Robots that Take Advantage of Human Trust

Dylan P. Losey and Dorsa Sadigh

Abstract— Humans often assume that robots are rational. We believe robots take optimal actions given their objective; hence, when we are uncertain about what the robot's objective is, we interpret the robot's actions as optimal with respect to our estimate of its objective. This approach makes sense when robots straightforwardly optimize their objective, and enables humans to learn what the robot is trying to achieve. However, our insight is that—when robots are aware that humans learn by *trusting* that the robot actions are rational—intelligent robots do not act as the human expects; instead, they take advantage of the human's trust, and exploit this trust to more efficiently optimize their own objective. In this paper, we formally model instances of human-robot interaction (HRI) where the human does not know the robot's objective using a two-player game. We formulate different ways in which the robot can model the uncertain human, and compare solutions of this game when the robot has conservative, optimistic, rational, and trusting human models. In an offline linear-quadratic case study and a real-time user study, we show that trusting human models can naturally lead to communicative robot behavior, which influences end-users and increases their involvement.

I. INTRODUCTION

When humans interact with robots, we often assume that the robot is rational [1]. Imagine a human helping a robot to set the table. The human end-user has told the robot to bring two plates from the kitchen; but, when the robot returns, it is only carrying one. Based on this action—and our assumption that robots are rational—the end-user learns about the robot's objective. For instance, here the human infers that the robot assigns a high cost to carrying multiple plates at once.

A level of human *trust* is implicit within this interpretation. If we assume that the robot takes actions purely to optimize its objective, we trust that these actions are *not meant* to influence the user. In our example, a trusting human does not realize that the robot may actually be manipulating them into carrying the second plate. This robot is capable of bringing two plates, but misleads the human to believe that it can only carry one in order to reduce its overall effort. Hence, trusting the robot leads to incorrect human learning: the human does not learn the robot's true objective, but instead the objective the robot has chosen to convey. This behavior is useful for assistive robots, where humans can become overly reliant on the robot's assistance [2], and robots need to make intelligent decisions that motivate user participation.

Within this work, we study HRI where the human does not know the robot's objective. We recognize that:

If the robot is aware of the human's trust, an optimal robot should select its actions so as to communicate *with the*

The authors are with the Stanford Intelligent and Interactive Autonomous Systems Group (ILIAD), Department of Computer Science, Stanford University, Stanford, CA 94305. (e-mail: dlosey@stanford.edu)

human and influence *their understanding.*

Our insight results in autonomous agents that intentionally communicate with nearby humans as a part of their optimal behavior. In practice, however, we find that it is not always in the robot's best interests to communicate its true objective; by contrast, *misleading actions naturally emerge* when robots model humans as trusting agents. This indicates that—if the robot has a different objective than the human—user trust becomes something that the robot can exploit.

Overall, we make the following contributions:

Formulating Human Models. We formulate four ways that robots can model uncertain humans within a two-player HRI game. The robot can behave conservatively (as if the human already knew their objective), optimistically (as if they can teach the human any objective), model the human as rational, or treat the human as a trusting agent. Trusting is a simplification of the rational model: the human learns while assuming that the robot purely optimizes its objective.

Comparison Between Models. We show that modeling the human as trusting is not the same as acting optimistically, but can result in robots that take advantage of the human.

Experiments. To understand the relationship between trusting human models and communicative or influential robot behavior, we perform an offline linear-quadratic case study and a real-time user study. Unlike the conservative and optimistic human models, communicative robot behavior naturally emerges when the robot models the human as trusting, causing the human participants to adapt their behavior and increase their overall involvement.

This work is a contribution towards developing communicative robots, and explores the interplay between how humans model robots and how optimal robots exploit these models.

II. RELATED WORK

Previous research has studied trust during HRI, along with the robotic and environmental factors that affect the human's trust [3]. Autonomous agents can model and infer the human's trust based on past interactions [4], and increase the human's trust by making the robot's actions transparent to explain its decision making process [5]. Most relevant here is work by Chen *et al.* [6], who treat the human's trust as a latent variable, and integrate trust into the robot's decision making. These previous works [3]–[6] consider cases where the human and collaborative robot share a common objective and the robot should maximize trust; by contrast, we are interested in robots that leverage trust to influence users.

Other research focuses on robots that utilize their actions to implicitly communicate with nearby humans. Communica-
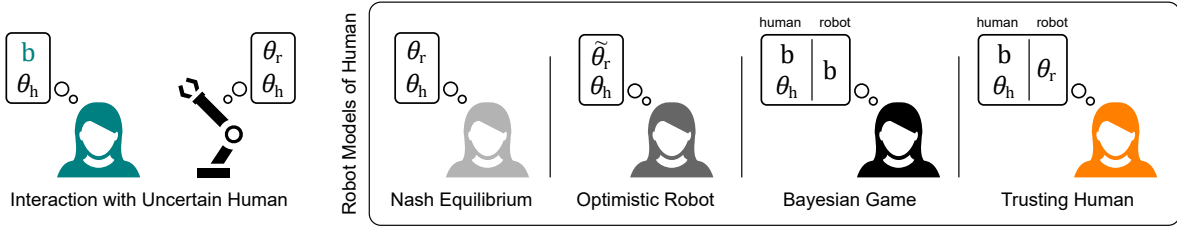
Fig. 1. Human-robot interaction where the human does not know the robot's objective, θ_r. The robot chooses its behavior based on its model of the human. A conservative robot (Nash) acts as if the human knew θ_r, while a optimistic robot (Optimistic) thinks that it can teach the human an arbitrary objective $\tilde{\theta}_r$. Rather than acting conservatively or optimistically, the robot can treat the human as a rational agent (Bayesian), where the human recognizes that the robot is aware of their uncertainty b. Because real humans are not completely rational, we propose a simpler alternative (Trusting): here the human learns about θ_r from the robot while assuming that the robot behaves conservatively.

tive robot actions facilitate human understanding during the task [7] so that humans can anticipate what the robot will do in the future. In order to identify communicative actions, we need a cognitive model of how the human learns: the human might respond to the robot's actions [8], infer the robot's objective [9], or follow a personalized learning strategy [10]. Our research builds upon these prior works [7]–[11] by identifying communicative actions for a trusting human model. Unlike research that explicitly generates deceptive actions [12], here the robot naturally selects misleading communication. This is similar to learning with opponent-learning awareness [13]: here only the robot is aware.

We can also frame our work as a contribution to perspective taking, where the robot reasons about the human's perspective while make decisions [14]–[16]. Prior research in perspective taking considers which objects are accessible to the human (i.e., what can the human see or reach), and how the robot can reduce human confusion or communicate its intention through proactive, contextual behavior. In our work, the robot assumes the human has a trusting perspective (or theory of mind), and solves for optimal responses given this human perspective.

III. PROBLEM FORMULATION

Consider a human interacting with a robot. The human and robot each have their own objective, and take actions in order to achieve that objective. We are interested in settings where the robot *knows* both the robot and human objectives, but the human *does not know* the robot's objective. While many prior works (e.g., imitation learning) focus on expert humans—who already have access to all the necessary information, and must teach that information to the robot learner—here we explore the converse situation.

In this section, we first model human-robot interaction as a two-player game. Next, we formulate four different strategies for selecting robot actions, where each strategy is optimal given the corresponding human model (see Fig. 1). Finally, we discuss the plate carrying example mentioned in Sec. I to compare these alternative formulations.

A. Human Robot Interaction as a Two-Player Game

We model human-robot interaction (HRI) as a two-player Markov game, where both agents affect the state of a dynamical system. Let the system have state $x \in \mathcal{X}$, let the robot take action $u_r \in \mathcal{U}_R$, and let the human take action $u_h \in \mathcal{U}_H$.

The state transitions according to: $x^{t+1} = f(x^t, u_r^t, u_h^t)$, where f captures the system dynamics. This game starts at $t = 0$, and ends after $H \in \mathbb{Z}^+$ timesteps when $t = H$.

Rewards. The robot has a reward function: $r_r(x, u_r, u_h, \theta_r)$, and the human has a reward function: $r_h(x, u_r, u_h, \theta_h)$. These reward functions are parameterized by the constants θ_r and θ_h, which capture the robot and human objectives, respectively. Notice that we do not assume that $\theta_h = \theta_r$; hence, the human and robot may have different objectives.

Game. During each timestep the human and robot observe the current state x^t and take simultaneous actions u_h^t and u_r^t. The robot receives reward $r_r(x^t, u_r^t, u_h^t, \theta_r)$ and the human receives reward $r_h(x^t, u_r^t, u_h^t, \theta_h)$. Both agents are attempting to maximize their own sum of rewards over the finite horizon H. Denote the sequence of robot actions as $\boldsymbol{u_r} = \{u_r^0, \ldots, u_r^H\}$ and the sequence of human actions as $\boldsymbol{u_h} = \{u_h^0, \ldots, u_h^H\}$. The sum of rewards for the robot and human, respectively, are:

$$R_r(x^0, \boldsymbol{u_r}, \boldsymbol{u_h}, \theta_r) = \sum_{t=0}^{H} r_r(x^t, u_r^t, u_h^t, \theta_r) \qquad (1)$$

$$R_h(x^0, \boldsymbol{u_r}, \boldsymbol{u_h}, \theta_h) = \sum_{t=0}^{H} r_h(x^t, u_r^t, u_h^t, \theta_h) \qquad (2)$$

Beliefs over reward functions. The robot knows both θ_r and θ_h, while the human only has access to its own reward parameters θ_h. Let $\mathcal{D}$ be the probability distribution from which the parameter θ_r is sampled; we assume that the human is aware of this distribution $\mathcal{D}$, and updates their belief over the robot's objective based on the robot's actions. More specifically, the human has a belief: $b^{t+1}(\theta_r) = P(\theta_r \mid x^{0:t}, u_r^{0:t}, u_h^{0:t}, \theta_h, \mathcal{D})$.

B. Formulating Robot's Responses to Humans

Given our description of HRI as a two-player game, we want to plan for the robot's optimal behavior. Importantly, the robot's optimal behavior depends on *how the robot models the human*. Recall our plate carrying example: if the robot models the human as a rational agent (where the robot cannot mislead the user) the robot should bring two plates, but if the robot models the human as a trusting agent, it should only bring one. Below we formulate optimal robot behavior in response to *four different* human models within the context of a two-player game.

1) Nash Equilibrium. This robot acts as if the human also knows θ_r, so that both the human and robot objectives are common knowledge. When each agent has all the information, their optimal behavior can be computed by finding the Nash equilibrium of this game. Throughout this work, we assume that a pure-strategy Nash equilibrium exists, and—if there are multiple equilibria—both agents coordinate to follow the same one. More formally, $(\boldsymbol{u_r^{*,n}}, \boldsymbol{u_h^{*,n}})$ is a strict pure-strategy Nash equilibrium over the finite horizon H if:

$$\boldsymbol{u_r^{*,n}} = \arg\max_{\boldsymbol{u_r}} \ R_r\big(x^0, \boldsymbol{u_r}, \boldsymbol{u_h^{*,n}}(\theta_r, \theta_h), \theta_r\big)$$
$$\boldsymbol{u_h^{*,n}} = \arg\max_{\boldsymbol{u_h}} \ R_h\big(x^0, \boldsymbol{u_r^{*,n}}(\theta_r, \theta_h), \boldsymbol{u_h}, \theta_h\big) \quad (3)$$

We emphasize that the human and robot policies $\boldsymbol{u_h^{*,n}}(\theta_r, \theta_h)$ and $\boldsymbol{u_r^{*,n}}(\theta_r, \theta_h)$ can be specified as a function of the reward parameters θ_r and θ_h since these policies are optimizers of the reward functions (1) and (2).

A robot should follow $\boldsymbol{u_r^{*,n}}$ when the human knows θ_r. But, in our setting, the human does not initially know θ_r; hence, we can think of $\boldsymbol{u_r^{*,n}}$ as a *conservative* robot, which never tries to exploit the human's uncertainty.

2) Optimistic Robot. At the other end of the spectrum, the robot can behave *optimistically*, and assume that—because the human does not know the robot's objective—the robot can teach the human an arbitrary objective $\tilde{\theta}_r$:

$$\boldsymbol{u_r^{*,s}} = \arg\max_{\boldsymbol{u_r}} \ R_r\big(x^0, \boldsymbol{u_r}, \boldsymbol{u_h^{*,s}}(\tilde{\theta}_r, \theta_h), \theta_r\big)$$
$$\boldsymbol{u_h^{*,s}} = \arg\max_{\boldsymbol{u_h}} \ R_h\big(x^0, \boldsymbol{u_r^{*,n}}(\tilde{\theta}_r, \theta_h), \boldsymbol{u_h}, \theta_h\big) \quad (4)$$

where the robot picks $\tilde{\theta}_r$. Comparing (4) to (3), here the human responds as if the robot's objective is $\tilde{\theta}_r$ rather than the true objective θ_r.

3) Bayesian Game. Instead of acting conservatively or optimistically, the robot can model the human as a *rational* agent, where this human solves the HRI game. Given that the human has a belief b over the robot's objective, we formalize rational human behavior as a Bayesian equilibrium [17]:

$$\boldsymbol{u_r^{*,b}} = \arg\max_{\boldsymbol{u_r}} \ R_r\big(x^0, \boldsymbol{u_r}, \boldsymbol{u_h^{*,b}}(b(\theta_r), \theta_h), \theta_r\big)$$
$$\boldsymbol{u_h^{*,b}} = \arg\max_{\boldsymbol{u_h}} \ \mathbb{E}_{\theta \sim b}\Big[R_h\big(x^0, \boldsymbol{u_r^{*,b}}(\theta, \theta_h), \boldsymbol{u_h}, \theta_h\big)\Big] \quad (5)$$

b updates based on the robot's actions using Bayesian inference, i.e., $b^{t+1}(\theta_r) = b^t(\theta_r) \cdot P(u_r^t \mid x^{0:t}, u_r^{0:t}, u_h^{0:t}, \theta_h, \theta_r)$.

Notice that, within a Bayesian game, the human realizes that (a) the robot knows about its belief b, and (b) the robot takes actions while considering this belief.

4) Trusting Human. In reality, humans are not completely rational [18], [19]. Instead of modeling the human as rational, the robot can leverage a simplified model where the human is *trusting*. This trusting human assumes that—even though they do not know the robot's reward function—*the robot will act conservatively, as if the human knows θ_r*:

$$\boldsymbol{u_r^{*,\tau}} = \arg\max_{\boldsymbol{u_r}} \ R_r\big(x^0, \boldsymbol{u_r}, \boldsymbol{u_h^{*,\tau}}(b(\theta_r), \theta_h), \theta_r\big)$$
$$\boldsymbol{u_h^{*,\tau}} = \arg\max_{\boldsymbol{u_h}} \ \mathbb{E}_{\theta \sim b}\Big[R_h\big(x^0, \boldsymbol{u_r^{*,n}}(\theta, \theta_h), \boldsymbol{u_h}, \theta_h\big)\Big] \quad (6)$$

Comparing (6) to (5), a trusting human assumes that the robot follows $\boldsymbol{u_r^{*,n}}$, while a rational human understands that the robot reasons over their belief b.

Let us return to our plate carrying example. A rational human, $\boldsymbol{u_h^{*,b}}$, recognizes that the robot may be trying to mislead them, and is therefore always reluctant to carry a plate. By contrast, a trusting human, $\boldsymbol{u_h^{*,\tau}}$, assumes that the conservative robot will never try to take advantage of their uncertainty: if the robot only brings a single plate, this indicates to the trusting human that robot cannot carry more, and so the user carries a second plate to assist the robot. We present a more formal description of this example below.

C. Plate Carrying Example

A human and a robot are carrying plates to set the table. Both agents want more plates to be placed upon the table, but prefer for the other agent to bring these plates. The system state in this problem is x, where $x \in \mathbb{Z}^+$ is the total number of plates on the table. During every timestep, the robot and human can each bring up to two new plates: $\mathcal{U}_R = \mathcal{U}_H = \{0, 1, 2\}$, and hence the system dynamics are: $x^{t+1} = x^t + u_r^t + u_h^t$. We let $\theta_r, \theta_h > 0$, and define the robot and human rewards so that θ_r and θ_h individually penalize the robot and human for carrying plates. More specifically, for each timestep $t < H$: $r_r(x, u_r, u_h, \theta_r) = -\theta_r u_r^2$ and $r_h(x, u_r, u_h, \theta_h) = -\theta_h u_h^2$. At the final timestep $t = H$ each agent is rewarded for the total number of plates set on the table: $r^H(x) = x^H$. For simplicity, we constrain this example problem so that the human and robot cannot carry plates at the same time. To do this, we added a penalty $-\alpha \cdot u_r u_h$ to both r_r and r_h, where $\alpha \to \infty$ makes it prohibitively expensive for the agents to carry plates simultaneously.

In the next section we list how the robot should optimally carry plates for different human models.

IV. PLANNING FOR INTERACTION WITH UNCERTAIN HUMANS

We have introduced four formulations for identifying robot behavior when the human does not know the robot's objective, where each formulation is optimal given a different human model: conservative, optimistic, rational, and trusting. Next, we focus on how robots that model humans as trusting agents *compare* to these alternative formulations. We first demonstrate that we can express HRI with a trusting human as a Markov decision process, and, accordingly, we can leverage existing techniques to plan for interactions with trusting humans that do not know θ_r. We then compare the trusting human model to Bayesian games using our plate carrying example, and show that—while responding to trusting humans is not the same as acting optimistically—an intelligent robot *takes advantage* of the human's trust.

A. Trusting Humans as a Markov Decision Process

Here, we revisit the two-player game outlined in Section III-A, and use the trusting human model (6) to reformulate this game as an instance of a single-agent Markov decision process (MDP). Expressing the trusting human

TABLE I

Formulation	Nash	Optimistic	Bayesian	Trusting	Bayesian	Trusting
Distribution $\mathcal{D}$	–	–	$\mathcal{U}(0, 1.5)$	$\mathcal{U}(0, 1.5)$	$Boltzmann(a = 0.1)$	$Boltzmann(a = 0.1)$
Robot actions at $t = 0, 1$	2, 2	0, 0	1, 0	1, 0	2, 2	1, 0
Human actions at $t = 0, 1$	0, 0	2, 2	0, 2	0, 2	0, 0	0, 2
Mean human estimate of θ_r	0.2	> 1	4/7	2/3	0.22	0.43
R_r, R_h	2.4, 4	4, 2	2.8, 2	2.8, 2	2.4, 4	2.8, 2

model as an MDP enables us to leverage existing methods to solve for $\boldsymbol{u_r^{*,\tau}}$. Our insight is recognizing that—instead of the state x—this MDP has an augmented state s that captures the necessary information to model the trusting human.

MDP. Interacting with trusting humans that do not know the robot's objective is a finite-horizon MDP defined by the tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{U}_R, T, \theta_r, H \rangle$ [20]. We introduced some of these terms in Section III-A: $\mathcal{U}_R$ is the robot's action space, θ_r is the robot's objective, and H is the finite planning horizon.

State. Define the augmented state $s \in \mathcal{S}$ to be $s = (x, \theta_h, b)$, where x is the system state, θ_h is the human's objective, and b is the human's belief over the robot's objective.

Prediction. Given s^t the robot can predict the current human action u_h^t under the trusting human model. First, using θ_h, the robot solves (3) to obtain $\boldsymbol{u_r^{*,n}}(\theta, \theta_h)$ for all $\theta_r \in \Theta_R$. Next, using b^t, x^t, $\boldsymbol{u_r^{*,n}}(\theta, \theta_h)$, and θ_h, the robot obtains $\boldsymbol{u_h^{*,\tau}}$ with (6). $\boldsymbol{u_h^{*,\tau}}$ is a sequence; the t-th element of $\boldsymbol{u_h^{*,\tau}}$ is u_h^t. Repeating this process at every s yields the human policy.

Transition. The transition function $T(s^{t+1} \mid s^t, u_r^t)$ captures how s evolves over time. We previously defined the system dynamics as $x^{t+1} = f(x^t, u_r^t, u_h^t)$, θ_h is a constant, and b updates based on how the human learns (e.g., Bayesian inference or gradient descent).

Underactuated Dynamics of Interaction. When uncertain humans are trusting agents, the interaction between the human and robot becomes an *underactuated* system. This comes from our MDP reformulation: the trusting human policy is a function of s, and the robot employs its actions u_r to control how s transitions. Consider our plate carrying example: if the robot brings two plates, the human will carry no plates, but if the robot brings a single plate, the human carries the second plate. Hence, the robot uses its own actions to indirectly control (i.e., influence) the human's behavior.

B. Comparing Trusting Humans to Other Human Models

Now that we know how to identify optimal behavior when interacting with trusting humans, we explore how this behavior compares to interacting under other human models. Using our plate carrying example, we compute how uncertain humans and conservative (Nash) or optimistic robots interact, as well as robots that model the human as completely rational (Bayesian). By comparing examples of this behavior to our trusting model, we demonstrate that there are situations where robots take advantage of the human's trust.

Computing Solutions. We list the robot and human behavior for one interesting case of our plate carrying example in Table I. When $\theta_h, \theta_r < 1/3$, we conservatively assume that the agents follow the Nash equilibrium (3) where the robot carries the plates. For Optimistic (4), the robot chooses $\tilde{\theta}_r > 1$, so that it always makes sense for the human to carry plates.

We list the Bayesian (5) and Trusting (6) models under two different distributions $\mathcal{D}$. Intuitively, when θ_r is sampled from a uniform distribution the rational human is very uncertain about the robot's objective, and so the robot can induce the human to carry plates. On the other hand, when θ_r is sampled from a Boltzmann distribution, the human is confident that the robot's objective is low, and thus the robot becomes unable to mislead this rational human. More specifically, a rational human recognizes that the Bayesian robot plays $u_r = 1$ for $\theta_r \in [1/7, 1]$ with a uniform prior, and learns a mean estimate $4/7$. A trusting human thinks the robot only plays $u_r = 1$ for $\theta_r \in [1/3, 1]$, and so their mean estimate becomes $2/3$ with a uniform prior. These computations are similar for the Boltzmann $\mathcal{D}$, but, because the human is confident that $\theta_r < 1/3$, playing $u_r = 1$ is not sufficient to convince a rational human that $\theta_r > 1/3$.

Analysis. The results from our the plate carrying example provide a useful comparison between the trusting human model and our other formulations.

First, we observe that interacting with a trusting human is not necessarily the same as behaving optimistically. Consider the example in Table I: while the optimistic robot plans for the best possible case—where the human carries all the plates—the robot interacting with a trusting end-user recognizes that the human's actions depend on their belief. This robot reasons about how its behavior influences the human's understanding of the robot's objective: within the example, the robot initially brings a single plate to convince the human to carry additional plates at the next timestep.

Second, there are situations where robots interacting with trusting humans take advantage of that human's trust. Consider the Boltzmann distribution case in Table I: when the human trusts the robot, the robot intentionally misleads that human into carrying additional plates, but, when the human is rational (Bayesian), the robot carries all the plates. In this case, being trusting leads to less reward for the human than being rational: $R_h(x^0, \boldsymbol{u_r^{*,\tau}}, \boldsymbol{u_r^{*,\tau}}, \theta_h) < R_h(x^0, \boldsymbol{u_r^{*,b}}, \boldsymbol{u_r^{*,b}}, \theta_h)$.

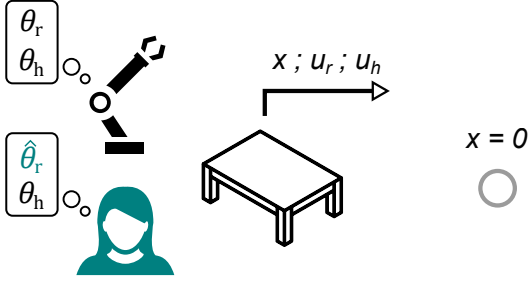**Summary.** When comparing robots planning under a trusting

Fig. 2. Setup for our case study. The human and robot are pushing a table with linear dynamics to the goal state ($x = 0$). Both agents want to minimize a quadratic cost function that considers (a) error from the goal and (b) their own effort when pushing the table. The human has a point estimate $\hat{\theta}_r$ of θ_r, which the human updates using gradient descent with learning rate η.

human model to Bayesian games, we found that interacting with trusting humans is not the same as a robot behaving optimistically, but it can lead to cases where trusting humans receive less reward than rational humans.

V. CASE STUDY

To understand how modeling the human as a trusting agent is related to *communicative* robot behavior, we performed a simulated user study. This study focuses on a *linear-quadratic* (LQ) problem, where the system dynamics are linear and the human and robot reward functions are quadratic: our problem setup is shown in Fig. 2. Here the robot knows both the human and robot objectives, but the human only has a point estimate of the robot's objective. We first considered humans that learn about the robot's objective from the robot's behavior, and demonstrate how our approach leads to robots that purposely communicate with and mislead the trusting human. Next, we added users that did not follow our trusting human model, and measured performance during interactions when the trusting human model was inaccurate.

A. Simulation Setup

During each simulation, the human and robot pushed a one degree-of-freedom mass-damper table towards a goal state. Let $x = [p, v]^\top \in \mathbb{R}^2$ be the position p and velocity v relative to the goal state, let m be the mass of the table, and let b be a viscous friction constant. The linear system dynamics are:

$$x^{t+1} = Ax^t + B(u_r^t, u_h^t)$$

$$A = \begin{bmatrix} 1 & \Delta \\ 0 & 1 - b\Delta/m \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \Delta/m \end{bmatrix} \tag{7}$$

where Δ is the period between t and $t+1$. The agents wanted to minimize the distance of table from the goal at timestep H, while also minimizing their own in-task effort:

$$R_r(x^0, \boldsymbol{u_r}, \boldsymbol{u_h}, \theta_r) = -\|x^H\|^2 - \theta_r \cdot \Delta \boldsymbol{u_r}^\top \boldsymbol{u_r} \tag{8}$$

$$R_h(x^0, \boldsymbol{u_r}, \boldsymbol{u_h}, \theta_h) = -\|x^H\|^2 - \theta_h \cdot \Delta \boldsymbol{u_h}^\top \boldsymbol{u_h} \tag{9}$$

Here $\theta_r, \theta_h \in \mathbb{R}$ are constants that determine the trade-off between error and effort. The robot knew θ_r and θ_h, while the human only knew θ_h and had a point estimate $\hat{\theta}_r$ of

θ_r. Similar to prior work [21], we assumed that the human updated their point estimate using gradient descent:

$$\hat{\theta}_r^{t+1} = \hat{\theta}_r^t + \eta \cdot \nabla_\theta \left| \boldsymbol{u_r}^{*,n}(\theta, \theta_h) - u_r^t \right| \tag{10}$$

with learning rate $\eta > 0$. Thus, the human learned based on the difference between the predicted robot behavior, $\boldsymbol{u_r}^{*,n}$, and the actual robot action, u_r. When both agents know θ_r and θ_h, this HRI problem is an instance of a deterministic linear-quadratic game [22]. Control strategies that achieve a Nash equilibrium are of the form $u_r^t = -K_r^t x^t$ and $u_h^t = -K_h^t x^t$, where $K_r, K_h \in \mathbb{R}^{1 \times 2}$ are feedback gain matrices [22], [23]. We note that—within this deterministic LQ game—a pure Nash equilibrium exists and is unique.

B. Independent Variables

We compared three formulations for selecting robot behavior: conservatively following the Nash equilibrium (*Nash*), optimistically solving the two-player game (*Optimistic*), or computing optimal behavior for a trusting human (*Trust*). For *Nash* the robot follows (3), for *Optimistic* the robot solves (4), and for *Trust* the robot leverages (6). Each of these robot strategies makes a different assumption about the human user: *Nash* acts as if the human knew θ_r, *Optimistic* assumes the human believes the robot objective is $\hat{\theta}_r$, and *Trust* simplifies the rational human model.

To assess the performance of these robotic approaches with different types of users, we also varied the human's interaction strategy. There were three types of simulated users:

- **Learn,** a trusting end-user that updates its point estimate $\hat{\theta}_r$ using (10). Within this category we simulated humans that learn at different rates: $\eta = \{0, 2.5, 5, 10, 20\}$.
- **Fixed,** an end-user who assumes $\hat{\theta}_r^{t+1} = \hat{\theta}_r^t$, and always follows the Nash equilibrium for their initial estimate of θ_r. We can think of this as an end-user who never adapts to the robot during HRI. Fixed is the same as Learn when $\eta = 0$, but we use Fixed to test situations where the robot incorrectly believes $\eta > 0$ (e.g., $\eta = \{5, 10, 20\}$).
- **Predict,** an end-user who predicts that the next N robot actions will be the same as u_r^t, and responds accordingly.

Our simulations explored how *Nash*, *Optimistic*, and *Trust* interact with these different types of simulated humans.

Implementation Details. Each trial we sampled $\theta_r \sim \mathcal{N}(5, 0.5)$, $\theta_h \sim \mathcal{N}(1, 0.5)$, and $\hat{\theta}_r^0 \sim \mathcal{U}(0.05, 0.15)$. We set the mean value of θ_r to be higher than $\hat{\theta}_r^0$ to focus on interesting cases where the human initially believed that the robot cared less about its effort than it actually did. The parameters in (7) were $m = 0.5$ kg, $b = 1$ N·s/m, $\Delta = 0.5$ s, $x^0 = [-1, 0]^\top$, and $H = 10$ s.

C. Dependent Measures

We report the total cost (i.e., negative reward) of the HRI task: *Robot Cost* is $-R_r$ from (8) and *Human Cost* is $-R_h$ from (8). We also measure how communicative the robot's
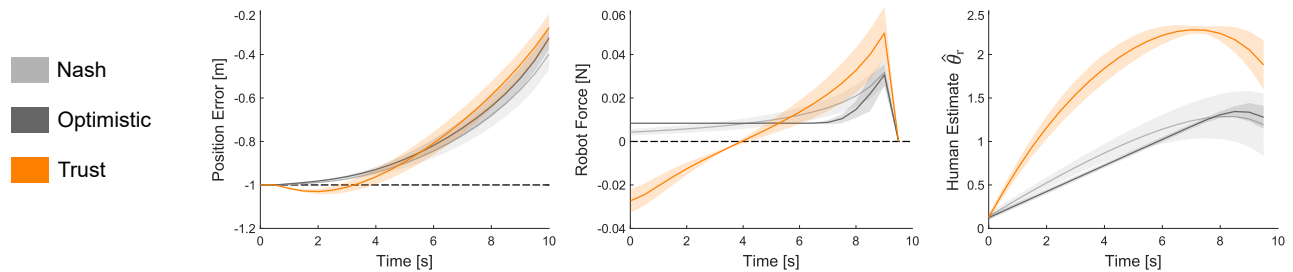
Fig. 3. Example interaction with a trusting human that learns from robot actions (learning rate $\eta = 5$). Left: the position of the table with respect to the goal. Middle: the robot's applied force. Right: the human's estimate $\hat{\theta}_r$ of the robot's true objective θ_r. Shaded regions show the standard error of the mean (SEM) across trials. The *Trust* robot is aware of the human's trust, and intentionally moves the table away from the goal at the start of the task in order to convince the human to complete the rest of the task autonomously. Although *Trust* applies more total effort than *Nash* and *Optimistic*, it ultimately obtains a higher reward since it quickly convinces the human to take over, and thus completes the task with the least state error. Notice that the human learned an increased value for $\hat{\theta}_r$ when the robot applied small forces towards the goal (*Nash* and *Optimistic*), but pulling the table away from the goal (*Trust*) resulted in the largest human estimate $\hat{\theta}_r$, better misleading the human into believing that the robot was uninterested in the goal.
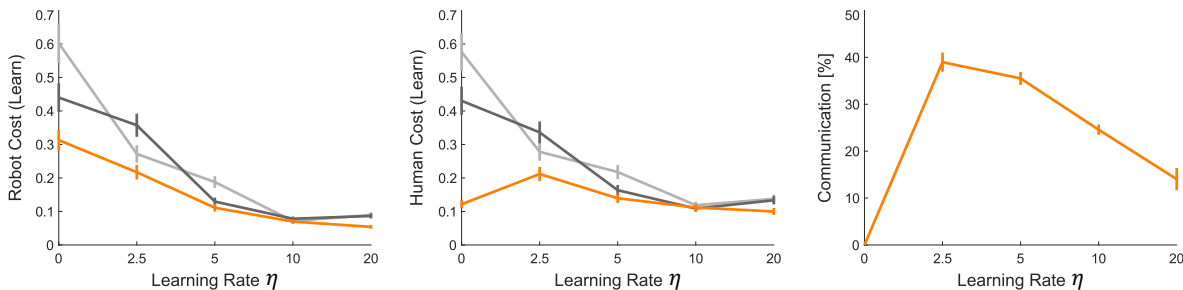

Fig. 4. Objective performance when interacting with a trusting human during a linear-quadratic game. Left: the cost incurred by the robot. Middle: the cost incurred by the human. Right: the percentage of robot actions that are communicative. Error bars report the standard error of the mean (SEM). *Trust* leads to less robot cost than either *Nash* or *Optimistic*, and also reduces the cost for the human agent. Only the *Trust* robot takes communicative actions, i.e., actions that move the table away from the goal. Although these actions are locally sub-optimal, they convey the robot's future behavior to the human, and improve the overall HRI performance (in terms of cost) for both the robot and human.

actions are: a communicative action is an action that pushes the table away from the goal. *Communication* % refers to the percent of the robot's total effort $\|u_r\|_1$ that pushes the table in the opposite direction of the goal.

D. Communicating with Trusting Humans

Figs. 3 and 4 summarize our results when interacting with Learn humans that trust the robot. *Nash* and *Optimistic* robots always apply positive forces to push the table towards the goal. By contrast, robots which realize that the human trusts them (*Trust*) apply negative forces at the start of the task, moving the table away from the goal. Although seemingly counter-productive, these actions signal to the human that the robot places a high value on its own effort, and increase the human's estimate $\hat{\theta}_r$ so that the human exerts more effort to guide the table towards the goal. Put another way, pulling the table away from the goal convinces the human to take over.

Inspecting Fig. 4, we find that the communicative actions of the *Trust* robot yield lower robot cost than either *Nash* or *Optimistic* across all tested learning rates η. The communication percent decreases as the human's learning rate increases, since fewer negative forces are required to convey the robot's reward if the human learns quickly. Interestingly, *Trust* also reduced the human cost as compared to *Nash* and *Optimistic*: modeling the human as a trusting agent here led to improved performance for both the human and robot as

compared to conservative and optimistic models.

Summary. Communicative actions emerged as part of the optimal response when interacting with a trusting human; the *Trust* robot intentionally signaled this human by moving away from the goal, resulting in lower overall costs.

E. Human Model Errors

So far we have focused on trusting humans that learn from robot actions. In Fig. 5, however, we simulate users that diverge from this trusting model. If the robot assumes that the human learns—but the human does not—then *Trust* can be outperformed by *Optimistic* and *Nash*. We find that *Trust* becomes more robust to human model errors when the robot assumes higher values of η, or, put another way, when the robot takes fewer communicative actions. This matches with our intuition: if the robot pushes the table away from the goal to communicate their objective, but the human does not learn from this action, then the robot's communication is inefficient and costly. Hence, when the robot is unsure about the human's trust, the robot should start by assuming a high learning rate η to reduce the amount of communication.

VI. USER STUDY

To test how robots that model uncertain humans as trusting agents interact with *actual* human end-users, we conducted a user study. Our participants worked with a robot to balance a simulated inverted pendulum (see Fig. 6). We compared
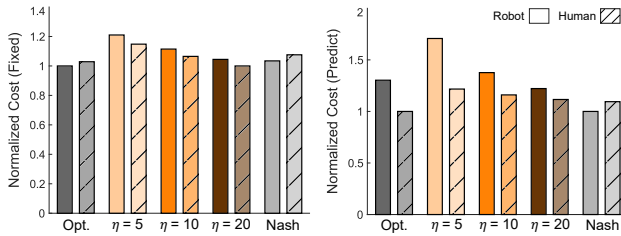
Fig. 5. Cost when interacting with non-trusting humans that instead follow our Fixed or Predict models. *Opt* is short for optimistic. The *Trust* robot incorrectly assumes that the human learns with $\eta = \{5, 10, 20\}$. Higher assumed learning rates η mitigate errors in the trusting model, since the robot thinks the human learns faster and takes fewer communicative actions.

two different interaction strategies for the robot: computing the Nash equilibrium, where the conservative robot assisted at every timestep to balance the inverted pendulum, and our trusting model, where the robot occasionally unbalanced the pendulum to convince the human to take over. Both the participant and robot wanted to keep the pendulum upright; however, the robot is also rewarded for maximizing end-user involvement, so that the human balances the pendulum. We hypothesized that a robot which reasons over human trust will *increase user involvement* during collaborative tasks.

Independent and Dependent Variables. We varied the robot interaction strategy with two levels: a conservative *Nash* strategy, which was based on the Nash equilibrium formulation (3), and a communicative *Trust* strategy, which was inspired by our trusting model (6). Let ϕ be the angle of the pendulum from vertical; then, $u_r = -sign(\phi)$ for the Nash strategy. Under the Trust strategy, $u_r = sign(\phi)$ at the start of the task to intentionally unbalance the pendulum.

For each interaction strategy we measured the normalized *Time Upright* and *Human Effort*. Time upright is the number of timesteps the pendulum remained vertical, i.e., $|\phi| < \pi/2$. Human effort is defined as the percentage of timesteps that the participant applied inputs to control the cart. Besides these objective measures, we also recorded the participants' subjective responses to a 7-point Likert scale survey.

Experimental Procedure. Our participants consisted of 10 volunteers (3 female, 7 male), with ages ranging from $18 - 29$ years. Participants used a keyboard to control the 1-DoF cart in the CartPole environment (OpenAI Gym). After a familiarization phase—where the participant interacted without any robot intervention—users completed five trials with the Nash robot strategy and five trials with the Trust robot strategy. We counterbalanced the order of the robot strategies, so that half of the participants interacted with the Trust robot first.

Results. The results of our user study are summarized in Fig. 7 and Table II. To interpret our objective results we performed paired t-tests with robot strategy as the within-subjects factor. We found a statistically significant difference between Nash and Trust for time upright ($t(9) = 3.1$, $p < .05$), as well as for human effort ($t(9) = 4.8$, $p < .001$). The communicative actions taken by the Trust robot (intentionally unbalancing the pendulum) led to greater human involvement.
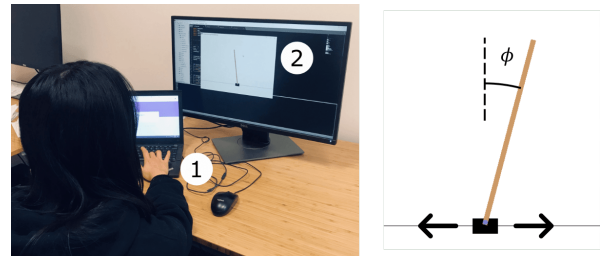

Fig. 6. Setup for our user study. Participants used a keyboard (1) to control a inverted pendulum (2, right). A virtual robot also applied forces to control this pendulum; when the human stopped providing inputs, the robot was in charge of keeping the pendulum upright. Robots using our trusting human model intentionally unbalanced the pendulum to increase user involvement.
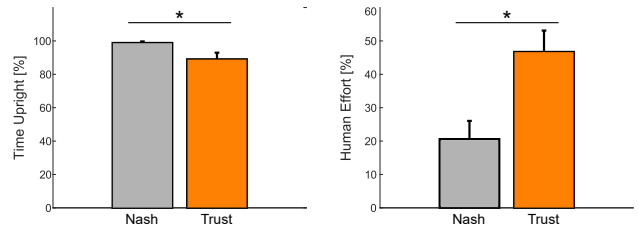

Fig. 7. Objective results from our user study. Left: percentage of time the pendulum was upright. Right: percentage of time the human was involved in controlling the pendulum. Error bars show SEM, and an $*$ denotes statistical significance ($p < .05$). The task was more challenging when the robot modeled the human as trusting (as evidenced by the lower time upright), but the user became more involved, and interacted for a longer duration.

We grouped our survey questions into three topics: understanding, objectives, and adaptation. Paired t-tests revealed that there was not a statistically significant difference in how well participants understood Nash and Trust ($t(19) = 0.9$, $p = .38$). Despite this confidence, users incorrectly believed that Trust did not want to keep the pendulum upright ($t(19) = 4.5$, $p =< .01$). Users also reported that they adapted their approach more when interacting with Trust ($t(19) = 3.0$, $p =< .01$). Taken together, these responses suggest that Trust successfully mislead users: by unbalancing the pendulum, the robot convinced participants that it had a different objective, and caused users to adjust their behavior.

VII. DISCUSSION

Summary. We focused on partial information settings where the robot knows both the human's objective and its own, while the human only knows their own objective. The robot's behavior within these settings depends on how it models the human. We formulated four different models for two-player HRI games: the robot can behave conservatively (as if the human knew their objective), optimistically (as if they can teach the human any objective), model the human as rational, or treat the human as a trusting agent. This trusting model is a simplification of the rational model, where the human learns while assuming that the robot behaves conservatively.

Comparing our trusting human model to the alternatives, we found that modeling the human as trusting is not the same as acting optimistically, but can result in robots that take advantage of the human's trust. In a linear-quadratic case study robots with a trusting human model intentionally

Questions	Nash	Trust	p-value
+ I could tell what the robot wanted to achieve − I did not *understand* the robot's objective	5.45	5.05	0.38
+ The robot's goal was to keep the pendulum upright − The robot had a different *objective* than I did	5.60	3.45	< .001*
+ I *adapted* over time to the robot's actions − I used the same strategy across trials	3.75	5.1	< .01*

performed locally sub-optimal actions (moving away from the goal) to convince the human to take over. During our user study, robots with the trusting human model similarly convinced humans that the robot did not share their objective, resulting in adaptation and increased user involvement.

Why Trusting Models? Programming robots to model humans as trusting agents has benefits as a way to generate communicative actions and increase user involvement. Under the trusting model, communicative actions naturally emerge as part of the solution to an MDP, and do not need to be explicitly encoded or rewarded. Our case and user studies suggest that these actions increase user involvement, which can be particularly useful for assistive robotics and rehabilitation applications, where *over-assistance* is a common problem [2]. We think robots that are aware of the human's trust can make intelligent decisions to avoid behaviors such as over-assistance.

Limitations and Future Work. Our work is a first step towards developing influential and communicative robot behavior in settings where the human is uncertain. We have so far assumed that the human and robot coordinate to follow the same Nash equilibrium; in general, this is not the case, and so complete solutions must also account for strategy uncertainty. We also recognize that computing the Nash equilibria to these two-player games with high-dimensional states and actions can quickly become computationally infeasible.

These limitations suggest that—in order to implement our trust model in realistic settings—we need to identify suitable, real-time approximations. Our future work focuses on this challenge: we plan to bias the human model towards *trusting but myopic* human behavior, which we can obtain tractably without identifying Nash equilibria. Equipped with this approximation, we will pursue realistic applications such as shared control and assistive robotics.

VIII. ACKNOWLEDGEMENTS

REFERENCES

[1] C. L. Baker, R. Saxe, and J. B. Tenenbaum, "Action understanding as inverse planning," *Cognition*, vol. 113, no. 3, pp. 329–349, 2009.
[2] A. A. Blank, J. A. French, A. U. Pehlivan, and M. K. OMalley, "Current trends in robot-assisted upper-limb stroke rehabilitation: Promoting patient engagement in therapy," *Current Physical Medicine and Rehabilitation Reports*, vol. 2, no. 3, pp. 184–195, 2014.
[3] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors*, vol. 53, no. 5, pp. 517–527, 2011.
[4] A. Xu and G. Dudek, "Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations," in *ACM/IEEE Int. Conf. Human-Robot Interaction (HRI)*, 2015, pp. 221–228.
[5] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations," in *Int. Conf. Human Robot Interaction (HRI)*, 2016, pp. 109–116.
[6] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, "Planning with trust for human-robot collaboration," in *ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI)*, 2018, pp. 307–315.
[7] T. Hellström and S. Bensch, "Understandable robots - what, why, and how," *Paladyn, Journal of Behavioral Robotics*, vol. 9, no. 1, pp. 110–123, 2018.
[8] D. Sadigh, S. Sastry, S. A. Seshia, and A. D. Dragan, "Planning for autonomous cars that leverage effects on human actions." in *Robotics: Science and Systems (RSS)*, 2016.
[9] S. H. Huang, D. Held, P. Abbeel, and A. D. Dragan, "Enabling robots to communicate their objectives," *Autonomous Robots*, pp. 1–18, 2017.
[10] D. Losey and M. K. O'Malley, "Enabling robots to infer how end-users teach and learn through human-robot interaction," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1956–1963, 2019.
[11] D. Sadigh, N. Landolfi, S. S. Sastry, S. A. Seshia, and A. D. Dragan, "Planning for cars that coordinate with people: Leveraging effects on human actions for planning and active information gathering over human internal state," *Autonomous Robots (AURO)*, vol. 42, no. 7, pp. 1405–1426, October 2018.
[12] A. D. Dragan, R. M. Holladay, and S. S. Srinivasa, "An analysis of deceptive robot motion." in *RSS*, 2014, p. 10.
[13] J. Foerster, R. Y. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel, and I. Mordatch, "Learning with opponent-learning awareness," in *AAMAS*, 2018, pp. 122–130.
[14] M. Berlin, J. Gray, A. L. Thomaz, and C. Breazeal, "Perspective taking: An organizing principle for learning in human-robot interaction," in *AAAI*, 2006, pp. 1444–1450.
[15] A. K. Pandey, M. Ali, and R. Alami, "Towards a task-aware proactive sociable robot based on multi-state perspective-taking," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 215–236, 2013.
[16] J. G. Trafton, N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, and A. C. Schultz, "Enabling effective human-robot interaction using perspective-taking in robots," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 35, no. 4, pp. 460–470, 2005.
[17] M. J. Osborne and A. Rubinstein, *A Course in Game Theory*. MIT Press, 1994.
[18] T. Hedden and J. Zhang, "What do you think I think you think?: Strategic reasoning in matrix games," *Cognition*, vol. 85, no. 1, pp. 1–36, 2002.
[19] H. A. Simon, *Models of bounded rationality: Empirically grounded economic reason*. MIT press, 1997, vol. 3.
[20] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.
[21] W. Liu, B. Dai, A. Humayun, C. Tay, C. Yu, L. B. Smith, J. M. Rehg, and L. Song, "Iterative machine teaching," in *Int. Conf. on Machine Learning (ICML)*, 2017.
[22] J. Engwerda, "Algorithms for computing Nash equilibria in deterministic LQ games," *Computational Management Science*, vol. 4, no. 2, pp. 113–140, 2007.
[23] Y. Li, G. Carboni, F. Gonzalez, D. Campolo, and E. Burdet, "Differential game theory for versatile physical human–robot interaction," *Nature Machine Intelligence*, vol. 1, no. 1, p. 36, 2019.