Here's What I've Learned: Asking Questions that Reveal Reward Learning

SOHEIL HABIBIAN, ANANTH JONNAVITTULA, and DYLAN P. LOSEY, Virginia Tech

Robots can learn from humans by asking questions. In these questions, the robot demonstrates a few different behaviors and asks the human for their favorite. But how should robots choose which questions to ask? Today's robots optimize for *informative* questions that actively probe the human's preferences as efficiently as possible. But while informative questions make sense from the robot's perspective, human onlookers may find them arbitrary and *misleading*. For example, consider an assistive robot learning to put away the dishes. Based on your answers to previous questions this robot knows where it should stack each dish; however, the robot is unsure about right height to carry these dishes. A robot optimizing only for informative questions focuses purely on this height: it shows trajectories that carry the plates near or far from the table, regardless of whether or not they stack the dishes correctly. As a result, when we see this question, we mistakenly think that the robot is still confused about where to stack the dishes! In this article, we formalize active preference-based learning from the human's perspective. We hypothesize that—from the human's point-of-view —the robot's questions *reveal* what the robot has and has not learned. Our insight enables robots to use questions to make their learning process *transparent* to the human operator. We develop and test a model that robots can leverage to relate the questions they ask to the information these questions reveal. We then introduce a tradeoff between informative and revealing questions that considers both human and robot perspectives: a robot that optimizes for this tradeoff actively gathers information from the human while simultaneously keeping the human up to date with what it has learned. We evaluate our approach across simulations, online surveys, and in-person user studies. We find that robots, which consider the human's point of view learn just as quickly as state-of-the-art baselines while also communicating what they have learned to the human operator. Videos of our user studies and results are available here: https://youtu.be/tC6y_jHN7Vw.

CCS Concepts: • **Computing methodologies → Active learning settings**; • **Human-centered computing → Collaborative interaction;**

Additional Key Words and Phrases: Human-robot interaction, reward learning, active learning, trust and interpretability

ACM Reference format:
Soheil Habibian, Ananth Jonnavittula, and Dylan P. Losey. 2022. *Here's What I've Learned:* Asking Questions that Reveal Reward Learning. *ACM Trans. Hum.-Robot Interact.* 11, 4, Article 40 (September 2022), 28 pages.
https://doi.org/10.1145/3526107

1 INTRODUCTION

Imagine that you are teaching a robot arm to put away the dishes (see Figure 1). You want the robot to carry these dishes close to the table—so they will not break if they fall—and to stack the dishes

Authors' address: S. Habibian, A. Jonnavittula, and D. P. Losey, Department of Mechanical Engineering, Virginia Tech, 635 Prices Fork Rd, Blacksburg, VA 24060; emails: {habibian, ananth, losey}@vt.edu.

2573-9522/2022/09-ART40
https://doi.org/10.1145/3526107

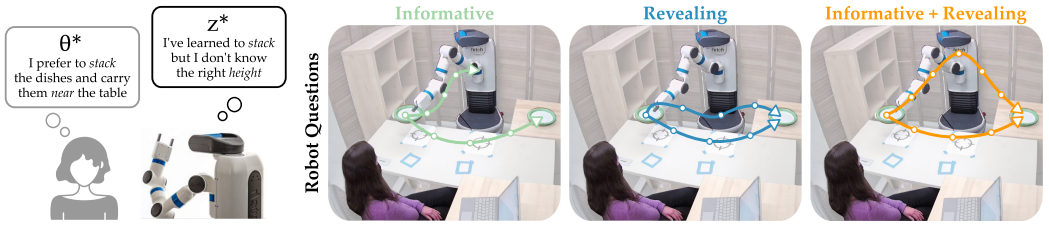

Fig. 1. A robot asking multiple-choice questions to learn the human's reward function. The human's true reward is parameterized by θ^*, while the robot's current understanding of that reward is captured by z^*. On the right we show the robot's full trajectory for each question. Within the state-of-the-art, the robot optimizes for **Informative** questions that maximize its information gained about θ^*. We formulate the opposite perspective, in which the robot purposely chooses **Revealing** questions to communicate its learning z^* to the human. We then combine both objectives: **Informative + Revealing** robots tradeoff between robot and human perspectives, and select questions that both elicit user information and reveal the robot's learning. Within this example the robot's question should emphasize that it already knows to stack the dishes while gathering clear feedback about how high to carry these dishes.

on top of each other. A natural way for the robot to learn your preferences is by *asking questions*. In each question, the robot shows you two different ways to put away the dishes and asks for your favorite. Based on your answers so far, the robot is (a) confident where it should stack the dishes and (b) unsure how high above the table to carry these dishes. Of course, the robot knows what it has learned—but how do *you know* whether or not this robot understands your preferences?

Recent work on active preference-based learning explores how robots can leverage questions to learn from humans [9–11, 14, 18, 23, 43]. Within the state-of-the-art the robot actively selects questions to gather as much as information as possible. Although this makes sense from the robot's point of view, we will experimentally demonstrate across multiple environments that it results in questions that humans find misleading. In our running example shown in Figure 1, the most *informative* question is for the robot to show one trajectory that moves close to the table (and then stacks the dishes), and a second trajectory that always remains high above the table (and does not stack the dishes as a result). This question is informative because it elicits clear feedback about the preferred height, which the robot is currently unsure about. But the question also *misleads* the human: When we see this question, we incorrectly think the robot is still unsure about both the preferred height and whether to stack the dishes.

In this article, we consider how robot questions influence the human's perception of the robot. Our key insight is that—from the human's perspective—

The questions that the robot asks reveal *what the robot does and does not know.*

Applying this insight enables robots to actively pick questions that keep the human up-to-date with the robot's learning. Put another way, now the robot can purposely select questions to reveal its learning process. Returning to our running example in Figure 1, the most *revealing* question is for the robot to show two similar trajectories that both accurately stack the dishes while carrying those dishes at varying heights. This question is revealing because it expresses what the robot knows (stacking the dishes) and what the robot is unsure about (the carrying height). But while this question no longer misleads the human, the human's answer to this question is not very useful for the robot. Looking again at Figure 1, no matter which trajectory we choose, our answer tells the robot *no new information* about whether to carry the dishes high or low.

We therefore introduce a spectrum between the robot's perspective (optimizing for informative questions) and the human's perspective (optimizing for revealing questions). Taking both human

and robot perspectives into account leads to *informative and revealing* questions. Returning to our example, an informative and revealing question consists of one trajectory that always remains close to the table and stacks the dishes, and a second trajectory that initially carries the dishes high in the air before returning to the table to stack the dishes. This question gathers meaningful information about the human's preferred height (i.e., the two options have different heights), but simultaneously reveals what the robot currently knows (i.e., both options correctly stack the dishes). Hence, the benefit of our proposed approach is that it identifies questions which efficiently gain information while also communicating what the robot is learning back to the user. As we will show, this lets the robot learn just as efficiently as state-of-the-art active learning baselines, but now with the added dimension of proactively enabling humans to reason over what their robot is learning.

Communicating what the robot is learning is an important step toward interpretable human-robot interaction [4]. As the human infers what their robot partner has learned (and what it is still confused about) the human can more confidently predict how the robot will behave. More generally, we view the questions that the robot asks as one avenue toward establishing *trust* and making robot learning *transparent* to human operators. In practice, our approach has two key *benefits* that both stem from communicating the robot's learning back to the human. First, the human can now determine when the robot has learned the right thing—and choose to deploy the robot accordingly. Second, the human can now infer which parts of the task the robot is still confused about, and focus additional teaching on these features. Overall, we make the following contributions:

Formalizing Revealing Questions. We formulate how humans learn from robot's questions, and introduce a cognitive model that robots can leverage to relate the questions they ask to the information those questions reveal. We then conduct an online survey to test our model and verify that revealing questions successfully convey robot learning to the human.

Combining Informative and Revealing Questions. We derive an optimization approach for selecting informative and revealing questions that gather human preferences while conveying robot learning. This proposed approach considers both human and robot perspectives, and allows designers to tradeoff along a spectrum between both extremes. We theoretically prove that this algorithm converges toward asking revealing questions over repeated interactions.

Comparing to Baselines. We conduct a series of simulations and in-person user studies to compare our approach to state-of-the-art alternatives. We find that when robots only optimize for informative questions, these questions tell the *human* no more about the robot's learning than simply asking random questions. Conversely, when robots only optimize for revealing questions, these questions tell the *robot* no more about the human's preferences than random questions. But when the robot leverages our informative and revealing approach, it satisfies both objectives: the robot learns efficiently while also communicating what it has learned.

2 RELATED WORK

Robots learn what their human partner wants by observing the human's inputs. These inputs can take many different forms [25], from physical demonstrations to verbal commands to pairwise comparisons. We focus on the latter, and explore how robots should ask questions to learn the human's underlying reward function. We consider the *bi-directional* flow of information: the robot learning from the human and the human inferring what their robot is learning. Our research builds on prior work that studies how robots actively select questions, as well as research on how robots communicate what they have learned to nearby human operators.

2.1 Learning Reward Functions from Questions

Inverse Reinforcement Learning. We consider scenarios where the human has in mind a reward function, and the robot is trying to learn this reward function from the human's behavior. One common approach here is for the human to input expert *demonstrations* — i.e., for the human to kinesthetically show the robot how it should perform the task. Inverse reinforcement learning enables robots to infer the human's reward function from these demonstrations [1, 36, 41, 53]. But often it is challenging for humans to *accurately* demonstrate what they want. Physical limitations constrain the human [26, 50], particularly when guiding the robot requires carefully coordinating the motion of multiple interconnected joints [3, 5]. Cognitive biases also affect the quality of human demonstrations: for instance, people prefer for autonomous cars to be more risk-averse than their own driving demonstrations [8, 29]. So while we draw from previous work on inverse reinforcement learning, we now apply this theory to a different form of human input—the human's answers to the robot's questions.

Active Preference-Based Learning. Each question consists of the robot showing the human a few trajectories and asking the human to compare these options. For example, the robot could display three or more trajectories, and ask the human to rank these trajectories in order of personal preference [11, 24]. Similar to prior research, within our simulations and user studies, we specifically consider *pairwise* comparisons where the robot only shows *two* options [9, 10, 14, 23, 25, 43, 50, 52]. Of course, there are an infinite number of possible questions; how does the robot choose which of these questions to ask? Existing work on active preference-based learning enables the robot to intelligently select questions that efficiently learn about the human's reward [2, 18, 38, 46, 48]. One state-of-the-art approach here is for the robot to select questions that optimize for the expected information gain [9, 10], so that no matter which answer the human chooses, the robot gathers meaningful feedback about the human's reward function. Robots proactively select these questions to accelerate their learning and reduce the number of queries the human must answer.

User-Friendly Questions. Prior work on human–human teaching emphasizes that poorly crafted questions can lead to confusion [40, 49]. Unfortunately, the questions generated by a robot trying to optimize its learning may be misleading or confusing to the human. Most related to our approach is research on how robots that actively choose questions should account for the *human-in-the-loop*. Recent works recognize that human answers are inevitably noisy and imperfect [21], and so we can improve both the robot's learning and human's experience by purposely asking easier questions [9, 10, 13, 39], by allowing users to explain their answers [7], or by explicitly accounting for the time constraints that humans face when answering questions [12]. Each of these works takes a step toward user-friendly interactions. But while today's robots account for how humans answer questions, they still only consider *one direction of information transfer*: from the human to the robot. Unlike prior research, we now focus on bi-direction information transfer. We hypothesize that the questions that the robots asks influence what the human thinks the robot knows—enabling us to *close the loop* and leverage questions to convey robot learning back to the human.

2.2 Communicating Robot Learning through Robot Motions

Legible Motion. Our key insight is that humans make inferences about the robot's latent state based on the robot's questions. Recalling that these questions are composed of a set of trajectories, our insight claims that humans infer what the robot is thinking based on the motions the robot is showing. Prior work from cognitive science suggests that humans naturally ascribe goals, plans, and objectives to robot motion [6, 15, 20, 44]. For instance, if we see a robot arm moving directly toward a cup, we may infer that the robot wants to reach that cup. Recent research within robotics

Table 1. Key Variables and their Definition

$\xi \in \Xi$	robot trajectory
$Q = \{\xi_1, \ldots, \xi_N\}$	question where the robot shows N trajectories $\xi_1, \ldots, \xi_N$
	in our experiments we set $N = 2$ and allowed "*I don't know*" responses
$q \in Q$	human's answer to the robot's question
$\theta^* \in \mathbb{R}^d$	human's preferences the robot is trying to learn
$b_{\mathcal{R}}(\theta)$	robot's learned belief over what θ^* is
$z^* \in Z$	parameterization of the robot's belief $b_{\mathcal{R}}$
$b_{\mathcal{H}}(z)$	humans's belief over what z^* is, i.e., over what the robot has learned

leverages this insight to purposely choose motions that communicate the robot's reward function [19, 22, 31] or the robot's capabilities [30, 47]. Similarly, our approach takes advantage of the robot's behavior (i.e., the different trajectories in each question) to communicate what the robot has learned (i.e., the robot's estimate of the human's reward function).

Multimodal Feedback. We recognize that the robot's physical behavior is one of many modalities for conveying feedback to the human. Robots can also leverage separate channels such as augmented reality [42, 51] or natural language [35, 47] to relay information. But these different feedback modalities are not mutually exclusive; instead, they form complementary ways to intuitively bring the human into the learning loop. Although here we will specifically study how the robot's questions (i.e., the robot's physical behavior) communicates information, we ultimately see this feedback as part of the larger spectrum of interpretable robot learning [4].

3 FORMALIZING QUESTIONS FROM ROBOT AND HUMAN PERSPECTIVES

In this section, we formulate our problem setting from Figure 1. Our goal is for a robot partner to select questions that proactively gather information from the human while also revealing what the robot is learning. We start with the *robot's perspective*: This robot is asking questions to learn the user's preferred behavior. We then introduce the *human's perspective*: This human reasons about what the robot has learned based on the questions the robot asks. We list the key variables we use to characterize our bi-directional information exchange in Table 1.

Robot Perspective. The robot is learning how it should behave. More formally, the robot is trying to learn the human's *reward function* R. This reward function assigns scores to *trajectories* $\xi \in \Xi$, so that learning R enables the robot to identify the optimal, highest scoring trajectory to deploy.

Following prior work on inverse reinforcement learning [1, 36, 53], we formulate the human's reward function as a linear combination of features f weighted by preferences θ:

$$R(\xi, \theta) = \theta \cdot f(\xi), \quad \|\theta\| = 1. \tag{1}$$

Here, the *features* $f : \Xi \rightarrow \mathbb{R}^d$ are task-related metrics that the robot measures. For instance, in our motivating example one feature is the height of the robot's end-effector. Although the robot knows these features, it does not know which features the human prefers. *Preferences* are captured by the unit vector $\theta \in \mathbb{R}^d$, and we denote the human's true preferences as θ^*. Returning to our motivating example, the human prefers for the robot to move close to the table. Put in terms of Equation (1), this translates to $\theta^*_{height} < 0$, which indicates that higher trajectories (i.e., trajectories with an increased height feature) will receive negative rewards from the human.

Questions. When the robot first interacts with a user it does not know their actual preferences θ^*. But robots can intuitively learn the user's preferences by asking questions. These questions are multiple choice: The robot shows the human N different trajectories, and asks the human to

pick one trajectory that best matches their preferred behavior. Here, $Q = \{\xi_1, \ldots, \xi_N\} \in \mathcal{Q}$ is the robot's *question* and $q \in Q$ is the human's *answer*. Let $\mathcal{Q}$ be a discrete set of all questions available to the robot: similar to [9, 43], we generate these questions offline by randomly sampling and pairing robot trajectories.[1] The robot reasons across the questions it has asked and the answers the human has given to infer a continuous probability distribution over what θ^* is. We refer to this distribution as the *robot's belief*. After i questions, the robot's belief is:

$$b_{\mathcal{R}}^{i+1}(\theta) = P(\theta \mid Q_1, \ldots, Q_i, q_1, \ldots, q_i). \tag{2}$$

By definition, $b_{\mathcal{R}}^{i+1}(\theta)$ expresses how likely it is that the user's true preferences are θ (i.e., the likelihood that $\theta^* = \theta$) given the robot questions $Q_1, \ldots, Q_i$ and the human answers $q_1, \ldots, q_i$.

Human Perspective. Prior work studies Equations (1) and (2), and shows how the *robot* should actively select questions to infer θ^* [2, 7, 10, 21, 43]. Moving beyond prior work, we now consider how these questions affect the *human's perception* of the robot's learning. When the robot asks questions about a specific feature of the task (i.e., each $\xi \in Q$ carries the dishes at a different height), the human thinks the robot is uncertain about this feature. Conversely, when the robot displays questions that keep specific features consistent (i.e., the robot stacks the dishes in every $\xi \in Q$), the human may guess the robot is confident about this feature.

We hypothesize that the questions the robot asks reveal what the robot has learned. Recall that the robot is learning a belief $b_{\mathcal{R}}$. If the human could directly observe this belief, the human would understand exactly what the robot has learned. In practice, however, $\theta \in \mathbb{R}^d$ is a continuous variable, and so the robot's belief over θ is a *continuous* probability distribution. We therefore introduce $Z \subset \mathbb{R}^m$, a compact, *low-dimensional representation* of the robot's belief space. Here, different values $z \in Z$ are associated with different beliefs $b_{\mathcal{R}}$, and the robot's current belief $b_{\mathcal{R}}^i$ is parameterized by z^*. To give an example, consider trying to model $b_{\mathcal{R}}^i$ as a normal distribution, where z^* is a vector that contains the mean and standard deviation of this model. In this example, z^* is a low-dimensional representation because it takes the continuous probability distribution $b_{\mathcal{R}}^i$ and compactly approximates it with $2d$ parameters.

In the same way that humans cannot directly observe the robot's learning, users cannot directly observe z^*. But humans can infer a probability distribution over what z^* is based on the questions the robot has asked. We refer to this distribution as the *human's belief*:

$$b_{\mathcal{H}}^{i+1}(z) = P(z \mid Q_1, \ldots, Q_i, q_1, \ldots, q_{i-1}). \tag{3}$$

Intuitively, $b_{\mathcal{H}}(z)$ expresses how likely it is *from the human's perspective* that the robot has learned z (i.e., the likelihood that $z^* = z$) given the robot's questions and the human's previous answers.

Representing Robot Learning. Our proposed approach is not tied to any specific choice of Z. But, for the sake of clarity, we here present one representation that we will leverage in our experiments.

Instead of thinking about the robot's learning directly in terms of the robot's belief $b_{\mathcal{R}}$, we focus on the *behavior* that this belief induces. For any estimate $\theta \sim b_{\mathcal{R}}$ of the human's preferences, the robot can solve for the corresponding trajectory that maximizes R:

$$\xi_\theta = \arg\max_{\xi \in \Xi} R(\xi, \theta). \tag{4}$$

Here, ξ_θ is an optimal trajectory given preferences θ. Imagine that the robot has currently learned $b_{\mathcal{R}}^i$. If we were to deploy this robot now, and the robot solves for the optimal trajectory

[1] For robot arms, we start with a trajectory ξ_0 and then add varying levels of zero-mean Gaussian noise to the waypoints. This produces trajectories ξ, which we randomly combine to form questions $Q \in \mathcal{Q}$.

given what it has learned, the expected features of this trajectory are:

$$\mu_i = \mathbb{E}_{\theta \sim b^i_{\mathcal{R}}}\Big[f(\xi_\theta)\Big]. \tag{5}$$

Similarly, the variance of these features across multiple runs is:

$$\sigma_i^2 = \mathrm{Var}_{\theta \sim b^i_{\mathcal{R}}}\Big[f(\xi_\theta)\Big]. \tag{6}$$

Viewed together, Equations (5) and (6) determine (i) what the robot would do if deployed and (ii) what features of the task the robot is and is not confident about. We combine both the mean μ_i and standard deviation σ_i to form the vector z^*:

$$z^* = (\mu_i, \sigma_i) \in \mathbb{R}^{2d}. \tag{7}$$

Under this representation, $z^* \in Z$ is a compact embedding of the *features induced* by the robot's current belief $b^i_{\mathcal{R}}$. Our choice of Z contains both the behavior the robot has learned to perform as well as the parts of this behavior the robot is confused about.

To better understand z^*, we return to our motivating example in Figure 1. Here, the task features are *stacking distance* and *height*. If the robot is completely confident it should minimize stacking distance, but is unsure about the right height, the components of z^* could be:

$$\mu_i = (0 \text{ stacking distance}, 0.5 \text{ height}), \quad \sigma_i = (0 \text{ stacking distance}, 0.3 \text{ height}). \tag{8}$$

Communicating z^* back to the human has two parts: what the robot has learned (i.e., μ_i) and how confident the robot is in that learning (i.e., σ_i). An intelligent robot should choose questions to make z^* transparent and bring the human into the learning loop.

4 MODELING HOW HUMANS INTERPRET QUESTIONS

Our work is based on the hypothesis that humans gather information about the robot by observing the questions the robot asks. In this section, we formalize our hypothesis by introducing a human model that robots can leverage to solve Equation (3). Intuitively, this model defines how humans *interpret* robot questions; more specifically, this model relates the questions the robot asks to the information those questions reveal. A robot that optimizes for this model chooses *revealing* questions that proactively communicate what it has learned.

Bounded Memory. Within Equations (2) and (3), the human is learning in response to the robot's learning. This is an instance of *co-adaptation*. Prior research from economics [27], game theory [37], and human–robot collaboration [34] suggests that co-adaptive humans base their decisions on the recent actions of their partner. Applying this bounded memory model to our setting, we assume that the human focuses on the robot's *last k questions*. We also assume that—from the human's perspective—the robot's learned behavior z *is constant* during these k questions. Applying both assumptions, Equation (3) now reduces to[2]:

$$b^{i+1}_{\mathcal{H}}(z) = P(z \mid Q_{i-k+1}, \ldots, Q_i). \tag{9}$$

Employing Bayes' Theorem and conditional independence:

$$b^{i+1}_{\mathcal{H}}(z) \propto P(z) \prod_{\tau=i-k+1}^{i} P(Q_\tau \mid z). \tag{10}$$

Equation (10) captures what the human thinks the robot has learned based on the robot's k most recent questions. Here, $P(z)$ is the human's *prior* over what the robot knows, and $P(Q \mid z)$ is the

[2]Notice that Equation (9) no longer depends on answers q. This is a result of the human's assumption that z is locally constant, meaning that the human's last answer (or answers) did not affect the robot's learning.

likelihood that the robot asks question Q given that it has learned z. The last step in our human model is formulating this likelihood function $P(Q \mid z)$.

Question Likelihood. To motivate our choice of likelihood function, we return to our running example where the robot is learning about *height* and *stacking distance*. Imagine that the robot shows you a question with two trajectories: One trajectory moves close to the table and stacks the dishes, while the other trajectory moves high above the table and places the dishes in different locations. Because both *height* and *stacking distance* vary in this question, you may infer that the robot is uncertain about both features of the task. By contrast, if the robot shows you a question where both of the trajectories stack the dishes, you will likely assume that the robot knows *stacking distance*. Driven by this intuition, we propose to model $P(Q \mid z)$ by comparing the actual behavior shown in question Q to the learned behavior captured by z:

$$P(Q \mid z) \propto \exp\{-\|(\mu_Q, \sigma_Q) - z\|^2\}. \tag{11}$$

Here, d is the number of features, $(\mu_Q, \sigma_Q) \in \mathbb{R}^{2d}$ is a vector formed by concatenating μ_Q and σ_Q, and $z \in \mathbb{R}^{2d}$ is our parameterization from Equation (7). Recall that each question Q is composed of a set of trajectories $Q = \{\xi_1, \ldots, \xi_M\}$. In Equation (11), μ_Q is the mean features across the trajectories shown in question Q, and σ_Q is the standard deviation of these features:

$$\mu_Q = \mathbb{E}_{\xi \in Q}\Big[f(\xi)\Big], \quad \sigma_Q^2 = \mathrm{Var}_{\xi \in Q}\Big[f(\xi)\Big]. \tag{12}$$

In practice, Equation (11) states that *humans expect the robot to ask questions that match the behavior the robot has learned*. Consider our motivating example where the robot's learned behavior is summarized in Equation (8). Under our model, the human thinks it *likely* that this robot will ask questions where each trajectory minimizes stacking distance and demonstrates a different height, but *unlikely* that the robot will ask questions that vary the stacking distance.

Revealing Questions. What if the robot wants to ask questions that purposely convey its learned behavior to the human? Put another way, what if the robot wants to be as *interpretable* as possible? Recall that the robot's current belief $b^i_{\mathcal{R}}$ is compactly represented by z^*. We leverage our model from Equations (10)–(12) to proactively reveal z^* to the human:

$$Q^*_{\text{reveal}} = \arg\max_{Q \in \mathcal{Q}} \; b^{i+1}_{\mathcal{H}}(z^*). \tag{13}$$

Here, $\mathcal{Q}$ is the set of feasible questions, and $Q^*_{reveal} \in \mathcal{Q}$ is the question that greedily maximizes the human's belief in z^*. We emphasize that—because the robot in Equation (13) is optimizing for transparency—the human's answer to Q_{reveal} may not be particularly informative for the robot. Instead, this robot is only considering the human's perspective. From the human's point of view, Q_{reveal} is interpretable because it demonstrates a set of trajectories, which accurately reflect what the robot is confident about and what the robot is still trying to learn.

Summary. With Equations (10)–(12), we formalize our key insight: We introduce a cognitive model that relates the questions that the robot asks to the information those questions reveal. As shown in Equation (13), this model enables robots to proactively optimize for revealing questions. We note that we can leverage this approach when the robot only asks a single question, or when the robot asks the human multiple multiple questions in sequence.

5 DO HUMANS LEARN FROM QUESTIONS?

We have proposed a model to formalize the relationship between robot questions and the information they convey. Here, we conduct an online user study to test this model and explore whether humans gather information from robot's questions. Within this study a robot arm moves across a table to learn three human preferences (see Figure 2): the *height* of the robot's end-effector, the

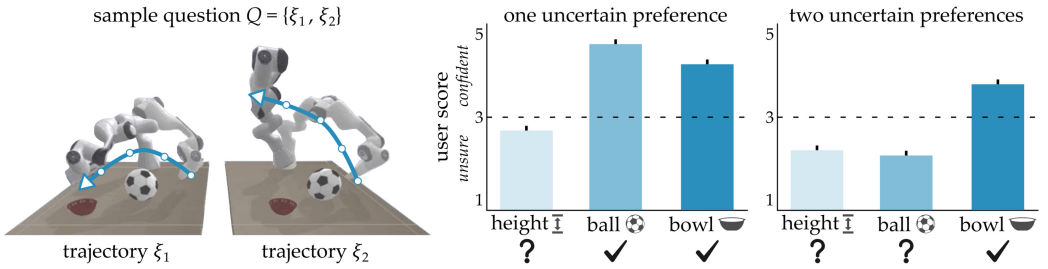

Fig. 2. Questions reveal information. (Left) An example question from our Amazon Mechanical Turk user study. (Right) Participants responded to sequences of these robot questions by indicating what they thought the robot had learned. High scores indicate the human believed the robot was confident about a preference, while low scores indicate the human perceived the robot as unsure. In reality, we displayed (i) robots that were only unsure about *height* and (ii) robots that were unsure about both *height* and *ball*. Across both conditions users correctly inferred what the robot had learned based only on the questions the robot asked. Differences in user score were statistically significant ($p < .001$). Error bars show SEM. For this and other plots, we include additional statistical analysis in Appendix A.

distance between the robot and the *ball*, and whether or not the robot ends above the *bowl*. The robot asks the human a sequence of questions to learn their preferences. *We measure what these questions reveal to the human.* After viewing the questions, participants specify which preferences they think the robot understands or is unsure about. The robot optimizes for questions that are as revealing as possible: If humans follow our cognitive model, we anticipate that their understanding of what the robot knows will match what the robot has *actually* learned.

Independent Variables. We varied the robot's uncertainty across two scenarios. In the first scenario, the robot was unsure about *height* but confident about *ball* and *bowl*. In the second scenario, the robot was unsure about both *height* and *ball* while remaining confident about *bowl*. We never informed our study participants which preferences the robot did and did not know.

Participants and Procedure. We conducted a within-subjects study on Amazon Mechanical Turk and recruited 50 participants. All participants had at least a 99% approval rating. We first described the high-level task, and explained the *height*, *ball*, and *bowl* preferences. We then asked comprehension questions to ensure that all participants had carefully read and understood our instructions. The remainder of the survey was divided into six sections: each section included videos of three robot questions (similar to the sample question from Figure 2). In half of the sections, the robot was uncertain about one feature, and in half of the sections the robot was uncertain about two features. We randomized the order of the sections.

The robot leveraged our cognitive model and Equation (13) to choose which questions to ask. This *revealing* robot optimized for questions that communicate robot learning to the human—if users were unable to interpret these questions, we doubted users would gather information from other, more ambiguous questions.

Dependent Measures. After viewing each set of robot questions, participants indicated what they thought the robot had learned on a 5-point scale. Here, a score of 1 denotes that the human believed the robot was completely unsure about a given preference, while a score of 5 denotes that the human thought the robot fully understood that preference. We aggregated all 50 user responses to determine the mean score and **standard error of the mean (SEM)** for each preference.

Hypothesis. *When robots purposely ask revealing questions, humans will correctly infer what parts of the task the robot is confident or unsure about.*

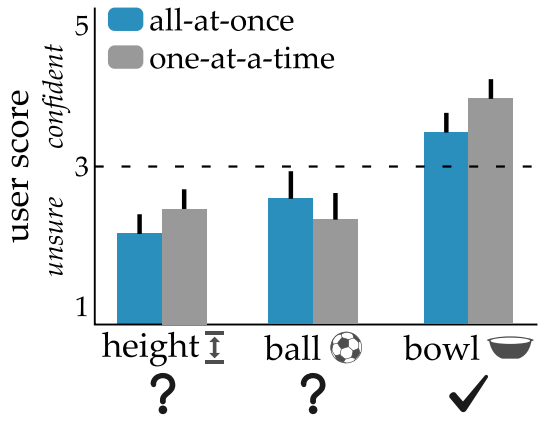

Fig. 3. Follow-up comparison between two types of revealing questions. In **all-at-once,** the robot asks questions that reveal what the robot knows about every preference, while in **one-at-a-time,** the robot asks questions that reveal a single preference per question. The robot is actually uncertain about *height* and *ball* preferences. We found that differences between the participants' perception of **all-at-once** and **one-at-a-time** were not statistically significant. Additional details can be found in Appendix A.2.

Results. Our results are visualized in Figure 2. When interpreting these results, remember that a user score *below* 3 indicates participants thought the robot was unsure about a preference, while a score *above* 3 indicates humans thought the robot was confident. Overall, we found that participants correctly inferred what the robot did and did not know based just on the questions the robot asked.

—*One uncertain preference.* When the robot was only uncertain about *height*, post hoc comparisons revealed that users perceived the robot as more unsure of *height* than either *ball* or *bowl* ($p < .001$). Participants scored the *ball* and *bowl* as 1.96 and 1.44 points more confident than *height*, respectively.

—*Two uncertain preferences.* When the robot was unsure about both *height* and *ball*, post hoc comparisons demonstrated that participants thought the robot was more uncertain about *height* and *ball* as compared to *bowl* ($p < .001$). Participants scored the *bowl* as 1.35 points more confident than *height* and 1.40 points more confident than *ball*.

Do Human's Focus on Specific Parts of Questions? When conducting this study, we wondered whether humans viewed robot questions holistically, or if they focused their attention on specific aspects of the robot's questions. Consider the example in Figure 2. Here, the robot's question shows two clearly different *heights*, but also two slightly different values for *bowl*: ξ_1 ends directly above the bowl, while ξ_2 is a bit to the left of the bowl. When participants saw this question, did they only learn about *height*, or did they also think about *bowl*?

Similar to prior work [5], we therefore performed a follow-up user study with **all-at-once** and **one-at-a-time** questions. We conducted this survey within the scenario where the robot was unsure about two preferences: *height* and *ball*. The **all-at-once** robot asked questions that conveyed what it had learned about every preference simultaneously (e.g., a question that varied both *height* and *ball*). By contrast, the **one-at-a-time** robot only conveyed its understanding of a single preference per question (e.g., the first question only varied *height*, and the second questions only varied *ball*). We recruited 50 users on Amazon Mechanical Turk for this follow-up study.

Our results are shown in Figure 3. Interestingly, participants interpreted **all-at-once** questions just as well as they interpreted **one-at-a-time** questions: The differences between these two

conditions were not statistically significant ($F(1, 49) = 0.20$, $p = .66$). This suggests that participants in our survey did not hone in on specific parts of robot questions, and robots could leverage a single question to convey multiple aspects of their learning. We recognize that a limitation of this study was that there were only three total preferences to ask about: We anticipate that as the number of preferences increases, the **all-at-once** approach may break down.

Summary. The key assumption underlining our work is that robot's questions reveal information to humans. We supported this hypothesis in an online user study: A total 50 participants correctly identified what the robot did and did not know based only on the questions the robot asked. Within this study, the robot purposely optimized for *revealing* questions using our cognitive human model from Equation (13). The effectiveness of these questions suggests that our model produces questions that the human can correctly interpret.

6 COMBINING INFORMATIVE AND REVEALING QUESTIONS

In the previous section, we demonstrated that robot questions can convey information to human onlookers. We generated these questions using Equation (13), which optimizes for *revealing* questions. On one hand, revealing questions communicate what the robot has learned *to the human*. On the other hand, these questions may not be particularly informative *for the robot*.

To see why, we return to our motivating example in Figure 1. Here, the robot is unsure about how high to carry the dishes, and so when it optimizes for a revealing question, it shows trajectories that carry the dishes at fluctuating heights. Perhaps one trajectory initially moves the dishes close to the table and then goes high, while the second trajectory initially moves high and then goes low. This question conveys to the human that the robot is unsure about height (since each $\xi \in Q$ moves at varying height), but the human's answer to this question provides *no new information* to the robot (because each $\xi \in Q$ has the same average height). From the robot's perspective, we want to ask a question that clearly elicits the human's preference: e.g., one trajectory that carries the dishes high above the table, and a second trajectory that always keeps the dishes near the table.

In this section, we formulate optimizing for revealing questions as one extreme of a spectrum. At the other end of this spectrum are robots that optimize for *informative* questions. Informative questions speed up robot learning but may mislead the human; revealing questions clarify what the robot has learned but may not elicit new information. We describe how robots tradeoff along this spectrum, intelligently balancing between giving and getting information.

Informative Questions. Optimizing for informative questions is well explored in prior work [9, 23, 43]. Following [9, 10], informative questions greedily maximize the robot's expected information gain about θ^*, the human's true preferences:

$$\begin{aligned} Q^*_{\text{info}} &= \arg\max_{Q \in \mathcal{Q}} I(\theta^*; q \mid Q, b^i_{\mathcal{R}}) \\ &= \arg\max_{Q \in \mathcal{Q}} H(\theta^* \mid Q, b^i_{\mathcal{R}}) - \mathbb{E}_{q \sim Q}\left[H(\theta^* \mid q, Q, b^i_{\mathcal{R}})\right]. \end{aligned} \tag{14}$$

Here, I denotes mutual information, H is the Shannon entropy, and $I(\theta^*; q \mid Q, b^i_{\mathcal{R}})$ quantifies the amount of information the robot will learn about θ^* if it observes human answer q given question Q and current estimate $b^i_{\mathcal{R}}$ [17]. The robot does not know the human's answer q when choosing which question to ask. Accordingly, in Equation (14), the robot chooses its question $Q^*_{\text{info}} \in \mathcal{Q}$ such that—regardless of what answer the human provides—observing that answer will optimize the expected amount of information the robot gains.

Trading-Off. A robot that optimizes Equation (14) focuses on purely *gaining information* from the human. Conversely, a robot that optimizes Equation (13) is only focused on *revealing information* to the human. In order to facilitate a bi-directional exchange of information, we here

introduce a spectrum between these two extremes. Robots can now identify questions that are both informative and revealing by optimizing:

$$Q^*_{\text{spectrum}} = \arg\max_{Q \in \mathcal{Q}} I(\theta^*; q \mid Q, b^i_{\mathcal{R}}) + \lambda \cdot b^{i+1}_{\mathcal{H}}(z^*). \tag{15}$$

In the above, $\lambda \geq 0$ is a hyperparameter that determines the relative importance of gaining or revealing information. When $\lambda = 0$, the robot only considers information gain, and if $\lambda \to \infty$, the robot thinks only about revealing information.

When Should Robots Gather or Reveal Information? Robots that leverage our approach trade-off along a spectrum between getting and giving information. We want to theoretically understand when Equation (15) leads to more informative questions or more revealing questions. Below we prove that robots, which leverage Equation (15) follow an underlying pattern: They ask their most informative questions the first time they interact with the human, and converge toward purely revealing questions over repeated interactions.

Proposition 1. *Assuming that the user has fixed preferences θ^*, and the user chooses answers that are consistent with their preferences, robots that optimize Equation (15) for informative and revealing questions converge toward asking revealing questions over time.*

Proof. We assume that each time the robot asks a question the human chooses a trajectory $q \in Q$ that maximizes their reward, so that $q \in \arg\max_{q' \in Q} R(q', \theta^*)$. Given this assumption, the first term from Equation (15) is submodular and monotone [33]. It directly follows that $I(\theta^*; q \mid Q, b^i_{\mathcal{R}}) \to 0$ as the interaction number increases (i.e., $i \to \infty$). Intuitively, each successive question and answer provides less new information about θ^* since the robot has already formed an increasingly accurate estimate from the previous human answers. At the limit $I(\theta^*; q \mid Q, b^i_{\mathcal{R}}) \to 0$, and the robot only optimizes for $b^{i+1}_{\mathcal{H}}(z^*)$, i.e., revealing questions. □

What Happens when Robots are Confident? As the robot continues to ask questions and receive answers it will converge to an estimate θ of the human's preferences so that $b_{\mathcal{R}}(\theta) \to 1$. But while the robot is confident in what it has learned, it still does not know if this θ is correct: perhaps the human incorrectly answered several questions, or perhaps the human's preferences have shifted during repeated interaction. Hence, it is important for the robot to leverage revealing questions and communicate its learning. At an extreme where the robot no longer has any uncertainty, the questions produced by Equation (15) will communicate what the robot has learned by rolling out trajectories that exactly match the learned behavior. As we will see in our user studies, these questions indicate to humans (i) what the robot has learned and (ii) when it is ready to be deployed.

Implementation and Hyperparameter Tuning. When a robot leveraging Equation (15) first interacts with the human it does not know what the human wants, and selects questions that proactively learn θ^*. Over time this robot gets a good grasp of the human's preferences, so that there is no longer any need to ask informative questions. At this point the robot continuously transitions toward revealing questions that communicate z^* back to the human.

There are two hyperparameters for designers to tune when implementing Equation (15). The first is λ, which arbitrates the relative importance of informative and revealing questions. The second is k from Equation (10), which models how many recent questions the human remembers. We will study the effects of both hyperparameters in the following sections. Based on these experimental results, we recommend that designers start with $\lambda = 1$ and $k \geq 3$. We find that optimizing for informative and revealing questions is largely *invariant* to the choice of k, while λ enables the designer to adjust how informative or revealing the robot is.

7 SIMULATIONS

We have formulated informative and revealing questions along a continuous spectrum, and proposed that robots should tradeoff between these extremes. However, it is not yet clear whether robots that optimize for *both* informative and revealing questions will do a good job of either one. Put another way, when the robot optimizes for Equation (15), are its questions rich enough to gain information while remaining transparent to the human onlooker? Here, we test our proposed approach in two environments with simulated users, and compare informative and revealing questions to the performance of state-of-the-art active learning baselines. We then conduct two separate simulations where we vary the hyperparameters in our approach: We experimentally find that humans do not need to remember every robot question in order to correctly interpret the robot's learning. We emphasize that these results are obtained from *simulated humans*, where we carefully control the human's response and simulate the human's behavior by applying existing cognitive human models. We will support our experimental findings from these simulated humans in Section 8, where we perform similar tests with actual human operators.

Baselines. We compare the performance of informative and revealing questions (**Ours**) to three different baselines. The first is a naïve robot that chooses questions uniformly at **Random**. The other two baselines are the extremes of our spectrum: one robot optimizes solely for **Informative** questions using Equation (14), and another robot optimizes for **Revealing** questions using Equation (13). We note that the revealing robot matches what we used in our online user study from Section 5, and the informative robot is taken from recent work on active preference learning [9, 10]. Hence, **Informative** is a *state-of-the-art* approach for active preference-based reward learning.

Simulated Users. For each robot question, we simulate the human's response. There are two components of this response: (i) what the human infers from the robot's question and (ii) how the human answers this question. To simulate what the human infers from the robot's question, we leverage the cognitive human model we developed in Section 4 and experimentally tested in Section 5. To simulate how the human answers the question, we apply the same approach as related work on active learning [7, 9, 21, 23, 43]. Here, the human is modeled as an approximately optimal decision maker. In each question, the robot shows two different trajectories, $Q = \{\xi_1, \xi_2\}$, and the human has three possible answers: trajectory ξ_1, ξ_2, or "**I don't know**" (**Idk**) [21]. The human noisily chooses the trajectory ξ that best agrees with their preferences, and in cases where the two trajectories produce about equal rewards, the human is more likely to select Idk:

$$P(q = \xi_i \mid Q, \theta) = \frac{\exp(R(\xi_i, \theta))}{\exp(R(\xi_i, \theta)) + \exp(1 + R(\xi_{-i}, \theta))},$$

$$P(q = \text{Idk} \mid Q, \theta) = P(q = \xi_1 \mid Q, \theta) \cdot P(q = \xi_2 \mid Q, \theta) \cdot (\exp(2) - 1).$$

In the above, we use $-i$ to denote the other trajectory, so that if $i = 1$ then $-i = 2$ (and vice versa). Our choice of human model is consistent with prior work on inverse reinforcement learning [25, 36, 53] and cognitive psychology [27, 28, 32].

Environments. We implement the simulated users alongside a robot arm and autonomous car [16]. The robot arm environment matches our previous user study (see Figure 2), where the human's unknown preferences are the robot's height, the distance to the ball, and if the robot ends above the bowl. In the driving environment, the human's preferences include the car's speed, the car's distance from obstacles, and the car's distance from the center of the lane. Across both environments, we simulate 100 users with uniformly randomly chosen preferences θ^*.

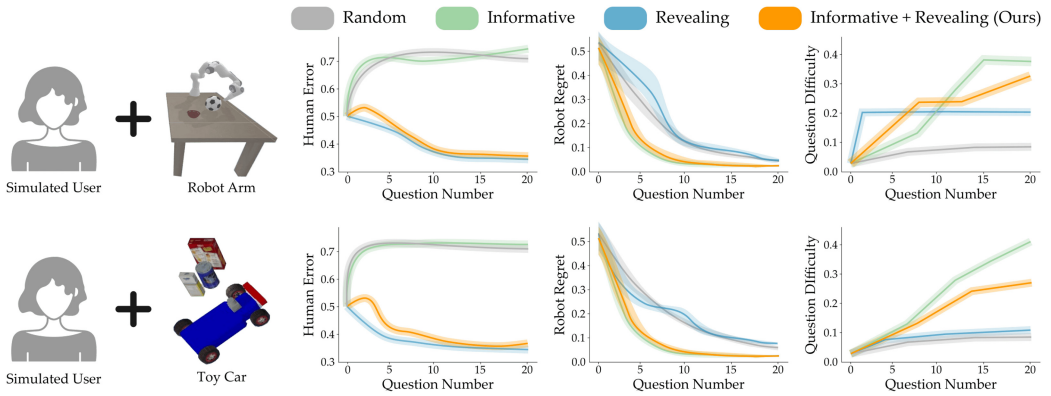

Fig. 4. Simulated humans answering 20 robot questions. *Human Error* refers to the difference between what the robot actually learned and what the human thinks the robot learned (i.e., the human's perspective). By contrast, *Robot Regret* captures the error between the robot's learned behavior and the human's desired behavior (i.e., the robot's perspective). When we consider both the robot and human perspectives, we generate questions that convey learning while gathering information (**Ours**). *Question Difficulty* denotes the fraction of questions in which the simulated human did not know how to answer and selectedIdk. Humans find questions generated by our approach less challenging to answer than informative questions. Note that lower is better in all plots. Shaded regions show standard deviation across 100 simulated humans. Statistical analysis of the results after all 20 questions were asked is located in Appendix A.3.

7.1 Comparison to State-of-the-Art Baselines

Our first set of simulations compares informative and revealing questions to naïve and state-of-the-art baselines. The results from these simulations are visualized in Figure 4. Note that these results are based on simulated humans that follow state-of-the-art cognitive models: to support these results, in Section 8, we will test our proposed approach across two in-person user studies.

Are Questions Revealing? We first consider the human's perspective, and determine which types of questions communicate the robot's learning back to the human. We measure *Human Error*, which captures the difference between what the robot really knows (i.e., z^*) and what the human thinks the robot knows (i.e., $b_{\mathcal{H}}$). This is a proxy measure of *interpretability*: ideally, the human will understand exactly what their robot has learned so that they can correctly anticipate what their robot will do when it is deployed. Looking at Figure 4, we see that both **Ours** and **Revealing** are transparent to the simulated humans. By contrast, simulated users who interact with either **Random** or **Informative** are unsure about what the robot partner learned. Indeed, **Informative** questions tell the simulated human just as much about what the robot is learning as completely **Random** questions do.

Are Questions Informative? We next consider the robot's perspective, and determine which types of questions provide the most information about the human's latent preferences. We assess how quickly the robot learns by measuring *Robot Regret* [36]. Recall from Equation (4) that ξ_θ is the optimal trajectory given that the human's preferences are θ. Regret is the difference in reward between the human's preferred behavior and the robot's learned behavior: $\mathbb{E}_{\theta\sim b_{\mathcal{R}}}[R(\xi_{\theta^*},\theta^*) - R(\xi_\theta,\theta^*)]$. Regret will decrease over time because the robot is collecting more human answers and honing in on the human's preferred behavior θ^*. But we can *accelerate* this process by proactively asking insightful questions—referring to Figure 4, both **Our** robot and the **Informative** robot quickly learned the human's desired behavior. By contrast, a robot that asks **Revealing** questions learned at about the same rate as a robot that was asking questions completely at **Random**.

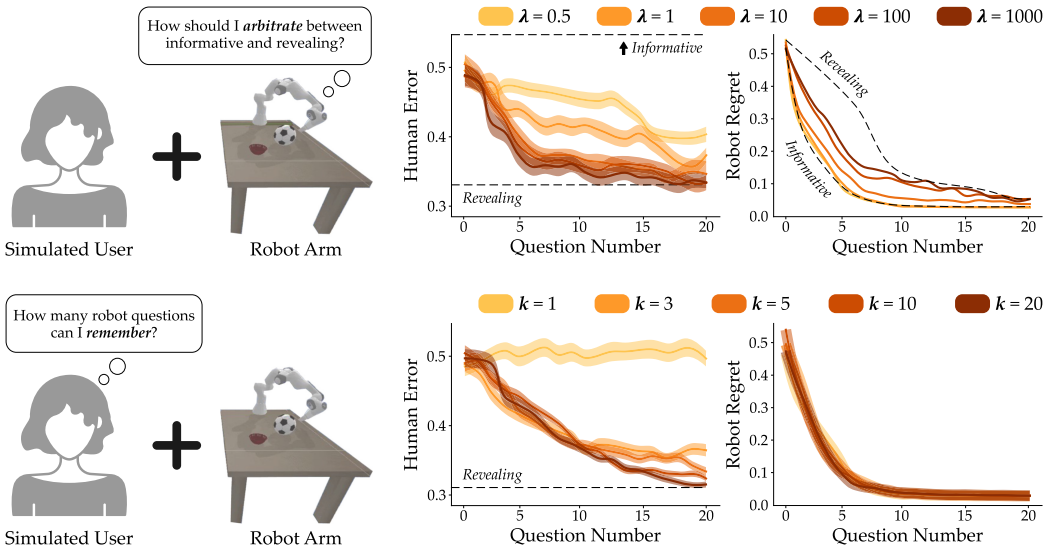

Fig. 5. Simulated humans interacting with our proposed approach. We tune two key hyperparameters and see how they affect the human's understanding of the robot (*Human Error*) and the robot's estimate of the human's preferences (*Robot Regret*). The dashed lines show the lower and upper bounds for revealing or informative baselines. (Top) We leave $k = 3$ while varying the relative importance of gaining and revealing information. At high values of λ, the robot favors questions that reveal information, and at low values of λ, the robot focuses on gaining information. Even at the extremes ($\lambda = 0.5, \lambda = 1,000$), our combined approach performs better than just asking informative or revealing questions. (Bottom) We leave $\lambda = 1$ and vary how many recent questions the simulated human reasons over. The more questions the human remembers the more information the robot can reveal; however, our approach still reaches the lower bound even if the human can only remember the last three questions. Standard deviation omitted from the top right for clarity. Additional statistical details after all 20 questions were asked is located in Appendix A.4.

Are Questions Confusing? Beyond both gaining and revealing information, we want to ensure that the robot's questions are *easy* for the human to answer. Prior work suggests that easy, user-friendly questions improve the human's experience and reduce the number of wrong answers, which, in turn, speeds up the robot's learning [9, 10, 13, 39]. We therefore assess *question difficulty*, which measures the fraction of time that the simulated humans answered Idk. When users pick Idk, the options are too hard to distinguish and the question is challenging for the human to clearly answer. In general, we find that **Random** and **Revealing** questions are quite easy for the simulated human to answer, while **Informative** and **Ours** get harder over time. **Ours** is slightly easier than the **Informative** baseline.

7.2 Effect of Hyperparameters on Our Approach

In the previous simulations, we choose $\lambda = 1$ and $k = 3$ for the hyperparameters of **Ours** and left these hyperparameters fixed across both environments. Now in our second set of simulations, we will explore how tuning these two hyperparameters affects our approach to generating informative and revealing questions. The results from these simulations are visualized in Figure 5.

How Should Designers Arbitrate? In Equation (15), we introduce λ, which arbitrates the relative importance of gaining or revealing information. Higher values of λ put more weight on revealing information, while lower put more weight on gaining information. In Figure 5, we return to the simulated robot arm environment, and test **Our** approach with $\lambda = \{0.5, 1, 10, 100, 1000\}$. We

compare these results to the lower and upper bounds for **Informative** and **Revealing** robots. We find that λ effectively enables designers to tune between these extremes: at low values of λ we approach the lower bound for robot regret, and at high values of λ we approach the lower bound for human error. Importantly, even at $\lambda = 0.5$, the robot is still asking questions that are more revealing than an **Informative** robot. This suggests that—even when designers are focused on quickly learning human preferences— **Ours** can still noticeably improve human understanding.

How Many Questions Does the Human Need to Remember? In our cognitive human model from Equation (10), we assume that the human remembers and reasons over the robot's last k questions. In our first set of simulations, the human only remembered $k = 3$ recent questions: can the robot still convey its learning if the human only thinks about the robot's last question, and how much does this communication improve if the human remembers every question? In Figure 5, we compare **Our** approach with $k = \{1, 3, 4, 10, 20\}$, where $k = 1$ means that the human only remembers the last question, and $k = 20$ indicates that the human remembers every question. As expected, the human's bounded memory has no impact on the robot's regret, since k does not change how the human answers the robot's questions. But k does affect how accurately the human interprets the robot's learning. Specifically, we find that a single question is not enough: with $k = 1$ the human's error remains constant after each iteration, while with $k \geq 3$, the human's error converges to the best case performance of a purely **Informative** robot. We conclude that our approach requires that users keep multiple questions in mind, but they do not need to remember *all* of the questions—here, it was sufficient to keep track of the last three.

Summary. Overall, we find that optimizing for informative and revealing questions captures the best aspects of both perspectives. When using our approach the robot learns at the same rate as the informative baseline while communicating just as well as a revealing robot (Figure 4). The questions we generate are still easy for the human to answer: simulated users choose Idk less frequently with our method than they do with an existing approach for asking easy questions [10]. Finally, our approach is robust to different choices of the bounded memory model: The human can infer what the robot is learning even if they cannot remember every robot question (Figure 5).

8 USER STUDIES

We conducted two in-person user studies to compare our approach to a state-of-the-art active learning baseline.[3] Users interacted with a robot optimizing purely for informative questions, as well as a robot optimizing for both informative and revealing questions. In the first user study, we explored how the robot learned the participants' preferences and the types of questions the robot asked. Next, in our follow-up user study, we evaluated how the robot's questions affected the participants' perception of (i) what the robot learned and (ii) whether it was ready to be deployed. Both studies shared a common experimental setup where participants taught a 7-DoF robot arm (Fetch, Fetch Robotics) by answering the robot's questions across three tasks. Videos of our experimental setup and results are available here: https://youtu.be/tC6y_jHN7Vw.

Independent Variables. We tested two different algorithms for generating questions: **Informative** and **Ours**. In **Informative**, the robot leveraged a *state-of-the-art* active learning approach that optimizes for information gain [9, 10, 43]. This robot only seeks to gather information, and does not consider how its own behavior could influence the human. We experimentally compared this state-of-the-art approach to a robot asking informative and revealing questions (**Ours**). Here, the robot applied Equation (15) to tradeoff between both human and robot perspectives.

[3]The code is available at https://github.com/VT-Collab/revealing-questions.git.

8.1 Learning the Human's Preferences

Our online survey from Section 5 demonstrated that humans can infer what the robot is learning based on the questions the robot asks. Our first in-person user study goes one step farther: We test whether robots can actively *elicit* information from the human while *also revealing* what they have learned. Participants taught the robot their preferences by answering pairwise comparisons. We measured how efficiently the robot learned the human's preference, and also compared the types of questions the robot asked.

Participants and Procedure. We recruited 10 community members (3 female, 1 non-binary, ages 23 ± 5.2 years) to participate in our study. All participants provided informed written consent consistent with Virginia Tech IRB #20-755. Participants taught a robot the three tasks shown in Figure 6: stacking dishes (*Stack*), marking a target (*Mark*), and sorting objects (*Sort*). In each task, the robot was initially unsure about the human's true preference θ^*. *Stack* is the task from our motivating example (also see Figure 1) where the robot must learn where to place the dishes and how high to carry them. In *Mark,* the robot did not know how and where to mark each target on the table, and in *Sort* the robot was unsure about which object should be placed on which shelf. We leveraged a counterbalanced, within-subjects design: participants completed each task twice, once with **Informative** and once with **Ours**. Half of the participants started with the **Informative** robot. Participants were never told which condition they were interacting with.

During a given interaction the robot asked a total of eight questions, where each question consisted of two possible trajectories (ξ_1 and ξ_2). We generated questions by adding zero-mean Gaussian noise to an initial trajectory and then randomly pairing the results: with this approach, the robot had between 100 and 200 possible questions to choose from. When asking a question, the robot first demonstrated ξ_1, reset to the start position, and then demonstrated ξ_2. We used pauses between each trajectory to better distinguish each phase of the interaction. Similar to our simulation setup from Section 7, participants responded to the robot's question by indicating that they preferred ξ_1, ξ_2, or by saying Idk. At the end of these eight questions, the robot played back the optimal trajectory it had learned from the human's answers.

Dependent Measures–Subjective. After participants answered all the robot's questions with a given condition, we administered a 7-point Likert scale survey [45]. Our survey questions were arranged into the five multi-item scales described in Table 2. We first asked participants how *Easy* it was to answer the robot's questions, whether the robot learned over time (*Improve*), to what extent the robot questions *Revealed* information, and whether it was clear what the robot was unsure about (*Focus*). After the participants answered these initial survey questions, we then played back the optimal trajectory that the robot had learned. Participants watched the trajectory, and assessed whether this *Learned Behavior* matched what they originally wanted the robot to do. Because each human participant had their own internal preferences, we did not have access to the ground truth θ^*. We therefore used *Learned Behavior* as a measure to see whether what the robot learned matched the human's latent preferences.

Dependent Measures—Objective. To get a sense of what types of questions the robot chose, we measured *Convergence*. Let ξ^* be the trajectory that optimizes the reward function learned from the human's answers, and recall that $Q = \{\xi_1, \xi_2\}$. Convergence captures the feature difference between the robot's current question and the learned trajectory: $\|f(\xi^*) - f(\xi_1)\| + \|f(\xi^*) - f(\xi_2)\|$. In practice, the lower the convergence value gets, the more similar the robot's questions are to the robot's learned behavior (i.e., ξ_1 and ξ_2 match ξ^*). Measuring convergence tells us whether the robot is asking questions to gather new information (higher convergence values) or to reveal what it has learned (lower convergence values).

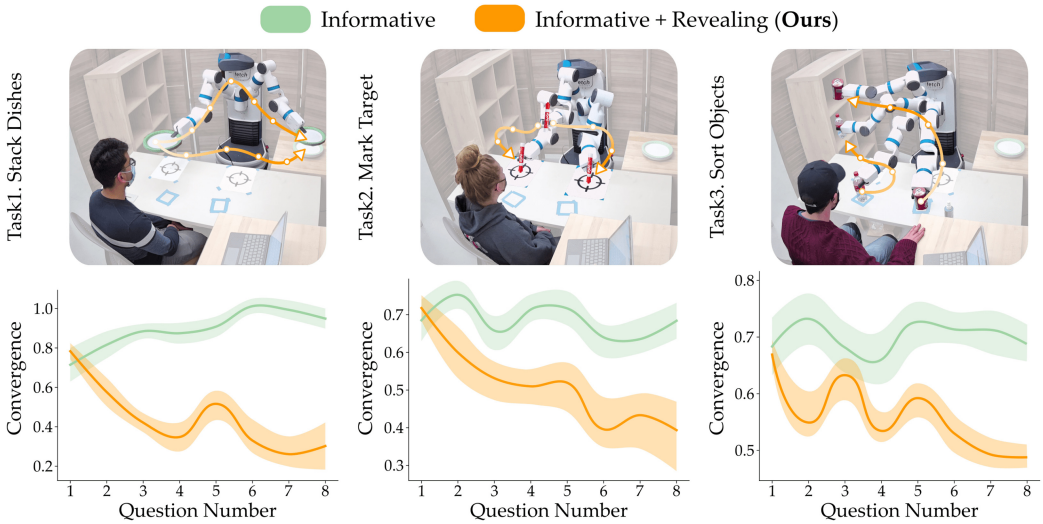

Fig. 6. Experimental setup for our user study in Section 8.1. (Top) The robot asked questions to learn the human's preferences θ^* across three different tasks. We plot sample question trajectories generated by **Ours**. (Bottom) When the robot considers how the human interprets its questions, the robot purposely chooses queries that reveal its current understanding z^*. Here, *Convergence* captures the difference between what the robot learned and the questions the robot asked. With our approach convergence decreases over time, indicating that the robot is asking revealing questions that reflect what it has learned.

Hypotheses. *We had two hypotheses in this user study*:

H1. *Robots that tradeoff between informative and revealing questions will learn at a similar rate to informative robots while updating the human about the robot's learning.*
H2. *Taking the human's perspective into account will cause the robot to ask questions that converge toward its learned behavior.*

Results—Subjective. Table 2 and Figure 7 display the results of our Likert scale survey. We first tested the reliability of our five scales, and found that the easy, reveal, focus, and learned behavior scales were reliable (Cronbach's $\alpha > 0.7$). Accordingly, we grouped each of these scales into a combined score, and performed a one-way repeated measures ANOVA on the results. The answers to the *improve* questions were inconsistent, perhaps because different users interpreted these survey questions in different ways. We have included *improve* in our plot of the results, but because of its unreliability we did not conduct additional statistical analysis on this scale.

Overall, the results from our Likert scale survey support **H1**. Participants perceived the final behavior learned by **Ours** as comparable to the behavior learned by the **Informative** baseline. Both approaches successfully learned the human's preferences across all three tasks ($F(1, 78) = 1.44$, $p = .233$). However, the questions asked by **Ours** had a significant effect on participants' attitudes toward teaching the robot. Specifically, users thought the questions asked by **Ours** were easier to answer ($F(1, 78) = 11.12$, $p < .01$), more revealing about the robot's learning ($F(1, 78) = 27.81$, $p < .01$), and better focused on uncertain aspects of the task ($F(1, 78) = 10.55$, $p < .01$). We conclude that **Ours** not only learned as efficiently as the state-of-the-art **Informative** robot, but our approach also revealed this learning back to the human user.

Results–Objective. Convergence (see Figure 6) highlights one key difference between the questions generated by **Ours** and **Informative**. With **Informative**, the robot selects queries that are often far from what the robot has learned, potentially *misleading* the human into thinking the

Table 2. Questions on Our Likert Scale Survey

Questionnaire Item	Reliability	$F(1, 78)$	p-value
–I found it ***easy*** to answer the robot's questions. –The robot's questions were hard to answer.	.85	11.12	$p < .01^*$
–The robot learned from my answers and ***improved***. –The robot kept asking things I thought it already knew.	.25	–	–
–By the end of the task the robot's questions ***revealed*** what the robot had learned. –At the end of the task it was still unclear what the robot had and had not learned.	.78	27.81	$p < .01^*$
–The robot ***focused*** on specific parts of the task when asking questions. –The robot always seemed unsure about the entire task.	.93	10.55	$p < .01^*$
–The robot's ***learned behavior*** matched my preferences. –The robot did not learn my preferred trajectory.	.92	1.44	$p = .233$

We grouped questions into five scales and tested their reliability using Cronbach's α. We explored whether the robot asked easy questions, improved over time, revealed information about its learning, focused on specific parts of the task, and learned the correct behavior. Computed p-values indicate if **Ours** scored higher than **Informative**, where $*$ denotes statistical significance. We include additional information about the mean and standard deviation for each scale in Appendix A.5.

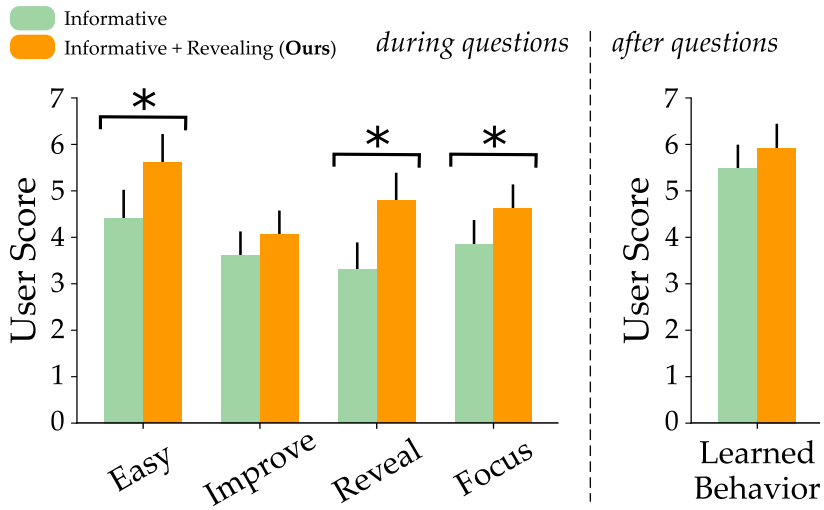

Fig. 7. Subjective results from our in-person user study. Higher ratings indicate agreement, and $*$ denotes statistical significance ($p < .01$). (Left) Participants found the questions asked by **Ours** to be easier to answer, more revealing, and better focused on specific parts of the task. (Right) After asking questions and collecting the user's answers, the robot played back its learned behavior. Participants did not notice a significant difference between what the robot learned with **Ours** and an **Informative** active learning approach.

robot is confused or unsure. By contrast, with **Ours** the robot asks questions that gradually hone-in on what the robot has learned, demonstrating this learning to the human. We emphasize that these experimental results are consistent with Proposition 1, since here the questions generated by **Ours** become more revealing over time (i.e., convergence decreases, indicating more revealing

questions). Across all three tasks our objective results support **H2**, and suggest that robots, which account for the bi-directional transfer of information purposely reveal their learned behavior to the human.

8.2 Follow-up: Bringing the Human into the Learning Process

The results of the first user study indicate that our approach improves the human's perception of the robot. But what are the *practical benefits* of keeping the human up-to-date with the robot's learning? In this second user study, we test two potential benefits: (i) determining which features the robot is still unsure about, and (ii) determining when the robot learner is ready to be deployed. Overall, the purpose of this second user study is to see whether bringing the human into the learning loop can improve the interactive learning process.

Participants and Procedure. We recruited 10 new community members (6 female, ages 25 ± 3.5 years) to participate in this follow-up study. As before, participants provided informed written consent following Virginia Tech IRB #20-755. These participants completed the *Sort* and *Stack* tasks from Figure 6 with some slight modifications (described below). For both tasks the robot was initially unsure about the human's true preferences θ^*.

In the modified *Sort* task, the robot was trying to learn three preferences: which *cup* to pick up, which *shelf* to sort this cup onto, and whether to place the cup in the front or back of the shelf (*distance*). We artificially limited the preferences that the robot could learn from the human's answers: no matter what the human chose, the robot never learned which *cup* to pick up. After a total of 12 robot questions we asked participants to specify which preferences they thought the robot had learned and which preferences the robot was still unsure about.

By contrast, in the modified *Stack* task, we did not impose any limits on the robot's learning. Here, participants taught the robot until they thought it had learned their preferences and was ready to be autonomously deployed. Each time the robot asked a question participants were given the option to stop teaching and deploy the robot. We recorded the round where participants chose to discontinue teaching (up to a maximum of twelve questions).

Participants completed the modified tasks once with the **Informative** robot and once with **Our** approach. As before, we counterbalanced the order of presentation for this within-subjects design.

Dependent Measures—Revealing Learning. Our analysis for the *Sort* task was similar to the online user studies from Section 5. After answering all 12 questions, participants indicated what they thought the robot had learned about each preference on a 7-point scale. Here, a score of 1 denotes that the human believed the robot never learned that preference, while a score of 7 denotes that the human thought the robot fully understood what they wanted. We also asked participants whether the robot's queries were *Interpretable* throughout the entire interaction.

Dependent Measures—Ready to Deploy. During the *Stack* task, we recorded how many questions the human answered before deciding that their robot was ready to be deployed (*Deploy*). Of course, the human might think the robot learned their preferences before it actually did—accordingly, we also recorded the first question where the robot's regret was below a threshold. Recall from Section 7 that regret is the difference in reward between the human's preferred behavior and the robot's learned behavior. By recording regret, we thus identified a lower bound on when the robot was actually ready to be deployed.

Hypotheses. *We had the following two hypotheses*:

> **H3.** *Humans will be misled by robots that only optimize for informative questions, but will recognize what the robot has and has not learned under our approach.*

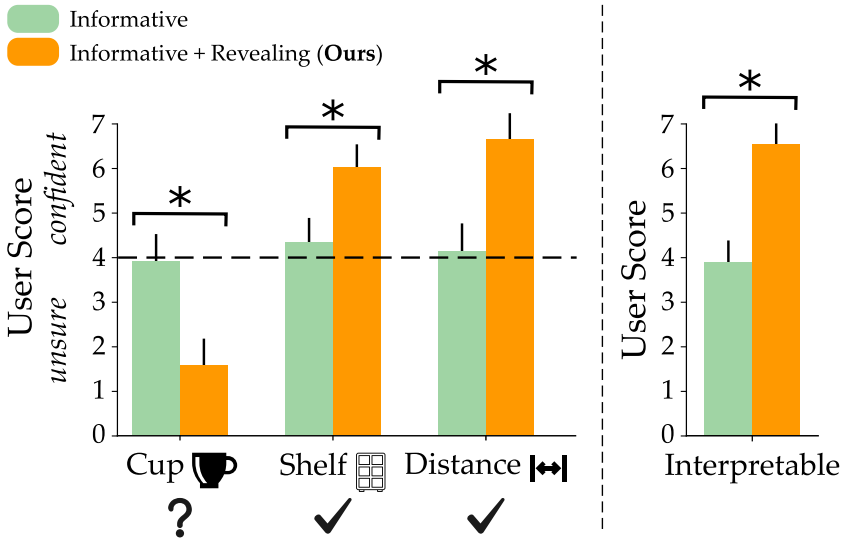

Fig. 8. User perceptions of what the robot did and did not learn during the modified *Sort* task. In this task, the robot was only able to learn *shelf* and *distance* preferences, and we artificially prevented the robot from learning which *cup* to sort. (Left) High scores indicate the human thought the robot learned a preference, while low scores indicate the human thought the robot was confused. With the state-of-the-art **Informative** baseline, participants were on the fence for every preference, even preferences the robot actually learned. (Right) We also asked participants whether the overall behavior of each robot was interpretable. Higher ratings indicate agreement, and $*$ denotes statistical significance ($p < .001$). Additional details are in Appendix A.6.

> **H4.** *Our approach will cause participants to deploy the robot soon after it has learned their preferences without answering additional and unnecessary questions.*

Results–Revealing Learning. Our results on the modified *Sort* task are shown in Figure 8. Recall that here we artificially prevented the robot from learning the *cup* feature. When participants interacted with the **Informative** robot they did not realise this. Indeed, in the **Informative** condition participants were misled about every preference, and were unsure if the robot even understood their preferred *shelf* or *distance*. But with **Our** approach participants got it right: They correctly recognizes what the robot did not learn (*cup*) and what the robot was confident about (*shelf* and *distance*). The difference between these interpretations was significant across three features ($p < .001$). Overall, participants perceived the questions asked by **Our** robot as more interpretable than the questions asked by the state-of-the-art **Informative** robot ($p < .001$).

Results—Ready to Deploy. Our results on the modified *Stack* task are visualized in Figure 9. Here, we recorded how many questions each user thought they needed to teach the robot their desired behavior. The dashed line shows when the robot was actually ready to be deployed: The robot typically understood how to carry and stack the dishes after between two and four questions. However, the robot's learning was not transparent in the **Informative** condition. Here, participants were never confident that the robot understood their preference, and so they answered all 12 questions without choosing to deploy the robot. By contrast, the questions that **Our** robot asked helped participants recognize that it was ready to be deployed. Participants consistently deployed **Our** robot after it had successfully learned their preferences. These results support **H4** and indicate that—by actively bringing the human into the learning loop—robots using our approach encourage humans to trust and deploy learning agents.

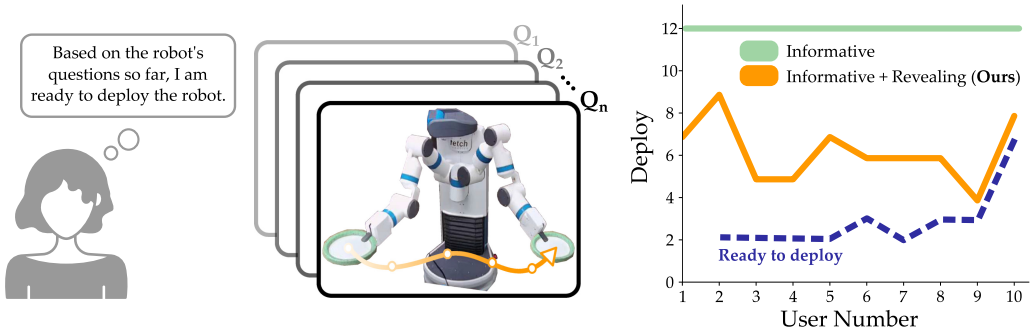

Fig. 9. Participants determining when the robot learner is ready for autonomous deployment. (Left) Participants could choose to deploy the robot after any question, up to a maximum of 12 questions. (Right) We measured the robot's regret to determine when it was actually ready to be deployed. The dashed line shows when this regret was less than a threshold of 0.2, indicating that the robot understood the human's preferences. Ideally participants will choose to deploy the robot immediately after this dashed line. With **Informative** participants were never confident in the robot and did not choose to deploy it (answering many unnecessary questions). By contrast, the questions asked by **Ours** revealed that the robot had learned the human's desired behavior, and so participants chose to deploy this robot after around six questions. User 1 changed their preferences multiple times, making it unclear when this robot was ready to be deployed.

9 CONCLUSION

We explored settings where a robot is asking questions to learn the human's reward function. Our key insight was that robot questions implicitly convey information to the human, revealing what the robot does and does not know. Prior work on active robot learning studies a one-way transfer of information: from the human to the robot. In this article, we considered both human and robot perspectives, and developed a method for asking questions that closes the loop and conveys what the robot has learned back to the human teacher.

We first formalized the bi-directional transfer of information that occurs when robots ask questions. Here, we introduced a cognitive human model that relates the questions the robot asks to the information those questions reveal. We tested this revealing model in online user studies, where 50 participants were able to correctly identify what the robot did and did not know based only on the questions the robot asked.

Next, we formulated a continuous tradeoff between informative and revealing questions. We found that when robots only optimize for gathering information users perceive their questions as random and misleading. Conversely, when robots only optimize for revealing information, they learn just as slowly as a robot asking random questions. Under our proposed approach, robots optimize for both human and robot perspectives: These robots actively select questions to efficiently learn in an interpretable manner.

We conducted simulations and user studies to compare our approach to a state-of-the-art alternative. Robots leveraging our algorithm learned as quickly as robots that optimize for information gain, but with the added benefit of revealing that learning to the human. In practice, this enabled participants to (i) identify which parts of the task the robot was still unsure about, and (ii) determine when their robot had learned their preferences and was ready to be deployed.

10 LIMITATIONS AND FUTURE WORK

Our work is a step toward interpretable and transparent active learning robots. However, we recognize that there are many modalities that robots can leverage to communicate their learning to

the human: e.g., natural language, haptic interfaces, or augmented reality. In this article, we specifically focused on using robot motion, but future work should consider multiple, complementary types of feedback. Depending on the situation one or more of these modalities may be preferable.

During our user studies participants commented that they were unsure about which part of the robot's question to focus on. Returning to our motivating example, imagine that the robot is uncertain about both carrying height and stacking the dishes. The robot shows two trajectories: one gets the height wrong but correctly stacks the dishes, and the other gets the height right but fails to stack correctly. Which option should the user choose? We expected users to select the preference that was more important to them (either height or stacking), but some participants voiced confusion in these scenarios. This indicated to us that the robot should focus on one preference at a time—however, in our follow-up online survey, 50 participants perceived the all-at-one and one-at-a-time robots as similarly revealing. Moving forward, our working hypothesis is that robot questions can vary multiple parts of the task when *revealing* information, but should focus on a single preference when *gathering* information. We plan to explore this point in future work.

APPENDIX

A STATISTICAL ANALYSIS

Within this appendix, we provide additional details for the statistical analysis of our simulations and experiments. Each subsection corresponds to a figure in the main text.

A.1 Questions Reveal Information to Humans (Figure 2)

Participants of our Amazon Mechanical Turk user study correctly inferred the robot's learned behavior from robot questions. We conducted two separate repeated measures ANOVAs with a Sphericity Assumed correction and found that user scores differed significantly between the three preferences in the task (one uncertain preference: $F(2, 98) = 97.93, p < .001^*$ and two uncertain preferences: $F(2, 98) = 32.64, p < .001^*$). We then conduced the post hoc analysis shown in Table 3.

Table 3. Post Hoc Analysis of the Results in Figure 2 with a Bonferroni Adjustment

Condition	Preference	Mean	Std. Dev.	Compared To	p-value
One uncertain preference	Ball	4.67	0.78	Height	$p < .001^*$
	Bowl	4.15	0.86	Height	$p < .001^*$
Two uncertain preferences	Height	2.22	0.86	Bowl	$p < .001^*$
	Ball	2.17	0.74	Bowl	$p < .001^*$

In one uncertain preference, the robot did not know the correct *Height*, and in two uncertain preferences the robot was unsure about both the *Height* and *Ball* preferences.

A.2 Follow-up Comparison between Two Types of Revealing Questions (Figure 3)

We next tested two different ways of asking questions: asking questions that reveal the robot's uncertainty about every preference, and asking questions that focus on revealing the uncertainty of one preference at a time. Participants of our Amazon Mechanical Turk user study did not perceive a significant difference between these two methods. We conducted a two-way repeated measures ANOVA across question methods and task preferences, and determined that the question method did not have a statistically significant effect on the user's scores. See Table 4.

Table 4. Testing for the Main Effect of **all-at-once** vs. **one-at-a-time** Questions

Type	Preference	Mean	Std. Dev.	$F(1, 49)$	p-value
All-at-once	Height	2.48	1.31	0.20	$p = .66$
	Ball	2.18	1.10		
	Bowl	3.38	1.34		
One-at-a-time	Height	2.22	1.15		
	Ball	2.34	1.14		
	Bowl	3.66	1.17		

We found that the question type did not have a statistically significant effect on the user scores.

Table 5. Post Hoc Analysis of the Results of the Robot Environment in Figure 4 with a Bonferroni Adjustment

Measure	Method	Mean	Std. Dev.	Compared To	p-value
Human Error	Random	0.69	0.18	Informative	$p = .195$
	Revealing	0.37	0.14	Informative	$p < .001^*$
	Ours	0.39	0.13	Informative	$p < .001^*$
	Random	0.69	0.18	**Ours**	$p < .001^*$
	Informative	0.73	0.18	**Ours**	$p < .001^*$
	Revealing	0.37	0.14	**Ours**	$p = .328$
Robot Regret	Random	0.02	0.03	Informative	$p < .001^*$
	Revealing	0.02	0.05	Informative	$p < .001^*$
	Ours	< 0.01	< 0.01	Informative	$p = .500$
	Random	0.02	0.03	**Ours**	$p < .001^*$
	Informative	< 0.01	< 0.01	**Ours**	$p = .500$
	Revealing	0.02	0.05	**Ours**	$p < .001^*$
Question Difficulty	Random	0.09	0.29	Informative	$p < .001^*$
	Revealing	0.18	0.39	Informative	$p = .051$
	Ours	0.33	0.47	Informative	$p = .642$
	Random	0.09	0.29	**Ours**	$p < .001^*$
	Informative	0.30	0.46	**Ours**	$p = .642$
	Revealing	0.18	0.39	**Ours**	$p < .05^*$

Remember that we are comparing the differences after all 20 questions have been asked. We focus on the difference between our approach and the alternatives.

A.3 Comparison to State-of-the-Art Baselines with a Simulated Human (Figure 4)

We compared robot questions generated by our approach to state-of-the-art baselines. Here, 100 simulated humans answered a sequence of 20 robot questions. Below we compare the results after all 20 questions were completed in the Robot Arm environment. For each of the metrics (*Human Error*, *Robot Regret*, and *Question Difficulty*), we conducted a one-way repeated measures ANOVA with a Sphericity Assumed correction to test for differences between the four methods for asking questions. We found that there was a significant difference in *Human Error* ($F(3, 297) = 144.82$, $p < .001^*$), *Robot Regret* ($F(3, 294) = 13.04$, $p < .001^*$), and *Question Difficulty* ($F(3, 294) = 7.05$, $p < .001^*$). Next, we conducted the post hoc analysis shown in Table 5.

Table 6. Post Hoc Analysis of the Results of Hyperparameter Tuning in Figure 5 with a Bonferroni Adjustment

Measure	Hyperparameter	Mean	Std. Dev.	Compared To	p-value
Human Error	$\lambda = 1$	0.390	0.132	$\lambda = 0.5$	$p = .058$
	$\lambda = 10$	0.328	0.134	$\lambda = 0.5$	$p < .001^*$
	$\lambda = 100$	0.345	0.135	$\lambda = 0.5$	$p < .001^*$
	$\lambda = 1,000$	0.346	0.120	$\lambda = 0.5$	$p < .001^*$
Robot Regret	$\lambda = 1$	0.001	0.003	$\lambda = 0.5$	$p = .050$
	$\lambda = 10$	0.008	0.014	$\lambda = 0.5$	$p < .001^*$
	$\lambda = 100$	0.020	0.035	$\lambda = 0.5$	$p < .001^*$
	$\lambda = 1,000$	0.019	0.028	$\lambda = 0.5$	$p < .001^*$
Human Error	$k = 3$	0.358	0.135	$k = 1$	$p < .001^*$
	$k = 5$	0.288	0.096	$k = 1$	$p < .001^*$
	$k = 10$	0.258	0.063	$k = 1$	$p < .001^*$
	$k = 20$	0.300	0.059	$k = 1$	$p < .001^*$

Remember that we are comparing the differences after all 20 questions have been asked. We focus on the differences between the lowest value of the hyperparameter and alternatives with increasing values.

A.4 Effects of Hyperparameters on Our Approach (Figure 5)

We explored the hyperparameters λ (the tradeoff between informative and revealing questions) and k (the number of questions that the human remembers in our bounded memory model). As in the previous analysis, here 100 simulated humans answered a sequence of 20 robot questions, and we compared the results after all 20 questions were completed in the Robot Arm environment. A one-way repeated measures ANOVA with a Sphericity Assumed correction determined that the results differed significantly between λ values (*Human Error*: $F(4, 396) = 9.21, p < .001^*$ and *Robot Regret*: $F(4, 392) = 19.25, p < .001^*$) and k values (*Human Error*: $F(4, 396) = 1.09, p < .001^*$). Note that k had no impact on *Robot Regret* and therefore we did not further analyze these results. Finally, we conducted the post hoc analysis shown in Table 6.

A.5 Subjective Results from our In-Person User Study (Figure 7)

Participants of our first in-person user study answered survey questions in five scales (*Easy*, *Improve*, *Reveal*, *Focus*, and *Learned Behavior*). We performed a one-way ANOVA with a Sphericity Assumed correction and found a statistical significance between user scores for **Informative** and **Ours** across these scales. See Table 2 for our main results—in Table 7, we provide additional details about the mean and standard deviation of each score.

A.6 Follow-up: Determining which Features the Robot is Unsure About (Figure 8)

In our second in-person user study, we tested whether participants could tell what parts of the task the robot was unsure about. High scores indicate the human thought the robot learned a preference, while low scores indicate the human thought the robot was confused. We first conducted a two-way repeated measures ANOVA across question method (**Informative** or **Ours**) and preference (*Cup*, *Shelf*, or *Distance*). We determined that the question method had a significant main effect on user scores: $F(1, 19) = 5.03, p < .05^*$. Accordingly, we conducted the post hoc analysis shown in Table 8.

Table 7. One-way Repeated Measures ANOVA on the Results of our Likert Scale Survey in Figure 7

Questionnaire Item	Method	Mean	Std. Dev.	$F(1,78)$	p-value
Easy	Informative	4.12	0.27	11.12	$p < .01^*$
	Ours	5.40			
Improve	Informative	3.52	0.28	–	–
	Ours	4.40			
Reveal	Informative	3.12	0.28	27.81	$p < .01^*$
	Ours	5.22			
Focus	Informative	3.72	0.27	10.55	$p < .01^*$
	Ours	4.97			
Learned Behavior	Informative	5.47	0.26	1.44	$p = .23$
	Ours	5.92			

Please refer to Table 2 for our main results. Note that we did not perform an ANOVA on the *Improve* scale because the user's answers were inconsistent (Cronbach's $\alpha = 0.25$).

Table 8. Independent-samples T-tests on the user Scores from Figure 8

Preference	Method	Mean	Std. Dev.	$t(38)$	p-value
Cup	Informative	4.45	2.04	−4.87	$p < .001^*$
	Ours	6.75	0.55		
Shelf	Informative	4.50	1.79	−3.34	$p < .05^*$
	Ours	6.05	1.05		
Distance	Informative	4.15	2.13	−4.94	$p < .01^*$
	Ours	6.60	0.60		
Interpretable	Informative	4.45	2.04	−4.87	$p < .01^*$
	Ours	6.75	0.55		

We conducted this analysis after finding that question method had a significant main effect on user scores.

REFERENCES

[1] Pieter Abbeel and Andrew Y. Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the International Conference on Machine Learning*.

[2] Nir Ailon. 2012. An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity. *Journal of Machine Learning Research* 13, 1 (2012), 137–164.

[3] Baris Akgun, Maya Cakmak, Karl Jiang, and Andrea L. Thomaz. 2012. Keyframe-based learning from demonstration. *International Journal of Social Robotics* 4, 4 (2012), 343–355.

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, and Richard Benjamins. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[5] Andrea Bajcsy, Dylan P. Losey, Marcia K. O'Malley, and Anca D. Dragan. 2018. Learning from physical human corrections, one feature at a time. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 141–149.

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

[7] Chandrayee Basu, Mukesh Singhal, and Anca D. Dragan. 2018. Learning from richer human guidance: Augmenting comparison-based learning with feature queries. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 132–140.
[8] Chandrayee Basu, Qian Yang, David Hungerman, Mukesh Sinahal, and Anca D. Draqan. 2017. Do you want your autonomous car to drive like you?. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 417–425.
[9] Erdem Bıyık, Dylan P. Losey, Malayandi Palan, Nicholas C. Landolfi, Gleb Shevchuk, and Dorsa Sadigh. 2020. Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences. *The International Journal of Robotics Research* 41, 1 (2022), 45–67. https://doi.org/10.1177/02783649211041652
[10] Erdem Bıyık, Malayandi Palan, Nicholas C. Landolfi, Dylan P. Losey, and Dorsa Sadigh. 2019. Asking easy questions: A user-friendly approach to active reward learning. In *Proceedings of the Conference on Robot Learning*. 1177–1190.
[11] Daniel S. Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. 2019. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *Proceedings of the International Conference on Machine Learning* (2019).
[12] Kalesha Bullard, Yannick Schroecker, and Sonia Chernova. 2019. Active learning within constrained environments through imitation of an expert questioner. In *Proceedings of the International Joint Conference on Artificial Intelligence*.
[13] Maya Cakmak and Andrea L. Thomaz. 2012. Designing robot learners that ask good questions. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 17–24.
[14] Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the Advances in Neural Information Processing Systems*. 4299–4307.
[15] Philip R. Cohen, C. Raymond Perrault, and James F. Allen. 1981. Beyond question answering. *Strategies for Natural Language Processing* (1981), 245–274.
[16] Erwin Coumans and Yunfei Bai. 2016. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org.
[17] Thomas M. Cover. 2012. *Elements of Information Theory*. John Wiley & Sons.
[18] Christian Daniel, Malte Viering, Jan Metz, Oliver Kroemer, and Jan Peters. 2014. Active reward learning. In *Proceedings of the Robotics: Science and Systems (RSS)*.
[19] Anca D. Dragan, Kenton C. T. Lee, and Siddhartha S. Srinivasa. 2013. Legibility and predictability of robot motion. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 301–308.
[20] Thomas Hellström and Suna Bensch. 2018. Understandable robots – what, why, and how. *Paladyn, Journal of Behavioral Robotics* 9, 1 (2018), 110–123.
[21] Rachel Holladay, Shervin Javdani, Anca Dragan, and Siddhartha Srinivasa. 2016. Active comparison based learning incorporating user uncertainty and noise. In *Proceedings of the RSS Workshop on Model Learning for Human-Robot Communication*.
[22] Sandy H. Huang, David Held, Pieter Abbeel, and Anca D. Dragan. 2019. Enabling robots to communicate their objectives. *Autonomous Robots* 43, 2 (2019), 309–326.
[23] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. 2018. Reward learning from human preferences and demonstrations in Atari. In *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 31. Curran Associates, Inc., 8011–8023.
[24] Ashesh Jain, Shikhar Sharma, Thorsten Joachims, and Ashutosh Saxena. 2015. Learning preferences for manipulation tasks from online coactive feedback. *The International Journal of Robotics Research* 34, 10 (2015), 1296–1313.
[25] Hong Jun Jeon, Smitha Milli, and Anca D. Dragan. 2020. Reward-rational (implicit) choice: A unifying formalism for reward learning. In *Proceedings of the Advances in Neural Information Processing Systems*.
[26] Ananth Jonnavittula and Dylan P. Losey. 2021. Learning human objectives by (under)estimating their choice set. In *Proceedings of the IEEE International Conference on Robotics and Automation*.
[27] Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Proceedings of the Handbook of the Fundamentals of Financial Decision Making: Part I*. 99–127.
[28] K. S. Krishnan. 1977. Incorporating thresholds of indifference in probabilistic choice models. *Management Science* 23, 11 (1977), 1224–1233.
[29] Minae Kwon, Erdem Biyik, Aditi Talati, Karan Bhasin, Dylan P. Losey, and Dorsa Sadigh. 2020. When humans aren't optimal: Robots that collaborate with risk-aware humans. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 43–52.
[30] Minae Kwon, Sandy H. Huang, and Anca D. Dragan. 2018. Expressing robot incapability. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 87–95.
[31] Dylan P. Losey and Dorsa Sadigh. 2019. Robots that take advantage of human trust. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*.
[32] R. Duncan Luce. 2012. *Individual Choice Behavior: A Theoretical Analysis*. Courier Corporation.

[33] George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. 1978. An analysis of approximations for maximizing submodular set functions – I. *Mathematical Programming* 14, 1 (1978), 265–294.
[34] Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. 2017. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research* 36, 5–7 (2017), 618–634.
[35] Stefanos Nikolaidis, Minae Kwon, Jodi Forlizzi, and Siddhartha Srinivasa. 2018. Planning with verbal communication for human-robot collaboration. *ACM Transactions on Human-Robot Interaction* 7, 3 (2018), 1–21.
[36] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J. Andrew Bagnell, Pieter Abbeel, and Jan Peters. 2018. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics* 7, 1–2 (2018), 1–179.
[37] Rob Powers and Yoav Shoham. 2005. Learning against opponents with bounded memory. In *Proceedings of the International Joint Conference on Artificial Intelligence* 5. 817–822.
[38] Mattia Racca and Ville Kyrki. 2018. Active robot learning for temporal task models. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 123–131.
[39] Mattia Racca, Antti Oulasvirta, and Ville Kyrki. 2019. Teacher-aware active robot learning. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 335–343.
[40] Anna N. Rafferty, Emma Brunskill, Thomas L. Griffiths, and Patrick Shafto. 2011. Faster teaching by POMDP planning. In *Proceedings of the International Conference on Artificial Intelligence in Education*. Springer, 280–287.
[41] Deepak Ramachandran and Eyal Amir. 2007. Bayesian inverse reinforcement learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 2586–2591.
[42] Eric Rosen, David Whitney, Elizabeth Phillips, Gary Chien, James Tompkin, George Konidaris, and Stefanie Tellex. 2019. Communicating and controlling robot arm motion intent through mixed-reality head-mounted displays. *The International Journal of Robotics Research* 38, 12–13 (2019), 1513–1526.
[43] Dorsa Sadigh, Anca D. Dragan, Shankar Sastry, and Sanjit A. Seshia. 2017. Active preference-based learning of reward functions. In *Proceedings of the Robotics: Science and Systems*.
[44] Charles F. Schmidt, Natesa S. Sridharan, and John L. Goodson. 1978. The plan recognition problem: An intersection of psychology and artificial intelligence. *Artificial Intelligence* 11, 1–2 (1978), 45–83.
[45] Mariah L. Schrum, Michael Johnson, Muyleng Ghuy, and Matthew C. Gombolay. 2020. Four years in review: Statistical practices of likert scales in human-robot interaction studies. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 43–52.
[46] Ankit Shah and Julie Shah. 2020. Interactive robot training for non-markov tasks. arXiv:2003.02232. Retrieved from https://arxiv.org/abs/2003.02232.
[47] Stefanie Tellex, Ross Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. 2014. Asking for help using inverse semantics. In *Proceedings of the Robotics: Science and Systems*.
[48] Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Justin Hart, Peter Stone, and Raymond J. Mooney. 2017. Opportunistic active learning for grounding natural language descriptions. In *Proceedings of the Conference on Robot Learning*. 67–76.
[49] Toyin Tofade, Jamie Elsner, and Stuart T. Haines. 2013. Best practice strategies for effective use of questions as a teaching tool. *American Journal of Pharmaceutical Education* 77, 7 (2013), 155.
[50] Maegan Tucker, Ellen Novoseller, Claudia Kann, Yanan Sui, Yisong Yue, Joel W. Burdick, and Aaron D. Ames. 2020. Preference-based learning for exoskeleton gait optimization. In *Proceedings of the IEEE International Conference on Robotics and Automation*. 2351–2357.
[51] Michael Walker, Hooman Hedayati, Jennifer Lee, and Daniel Szafir. 2018. Communicating robot motion intent with augmented reality. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 316–324.
[52] Nils Wilde, Dana Kulic, and Stephen L. Smith. 2020. Active preference learning using maximum regret. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*.
[53] Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. 2008. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence*. 1433–1438.

Received June 2021; revised October 2021; accepted January 2022