Enabling Robots to Infer How End-Users Teach and Learn Through Human-Robot Interaction

Dylan P. Losey ⬤, *Student Member, IEEE*, and Marcia K. O'Malley ⬤, *Senior Member, IEEE*

Abstract—During human-robot interaction, we want the robot to understand us, and we want to intuitively understand the robot. In order to communicate with and understand the robot, we can leverage interactions, where the human and robot observe each other's behavior. However, it is not always clear how the human and robot should interpret these actions: a given interaction might mean several different things. Within today's state of the art, the robot assigns a *single* interaction strategy to the human, and learns from or teaches the human according to this fixed strategy. Instead, we here recognize that different users interact in different ways, and so one size does not fit all. Therefore, we argue that the robot should maintain a *distribution* over the possible human interaction strategies, and then infer how each individual end-user interacts during the task. We formally define learning and teaching when the robot is uncertain about the human's interaction strategy, and derive solutions to both problems using Bayesian inference. In examples and a benchmark simulation, we show that our personalized approach outperforms standard methods that maintain a fixed interaction strategy.

Index Terms—Cognitive Human-Robot Interaction, Learning from Demonstration, Human Factors and Human-in-the-Loop.

I. Introduction

HUMAN-ROBOT interaction (HRI) provides an opportunity for the human and robot to exchange information. The robot can learn from the human by observing their behavior [1], or teach the human through its own actions [2]. In applications such as autonomous cars, personal robots, and collaborative assembly, fluent human-robot communication is often necessary.

In order to learn from and teach with interactions, however, the human and robot must correctly interpret the meaning of each other's behavior. Consider an autonomous car following behind a human driven car. If the human car slows down, what should the robotic car infer: is the human teaching the robot to also slow down, or signaling that the robot should pass? When learning from an end-user, the robot needs a model of that end-user's *teaching strategy*, i.e., how the human's actions relate to the information that human wants to convey. Conversely, when teaching the end-user, the robot must model that end-user's *learning strategy*, i.e., how the human interprets the robot's

actions. Together, we define the end-user's teaching and learning strategies as their *interaction strategy*.

In the state-of-the-art, the robot assigns a pre-programmed, fixed interaction strategy to every human; each individual end-user is assumed to teach or learn in the same way. Instead:

We here recognize that different users have different interaction strategies, and we should infer the current end-user's interaction strategy based on their actions.

Rather than a single fixed estimate of the human's interaction strategy, we argue that the robot should maintain a distribution (i.e., belief) over the possible human interaction strategies, and update this belief during the task. By reasoning over this belief, the robot can adapt to everyday end-users, instead of requiring each human to comply with its single pre-defined strategy.

Overall, we make the following contributions:

Learning and Teaching with Strategy Uncertainty. We introduce and formulate two novel problems in human-robot interaction, where the robot must optimally communicate with the human, but the robot is unsure about how the current end-user teaches or learns.

Solution with Bayesian Inference. We derive methods for the robot to learn and teach under strategy uncertainty. We show that—when the robot does not know the end-user's interaction strategy—optimal solutions infer and update a belief over that interaction strategy, resulting in personalized interactions.

Simulated Comparison to Current Methods. Using didactic examples and an inverse reinforcement learning simulation, we compare our proposed approach to robots that reason over a fixed interaction strategy. We also consider practical challenges such as noisy and unmodeled interaction strategies.

II. Related Work

When a human expert is using interactions to teach a robot, the robot can leverage learning from demonstration (LfD) to understand how it should behave [1]. Most similar to our setting is inverse reinforcement learning (IRL), an instance of LfD where the robot learns the correct reward function from human demonstrations [3], [4]. Prior works on IRL generally assume that every human has a single, fixed teaching strategy [5]: the human teaches by providing optimal demonstrations, and any sub-optimal human behavior is interpreted as noise [6]–[9]. Alternatively, robots can also learn about the human while learning from that human. In Nikolaidis *et al.* [10], for instance, the robot learns about the end-user's adaptability in addition to their reward function. Building on these works, we will infer the end-user's teaching strategy, so that the robot can more accurately learn from human interactions.

Machine teaching—also known as algorithmic teaching—identifies the best way for an expert robot to teach the novice human [2]. In order to teach optimally, however, the robot must know how the human learns. Recent machine teaching works [11]–[13] have addressed this problem by using human feedback to resolve mismatches between the assumed human learning strategy and the user's actual learning strategy. Most related to our research is work by Huang *et al.* [14], which compares the performance of different models of human learning. These authors generated the optimal teaching examples for each proposed learning strategy, and then used human feedback to identify the single best teaching strategy across all users. Like Huang *et al.* [14], we here reason over multiple learning strategies, but now we want to infer each user's specific learning strategy based on their individual responses.

III. PROBLEM STATEMENT

Consider a human who is interacting with a robot. In this setting, both the human and the robot are agents. Let us assume that one of these agents has a target model, $\theta^* \in \Theta$, which they want to teach to the other agent. Here Θ is the space of possible target models, and θ^* is particular behavior that the teacher wants to convey to the learner. For example, the teacher may want to show the learner a better way to complete the current task, or communicate how it will interact in the future. We are interested in how the robot should behave when it is (a) *learning* θ^* *from* or (b) *teaching* θ^* *to* a human agent.

Let us denote the robot state as x. The human takes action u, and the robot takes action a; these actions and the state x are observed by both the robot and the human. We use a superscript t to denote the current timestep, so that x^t is the state at time t, and $x^{0:t}$ is the sequence of states from the start of the task to the current time t. In the context of supervised learning, we can think of x as the input features, and u and a as the output labels assigned by the human and robot, respectively [15]. Here Θ is a hypothesis space, and θ^* defines the correct mapping from features to labels.

A. Learning from the Human

When the human is the expert—i.e., the human knows θ^*, but the robot does not—the robot should learn from the human. The human wants to teach the robot θ^*, and has a *teaching strategy* ϕ^*, which determines what actions the human selects to convey θ^* to the robot. More formally, a teaching strategy $\phi \in \Phi$ relates the setting (x, θ) to the human action u:

$$\pi(u \mid x, \theta, \phi) \tag{1}$$

Here $\pi \in [0, 1]$ is the probability that the human will take action u given x, θ, and ϕ. We point out that (1) is also the human's policy when teaching the robot, and that this policy is parameterized by ϕ. In other words, if the robot knows the teaching strategy ϕ^*, then it can leverage (1) to correctly interpret the meaning behind the human's actions.

In practice, however, the robot *does not know* what teaching strategy an end-user will employ. Hence, we argue that the robot should maintain a probability distribution over ϕ as it learns from the human. We refer to this problem of learning from the end-user when uncertain about their teaching strategy as *learning with strategy uncertainty*:

Definition 1: (Learning with Strategy Uncertainty). Given a discrete or continuous set of possible teaching strategies Φ and target models Θ, infer an optimal estimate of θ^* based on the history of states $x^{0:t}$ and human actions $u^{0:t}$.

B. Teaching the Human

Next we consider the opposite situation, where the robot is the expert, and is trying to teach θ^* to the human. Here the human agent has some *learning strategy* ψ^*, which determines how the human interprets the robot's actions a. A learning strategy $\psi \in \Psi$ expresses the relationship (from the human's perspective) between the setting (x, θ) and the robot action a:

$$\pi(a \mid x, \theta, \psi) \tag{2}$$

In the above, π is the human's model of the robot's policy—not necessarily the robot's actual policy—and this model is parameterized by ψ. So now, if the robot knows the user's true learning strategy ψ^*, the robot can leverage (2) to anticipate how its actions will alter the human's understanding of θ^*.

But, when teaching an actual end-user, the robot *does not know* what learning strategy that specific user has. Similar to before, we therefore argue that the robot should maintain a distribution over the learning strategies ψ when teaching the human. We refer to this problem, where the robot is teaching a user but is unsure about that end-user's learning strategy, as *teaching with strategy uncertainty*:

Definition 2: (Teaching with Strategy Uncertainty). Given a discrete or continuous set of possible learning strategies Ψ and target models Θ, select the robot action a^t that optimally teaches θ^* based on the history of states $x^{0:t}$, robot actions $a^{0:t-1}$, and human actions $u^{0:t}$.

C. Assumptions

Throughout this work, we will assume that the interaction strategies ϕ^* and ψ^* for each individual user are constant, and are not affected by the robot's behavior. Put another way, the robot cannot influence the human's interaction strategy by selecting different actions. This assumption is consistent with prior HRI research [2], [5]: however, we can also extend our proposed approach to address cases where the human's interaction strategy *does change* by incorporating a forgetting factor or transition model within the Bayesian inference.

IV. ROBOT LEARNING WITH STRATEGY UNCERTAINTY

Within this section we focus on learning from the human, where the robot does not initially know the human's teaching strategy ϕ^*. Learning here is challenging, because the robot is uncertain about how to interpret the human's actions. First, we demonstrate how the robot can learn from multiple models of the human's teaching strategy. Second, we enable the robot to update its joint distribution over ϕ and θ, and simultaneously learn both the human's teaching strategy and target model. We provide an example which compares learning this joint distribution to learning with a single fixed estimate of ϕ^*.

A. Multiple Teaching Strategies

The robot starts with a prior $b^0(\theta)$ over what θ^* is, and updates that belief at every timestep t based on the observed states and

actions. The robot's belief over target models is:

$$b^{t+1}(\theta) = P(\theta \mid u^{0:t}, x^{0:t}) \qquad (3)$$

In other words, $b^{t+1}(\theta)$ is the probability that $\theta = \theta^*$ given the history of observed states and human actions up to timestep t. Applying Bayes' rule, and recalling from (1) that the human's actions u are conditionally independent, the robot's Bayesian belief update becomes [16]:

$$b^{t+1}(\theta) = \frac{b^t(\theta) \cdot P(u^t \mid x^t; \theta)}{\int_\Theta b^t(\xi) \cdot P(u^t \mid x^t; \xi) \, d\xi} \qquad (4)$$

We here used a semicolon to separate the observed variables from the hidden variables. The denominator—which integrates over all possible target models—is a normalizing constant. Omitting this constant, we can more succinctly write (4) as:

$$b^{t+1}(\theta) \propto b^t(\theta) \cdot P(u^t \mid x^t; \theta) \qquad (5)$$

where $P(u \mid x; \theta)$ is the robot's *observation model*, i.e., the likelihood that the human takes action u given x and θ.

To correctly learn from the end-user, the robot needs an accurate observation model. We saw in Section III-A that the most accurate observation model is the user's policy π, which is parameterized by the true teaching strategy ϕ^*. Within the state-of-the-art, the robot often assumes that the user's policy is parameterized by ϕ^0, where ϕ^0 is some estimate of ϕ^*:

$$P(u^t \mid x^t; \theta) = \pi(u^t \mid x^t; \theta, \phi^0) \qquad (6)$$

Rather than a *constant point estimate* of the human's teaching strategy, we argue that the robot should maintain a *belief over multiple teaching strategies*. In the simplest case, the robot has a prior $b^0(\phi)$ over what ϕ^* is, but does not update this belief between timesteps. Here the observation model becomes:

$$P(u^t \mid x^t; \theta) = \int_\Phi \pi(u^t \mid x^t; \theta, \phi) \cdot b^0(\phi) \, d\phi \qquad (7)$$

Note that (6) is a special case of (7) where $b^0(\phi^0) = 1$. When learning with (7), the robot does not interpret human actions in the context of just one teaching strategy. Instead, the robot considers what the action u implies for each possible teaching strategy, and then learns across these strategies. We can think of (7) as the best fixed learning strategy when b^0 is known.

B. Inferring a Joint Belief

Now that we have introduced learning with multiple teaching strategies, we can solve learning with strategy uncertainty (Definition 1). Here we not only want to learn the target model θ^*, but we also recognize that the robot is uncertain about ϕ^*. Let us define the robot's joint belief $b(\theta, \phi)$ over the target models $\theta \in \Theta$ and teaching strategies $\phi \in \Phi$ to be:

$$b^{t+1}(\theta, \phi) = P(\theta, \phi \mid u^{0:t}, x^{0:t}) \qquad (8)$$

Again leveraging Bayes' rule and conditional independence:

$$b^{t+1}(\theta, \phi) \propto b^t(\theta, \phi) \cdot P(u^t \mid x^t; \theta, \phi) \qquad (9)$$

where $P(u \mid x; \theta, \phi)$ is the conditional probability of human action u given x, θ, and ϕ. But this is the same as (1), so that:

$$b^{t+1}(\theta, \phi) \propto b^t(\theta, \phi) \cdot \pi(u^t \mid x^t; \theta, \phi) \qquad (10)$$
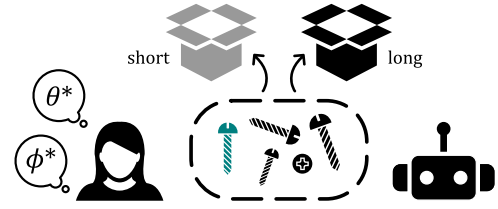

Fig. 1. A human and robot are interacting to sort screws of different lengths. The human teacher indicates one screw that they consider short (highlighted), and then the robot learner sorts the screws into the short and long boxes. The robot does not know which screws are short *a priori*, and also does not know the end-user's teaching strategy. For instance, the human could be indicating a short screw at random, or purposely selecting the longest of the short screws.

Using (10), we learn about *both the human's target model and the human's teaching strategy* from $x^{0:t}$ and $u^{0:t}$.

Let us compare the observation model for this joint learning rule to the observation models from (6) and (7). If we rewrite (10) into the form of (5), we obtain the observation model:

$$P(u^t \mid x^t; \theta) = \int_\Phi \pi(u^t \mid x^t; \theta, \phi) \cdot b^t(\phi \mid \theta) \, d\phi \qquad (11)$$

where the belief over teaching strategies given θ is:

$$b^t(\phi \mid \theta) = P(\phi \mid u^{0:t-1}, x^{0:t-1}; \theta) \qquad (12)$$

Intuitively, a robot implementing (11) and (12) reasons across multiple teaching strategies when learning from the end-user, and also updates its belief over these teaching strategies every timestep. We find that the observation model (7) is a special case of (11) when the robot never updates $b^0(\phi \mid \theta) = b^0(\phi)$, i.e., if the robot's belief over teaching strategies is constant. Accordingly, (6) is a special case of (11) by extension. Our analysis shows that inferring a joint belief over θ and ϕ both generalizes prior work and is an optimal learning rule.

C. Learning Example

To demonstrate how the proposed observation models affect the robot's learning, we here provide an example simulation. Consider the sorting task in Fig. 1, where the robot is attempting to learn the right threshold classifier from the human. At each timestep t, the human action u indicates one screw that should be classified as short; the robot then classifies the remaining screws without additional guidance. Let θ^* be the correct decision boundary, and let the robot's reward equal the total number of screws classified correctly. We can think of this example as an instance of inverse reinforcement learning [4], where the robot learns the true objective θ^*.

Importantly, we include two different teaching strategies $\phi \in \Phi$ that the human might use. Within the first strategy, ϕ_1, the human noisily indicates the short screw closest to θ^*, so that $\pi(\phi_1) \propto \exp\{-\frac{1}{2} \cdot |\theta^* - u|\}$. Within the second strategy, ϕ_2, the user indicates a short screw uniformly at random, so that $\pi(\phi_2) \propto 0.9$ if $u \leq \theta^*$ or $\pi(\phi_2) \propto 0.1$ otherwise. Each end-user leverages one of these two teaching strategies; however, the robot does not know which.

Observation Models. We compare (6), (7), and (11). Let ϕ_1 denote a robot that learns with (6), and assumes $\phi_1 = \phi^*$ for all users. Similarly, ϕ_2 is a robot that assumes $\phi_2 = \phi^*$. *Prior* denotes a robot with observation model (7), and *Joint* leverages
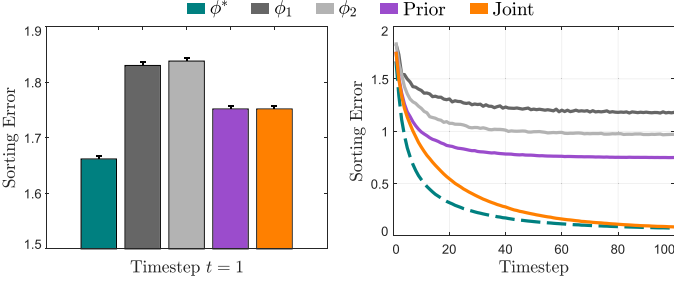
Fig. 2.   Learning when uncertain about the human's teaching strategy. Here two teaching strategies are equally likely. Left: number of screws incorrectly sorted after a single interaction. Right: number of screws incorrectly sorted after t interactions. Maintaining a distribution over both teaching strategies results in more optimal robot behavior than always assuming one strategy. Notice that *Prior* and *Joint* are equal when $t = 1$. Error bars indicate SEM.

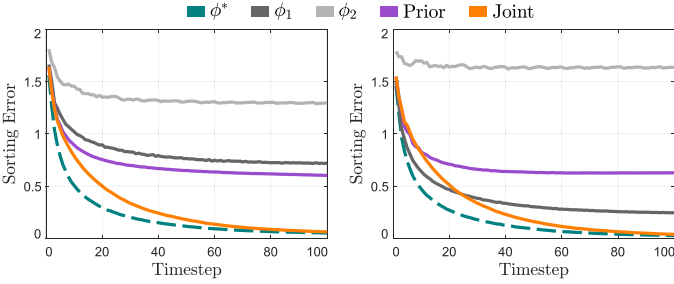

Fig. 3.   Learning when one human teaching strategy is more likely. Left: 70% of humans use teaching strategy ϕ_1. Right: 90% of humans use ϕ_1, but the robot incorrectly believes both strategies are equally likely *a priori*, i.e., $b^0(\phi_1) = 0.5$. Even when the robot's prior is wrong, learning with strategy uncertainty (*Joint*) leads to better performance over time than assuming ϕ_1.

our proposed approach (11). Finally, ϕ^* is an ideal robot that knows the teaching strategy for each individual user.

Simulation. At timestep t the robot observes the action u^t and updates its belief $b^{t+1}(\theta)$ with (5). Next, the robot optimally sorts 10 screws based on its current belief [8]. At timestep $t + 1$ the task is repeated with the same end-user (who has a constant θ^* and ϕ^*). The results of these simulations averaged across 10^5 end-users are shown in Figs. 2 and 3.

Analysis. Using our proposed *Joint* observation model resulted in fewer errors than learning under ϕ_1, ϕ_2, or *Prior*. With *Joint* the robot was able to personalize its learning strategy to the current end-user across multiple iterations, and more accurately learn what the user was communicating (see Fig. 2). As expected, *Prior* outperformed other fixed strategies when b^0 was correct; however, if the robot did not have an accurate prior over teaching strategies, the *Prior* observation model (7) was less optimal than ϕ_1 (see Fig. 3, right). We found that our proposed approach was robust to this practical challenge: despite having the wrong prior, *Joint* still caused the robot's behavior to converge to the ideal learner, ϕ^*.

V. ROBOT TEACHING WITH STRATEGY UNCERTAINTY

Within this section we consider the opposite problem, where the expert robot is teaching the human about θ^*, but does not know the end-user's learning strategy ψ^*. Teaching here is challenging because the robot is not certain what the user will learn from its actions. We first outline a specific instance of robot teaching, where the human learns through Bayesian inference. Next, we demonstrate how the robot can teach with multiple

models of the human's learning strategy, and derive one solution to teaching with strategy uncertainty. In a simulated example, we compare these methods to robots that teach with a constant point estimate of ψ^*. We also describe how the robot can trade-off between teaching θ^* to and learning ψ^* from the human via active teaching.

A. Teaching Bayesian Humans

Similar to previous works in machine teaching [14], [17] and cognitive science [18], [19], we assume that the human learns by performing Bayesian updates. Thus, the human's belief over the target models after robot action a^t becomes:

$$b^{t+1}(\theta) = \frac{b^t(\theta) \cdot \pi(a^t \mid x^t; \theta; \psi^*)}{Z(\psi^*)} \qquad (13)$$

where we use ; to denote that the human observes ψ^* but not θ^*. The denominator is again the normalizing constant:

$$Z(\psi) = \int_\Theta b^t(\xi) \cdot \pi(a^t \mid x^t; \xi; \psi) \, d\xi \qquad (14)$$

We point out that π in (13) and (14) is (2), the policy that the human assigns to the robot. The human interprets the robot's actions—and updates its belief—based on this policy, which is parameterized by the human's true learning strategy ψ^*. Here b^t is also the *state* of the human at timestep t, and (13) defines the *state dynamics* (i.e., the human's transition function).

The robot should select actions so that this state transitions to $b(\theta^*) = 1$. Let us define the ideal robot action as:

$$a^t = \arg\max_a b^{t+1}(\theta^*) \qquad (15)$$

where a^t will greedily maximize the human's belief in θ^* at the subsequent timestep. The human takes an action u^t based on what they have previously learned; the human actions are therefore observations on the human's state, i.e., $u^t = h(b^t)$. For example, the human's action could be completing a test about the target models, or performing the task themselves. Here we consider the simplest case, where:

$$u^t = h(b^t) = b^t \qquad (16)$$

Hence, the human feedback u^t provides their actual belief over the target models at the current timestep. The robot observes the human state b^t from (16), and selects action a^t with (15) to shift the human towards the desired state b^{t+1}.

B. Multiple Learning Strategies

Consider cases where the robot is teaching this Bayesian human, but does not know the human's learning strategy ψ^*. When ψ^* (and therefore the future state b^{t+1}) is unknown, teaching is analogous to controlling an agent with unknown state dynamics [20]. Define $\hat{b}$ as the robot's prediction of the human's state given the history of actions and world states:

$$\hat{b}^{t+1}(\theta) = P(\theta \mid u^{0:t}, a^{0:t}, x^{0:t}) \qquad (17)$$

Since the human performs Bayesian inference (13), and recalling that the robot observes $b^t(\theta)$, we equivalently have:

$$\hat{b}^{t+1}(\theta) = \int_\Psi \frac{u^t(\theta) \cdot \pi(a^t \mid x^t; \theta, \psi)}{Z(\psi)} \cdot b^t(\psi) \, d\psi \qquad (18)$$

Within the above, $b(\psi)$ is the robot's belief over the human's learning strategies. For the state-of-the-art, the robot *estimates* the human's learning strategy as ψ^0, so that (18) reduces to:

$$\hat{b}^{t+1}(\theta) = \frac{u^t(\theta) \cdot \pi(a^t \mid x^t; \theta, \psi^0)}{Z(\psi^0)} \quad (19)$$

Instead, we here argue that the robot should teach with a *belief over multiple learning strategies*. Let $b^0(\psi)$ be the prior over what ψ^* is. If the robot never updates this initial belief, then the predicted human state after action a^t becomes:

$$\hat{b}^{t+1}(\theta) = \int_\Psi \frac{u^t(\theta) \cdot \pi(a^t \mid x^t; \theta, \psi)}{Z(\psi)} \cdot b^0(\psi) \, d\psi \quad (20)$$

Comparing (20) to (19), now the robot reasons about how its actions are interpreted by each learning strategy. When selecting the action a^t with (15)—where we replace b^{t+1} with prediction $\hat{b}^{t+1}$—this robot teaches across multiple strategies.

C. Inferring the Learning Strategy

Because the robot is getting feedback from the user, however, we can also infer that specific user's learning strategy, ψ^*. Learning about ψ^* provides a solution to teaching with strategy uncertainty (Definition 2), and results in robots that adapt their teaching to match the human. Let us formally define the robot's belief over learning strategies as:

$$b^t(\psi) = P(\psi \mid u^{0:t}, a^{0:t}, x^{0:t}) \quad (21)$$

We use the subscript t instead of $t+1$ since $b^t(\psi)$ does not actually depend on a^t, as we will show. Applying Bayes' rule:

$$b^t(\psi) \propto P(u^{0:t} \mid a^{0:t}, x^{0:t}; \psi) \cdot P(\psi \mid a^{0:t}, x^{0:t}) \quad (22)$$

Recalling that the human's learning strategy is not altered by the robot, $P(\psi \mid a^{0:t}, x^{0:t}) = P(\psi)$. Moreover, because the human is a Bayesian learner, and $u^t = b^t$, here u^t depends on u^{t-1}, a^{t-1}, x^{t-1}, and ψ (13). Hence, (22) simplifies to:

$$b^t(\psi) \propto b^{t-1}(\psi) \cdot P\left[u^t \middle| \frac{u^{t-1} \cdot \pi(a^{t-1} \mid x^{t-1}; \theta, \psi)}{Z(\psi)}\right] \quad (23)$$

Intuitively, (23) claims that the belief over learning strategies is updated based on the differences between the human's actual state (left side of the likelihood function) and the predicted human state given ψ (right side of the likelihood function)[1]. By observing u, we can use (23) to infer the human's learning strategy. By then substituting (23) back into (18), the robot *learns about ψ^* while teaching the human θ^**: thus, using (18) with (23) addresses teaching with strategy uncertainty.

D. Teaching Example

Here we provide an illustration of how reasoning over multiple learning models can improve teaching with uncertainty. As shown in Fig. 4, the robot is moving towards goal position θ^*, and wants to teach that goal to the nearby human. At each timestep t, the robot's action a is an incomplete trajectory (e.g., see the three trajectory segments in Fig. 4). After observing this robot trajectory, the human updates their belief over θ^*; specifically, the human applies Bayesian inference to determine

---

[1] We used Kullback-Leibler (KL) divergence [21] to define the likelihood of u^t given the right side of (23), but other options are possible.
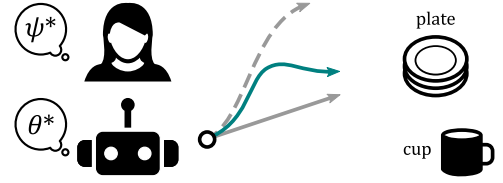

Fig. 4. A robot is moving towards a goal position, and is trying to convey that goal to the human. Here the robot teacher is moving to the plate, θ^*, but does not know the human's learning strategy, ψ^*. For example, the human may learn best from goal-directed trajectories (solid) or exaggerated trajectories (dashed). When the robot teacher reasons over each of these human learning strategies, it selects an action that teaches both types of learners (highlighted).
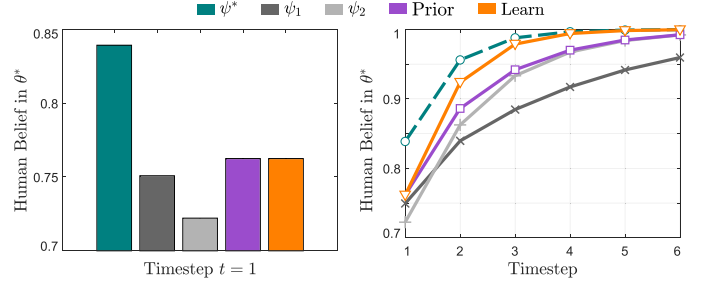

Fig. 5. Teaching when uncertain about the human's learning strategy. Left: the human's confidence that the robot's goal is θ^* after one interaction. When using *Prior* or *Learn*, the robot selects a slightly exaggerated trajectory, which teaches both groups of learners. Right: the human's belief after t iterations. When the robot reasons about a distribution over learning strategies, it teaches θ^* more quickly than when focusing on a single type of learner (ψ_1 and ψ_2).

whether the robot's goal is the cup or the plate. The robot uses (15) with prediction $\hat{b}^{t+1}$ to select the trajectory a which will teach the human the most about θ^*.

We consider two possible learning strategies $\psi \in \Psi$ for the simulated end-users. Humans with ψ_1 learn best from *legible* (i.e., exaggerated) trajectories [22]: $\pi(\psi_1) = (0.1, 0.3, 0.45)$ if the robot moves directly towards θ, slightly exaggerates, or fully exaggerates, respectively. By contrast, under ψ_2 the user learns best from *predictable* (i.e., goal-directed) trajectories, such that $\pi(\psi_2) = (0.35, 0.2, 0.15)$ if the robot moves directly towards θ, slightly exaggerates, or exaggerates, respectively. The robot does not know which strategy a given user selects.

Prediction Method. We compare (18), (19), and (20). Let ψ_1 denote a robot which predicts that every user learns with (19), where $\psi^0 = \psi_1$. Likewise, ψ_2 is a robot that estimates $\psi^0 = \psi_2$. The *Prior* robot reasons over both learning strategies using (20), and our proposed *Learn* robot solves teaching with strategy uncertainty by leveraging (18) with (23).

Simulation. The robot observes the human action u—i.e., the human's current belief—and selects an action a using (15) and its prediction method. The robot can select between 6 different legible or goal-directed trajectory segments (3 for each goal θ). The human is a Bayesian learner. Our results (averaged across 10^5 simulated users) are depicted in Figs. 5 and 6.

Analysis. Robots using our proposed *Learn* approach more quickly taught θ^* than with the fixed teaching methods ψ_1, ψ_2, or *Prior*. Reasoning over human learning strategies led to better teaching during a single interaction (see Fig. 5). For multiple iterations, we tested practical scenarios where the robot has the wrong prior: in every case, *Learn* yielded the fastest convergence, and taught as well as the ideal teacher after ≈ 5 timesteps (see Fig. 6). Intuitively, the *Learn* robot gradually
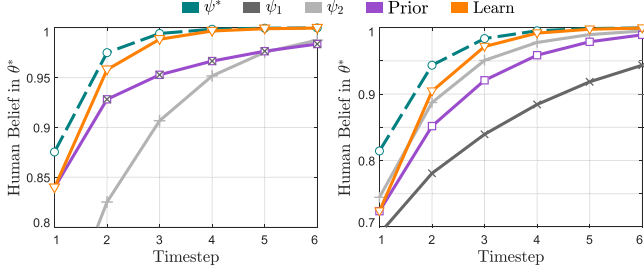
Fig. 6.   Teaching when one learning strategy is more likely. Left: the robot knows that 80% of users learn with ψ_1, and so the *Prior* robot greedily selects legible actions (*Prior* $= \psi_1$). Right: the robot initially thinks both strategies are equally likely, but 70% of users have ψ_2. Learning about ψ^* while teaching θ^* (*Learn*) outperforms ψ_2 over time, even with the wrong prior.
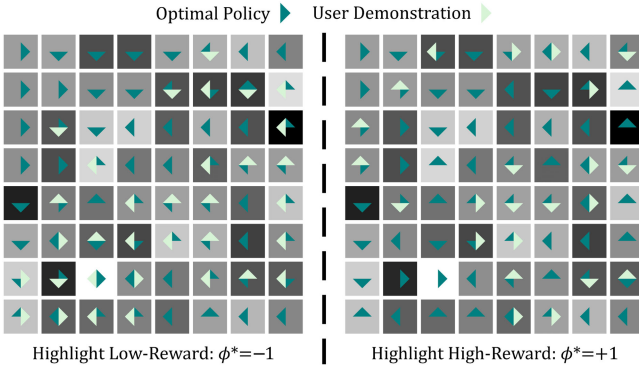

Fig. 7.   Sample 8-by-8 gridworld labeled by two different simulated users. Both users have the same reward parameters θ^*, and the grid cells are colored based on this reward (lighter cells have higher reward). The optimal policy for θ^* is shown by the darker arrows. Left: the human leverages teaching strategy $\phi^* = -1$, and teaches the robot by biasing their demonstration to highlight low-reward states. Right: this user instead uses teaching strategy $\phi^* = +1$, and favors states with locally higher reward. The robot attempts to learn θ^* given a demonstration (like the ones shown above), but does not know ϕ^*.

shifted to teaching with either legible or predictable trajectories, while the *Prior* robot continued to compromise between both strategies instead of adapting to the specific user.

E. Active Teaching

Like we saw in the previous example, learning about the human's learning strategy ψ^* can improve the robot's teaching. Hence, we here focus on selecting robot actions which *actively* gather information about ψ^*, so that the robot more quickly adapts its teaching to the end-user. Let us formulate teaching with strategy uncertainty as a partially observable Markov decision process (POMDP) [16]: the state is $\left(b^t(\theta), \theta^*, \psi^*\right)$, the action is (a^t, x^t), the observation is u^t, the state transitions with (13)—where θ^* and ψ^* are constant—the observation model is (23), and the reward is $b^t(\theta^*)$. Solving this POMDP causes the robot to optimally trade-off between exploring for more information about ψ^* and exploiting that information to maximize the human's belief in θ^*. When solving this POMDP is intractable, we can more simply perform active teaching by favoring actions that gather information about ψ^* [23]:

$$a^t = \arg\max_a \left\{ b^{t+1}(\theta^*) - \lambda \cdot H(b^{t+1}(\psi)) \right\} \quad (24)$$

In the above, $\lambda \geq 0$, and H is the Shannon entropy. Comparing (24) to (15), now the robot selects actions to disambiguate between the possible learning strategies (i.e., reduce the entropy of the robot's belief over ψ). Intuitively, we expect a robot that is actively teaching with (24) to select actions, a, which *cause users with different learning strategies to respond in different ways*, allowing that robot to more easily infer ψ^*.

VI. Robot Learning Simulations

To compare our *learning with strategy uncertainty* against the state-of-the-art in a realistic problem setting, we performed a simulated user study. We here consider an instance of inverse reinforcement learning (IRL): the human demonstrates a policy, and the robot attempts to infer the human's reward function from that demonstrated policy [3]–[5]. Unlike the example in Section IV-C, now θ^* (the human's reward parameters) and ϕ^* (the human's demonstration strategy) lie in continuous spaces. We compared robots that learn θ^* with a constant point estimate of ϕ^* to our proposed method, where the robot learns about both θ^* and ϕ^* from the human. To test the robustness of our method within more complex and challenging scenarios, we also introduced noisy end-users, who did not follow any of the modeled teaching strategies.

A. Setup and Simulated Users

Within each simulation the human and robot were given an 8-by-8 gridworld (64 states). The state reward, $R(x, \theta)$, is the linear combination of state features $f(x)$ weighted by θ, so that $R(x, \theta) = \theta \cdot f(x)$. The human knows θ^*, and provides a demonstration $\pi(u \mid x, \theta^*, \phi^*)$. This demonstration is a policy, where the human labels each state x with action u; actions deterministically move in one of the four cardinal directions. The discount factor—which defines the relative importance of future and current rewards—was fixed at $\gamma = 0.9$.

Our setting is based upon previous IRL works [5], where this problem is more formally introduced as a Markov decision process (MDP). These prior works typically assume that the human's demonstrated policy approximately solves the MDP, i.e., maximizes the expected sum of discounted rewards [8], [9]. By contrast, *we here considered users with a spectrum of demonstration strategies*. Let $Q(x, u, \theta)$ be the reward for taking action u in state x, and then following the optimal policy for reward parameters θ. We define the probability that the simulated user takes action u given x, θ^*, and ϕ^* as:

$$\pi \propto \exp\left\{ \alpha \left[Q(x, u, \theta^*) + \phi^*\left(R(x', \theta^*) - R(x, \theta^*)\right) \right] \right\} \quad (25)$$

where $\phi^* \in [-1, 1]$, and x' is the state reached after taking action u in state x. When $\phi^* = 0$, (25) is the same as the observation model from [8], [9]. As $\phi^* \to +1$, the human biases their demonstration towards states that have locally higher rewards; conversely, when $\phi^* \to -1$, the human favors states with lower rewards. Sample user demonstrations with different teaching strategies are shown in Fig. 7.

B. Independent Variables

We compared four different approaches for learning θ^* from the user's demonstration: ϕ^*, $\phi = -1$, $\phi = +1$, and *Joint*.
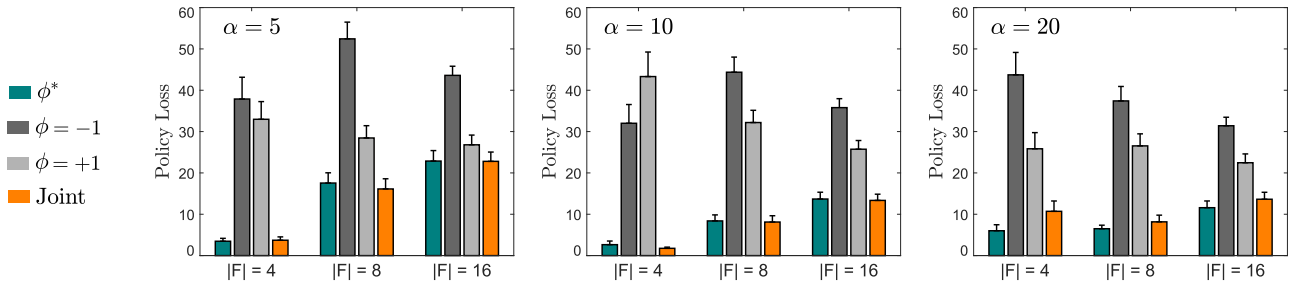
Fig. 8. Learning the human's teaching strategy in addition to learning the human's reward parameters (*Joint*) reduces the robot's *Policy Loss*. Left: $\alpha = 5$. Middle: $\alpha = 10$. Right: $\alpha = 20$. Recall that ϕ^* is an ideal learner, and lower values of α indicate increasingly random users. Error bars denote SEM.
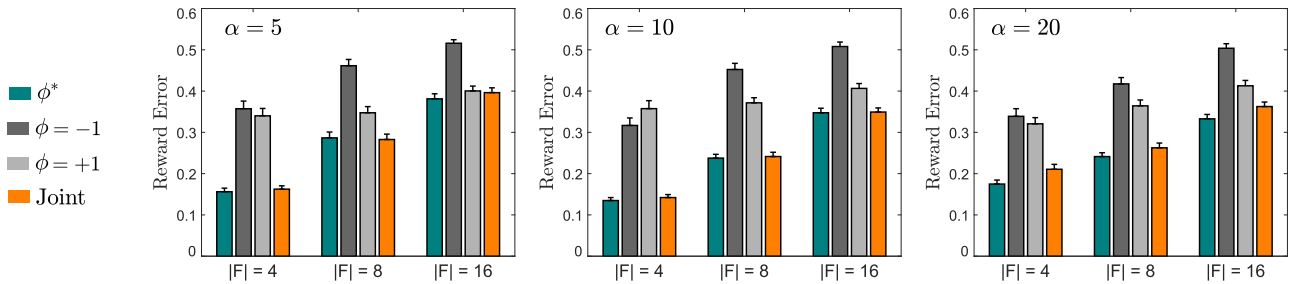

Fig. 9. The robot more accurately learns the human's reward parameter θ^* when maintaining a distribution over the teaching strategies. Left: $\alpha = 5$. Middle: $\alpha = 10$. Right: $\alpha = 20$. Note that $\phi = +1$ outperformed $\phi = -1$ because it could correctly interpret the demonstrations from a wider range of ϕ^*.

Under ϕ^* the ideal robot knows the human's true teaching strategy, while $\phi = -1$ and $\phi = +1$ indicate robots which assume that the human's demonstration is biased towards low-reward or high-reward states, respectively. *Joint* refers to a robot which attempts to learn both ϕ^* and θ^* from the human's demonstration, as discussed in Section IV-B.

To see how these approaches scale with the length of the feature vector, $f \in F$, we performed simulations with $|F| = 4$, 8, and 16 features. In practice, each state x was randomly assigned a feature vector with $|F|$ binary values, indicating which features were present in that particular gridworld state.

Finally, to test how well the robot learned when the human demonstrations were imperfect, we varied the value of α in (25). Parameter α represents how close to optimal the human is: as $\alpha \to 0$, the human becomes increasingly random, while the human always chooses the best action when $\alpha \to \infty$.

We simulated 100 users for each combination of $|F|$ and α, where the users' teaching strategies were uniformly distributed in the continuous interval $\phi^* \in [-1, 1]$. The gridworld and θ^* were randomly generated for each individual user.

C. Dependent Measures

For each simulation we measured the robot's learning performance in terms of *Reward Error*, *Strategy Error*, and *Policy Loss*. *Reward Error* is the difference between the robot's mean estimate of θ^* and the correct reward parameters: $\|\theta^* - \hat{\theta}\|_1$. Similarly, *Strategy Error* is the error between the robot's mean estimate of ϕ^* and the user's actual teaching strategy: $|\phi^* - \hat{\phi}|$. *Policy Loss* measures how much reward is lost by following the robot's learned policy (which maximizes reward under $\hat{\theta}$) as compared to the optimal policy for θ^* [8].

The code for our examples and simulations can be found at https://github.com/dylanplosey/iact_strategy_learning.

D. Results and Discussion

We performed a mixed ANOVA with the number of features and value of α as between-subjects factors, and the learning approach as a within-subjects factor, for both *Policy Loss* and *Reward Error* (see Figs. 8 and 9). Since we found a statistically significant interaction for both dependent measures ($p < .05$), we next determined the simple main effects.

Simple main effects analysis showed that *Joint* resulted in significantly less *Policy Loss* than either $\phi = -1$ or $\phi = +1$ for each different combination of $|F|$ and α ($p < .05$). We similarly found that *Joint* resulted in significantly less *Reward Error* ($p < .001$) for every case except $|F| = 16$, $\alpha = 5$; here there was no statistically significant difference between *Joint* and $\phi = +1$ ($p = .498$). These results from Figs. 8 and 9 suggest that learning while maintaining a distribution over ϕ results in objectively better performance than learning with a fixed point estimate of ϕ^*.

Next, we investigated how well the *Joint* method learned the individual users' teaching strategies. We performed a two-way ANOVA to find the effects of $|F|$ and α on the *Joint* robot's *Strategy Error* (see Fig. 10). We found that the number of features ($F(2, 891) = 23.813, p < .001$) and the human's α ($F(2, 891) = 22.679, p < .001$) had a significant main effect. Post-hoc analysis with Tukey HSD revealed that $|F| = 16$ and $\alpha = 20$ led to significantly higher *Strategy Error* than the other values of $|F|$ and α, respectively. As shown in Fig. 10, the robot had larger *Strategy Error* for higher values of α because it was unable to distinguish between teachers with $\phi^* > 0$; i.e., these
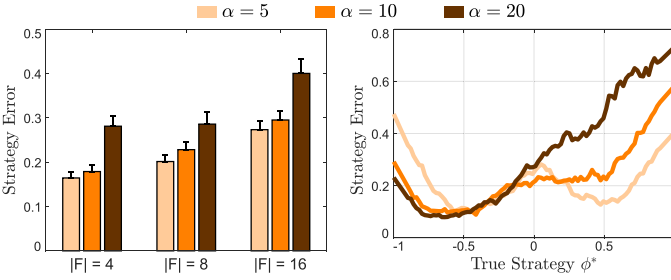
Fig. 10. Error between the learned and true teaching strategies under *Joint*. Left: *Strategy Error* for different numbers of features and values of α. Right: *Strategy Error* (averaged across features) as a function of the user's teaching strategy. Teachers with $\phi^* < 0$ were easier to distinguish than with $\phi^* > 0$.
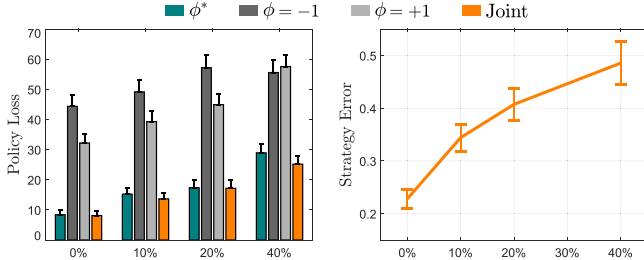

Fig. 11. Learning from users who provided demonstrations with unmodeled noise. Left: *Policy Loss* for each learning approach as the noise ratio increased. Right: *Strategy Error* (for *Joint*) as a function of the noise ratio. Although the *Joint* robot's *Strategy Error* increased in proportion to the noise ratio during the user demonstrations, its *Policy Loss* was on par with the ideal learner, ϕ^*.

different teachers provided similar policy demonstrations when $\alpha = 20$.

Finally, we conducted a followup simulation in which we introduced unmodeled noise (see Fig. 11). Here $|F| = 8$ and $\alpha = 10$, but we now increased the ratio of the human taking completely random actions, which were not modeled in (25). *Joint* resulted in significantly less *Policy Loss* than $\phi = -1$ or $\phi = +1$, even as the ratio of unmodeled user noise increased. Hence, reasoning over multiple strategies still improved performance for cases where the noisy end-user did not comply with any of the modeled teaching strategies.

E. Challenges and Limitations

Although this simulated user-study supports learning with strategy uncertainty, there are still practical challenges that may limit our proposed approach. In particular, the end-user's actual interactions may not match any of the learning or teaching models, such that $\phi^* \notin \Phi$ or $\psi^* \notin \Psi$. Having the wrong hypothesis space is often unavoidable when using models to learn from humans: but, as we show in Fig. 11, our proposed approach does remain robust to some errors in the hypothesis space (such as noisy users, that do not follow any of the included models). In practice, designers could leverage data from previous trials to construct a richer space of possible interaction strategies, so that Φ is updated to include ϕ^*.

VII. CONCLUSION

Because the human's interaction strategy during HRI varies from end-user to end-user, robots that assume a fixed, predefined interaction strategy may result in inefficient, confusing interactions. Thus, we proposed that the robot should maintain

a distribution over the human interaction strategies, and *exchange information while reasoning over this distribution*. We here introduced robot (a) learning with strategy uncertainty and (b) teaching with strategy uncertainty, and derived solutions to both novel problems. We performed learning and teaching examples—as well as learning simulations—and compared our approach to the state-of-the-art. Unlike standard approaches that assume every user interacts in the same way, we found that attempting to infer each individual end-user's interaction strategy led to improved robot learning and teaching, while remaining robust to unmodeled strategies.

REFERENCES

[1] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robot. Auton. Syst.*, vol. 57, no. 5, pp. 469–483, 2009.

[2] X. Zhu, "Machine teaching: An inverse problem to machine learning and an approach toward optimal education," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 4083–4087.

[3] A. Y. Ng *et al.*, "Algorithms for inverse reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2000, pp. 663–670.

[4] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2004.

[5] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters, "An algorithmic perspective on imitation learning," *Found. Trends Robot.*, vol. 7, no. 1-2, pp. 1–179, 2018.

[6] J. Choi and K.-E. Kim, "MAP inference for Bayesian inverse reinforcement learning," in *Proc. Int. Conf. Neural Inform. Process. Syst.*, 2011, pp. 1989–1997.

[7] D. P. Losey and M. K. O'Malley, "Including uncertainty when learning from human corrections," in *Proc. Conf. Robot Learn.*, 2018, pp. 123–132.

[8] D. Ramachandran and E. Amir, "Bayesian inverse reinforcement learning," *Urbana*, vol. 51, no. 61801, pp. 1–4, 2007.

[9] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning," in *Proc. AAAI Conf. Artif. Intell.*, 2008, vol. 8, pp. 1433–1438.

[10] S. Nikolaidis, Y. X. Zhu, D. Hsu, and S. Srinivasa, "Human-robot mutual adaptation in shared autonomy," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, 2017, pp. 294–302.

[11] W. Liu, B. Dai, X. Li, J. M. Rehg, and L. Song, "Iterative Machine Teaching," in *Proc. Int. Conf. Machine Learning (ICML)*, 2017, arXiv:1710.07742.

[12] F. S. Melo, C. Guerra, and M. Lopes, "Interactive optimal teaching with unknown learners," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 2567–2573.

[13] A. N. Rafferty, E. Brunskill, T. L. Griffiths, and P. Shafto, "Faster teaching via POMDP planning," *Cogn. Sci.*, vol. 40, no. 6, pp. 1290–1332, 2016.

[14] S. H. Huang, D. Held, P. Abbeel, and A. D. Dragan, "Enabling robots to communicate their objectives," in *Proc. Auton. Robots*, 2017, pp. 1–18.

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available:http://www.deeplearningbook.org

[16] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 2009.

[17] X. Zhu, "Machine teaching for Bayesian learners in the exponential family," in *Proc. Int. Conf. Neural Inform. Process. Syst.*, 2013, pp. 1905–1913.

[18] N. Chater and M. Oaksford, *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*. London, U.K.: Oxford Univ. Press, 2008.

[19] M. Jones and B. C. Love, "Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition," *Behavioral Brain Sci.*, vol. 34, no. 4, pp. 169–188, 2011.

[20] K. J. Åström and B. Wittenmark, *Adaptive Control*, 2nd ed. Mineola, NY, USA: Dover, 2008.

[21] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[22] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, 2013, pp. 301–308.

[23] D. Sadigh, S. S. Sastry, S. A. Seshia, and A. Dragan, "Information gathering actions over human internal state," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 66–73.