where the robot knows exactly what the human’s choice set is ($\Xi_{\mathcal{R}} = \Xi_{\mathcal{H}}$). Next, we test Bayesian inverse reinforcement learning (**BIRL**) where the robot assumes the human is capable of providing any demonstration [12]. This leads to $\Xi_{\mathcal{R}} \supseteq \Xi_{\mathcal{H}}$. Finally, we implement a robot that compares human demonstrations to noisy alternatives (**Noise**). This method is a modified version of D-REX [27] without ranking feedback, and can be seen as an instance of our formalism where the only counterfactuals are noisy deformations.

Environments. We simulate humans with various limitations and test our algorithm across four learning environments (see Fig. 2). In *Lavaworld* the human teleoperates a 2D point mass towards a goal. The human wants to avoid lava, but has a limited range of visibility around the point mass. In *Cartpole* the human tries to keep an inverted pendulum upright, and in *Lander* the human lands a craft on the surface of the moon (OpenAI Gym [33]). Within both *Cartpole* and *Lander* we limit how frequently the simulated human can change their input to mimic the response time of actual users. Finally, in *Swarm* the human navigates an obstacle course with three race cars (developed in Pybullet [34]). All three cars move simultaneously, but we limit the user to only control one car at a time. For each environment we hand-coded a discrete set of possible reward functions.

Results. In Fig. 2 we compute the robot’s belief using Equation (4) and then evaluate the Shannon entropy of that belief. We find that robots using **Ours** extrapolate what the human really wanted from their suboptimal demonstrations. Across a spectrum of different environments and different types of simulated users, we also find that robots which overestimate the human’s choice set (**BIRL**) are risk-seeking, while robots which underestimate the human’s choice set (**Noise**, **Ours**) are risk-averse.

VII. USER STUDY

Motivated by the application of assistive robotics we designed a user study with three manipulation tasks. Participants teleoperated a 7-DoF robot (Franka Emika) using a 2-DoF joystick. Users were unable to fully demonstrate what

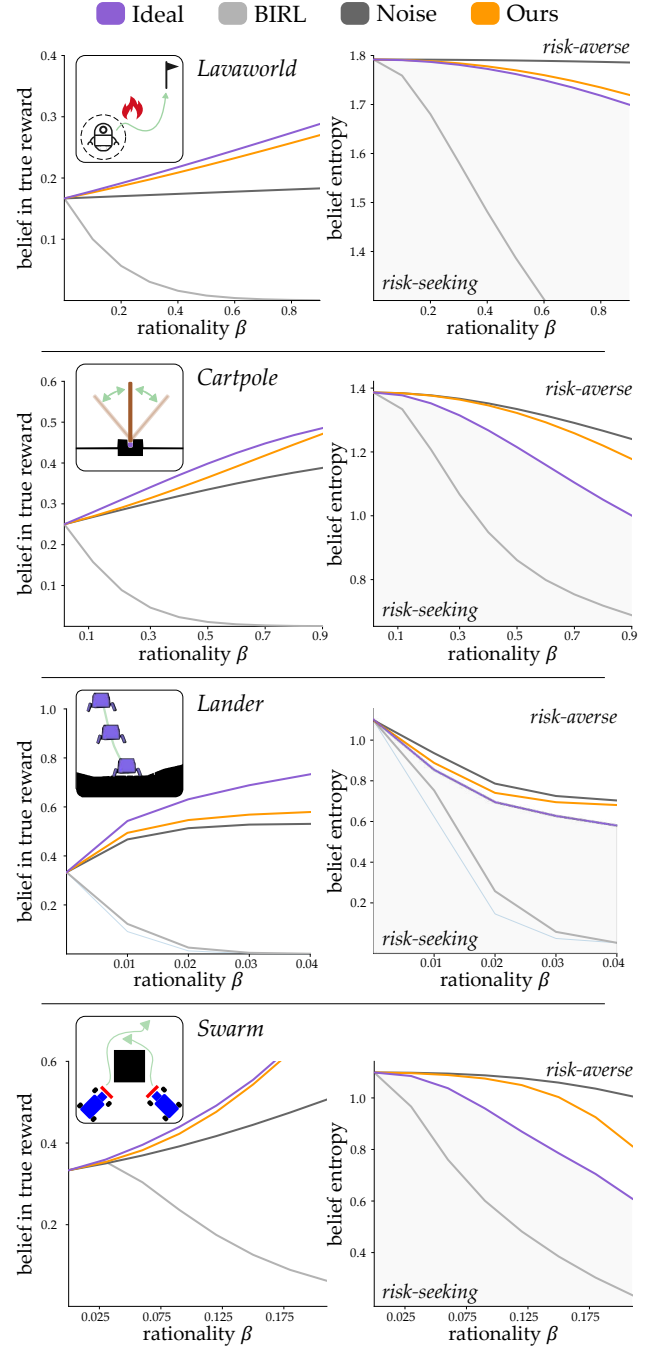


Fig. 2. Simulated users with different levels of rationality (β) interact in four environments. (Left) The robot’s learned belief in the human’s true reward. Higher values indicate the robot has learned what the human really meant. (Right) The robot’s confidence over its learned belief. Robots with less entropy than **Ideal** are overconfident in their prediction (risk-seeking), while robots that have more entropy are overly cautious (risk-averse).

they wanted because of challenges and limitations when controlling the robot (Fig. 3). Here the robot had a *continuous* space of rewards $r_{\theta}(\xi) = \theta \cdot \Phi(\xi)$, and our proposed approach leveraged Algorithm 1 with Metropolis–Hastings sampling to estimate the human’s true reward weights $\theta \in \mathbb{R}^k$.

Experimental Setup. Each participant attempted to demonstrate three manipulation tasks. In *Constraint*, users needed to drop-off trash in a waste bin. The bin was placed behind an obstacle that the robot must rotate around, but users could

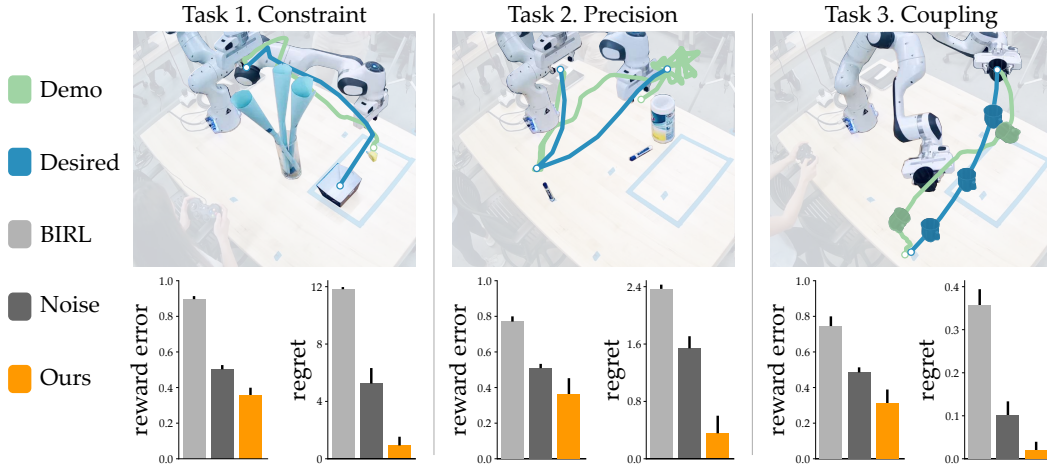


Fig. 3. Human demonstrations and objective results from our user study. (Top) Example user demonstrations and desired behavior for three different tasks. In each task the user is faced with a different challenge that prevents them from inputting their desired behavior. (Bottom) When the robot errs towards underestimating the human’s capabilities (**Ours**), it extrapolates what the human really wants from their limited demonstrations. This results in a more accurate estimate of the human’s true reward and lower regret across the robot’s learned behavior. Error bars indicate standard error about the mean.

only control the end-effector’s position. In *Precision*, users needed to carefully stack a marker. This was challenging because noise was injected into the robot’s motion. Finally, *Coupling* is our motivational task from Fig. 1, where users teach the robot to carry coffee upright. Here the robot’s translation was coupled with its rotation, so that moving the robot across the table caused the cup to inadvertently tilt.

Independent Variables. Participants first provided 3-5 tele-operated demonstrations \mathcal{D} for each task. We then compared what the robot learned with three different methods: **BIRL** [12], **Noise** [27], and **Ours** (Algorithm 1). **BIRL** assumes the human can provide any demonstration, while **Noise** and **Ours** err towards underestimating the human’s capabilities.

Dependent Measures – Objective. We obtained the *Error* $\|\theta - \hat{\theta}\|$ between the true reward weights θ and the mean of the estimated reward weights $\hat{\theta}$. We also computed *Regret*: $r_{\theta}(\xi^*) - r_{\theta}(\xi_{\mathcal{R}})$. Here ξ^* is the trajectory that maximizes r_{θ} , and $\xi_{\mathcal{R}}$ is the robot’s learned trajectory which maximizes $r_{\hat{\theta}}$. If the robot learns what the human really wants, $\xi_{\mathcal{R}} = \xi^*$.

Dependent Measures – Subjective. We administered a 7-point Likert scale survey after showing users what the robot learned (see Fig. 4). Questions were organized along five scales: how confident users were that the robot *Learned* their objective, how *Intuitive* the robot’s behavior was, whether the robot *Extrapolated* from their demonstrations, how trustworthy users thought the robot was (*Deploy*), and whether they would use the shown method again (*Prefer*).

Participants and Procedure. We recruited 10 subjects from the Virginia Tech student body to participate in our study (4 female, average age 22 ± 3 years). All subjects provided informed written consent prior to the experiment. We used a within-subjects design, where we counterbalanced the order of learning algorithms.

Hypothesis. *When human teachers face limitations, robots that learn from similar and simpler alternatives best extract what the human wants, and are preferred by users.*

Results – Objective. Our objective results for each task are

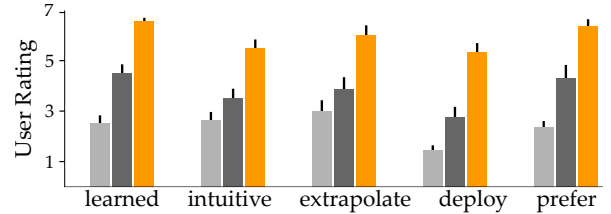


Fig. 4. Survey results from our user study. Higher ratings indicate user agreement. Posthoc comparisons revealed that the differences between our method and the baselines were statistically significant ($p < .01$).

displayed in Fig. 3. Lower errors indicate the robot’s reward estimate approaches the true reward, while lower regret shows that the robot’s resulting behavior matches the desired trajectory. With **BIRL**, the robot learned to replicate what the human demonstrated. But because of their limitations, users were unable to demonstrate their desired trajectory — and thus **BIRL** learned suboptimal behavior.

Under **Noise**, the robot compared the human’s demonstrations to similar alternatives (i.e. noisy perturbations). But just considering similar alternatives was not enough to reveal what the human meant. **Ours** outperformed **Noise** because it reasoned over both similar and simpler trajectories. Equipped with this choice set, **Ours** inferred the human’s reward (low error), and output near optimal trajectories (low regret).

Results – Subjective. We report the results of the user survey in Fig. 4. We received comments such as “*Method C (Ours) was by far the best and the only one I would trust,*” which highlight the perceived benefits of our approach.

VIII. CONCLUSION

We analyzed the consequences of over- and underestimating the human’s capabilities when learning rewards from demonstrations. Both our risk-sensitivity analysis and experimental results suggest that erring towards underestimating the human’s choice set results in safer and more inclusive learning. **Limitations:** we recognize that our properties for generating similar and simpler alternatives require hyperparameter tuning, which may depend on the environment.

REFERENCES

- [1] D. P. Losey, K. Srinivasan, A. Mandlekar, A. Garg, and D. Sadigh, "Controlling assistive robots with learned latent actions," in *IEEE International Conference on Robotics and Automation*, 2020, pp. 378–384.
- [2] L. V. Herlant, R. M. Holladay, and S. S. Srinivasa, "Assistive teleoperation of robot arms via automatic time-optimal mode switching," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2016, pp. 35–42.
- [3] B. D. Argall, "Autonomy in rehabilitation robotics: An intersection," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, pp. 441–463, 2018.
- [4] H. J. Jeon, S. Milli, and A. D. Dragan, "Reward-rational (implicit) choice: A unifying formalism for reward learning," *arXiv preprint arXiv:2002.04833*, 2020.
- [5] R. Freedman, R. Shah, and A. Dragan, "Choice set misspecification in reward inference," in *CEUR Workshop Proceedings*, 2020.
- [6] O. Evans, A. Stuhlmüller, and N. D. Goodman, "Learning the preferences of ignorant, inconsistent agents," in *AAAI*, 2016.
- [7] D. Kahneman and A. Tversky, "Prospect theory: An analysis of decision under risk," in *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, 2013, pp. 99–127.
- [8] S. Choi, S. Kariv, W. Müller, and D. Silverman, "Who is (more) rational?" *American Economic Review*, vol. 104, no. 6, pp. 1518–50, 2014.
- [9] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters, "An algorithmic perspective on imitation learning," *Foundations and Trends in Robotics*, vol. 7, no. 1-2, pp. 1–179, 2018.
- [10] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *International Conference on Machine Learning*, 2004.
- [11] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning," in *AAAI*, vol. 8, 2008, pp. 1433–1438.
- [12] D. Ramachandran and E. Amir, "Bayesian inverse reinforcement learning," in *IJCAI*, vol. 7, 2007, pp. 2586–2591.
- [13] C. Finn, S. Levine, and P. Abbeel, "Guided cost learning: Deep inverse optimal control via policy optimization," in *International Conference on Machine Learning*, 2016, pp. 49–58.
- [14] D. M. Taylor, *Americans With Disabilities: 2014*. US Census Bureau, 2018.
- [15] D. Gopinath, M. N. Javaremi, and B. D. Argall, "Customized handling of unintended interface operation in assistive robots," *arXiv preprint arXiv:2007.02092*, 2020.
- [16] R. M. Aronson, T. Santini, T. C. Kübler, E. Kasneci, S. Srinivasa, and H. Admoni, "Eye-hand behavior in human-robot shared manipulation," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 4–13.
- [17] K. Muelling, A. Venkatraman, J.-S. Valois, J. E. Downey, J. Weiss, S. Javdani, M. Hebert, A. B. Schwartz, J. L. Collinger, and J. A. Bagnell, "Autonomy infused teleoperation with application to brain computer interface controlled manipulation," *Autonomous Robots*, vol. 41, no. 6, pp. 1401–1422, 2017.
- [18] D. P. Losey and M. K. O'Malley, "Enabling robots to infer how end-users teach and learn through human-robot interaction," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1956–1963, 2019.
- [19] A. Bajcsy, D. P. Losey, M. K. O'Malley, and A. D. Dragan, "Learning from physical human corrections, one feature at a time," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 141–149.
- [20] S. Reddy, A. Dragan, and S. Levine, "Where do you think you're going?: Inferring beliefs about dynamics from behavior," in *Advances in Neural Information Processing Systems*, 2018, pp. 1454–1465.
- [21] A. Bobu, A. Bajcsy, J. F. Fisac, S. Deglurkar, and A. D. Dragan, "Quantifying hypothesis space misspecification in learning from human-robot demonstrations and physical corrections," *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 835–854, 2020.
- [22] D. H. Grollman and A. Billard, "Donut as I do: Learning from failed demonstrations," in *IEEE International Conference on Robotics and Automation*, 2011, pp. 3804–3809.
- [23] K. Shiarlis, J. Messias, and S. Whiteson, "Inverse reinforcement learning from failure," in *International Foundation for Autonomous Agents and Multiagent Systems*, 2016.
- [24] E. Bıyık, D. P. Losey, M. Palan, N. C. Landolfi, G. Shevchuk, and D. Sadigh, "Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences," *arXiv preprint arXiv:2006.14091*, 2020.
- [25] D. Sadigh, A. D. Dragan, S. Sastry, and S. A. Seshia, "Active preference-based learning of reward functions," in *Robotics: Science and Systems*, 2017.
- [26] S. Choi, K. Lee, and S. Oh, "Robust learning from demonstrations with mixed qualities using leveraged gaussian processes," *IEEE Transactions on Robotics*, vol. 35, no. 3, pp. 564–576, 2019.
- [27] D. S. Brown, W. Goo, and S. Niekum, "Better-than-demonstrator imitation learning via automatically-ranked demonstrations," in *Conference on Robot Learning*, 2020, pp. 330–359.
- [28] M. Kalakrishnan, P. Pastor, L. Righetti, and S. Schaal, "Learning objective functions for manipulation," in *IEEE International Conference on Robotics and Automation*, 2013, pp. 1331–1336.
- [29] A. Boularias, J. Kober, and J. Peters, "Relative entropy inverse reinforcement learning," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 182–189.
- [30] C. L. Baker, R. Saxe, and J. B. Tenenbaum, "Action understanding as inverse planning," *Cognition*, vol. 113, no. 3, pp. 329–349, 2009.
- [31] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.
- [32] D. P. Losey and M. K. O'Malley, "Trajectory deformations from physical human-robot interaction," *IEEE Transactions on Robotics*, vol. 34, no. 1, pp. 126–138, 2017.
- [33] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI Gym," *CoRR*, vol. abs/1606.01540, 2016.
- [34] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," <http://pybullet.org>, 2016–2019.