Game-Theoretic Modeling of Human Adaptation in Human-Robot Collaboration

Stefanos Nikolaidis*, Swaprava Nath†, Ariel D. Procaccia†, and Siddhartha Srinivasa*
{*Robotics Institute, †Computer Science Department}, Carnegie Mellon University
snikolai@cmu.edu, {swapravn, arielpro}@cs.cmu.edu, siddh@cmu.edu

ABSTRACT

In human-robot teams, humans often start with an inaccurate model of the robot capabilities. As they interact with the robot, they infer the robot's capabilities and *partially adapt* to the robot, i.e., they might change their actions based on the observed outcomes and the robot's actions, without replicating the robot's policy. We present a *game-theoretic model* of human partial adaptation to the robot, where the human responds to the robot's actions by maximizing a reward function that changes stochastically over time, capturing the evolution of their expectations of the robot's capabilities. The robot can then use this model to decide optimally between taking actions that reveal its capabilities to the human and taking the best action given the information that the human currently has. We prove that under certain observability assumptions, the optimal policy can be computed efficiently. We demonstrate through a human subject experiment that the proposed model significantly improves human-robot team performance, compared to policies that assume complete adaptation of the human to the robot.

1. INTRODUCTION

A lot of work in robotics has focused on enabling robots to perform useful tasks for and with people. One of the main goals has been to make robots part of our everyday life, helping people as effective members of human-robot teams. In order to leverage recent advances in robot capabilities, human teammates should know what the robot can and cannot do: the robot's perceived capability should match its true capability.

Prior work has shown that there is often a disconnect between users' perceptions and a robot's true capability, mainly due to lack of experience with working with robots and to the influence of popular culture [1–3]. This gap in expectation can significantly reduce human-robot team performance [4].

For example, we consider the table-clearing task illustrated in Fig. 1. The user and the robot are tasked with

Figure 1: Top: User performs a repeated table-clearing task with the robot. The robot fails intentionally in the beginning of the task, in order to reveal its capabilities to the human teammate. Bottom-left: The robot drops the blue bin off the table while moving towards the left bottle. Bottom-right: The torques applied exceed their limits when the robot attempts a grasp at an extended configuration, and the robot stops moving.

clearing the table by placing items in the bins. The clearing task is repeated a number of times. We call each repetition a *round*. The user lacks the following information about the robot:

- The robot does not know where the green bin is. If the robot moves, it may collide with the green bin, inadvertently pushing the adjacent blue bin off the table.
- The robot cannot lift the bottle that is farthest away from its base: the bottle is filled with water and the torques required for a lifting motion exceed the robot's motor torque limits. If the robot attempts to lift the bottle, the robot's control software will abort and the robot will stop moving.

Nikolaidis et al. [5] proposed a human-robot mutual adaptation formalism, where the robot builds a model of human adaptation to guide the user towards an optimal — with re-

spect to some objective performance metric — way of completing the task. Every time the human and the robot take an action, the user was modeled as either *completely adopting the robot policy as her[1] own* with some probability, or keeping her current policy. This probability was defined as the user's adaptability, which indicated her willingness to adapt to the robot. The formalism allowed the robot to infer the adaptability of its teammate through interaction, and guide the user towards an optimal policy unknown to them in advance.

While in [5], the human was modeled as completely adopting the robot's optimal policy with some probability, in many collaborative settings human adaptation can be more subtle. We use the table-clearing task described above as an example, and we let the robot attempt to grasp the bottle that is closest to its base, dropping the blue bin in the process. This will likely cause the human teammate to change her actions: in the next round, she will move the green or blue bin out of the robot's way. However, without observing the robot fail in lifting the other bottle, she still has no information about which action to take (i.e. emptying the bottle of water), if the robot attempts to lift the bottle.

This is an example, where *the human may change their actions based on the robot actions, while not completely adopting the robot's optimal policy.*

In this paper, we propose a game-theoretic model of human partial adaptation to the robot. We assume that the robot knows a "true" objective metric of team performance in the form of a reward matrix. We base this assumption on insights from early work on Stackelberg security games, which used domain expert knowledge to specify the reward of the defender/leader (AI agent) and the attacker/follower (human), showing remarkable results [6].

We model the human as following a best-response strategy to the robot action, based on their own, possibly distorted, reward function. The human reward function changes over time, as the human observes the outcomes of the robot and her own actions.

The model allows the robot to *reason over how the human expectations of the robot capabilities will change based on its own actions.* The robot uses this model to compute an optimal policy, which enables it to decide optimally between *revealing information* to the human and *choosing the best action given the information that the human currently has.*

We prove that, if the robot can observe whether the user has learned at each round, the computation of the optimal policy is simple (Lemmas 1 and 2), and can be done in time polynomial in the number of robot actions and the number of rounds (Theorem 1).

We show through a human subject experiment in a table-clearing task that the proposed model significantly improves human-robot team performance, compared to policies that assume complete human adaptation to the robot. Additionally, we show through simulations that the proposed model performs well for a variety of randomly generated tasks. This is the first step towards modeling the change of human expectations of the robot capabilities through interaction, and integrating the model into robot decision making in a principled way.

---

2. RELEVANT WORK

A lot of research in robotics has focused on one-way robot adaptation to the human, where the robot learns a human skill or preference [7–13]. Other approaches enable robots to reduce the uncertainty over human intention through information-seeking actions [14–18], through negotiation with the human [19], or through decomposition of a complex task into subtasks [20]. There has also been work in human adaptation to the robot in social [21–24], and physical human-robot interaction [25].

Li et al. [26] suggest that the human-robot collaboration problem in physical human-robot interaction can be modeled as a two-player game. They assume that the human partner exerts a force by optimizing an unknown cost function; the robot's cost-function is then updated based on the gradient of the error between the actual force applied by the human and the force predicted by the robot's cost function, until an equilibrium is achieved. Menell et al. [27] define a cooperative inverse reinforcement learning (CIRL) problem as a partial information two-player game, where the robot maximizes the unknown human reward in a cooperative setting. They show that solving the game results in active teaching and active learning behaviors. The framework has yet to be evaluated in a human subject experiment. In contrast to both papers, in our work the roles are reversed, since the human learns the robot reward through interaction. In a repeated collaborative task with different actions, human adaptation can be more subtle than in a force exchange scenario. Additionally, the learner (human) does not run an inverse reinforcement learning algorithm. Instead, we model the human as learning with some probability the best-response to the robot action observed. This captures how human actions change over time based on their updated expectations of the robot capabilities, and it enables the robot to decide optimally between communicating the true rewards to the human and maximizing the immediate reward given the current human strategy.

There is also relevant work in the social navigation domain: In the manuscript by Trautman et al. [28], human and robot trajectories are jointly planned as the optimum of a reward function that combines goal completion and collision avoidance. Sadigh et al. [29] model the interaction of a human driver with an autonomous car as a dynamical system, where the human follows a best-response strategy to the robot actions. By contrast, we focus on a repeated task in a collaborative setting where the human reward function may change over time, as the human observes the outcomes of the robot and her own actions.

We draw upon insights from previous work on a particular class of Stackelberg games [30], the *repeated Stackelberg security games* [31]. In this setting, the follower observes the leader's possibly randomized strategy, and chooses a best-response. We extend this model to a human-robot collaboration setting, where the leader is the robot and the follower is the human, and we model human adaptation by having the follower's reward stochastically changing over time.

3. FORMAL MODEL

Consider a two-player game represented by the player set $N = \{R, H\}$, where player R is the `Robot` and player H is the `Human`. Each of them has a *finite* set of actions denoted by $A^R = \{a_1^R, \ldots, a_m^R\}$ and $A^H = \{a_1^H, \ldots, a_n^H\}$ respectively. The payoff associated with each pair of actions is uniquely

identified by a matrix $R = [r_{i,j}], (i,j) \in [m] \times [n]$, where the entry $r_{i,j}$ denotes the reward[2] for the action pair $(a_i^\text{R}, a_j^\text{H})$ chosen by these two players. We denote the reward vector corresponding to row i by r_i, i.e., $r_i = (r_{i,1}, \ldots, r_{i,n})$. Importantly, the *same reward* is experienced together by both players. Therefore this is an *identical payoff* game where the goal is to maximize the total reward obtained in T (finite) rounds of playing this repeated game. If the reward matrix was perfectly known to both the agents, they would have played the action pair that gives the maximum reward in each round.

However, we assume that in the beginning of the game, the robot has perfect information about the reward matrix, whereas the human has possibly incorrect information (captured by a reward matrix R^H which the human *believes* to be the true reward matrix). In different rounds of the game, the human probabilistically learns different entries of this matrix and picks action accordingly. We will assume that the human is capable of taking the optimal action given her knowledge of the payoffs, e.g., if a specific row of this matrix is completely known to the human and the robot plays the action corresponding to this row,[3] the human will pick the action that maximizes the reward in this row. However, if the entries of a row are yet to be learned by the human, the human picks an action according to $\arg\max r_i^H$, where r_i^H is the i-th row of R^H.

The only aspect of this game that may change over time is the state of the human, which we denote by $x_t, t \in [T]$. Therefore, the state of the game is simply the state of the human agent. We denote the state space of the game as $\mathcal{X}$; it will be instantiated below in different models of information dissemination.

A policy $\pi = (\pi_1, \ldots, \pi_T)$ is a sequence of robot action functions $\pi_t : \mathcal{X} \to A^\text{R}$, $t \in [T]$. The decision problem of the robot is to find the optimal policy $\pi^* = (\pi_1^*, \ldots, \pi_T^*)$ that maximizes the expected payoff U_1 starting from round 1, defined as follows. Denoting the strategy of the human by $s^\text{H} : A^\text{R} \times \mathcal{X} \to A^\text{H}$.

$$U_1(\pi|x_1) \triangleq \mathbb{E}\left[\left.\sum_{t=1}^{T} R(\pi_t(x_t), s^\text{H}(\pi_t(x_t), x_t))\right| x_1\right]$$
$$\pi^* \in \operatorname*{argmax}_{\pi} U_1(\pi|x_1) \tag{1}$$

4. APPROACH

We consider a setting where, in each round, the robot plays first by choosing a row. We model the strategy of the human $s^\text{H} : A^\text{R} \times \mathcal{X} \to A^\text{H}$ as maximizing a human reward function R^H. In other words, the human *best responds* to the robot action, according to the (possibly erroneous) way she currently perceives the payoffs. The human reward matrix R^H evolves over time, as the human learns the "true" reward R through interaction with the robot. We propose a model of human *partial adaptation*, where the human learns with probability α the entries of row r_i that correspond to the robot action a_i^R played, and with probability $(1 - \alpha)$ none of the entries. We consider the following models, based on when the human learning occurs, and on whether the robot directly observes if the human has learned.
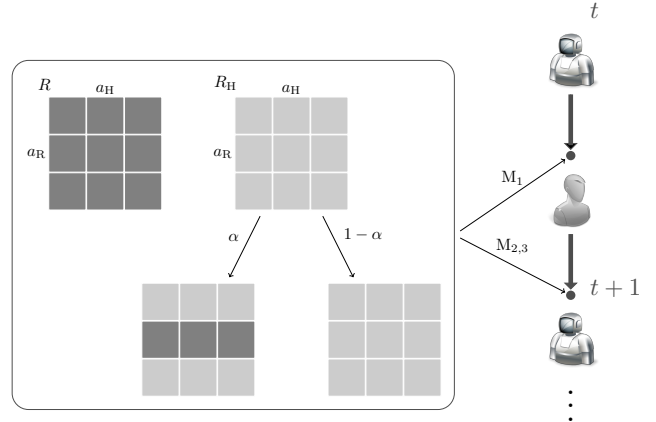

Figure 2: Models of human partial adaptation, described in Sec. 4. The human learns with probability α the entries of row r_i that correspond to the robot action a_i^R played, and with probability $1 - \alpha$ none of the entries. The learning occurs before her action (*learning from robot action* – $\mathbf{M_1}$), or after her action (*learning from experience* – full observability ($\mathbf{M_2}$) or partial observability ($\mathbf{M_3}$)).

$\mathbf{M_1}$. The human learns the payoffs immediately after the robot plays a row, and before she takes her own action. The robot can infer whether the human has learned the row, by observing the reward after the human has played in the same round. We call this *learning from robot action*, where the robot has *full observability* of the human internal state. This model is studied in Sec. 6.1.

$\mathbf{M_2}$. The human learns the payoffs associated with a row after she plays in response to the robot's action. The robot can observe whether the human has learned before the start of the next round, for instance by directly asking the human, or by interpreting human facial expressions and head gestures [32]. We call this model *learning from experience*, where the robot has *full observability* of the human internal state. This model is studied in Sec. 6.2.

$\mathbf{M_3}$. Identically to model $\mathbf{M_2}$, the human learns a row after her action in response to the robot action. However, the robot does not immediately observe whether the human has learned, rather infers it through the observation of human actions in subsequent rounds of the game. This is a case of *learning from experience*, *partial observability*.

We note that we do not define a model for *learning from robot action*, *partial observability* case, since the robot can always directly observe whether the human has learned, based on the reward resulting from the human action in the same round.

Figure 2 shows the different models. In Section 5, we discuss the general case of partial observability (Model $\mathbf{M_3}$) and formulate the problem as a Markov Decision Process [33]. Computing the optimal policy in this case is exponential in the number of robot actions m. However, when the robot has full observability of the human state (Models $\mathbf{M_1}, \mathbf{M_2}$), the optimal policy has a special structure and can be computed in time polynomial in m and T (Section 6).

5. THEORY: PARTIAL OBSERVABILITY

In this section we examine the hardest case, where the human learns the payoffs associated with the row after their choice of actions, and the robot cannot directly observe

whether the human has learned the payoffs (model $\mathbf{M_3}$). Instead, the robot infers whether the human has learned the row by observing the human response in subsequent rounds of the game.

While the human state is partially observable, we can exploit the structure of the problem and reduce it to a Markov Decision Process based on the following observation: the probability of the human having learned a row is either 0 when it is played for the first time; α after it is played by the robot and the human responds sub-optimally; and 1 after the the human has played the actual best-response strategy (according to R) for that row (which means she has learned the true rewards in the previous round).

We define a Markov decision process in this setting as a tuple $\{\mathcal{X}, A^{\mathrm{R}}, P, R, T\}$, where:
- $\mathcal{X} \in \{0, \psi, 1\}^m$ is a finite set of human states. A state x is represented by a vector $(x_1, x_2, ..., x_m)$, where $x_i \in \{0, \psi, 1\}$ and i is the corresponding row in the matrix. The starting state is $x_i = 0$ for each row i. $x_i = \psi$ indicates that the robot does not know whether human has learned row i or not. In this state, the human plays the best response in that row with probability α, or an action defined by the strategy s^{H} of the human with probability $(1 - \alpha)$. If the human plays best-response, then the robot knows that human has learned row i, thus the entry for that row is $x_i = 1$.
- $A^{\mathrm{R}} = \{a_1^{\mathrm{R}}, \ldots, a_m^{\mathrm{R}}\}$ is a finite set of robot actions.
- $P : \mathcal{X} \times A^{\mathrm{R}} \longrightarrow \Pi(\mathcal{X})$ is the state transition function, indicating the probability of reaching a new state x' from state x and action a_i^{R}. State x transitions to a new state x' with all vector entries identical, apart from the element x_i corresponding to the row played. If the robot plays i for the first time ($x_i = 0$), the corresponding entry in the next state x' deterministically becomes $x_i' = \psi$, since the robot no longer knows whether the human has learned the payoffs for that row. If $x_i = \psi$, the human may have learned that row in the past and play the best-response strategy, leading to a transition to $x_i' = 1$ with probability α. If the human does not play the best-response strategy, the robot still does not know whether they will have learned the payoffs after the current round, thus $x_i' = \psi$ with probability $(1 - \alpha)$. If $x_i = 1$, the corresponding entry in all subsequent states will be $x_i' = 1$, i.e., if the human learns a row, we assume that she remembers the row in the future.
- $R : A^{\mathrm{R}} \times A^{\mathrm{H}} \longrightarrow \mathbb{R}$ is the reward function, giving the immediate reward gained by performing a human and robot action. Note that if action i is played and the state has $x_i = \psi$, the reward will be based on the best response in row i of R with probability α, and on row i of R^{H} with probability $(1 - \alpha)$ — we consider the *expected* reward.
- T is the number of rounds.

The robot's decision problem is to find the optimal policy $\pi^* = (\pi_1^*, \ldots, \pi_T^*)$ to maximize the expected payoff U_1 as defined in Eq. 1.

We observe that in the current formalism, the size of the state-space is $|\mathcal{X}| = 3^m$, where m is the number of robot actions. Therefore, the computation of the optimal policy requires time exponential in m. In Section 6, we show that for the case where the robot can observe whether the human has learned the payoffs, the optimal policy can be computed in time polynomial in m and T.

6. THEORY: FULL OBSERVABILITY

In this section, we assume that the robot can observe whether the human has learned the payoffs. We instantiate state x_t as a vector $(x_{t,1}, x_{t,2}, \ldots, x_{t,m})$, where each $x_{t,i}$ is now a binary variable in $\{0, 1\}$ denoting the robot's knowledge in round t of whether row i is learned by the human. In contrast to Sec. 5, there is no uncertainty about whether the human has learned or not (therefore no ψ state).

6.1 Learning from Robot Action

This is the scenario where the human might learn the payoffs immediately after the robot plays a row, and before she takes her own action (Model $\mathbf{M_1}$ in Sec. 4). Clearly, the robot can figure out if the human learned the row by observing the reward for that round. Our algorithmic results in this model strongly rely on the following lemma.

LEMMA 1. *In model* $\mathbf{M_1}$, *if, under the optimal policy* π^*, *there exists* $\tau \in \{2, \ldots, T\}$ *and* $i \in [m]$ *such that* $x_{\tau,i} = 1$ *and* $\max r_i \geqslant \max r_j$ *for all* j *such that* $x_{\tau,j} = 1$, *then* $\pi_t^*(x_t) = a_i^{\mathrm{R}}$ *for all* $\tau \leqslant t \leqslant T$ *and for all* $x_t = x_\tau$.

This lemma says that the optimal policy for the robot is to pick the action a_i^{R} when i is the row that yields the maximum reward among the rows already learned by the human. As we will show in detail later, this directly leads to a computationally efficient algorithm, via the following insight: *if the robot plays a row and this row is successfully revealed to the human, the optimal policy for the robot is to keep playing that row until the end of the game.*

The main idea behind the proof below is: if at round $t-1$ the optimal policy plays row 2, and that row is revealed, then it will not explore the unrevealed (higher rewarding) row 1 afterwards. The reason is that if the optimal policy chose to explore row 1 at some time in the future — which is a contradiction to the lemma — then playing row 1 at round $t-1$ would have been optimal, therefore an optimal policy would not have played row 2 at round $t-1$.

PROOF OF LEMMA 1. Assume for contradiction that the lemma does not hold, and let t be the *last* round in which the optimal policy violates the lemma, i.e., the last round in which there are $i, j \in [m]$ such that $x_{t,i} = 0$ and $x_{t,j} = 1$, but the optimal policy plays row i. Without loss of generality assume that these i and j are rows 1 and 2, respectively. For all rounds from $t + 1$ to T, it holds (by the choice of t) that if row i is revealed to the human, the optimal policy will continue playing a_i^{R} (if there are multiple such rows, it plays the one with highest reward).

Let the maximum rewards corresponding to rows 1 and 2 be R_1 and R_2, respectively, i.e., $R_k = \max r_k$. We assume w.l.o.g. that row 2 has the highest maximum reward among all revealed rows. We can also assume that $R_1 > R_2$, since a policy that moves away from a row that is simultaneously known and more rewarding is clearly suboptimal.

If a row is not learned, the reward associated with actions a_1^{R} and a_2^{R} are C_1 and C_2, where $C_k = r_k[\arg\max r_k^H]$. Clearly, $C_1 \leqslant R_1$ and $C_2 \leqslant R_2$. Since the optimal policy chose a_1^{R} in round t over a_2^{R}, the expected payoff of choosing a_1^{R} in round t must be larger than that of a_2^{R}, i.e.,

$$\alpha(R_1 + U_{t+1}(\pi^*|(1, 1, \ldots))) + (1 - \alpha) \cdot (C_1 + U_{t+1}(\pi^*|(0, 1, \ldots)))$$
$$> R_2 + U_{t+1}(\pi^*|(0, 1, \ldots)),$$

where the first term on the LHS shows the expected payoff if row 1 is learned in round t, and the second term shows

the payoff when it is not. It follows that

$$\alpha(R_1 + R_1 \cdot (T - t - 1)) + (1 - \alpha) \cdot (C_1 + R_2 \cdot (T - t - 1))$$
$$> R_2 + R_2 \cdot (T - t - 1). \quad (2)$$

The implication holds because from round $t+1$, we assume (by the choice of t) that the optimal policy continues playing the best action among the revealed rows. We make the above inequality into an equality by adding a slack variable $\epsilon > 0$ as follows.

$$\alpha R_1 \cdot (T - t) + (1 - \alpha)(C_1 + R_2 \cdot (T - t - 1))$$
$$= R_2 + R_2 \cdot (T - t - 1) + \epsilon. \quad (3)$$

Denote the LHS of the above equality as ρ_1. Note that this is the assumed optimal value of the objective function at round t when the state x_t is $(0, 1, \ldots)$, i.e., $U_t(\pi^*|(0, 1, \ldots)) = \rho_1$. Rearranging the expressions above, we get,

$$\alpha R_1 \cdot (T - t) + (1 - \alpha)C_1 = R_2 + \alpha R_2 \cdot (T - t - 1) + \epsilon. \quad (4)$$

We claim that if the optimal policy chooses the action a_1^R at round t, then the expected payoff in round $t - 1$ from choosing the action a_1^R would have been larger than that of the action a_2^R. If our claim is true, then the current policy, which chose a_2^R at $t - 1$, cannot be optimal, and we reach a contradiction. To analyze the decision problem in round $t - 1$, we need to consider two possible states of the game in this round.

Case 1: $x_{t-1} = (0, 0, \ldots)$. In this state, playing a_1^R gives an expected payoff of

$$\alpha(R_1 + U_t(\pi^*|(1, 0, \ldots))) + (1 - \alpha)(C_1 + U_t(\pi^*|(0, 0, \ldots)))$$
$$\geqslant \alpha(R_1 + R_1(T - t)) + (1 - \alpha)(C_1 + U_t(\pi^*|(0, 0, \ldots))). \quad (5)$$

The inequality holds because in state $(1, 0, \ldots)$, playing a_1^R yields at least R_1 in every subsequent round. Playing a_2^R in round $t - 1$ yields,

$$\alpha(R_2 + \rho_1) + (1 - \alpha)(C_2 + U_t(\pi^*|(0, 0, \ldots))). \quad (6)$$

This expression is similar to the RHS of Equation (5), except that the expected payoff at $x_t = (0, 1, \ldots)$ is assumed to be ρ_1. We claim that the expression on the RHS of Eq. (5) is larger than the expression in Eq. (6), for which we need to show that

$$\alpha(R_1 + R_1(T - t)) + (1 - \alpha)C_1$$
$$> \alpha(R_2 + \rho_1) + (1 - \alpha)C_2$$
$$\iff \alpha R_1 + R_2 + \alpha R_2 \cdot (T - t - 1) + \epsilon$$
$$> \alpha(R_2 + R_2 + R_2 \cdot (T - t - 1) + \epsilon) + (1 - \alpha)C_2$$
$$\iff \alpha R_1 + R_2 + \epsilon > \alpha R_2 + \alpha R_2 + (1 - \alpha)C_2 + \alpha\epsilon.$$

In the first equivalence, we substitute the expression from Eq. (4) on the LHS and the expression of ρ_1 from Eq. (3) on the RHS. The second equivalence holds by canceling out one term. We see that the final inequality is true since $R_2 \geqslant C_2, R_1 > R_2$, and $0 < \alpha < 1$.[4]

Case 2: $x_{t-1} = (0, 1, \ldots)$, in this state playing the action a_1^R gives an expected payoff of at least

$$\alpha(R_1 + R_1 \cdot (T - t)) + (1 - \alpha)(C_1 + U_t(\pi^*|(0, 1, \ldots)))$$
$$= \alpha(R_1 + R_1 \cdot (T - t)) + (1 - \alpha)(C_1 + \rho_1). \quad (7)$$

---

[4]If $\alpha = 1$, playing the row $\arg\max R_i$ is optimal and the lemma holds trivially. For $\alpha = 0$, the lemma is vacuously true. So, we assume $0 < \alpha < 1$ w.l.o.g.

This is similar to the RHS of Eq. (5) except that now we can replace $U_t(\pi^*|(0, 1, \ldots))$ with ρ_1. On the other hand, the expected payoff of the action a_2^R in round $t - 1$ is given by $R_2 + \rho_1$ — because at state $(0, 1, \ldots)$ in round $t-1$, action a_2^R gives R_2 deterministically, since the human knows row 2. The state remains the same even after reaching round t. The expected payoff at this round for this state is assumed to be ρ_1. Now to show that the expression in Eq. (7) is larger than $R_2 + \rho_1$, we need to show that

$$\alpha(R_1 + R_1 \cdot (T - t)) + (1 - \alpha)(C_1 + \rho_1) > R_2 + \rho_1$$
$$\iff \alpha R_1 + \alpha R_1 \cdot (T - t) + (1 - \alpha)C_1 > R_2 + \alpha\rho_1$$
$$\iff \alpha R_1 + R_2 + \alpha R_2 \cdot (T - t - 1) + \epsilon$$
$$> R_2 + \alpha R_2 + \alpha R_2 \cdot (T - t - 1) + \alpha\epsilon$$
$$\iff \alpha R_1 + \epsilon > \alpha R_2 + \alpha\epsilon$$

The first equivalence comes from reorganizing the inequality. The second equivalence is obtained through substitution using Eqs. (3) and (4). The third equivalence follows by canceling out two terms. The last inequality is true since $R_1 > R_2$ and $0 < \alpha < 1$.

To summarize, we have reached a contradiction in both cases, which are exhaustive. This proves the lemma. $\square$

6.2 Learning from Experience

Recall that in model $\mathbf{M_2}$, the human learns with probability α all payoffs associated with a row *after* she plays her action in response to the robot playing an unrevealed row. She does not learn with probability $1 - \alpha$. This model is the same as model $\mathbf{M_3}$ of Sec. 5, with an additional assumption: before the robot takes its next action, it can observe the current state.

We show that in this setting too, the optimal policy has a special structure similar to that under model $\mathbf{M_1}$ (Sec. 6.1), which can be computed in time polynomial in m and T.

LEMMA 2. *In model $\mathbf{M_2}$, if, under the optimal policy π^*, there are $\tau \in \{2, \ldots, T\}$ and $i \in [m]$ such that $x_{\tau,i} = 1$ and $\max r_i \geqslant \max r_j$ for all j such that $x_{\tau,j} = 1$, then $\pi_t^*(x_t) = a_i^R$ for all $\tau \leqslant t \leqslant T$ and for all $x_t = x_\tau$.*

The proof is similar to the proof of Lemma 1. However, the expected payoffs and the corresponding inequalities are different. Therefore, we provide a proof sketch that identifies the differences from the previous proof.

PROOF OF LEMMA 2 (SKETCH). As before, the idea of the proof is to show that if the optimal policy changes its action from playing the revealed row that yields maximum reward, a_2^R, to playing an unrevealed row of higher maximum reward, a_1^R, for the last time in round t, then it must have done so in its previous round, leading to a contradiction. In model $\mathbf{M_2}$, the human does not observe the payoffs of the row played by the robot before she plays her own action. Therefore, we can assume w.l.o.g. that when an unrevealed row is played, its reward is no larger than the maximum reward of that row, e.g., $C_1 \leqslant R_1$ if row 1 is played. Hence, if the optimal policy changes its action from a_2^R to a_1^R in round t when $x_t = (0, 1, \ldots)$, the inequality equivalent to Eq. (2) must be

$$C_1 + \alpha R_1 \cdot (T - t - 1) + (1 - \alpha)R_2 \cdot (T - t - 1)$$
$$> R_2 + R_2 \cdot (T - t - 1). \quad (8)$$

After adding the slack variable, we get,

$$\rho_1 \triangleq C_1 + \alpha R_1 \cdot (T - t - 1) + (1 - \alpha)R_2 \cdot (T - t - 1)$$
$$= R_2 + R_2 \cdot (T - t - 1) + \epsilon$$
$$\implies C_1 + \alpha R_1 \cdot (T - t - 1) = R_2 + \alpha R_2 \cdot (T - t - 1) + \epsilon.$$

In *Case 1*, the expected payoff of playing a_1^{R} is at least: $C_1 + \alpha R_1 \cdot (T - t) + (1 - \alpha)U_t(\pi^*|(0, 0, \ldots))$. The expected payoff of playing a_2^{R} is: $C_2 + \alpha\rho_1 + (1 - \alpha)U_t(\pi^*|(0, 0, \ldots))$. We show that the first expression is larger than the second, i.e.,

$$C_1 + \alpha R_1 \cdot (T - t) > C_2 + \alpha\rho_1$$
$$\iff \alpha R_1 + R_2 + \alpha R_2 \cdot (T - t - 1) + \epsilon$$
$$> C_2 + \alpha R_2 + \alpha R_2 \cdot (T - t - 1) + \alpha\epsilon$$
$$\iff \alpha R_1 + R_2 + \epsilon > C_2 + \alpha R_2 + \alpha\epsilon.$$

The final inequality holds since $R_1 > R_2 \geqslant C_2$ and $0 < \alpha < 1$.

Similarly for *Case 2*, the expected payoff of playing a_1^{R} is at least:

$$C_1 + \alpha R_1 \cdot (T - t) + (1 - \alpha)U_t(\pi^*|(0, 1, \ldots))$$
$$\geqslant C_1 + \alpha R_1 \cdot (T - t) + (1 - \alpha)R_2 \cdot (T - t).$$

On the other hand, the expected payoff of playing a_2^{R} is $R_2 + \rho_1$. We again show that the RHS of the first expression is larger than the second, i.e.,

$$C_1 + \alpha R_1 \cdot (T - t) + (1 - \alpha)R_2 \cdot (T - t) > R_2 + \rho_1$$
$$\iff C_1 + \alpha R_1 \cdot (T - t - 1) + \alpha R_1 + (1 - \alpha)R_2 \cdot (T - t - 1)$$
$$+ (1 - \alpha)R_2 > R_2 + R_2 + R_2 \cdot (T - t - 1) + \epsilon$$
$$\iff R_2 + R_2 \cdot (T - t - 1) + \epsilon + \alpha R_1 + (1 - \alpha)R_2$$
$$> R_2 + R_2 + R_2 \cdot (T - t - 1) + \epsilon$$
$$\iff \alpha R_1 > \alpha R_2,$$

which holds since $R_1 > R_2$ and $0 < \alpha < 1$. $\square$

Algorithm 1 Optimal Policy: Full Observability

Input: matrix R, time horizon T, parameter α
Output: optimal action a_t^* in each round t
$U_t(x_t), a_t^*(x_t) = \mathtt{OptPolicy}(x_t, t)$
procedure $\mathtt{OptPolicy}(x_t, t)$
    if $t > T$ **then**
        return (0, None)
    else
        if x_t has at least one 1 **then**
            find a row k^* s.t. $k^* \in \underset{k:x_{t,k}=1}{\mathrm{argmax}} \max r_k$
            return $(\max r_{k^*} \times (T - t), k^*)$
        else
            find a row
            $i^* \in \underset{k\in[m]}{\mathrm{argmax}} [\alpha(R_k + U_{t+1}(\mathbf{e}_k)) + (1 - \alpha)(C_k$
                      $+ U_{t+1}(\mathbf{0}))]$
            and its value u_{i^*} (*for model* $\mathbf{M_1}$)
            OR
            find a row
            $i^* \in \underset{k\in[m]}{\mathrm{argmax}} [C_k + \alpha U_{t+1}(\mathbf{e}_k) + (1 - \alpha)U_{t+1}(\mathbf{0})]$
            and its value u_{i^*} (*for model* $\mathbf{M_2}$)
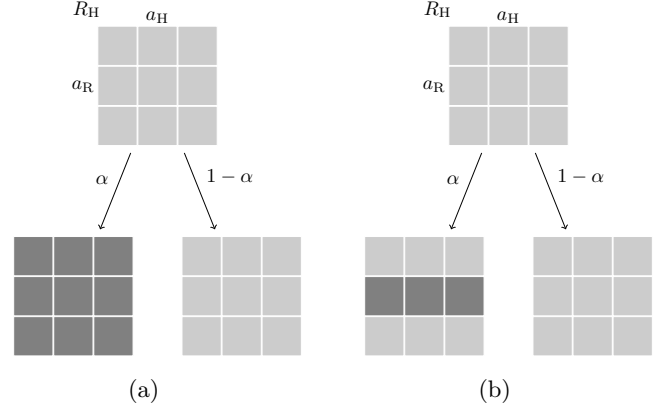   **return** (u_{i^*}, i^*)

Figure 3: The robot reward matrix R is in dark shade and the human reward matrix R^{H} in light shade. (a) The robot reveals its whole reward matrix with probability α. (b) The robot reveals the row played (in this example row 2) with probability α.

6.3 Design of an efficient algorithm

As advertised, using Lemmas 1 and 2, we can easily prove the following theorem.

THEOREM 1. *In models* $\mathbf{M_1}$ *and* $\mathbf{M_2}$, *an optimal policy can be computed in polynomial time.*

Indeed, the algorithm is specified as Algorithm 1. Here $\mathbf{e}_k$ denotes the m-dimensional standard unit vector in direction k. This algorithm runs in time polynomial in m and T since the inner **else** condition does not branch into two independent computations. This is because when at least one coordinate of x_t is 1, the inner **if** condition is met and the expected payoff in that case is computed without recursion. Therefore, in every round the number of computations is $O(m)$, and the algorithm has complexity $O(mT)$.

7. FROM THEORY TO USERS

We conduct a human subject experiment to evaluate the proposed model in a table-clearing task (Fig. 1). We focus on the case where the human *learns from experience* (Models $\mathbf{M_2}, \mathbf{M_3}$). We are interested in showing that the policies computed using the partial adaptation model will result in better performance than policies that model the human as completely adapting to the robot, which is an assumption used in previous work on human-robot mutual adaptation [5].

7.1 Manipulated Variables

Observability. We used two settings — one where the robot does not directly observe whether the human has learned (Sec. 5), and one where the robot observes directly whether the human has learned (Sec. 6.2).
Adaptation. We compared the proposed partial adaptation model with a model of complete adaptation, where the robot models the human as learning all rows of the payoff matrix with probability α after a row is played, instead of learning only the row played (Fig. 3a).

7.2 Hypothesis

We hypothesize that the robot policies that model the human as partially adapting to the robot will perform better than the policies that assume complete adaptation of the human to the robot.

7.3 Experiment Setting

Table-clearing task. We test the hypothesis in the table-clearing task of Fig. 1, where a human and a robot collaborate to clear the table from objects. In this task, the human can take any of the following actions: {pick up any of the blue cups and place them on the blue bin, change the location of any of the bins, empty any of the bottles of water}. The robot can either remain idle or pick up any of the bottles from the table and move them to the red bin. The goal is to maximize the number of objects placed in the bins.

The human does not have in advance the following information about the robot: (1) the robot does not know the location of the green bin. Therefore, when the robot attempts to grab one of the bottles, it may push the green bin, dropping the blue bin off the table. (2) The robot will fail if it picks up the bottle that is farthest away from it, if that bottle has water in it. This is because of its motor torque limits.

Model parameters. This information is represented in the form of a payoff matrix R. The entries correspond to the number of objects in the bins after each human and robot action. Table 1 shows part of R; it includes only the subset of human actions that affect the outcome. For instance, if the robot starts moving towards the bottle that is closest to it (action 'Pick up closest') and the human does not move the green or blue bin out of the way, the robot will drop the blue bin off the table, together with any blue cups that the human has placed. Therefore, at the end of the task only the bottle will be cleared from the table, resulting in a reward of 1. If the robot attempts to pick up both bottles (action "pick up both") and the human does not empty the bottle of water before the robot grasps it, the robot will fail, resulting in a reward of 0. If the human has emptied the bottle and moved the blue bin (action "Clear cups & move bin & empty bottle"), the robot will successfully clear both bottles without dropping the bin, resulting in a reward of 4 (2 bottles in the red bin and 2 cups in the blue bin).

In the beginning of the task, we assume that the human response to all robot actions will be "Clear cups"; since the human has not observed the robot dropping the bin or failing to pick up the bottle, she has no reason to move the bin or empty the bottle of water. We also assume that she does not learn any payoffs if the robot remains idle ("Noop" action). We set the probability of learning $\alpha = 0.9$, since we expected most participants to learn the best-response to the robot actions after observing the outcome of their actions.

Procedure. The experimenter first explained the task to the participants and informed them about the actions that they could take, as well as about the robot actions. Participants were told that the goal was to maximize the number of objects placed in the bins at each round. They performed the task three times ($T = 3$). In the full observability setting, the experimenter asked the participants after each round, what would their action be if the robot did the same action in the next round. The experimenter then inputted their response (learned or not learned) into the program that executed the policy. When the robot failed to pick up the bottle, the experimenter informed them that the robot

had failed. Participants were told that the error message displayed in the terminal was: "The torque of the robot motors exceeded their limits." This is the generic output of our ROS-based hardware interface, when the measured torques exceed the manufacturer limits. We added a short, general explanation about how torque is related to distance and applied force. At the end, participants answered open-ended questions about their experience in the form of a video-taped interview.

	Clear cups	Clear cups & move bin	Clear cups & move bin & empty bottle
Noop	2	2	2
Pick up closest	1	3	3
Pick up both	0	0	4

Table 1: Part of payoff matrix R for table-clearing task. The table includes only the subset of human actions that affect performance.

7.4 Subject Allocation

We recruited 60 participants from a university campus. We chose a between-subjects design in order to avoid biasing users towards policies from previous conditions.

8. RESULTS AND DISCUSSION

Analysis. We evaluate team performance by the accumulated reward over the three rounds of the task (Fig. 4-left). We observe that the mean reward in the partial adaptation policy was 42% higher than that of the complete adaptation policy in the partial observability setting, and 52% higher than that of the complete adaptation policy in the full observability setting. A factorial ANOVA showed no significant interaction effects between the observability and adaptation factors. The test showed a statistically significant main effect of adaptation ($F(1, 56) = 18.58, p < 0.001$), and no significant main effect of observability. These results support our hypothesis.

The difference in performance occurred because in the complete adaptation model the robot erroneously assumed that the human had learned the best-response to the "Pick up both" action, after the robot played the row "Pick up closest". In this section, we examine the partial and complete adaptation policies in the *partial-observability* setting. The interpretation of the robot actions in the *full-observability* setting is similar, and we omit it because of space limitations. The robot chooses the action "Pick up both" for round $T = 1$ (as well as for $T = 2, 3$) in the partial adaptation condition[5], since the loss of receiving zero reward at $T = 1$ is outweighed by the rewards of 4 in subsequent rounds, if the human learns the best-response to that action, which occurs with high probability ($\alpha = 0.9$). On the other hand, the robot in the complete adaptation condition takes the action "Pick up closest" at $T = 1$ and "Pick up both" at $T = 2$ and $T = 3$. This is because the model assumes that the human will learn the best-response for all robot actions if the robot

---

[5]Unless specified otherwise, for the rest of this section we refer to the partial observability level of the observability factor.
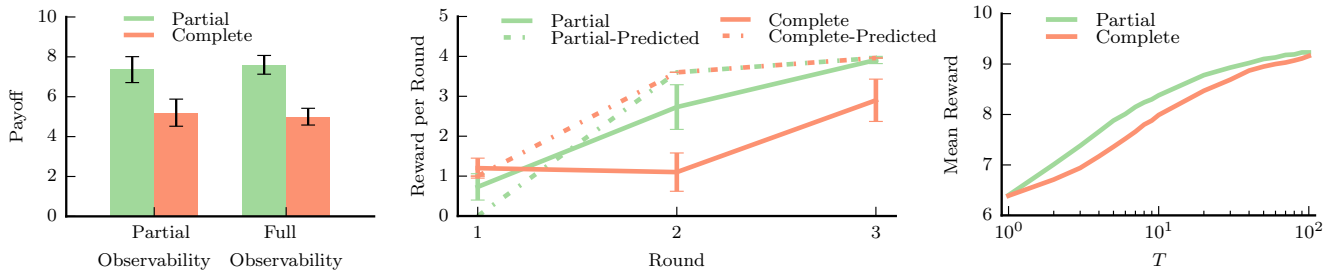
Figure 4: Left: Accumulated reward over 3 trials of the table-clearing task for all four conditions. Center: Predicted and actual reward by the partial and complete adaptation policies in the partial observability setting. Right: Mean reward over time horizon T for simulated runs of the complete and partial adaptation policies in the partial observability setting. The gain in performance from the partial adaptation model decreases for large values of T. The x-axis is in logarithmic scale.

plays either "Pick up closest" or "Pick up both", and the predicted reward of 1 for the action "Pick up closest" is higher than the reward of 0 for "Pick up both" at $T = 1$.

Fig. 4 (center) shows the expected immediate reward predicted by the partial and complete adaptation model for each round in the partial observability setting, and the actual reward that participants received. We see that the immediate reward in the complete adaptation condition at $T = 2$ was significantly lower than the predicted one. The reason is that six participants out of 10 in that condition did not infer at $T = 1$ that the robot was unable to pick up the second bottle and did not empty the bottle at $T = 2$, which was the best-response action. From the four participants that emptied the bottle, two of them justified their action by stating that "there was enough time to empty the bottle" before the robot would grab it. The same justification was given by three participants out of eleven in the partial adaptation condition, who emptied the bottle at $T = 1$ without knowing that this was required for the robot to be able to pick it up. This caused the actual reward to be higher than its predicted value of 0. Additionally, the actual reward at $T = 2$ was lower than the predicted value. We attribute this to the fact that 73% of participants learned the best-response for the robot action (emptying the bottle that was farthest away) in that round, whereas the predicted value assumed a probability of learning $\alpha = 0.9$. In $T = 3$, the actual reward matched the prediction closely, since all participants eventually learned that they should empty the bottle.

Generalizability of the results. The results discussed above are compelling in that they arise from an actual human-subject experiment, but they are limited to one task. We are interested in showing — via simulations — that the proposed model performs well for a variety of tasks. We randomly generated instances of the reward matrix R and α values and simulated runs of the partial and complete adaptation policies for increasing time horizons T. The simulated humans partially adapted to the robot, and the robot did not observe whether they learned. For each value of T, we randomly sampled 1000 reward matrices R and simulated 100 policy runs for each sampled instance of R. Fig. 4 (right) shows the reward averaged over the number of rounds T, policy runs and instances of R. For $T = 1$, the mean reward is the same for both models, since there is no adaptation. The partial adaptation policies consistently outperform the complete adaptation ones for a large range of T. For large values of T the performance difference decreases. This is because the human eventually learns the true payoffs and

the gain from playing the true best response outweighs the initial loss caused by the complete adaptation model.

Selection of α. We note that the α value, which represents the probability that the human learns the true robot capabilities, is task and population-dependent. In our experiment, participants were recruited from a university campus, and most of them were able to infer that they should empty the bottle, after observing the robot failing and being notified that "the robot exceeded its torque limits." Different participant groups may require a different α value. The value of α could also vary for different robot actions; we conjecture that our theoretical results hold also when there is a different adaptation probability α_i for each row i of the payoff matrix, which we leave as future work.

9. CONCLUSION

We presented a game-theoretic model of human partial adaptation to the robot. The robot used this model to decide optimally between taking actions that reveal its capabilities to the human and taking the best action given the information that the human currently has. We proved that under certain observability assumptions, the optimal policy can be computed efficiently. Through a human subject experiment, we demonstrated that policies computed with the proposed model significantly improved human-robot team performance, compared to policies that assume complete adaptation of the human to the robot.

While our model assumes that the human may learn only the entries of the row played by the robot, there are cases where a robot action may affect entries that are associated with other actions, as well. For instance, Cha et al. [2] have shown that conversational speech can affect human perception of robot capability in physical tasks. We are excited to explore the structure of probabilistic graphical models of human adaptation, and use the theoretical insights from this work to develop efficient algorithms for the robot.

Acknowledgments

10. REFERENCES

[1] J. Forlizzi, "How robotic products become social products: an ethnographic study of cleaning in the home," in *Proceedings of the ACM/IEEE international conference on Human-robot interaction.* ACM, 2007, pp. 129–136.

[2] E. Cha, A. D. Dragan, and S. S. Srinivasa, "Perceived robot capability," in *Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on.* IEEE, 2015, pp. 541–548.

[3] A. Powers and S. Kiesler, "The advisor robot: tracing people's mental model from a robot's physical attributes," in *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction.* ACM, 2006, pp. 218–225.

[4] V. Groom and C. Nass, "Can robots be teammates?: Benchmarks in human–robot teams," *Interaction Studies*, vol. 8, no. 3, pp. 483–500, 2007.

[5] S. Nikolaidis, A. Kuznetsov, D. Hsu, and S. Srinivasa, "Formalizing human-robot mutual adaptation: A bounded memory model," in *The Eleventh ACM/IEEE International Conference on Human Robot Interation.* IEEE Press, 2016, pp. 75–82.

[6] M. Tambe, *Security and game theory: algorithms, deployed systems, lessons learned.* Cambridge University Press, 2011.

[7] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robot. Auton. Syst.*, 2009.

[8] C. G. Atkeson and S. Schaal, "Robot learning from demonstration," in *ICML*, 1997.

[9] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *ICML*, 2004.

[10] M. N. Nicolescu and M. J. Mataric, "Natural methods for robot task learning: Instructive demonstrations, generalization and practice," in *AAMAS*, 2003.

[11] S. Chernova and M. Veloso, "Teaching multi-robot coordination using demonstration of communication and state sharing," in *AAMAS*, 2008.

[12] B. Akgun, M. Cakmak, J. W. Yoo, and A. L. Thomaz, "Trajectories and keyframes for kinesthetic teaching: a human-robot interaction perspective," in *HRI*, 2012.

[13] S. Nikolaidis and J. Shah, "Human-robot cross-training: computational formulation, modeling and evaluation of a human team training strategy," in *HRI*, 2013.

[14] O. Lemon and O. Pietquin, *Data-Driven Methods for Adaptive Spoken Dialogue Systems: Computational Learning for Conversational Interfaces.* Springer Publishing Company, Incorporated, 2012.

[15] F. Broz, I. Nourbakhsh, and R. Simmons, "Designing pomdp models of socially situated tasks," in *RO-MAN*, 2011.

[16] T. Bandyopadhyay, K. S. Won, E. Frazzoli, D. Hsu, W. S. Lee, and D. Rus, "Intention-aware motion planning," in *WAFR.* Springer, 2013.

[17] O. Macindoe, L. P. Kaelbling, and T. Lozano-Pérez, "Pomcop: Belief space planning for sidekicks in cooperative games." in *AIIDE*, 2012.

[18] S. Nikolaidis, R. Ramakrishnan, K. Gu, and J. Shah, "Efficient model learning from joint-action demonstrations for human-robot collaborative tasks," in *HRI*, 2015.

[19] E. Karpas, S. J. Levine, P. Yu, and B. C. Williams, "Robust execution of plans for human-robot teams," in *ICAPS*, 2015.

[20] T.-H. D. Nguyen, D. Hsu, W. S. Lee, T.-Y. Leong, L. P. Kaelbling, T. Lozano-Perez, and A. H. Grant, "Capir: Collaborative action planning with intention recognition." in *AIIDE*, 2011.

[21] M. A. Goodrich and A. C. Schultz, "Human-robot interaction: a survey," *Foundations and trends in human-computer interaction*, 2007.

[22] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive robots as social partners and peer tutors for children: A field trial," *Human-computer interaction*, 2004.

[23] B. Robins, K. Dautenhahn, R. Te Boekhorst, and A. Billard, "Effects of repeated exposure to a humanoid robot on children with autism," in *Designing a more inclusive world*, 2004.

[24] A. Green and H. Huttenrauch, "Making a case for spatial prompting in human-robot communication, in multimodal corpora: From multimodal behaviour theories to usable models," in *workshop at LREC*, 2006.

[25] S. Ikemoto, H. B. Amor, T. Minato, B. Jung, and H. Ishiguro, "Physical human-robot interaction: Mutual learning and adaptation," *IEEE Robot. Autom. Mag.*, 2012.

[26] Y. Li, K. P. Tee, W. L. Chan, R. Yan, Y. Chua, and D. K. Limbu, "Role adaptation of human and robot in collaborative tasks," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 5602–5607.

[27] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell, "Cooperative inverse reinforcement learning," 2016.

[28] P. Trautman and A. Krause, "Unfreezing the robot: Navigation in dense, interacting crowds," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on.* IEEE, 2010, pp. 797–803.

[29] D. Sadigh, S. Sastry, S. A. Seshia, and A. D. Dragan, "Planning for autonomous cars that leverages effects on human actions," in *Proceedings of the Robotics: Science and Systems Conference (RSS)*, 2016.

[30] V. Conitzer and T. Sandholm, "Computing the optimal strategy to commit to," in *Proceedings of the 7th ACM conference on Electronic commerce.* ACM, 2006, pp. 82–90.

[31] M.-F. Balcan, A. Blum, N. Haghtalab, and A. D. Procaccia, "Commitment without regrets: Online learning in stackelberg security games," in *Proceedings of the Sixteenth ACM Conference on Economics and Computation.* ACM, 2015, pp. 61–78.

[32] R. El Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Real-time vision for human-computer interaction.* Springer, 2005, pp. 181–200.

[33] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach.* Pearson Education, 2003.