Robot Planning with Mathematical Models of Human State and Action

Anca D. Dragan (anca@berkeley.edu)
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Summary of work in collaboration with P. Abbeel, R. Bajcsy,
A. Bestick, J. Fisac, T. Griffiths, JK. Hedrick, J. Heimrick, N. Landolfi, C. Liu,
D. Hadfield Menell, S. Milli, A. Nagabaudi, S. Russell, D. Sadigh, S. Sastry, S. Seshia, S. Srinivasa, A. Zhou

***Abstract*—Robots interacting with the physical world plan with models of physics. We advocate that robots interacting with people need to plan with models of cognition. This writeup summarizes the insights we have gained in integrating computational cognitive models of people into robotics planning and control. It starts from a general game-theoretic formulation of interaction, and analyzes how different approximations result in different useful coordination behaviors for the robot during its interaction with people.**

I. Introduction

Robots act to maximize their utility. They reason about how their actions affect the state of the world, and try to find the actions which, in expectation, will accumulate as much reward as possible. We want robots to do this well so that they can be useful to us – so that they can come in support of real people. But supporting people means having to work with and around them. We, the people, are going to have to share the road with autonomous cars, share our kitchens with personal robots, share our control authority with prosthetic and assistive arms.

Sharing is not easy for the robots of today. They know how to deal with obstacles, but people are more than that. We reason about the robot, we make decisions, we act. This means that the robot needs to make predictions about what we will think, want, and do, so that it can figure out actions that coordinate well with ours and that are helpful to us. Much like robots of today have a theory of *physics* (be it explicitly as an equation or implicitly as a learned model), the robots of tomorrow will need to start having a theory of *mind*.

Our work for the past few years has focused on integrating mathematical theories of mind, particularly about human *future actions* and *beliefs*, into the way robots plan their *physical*, task-oriented actions.

This required a change from the robotics problem formulation (Fig.1, left), to an *interaction* problem formulation (Fig.1, right). Interaction means there is not a single agent anymore: the robot and human are both agents in a two player game, and they take actions according to utility functions that are not necessarily identical or known to each other. The paper outlines this formally in Sec. II, and then summarizes the different approximations we've explored and what we've learned from them [3, 6–10, 12–14, 19–21, 24]:

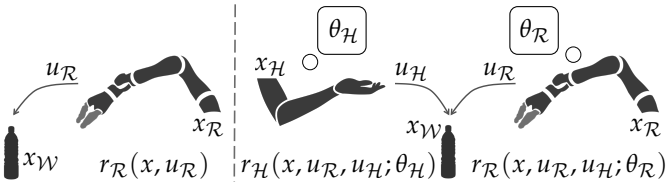


Fig. 1: Left: Traditional robotics formalism: the robot takes actions $u_\mathcal{R}$ to optimize a reward or cost function $r_\mathcal{R}$. Right: Interaction formalism: the robot is not acting in isolation; the human takes actions $u_\mathcal{H}$ to optimize a reward function $r_\mathcal{H}$, possibly different from that of the robot's, and parametrized by human internal/hidden state $\theta_\mathcal{H}$; the human does not know the robot internal state θ_R, which parametrizes the robot's reward function; and both functions depend on both agents' actions.

Accounting for the *physical human behavior* during interaction (Sec. III).

> *One important insight for us has been that people can't be treated as just obstacles that move: they will* ***react*** *to what the robot does.*

If a car starts merging in front of you, you break. If the robot helping you assemble a part employs a different strategy than you expected, you adapt. It took more and more sophisticated approximations to the game above to account for this.

Our first approximation to the game started by assuming a shared utility function and treating the person as a perfect collaborator, but replanning at every step to adapt to when the person deviates from the collaborative plan [13]; we then relaxed this to an imperfect collaborator model, showing that the robot can leverage its actions to guide the person to perform better in the task [3]; finally, we investigated a model of the person as optimizing a different utility function, but simply

computing a best response to the robot's actions (as opposed to solving the full dynamic game) [20] – this model enables the robot to account for how people will react to its actions, and thus perform better at its task.

Using the human behavior to infer *human internal states* (Sec. IV). The models above were a first step in coordinating with people, but they were disappointing in that they still assumed perfect information, i.e. everything is known to both parties. It is simply not true that we will be able to give our robots up front a perfect model of each person they will interact with. Next, we studied how robots might be able to estimate internal, hidden, human states, online, by taking human behavior into account as evidence about them.

> *Another important insight has been that robots should not take their objective functions for granted: they are easy to misspecify and change from person to person. Instead, robots should optimize for what the person wants internally, and use human guidance to estimate what that is.*

Inverse Reinforcement Learning [15, 17, 25] already addresses this for a *passive* observer analyzing *offline* human *demonstrations* of approximately optimal behavior *in isolation*. Our work builds on three goals: 1) making these inferences *actively* and *online*; we leverage not just queries [21], but also the robot's physical actions [19]; 2) accounting for the fact that if the person knows the robot is trying to learn, they will act differently from what they do in isolation and *teach* [9]; and, perhaps most importantly, 3) leveraging richer sources of human behavior beyond demonstrations, like physical corrections [1], orders given to the robot [14], and even the reward function specified [10] – all of these are observations about the actual desired reward.

We argue that interpreting the reward function designed for the robot as useful information about the true desired reward, by leveraging the context in which this function was designed to begin with, can make robots less sensitive to negative consequences of misspecified rewards [10]. Similarly, we find that interpreting orders as useful information about the desired reward give robots a provable incentive to accept human oversight as opposed to bypass it in pursuit of some misspecified objective [8]. Overall, we find that accepting that the robot's reward function is not given, but part of the person's internal state, is key to safe and customizable robots.

Accounting for *human beliefs* about robot internal states (Sec. V).

> *Robots need to make inferences about people during interaction, but people, too, need to make inferences about robots. Robot actions influence whether people can make the correct inferences.*

The third part of our work focuses on getting robots to produce behavior that enables these human inferences to happen correctly, whether they are about the robot's behavior [7], or about the robot's internal states (like utility [12], goals [6], or even level of uncertainty [24]). Although these increases the robot's transparency, we have been encoding the need for that in the objective directly, whereas really it should be a consequence of solving the interaction problem well. This is something we are actively looking into, but which increases the computational burden.

All this research stands on the shoulders of inspiring works in computational human-robot interaction, nicely summarized in sections 6 and 7 of [23] and not repeated here. What this writeup contributes a summary of our own experiences in this area, particularly focusing on physical, task-oriented interaction. It provides a common formalism that, in retrospect, can be seen as the general formulation that seeded these works, along with a quasi-systematic analysis of the different ways to approximate solutions and the sometimes surprisingly interesting and powerful behavior that emerges when we do that. This enables robots to *tractably and autonomously generate* different kinds of behavior for interaction, often *in spite of the continuous state and action spaces* they have to handle – from arms that guide and improve human performance in handovers, to cars that negotiate at intersections, to robots that purposefully make their inner-workings more transparent to their end-users.

II. General Interaction as a Game

In general, we can formulate interaction as a 2-player dynamic game between the human and the robot. The state x contains the world state along with the robot and human state:

$$x = (x_{\mathcal{W}}, x_{\mathcal{R}}, x_{\mathcal{H}})$$

Each agent can take actions, and each agent has a (potentially different) reward function:

$$r_{\mathcal{R}}(x, u_{\mathcal{R}}, u_{\mathcal{H}}; \theta_{\mathcal{R}})$$

for the robot and

$$r_{\mathcal{H}}(x, u_{\mathcal{R}}, u_{\mathcal{H}}; \theta_{\mathcal{H}})$$

for the human, each with parameters θ. The two do not necessarily know each other's reward functions (or equivalently, each other's parameters θ).

Let T be the time horizon, and $\mathbf{u}_{\mathcal{R}}$ and $\mathbf{u}_{\mathcal{H}}$ be a robot and human, respectively, control sequence of length T. We denote by $R_{\mathcal{R}}$ the cumulative reward to the robot

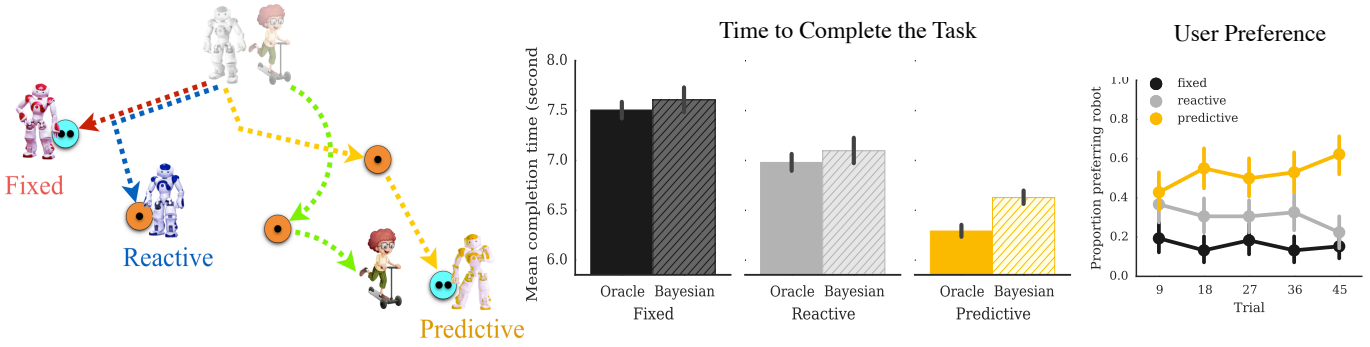


Fig. 2: The Human as a Collaborator: On online study on collaborative Traveling Salesman Problems. The robot models the person's behavior as rational w.r.t. to the same reward function as its own reward function. In the "Fixed" condition, the robot plans an optimal centralized plan for both the human and the robot, and executes its portion. If the human has a different plan, they end up needing to adapt. In the "Reactive" condition, the robot adapts its plan after every target the human reaches, recomputing the new optimal centralized plan from that new state. This performs significantly better than the human plan in objective and subjective task performance measures. Finally, in the "Predictive" condition, the robot uses Bayesian Inference to predict the human's next target via an observation model based on the rationality assumption, and can proactively replan before the human reaches their target.

from the starting state x^0:

$$R_{\mathcal{R}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}_{\mathcal{H}}; \theta_{\mathcal{R}}) = \sum_{t=0}^{T} r_{\mathcal{R}}(x^t, u^t_{\mathcal{R}}, u^t_{\mathcal{H}}; \theta_{\mathcal{R}})$$

and by $R_{\mathcal{H}}$ the cumulative reward for the human:

$$R_{\mathcal{H}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}_{\mathcal{H}}; \theta_{\mathcal{H}}) = \sum_{t=0}^{T} r_{\mathcal{H}}(x^t, u^t_{\mathcal{R}}, u^t_{\mathcal{H}}; \theta_{\mathcal{H}})$$

One way to model what the person will do is to model them as rationally solving this game. There are several issues with this. The first is that it is computationally intractable. The second is that if $r_{\mathcal{H}} \neq r_{\mathcal{R}}$, the game will have many equilibria, so even if we could compute all of them we'd still not be sure which the person is using. The third is that even without the first two issues, this would still not be a good model for how people make decisions in day to day tasks [11, 18].

Our work has thus explored different approximations to this problem that might better match what people do, while enabling robots to actually generate their actions in realistic tasks in (close to) real time.

III. Human Behavior During Interaction

Because in interaction the robot's reward depends on what the person does, the ability to anticipate human actions becomes crucial in deciding what the robot should do. Rather than modeling the person as solving the game above, we explored several approximations that each led to different yet useful behaviors in interaction.

The Perfect Collaboration Model. The easiest simplifying assumption is actually that the person is optimizing the same reward function:

$$r_{\mathcal{H}} = r_{\mathcal{R}}$$

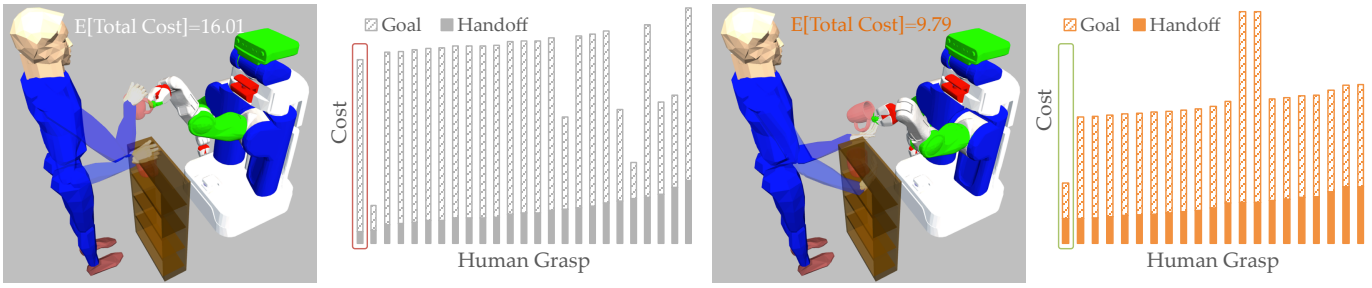
and both agents know this (no more partial information). This assumption turns planning for interaction into a much easier problem, analogous to the original robotics problem: it is pretending like the person is just some additional degrees of freedom that the robot can actuate – their actions will follow the optimal centralized plan:

$$(\mathbf{u}^*_{\mathcal{R}}, \mathbf{u}^*_{\mathcal{H}}) = \underset{\mathbf{u}_{\mathcal{R}}, \mathbf{u}_{\mathcal{H}}}{\operatorname{argmax}} R_{\mathcal{R}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}_{\mathcal{H}})$$

Despite its simplicity, we have found that this can be very useful so long as the robot *replans* at every time step. People inevitably deviate from from the optimal centralized plan even from the first step, ending up in some new state – because they don't actually optimize the same reward, because they don't know that the robot is optimizing the same reward, or because they are not perfect optimizers. But the robot can recompute the centralized optimal plan from that new state, and proceed with the first action from the new $\mathbf{u}^*_{\mathcal{H}}$.

Fig.2 shows a comparison from [13] between a "Fixed" robot strategy, where the robot executes the originally planned $\mathbf{u}^*_{\mathcal{R}}$ regardless of what the person does, and a "Reactive" strategy, in which the robot keeps updating $\mathbf{u}_{\mathcal{R}}$ based on the new centralized optimal plan from the current state at every step. Here, $R_{\mathcal{H}} = R_{\mathcal{R}}$ and equates to the total time to solve the task. We recruited 234 participants on Amazon Mechanical Turk who played collaborative Traveling Salesman Problems with a robot avatar in a within-subjects design that included a 3rd condition, discussed later, and a randomized trial order. We found that the Reactive condition led to the task being completed significantly faster by the human-robot team, and that participants preferred the Reactive robot especially in the beginning.

Overall, online planning with a perfect collaborative model of human behavior can be useful, already enabling the robot to continually adapt to the person's plan even though it does not get it right a priori.



(a) R assumes global H; H chooses grasp myopically and incurs high cost at the goal

(b) Cost histogram of H's choices; the best myopic choice has high global cost

(c) R assumes myopic H; H chooses grasp myopically but still incurs low cost at the goal

(d) Cost histogram of H's choices; the best myopic choice also has low global cost

Fig. 3: People can be myopic about their decision, greedily optimizing reward as opposed to looking ahead (a), e.g. choosing the most comfortable way to grab a mug even though it they would have to regrasp it in order to set it down in a shelf. This might mean that the best action locally has poor reward globally (b). When accounting for human myopia, the robot can select actions such that the myopic response is still good globally (c,d).

Collaborative but Approximately Optimal. An improvement upon the perfectly rational collaborator human model is to recognize that people are not actually perfectly rational. In [3], we model people as collaborative still, but no longer assume $\mathbf{u}^*_{\mathcal{H}}$ is perfect. In particular, we assumed that people are *greedy* or *myopic* in their decisions. Their action at time t will be the one that looks locally good, not the one that is optimal over the full time horizon:

$$u^{t*}_{\mathcal{H}}(u^t_{\mathcal{R}}) = \arg\max_{u^t_{\mathcal{H}}} r_{\mathcal{R}}(x^t, u^t_{\mathcal{R}}, u^t_{\mathcal{H}})$$

The robot can then choose its own actions such that, when coupled with the person's myopic response to them, the combined plan is as high-reward as possible:

$$\mathbf{u}^*_{\mathcal{R}} = \arg\max_{\mathbf{u}_{\mathcal{R}}} R_{\mathcal{R}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}^*_{\mathcal{H}}(\mathbf{u}_{\mathcal{R}}))$$

This results in the robot *guiding* the person's actions, helping them overcome cognitive, bounded-rationality limitations as much as possible.

Interaction becomes an *underactuated system*: the robot no longer assumes it can directly actuate $\mathbf{u}_{\mathcal{H}}$, but accounts for how $\mathbf{u}_{\mathcal{R}}$ will influence $\mathbf{u}_{\mathcal{H}}$ and takes that into account when planning.

In particular, we investigated a handover task in which participants had to take an object from the robot and place it at a goal location. We used as $r_{\mathcal{H}} = r_{\mathcal{R}}$ as negative ergonomic cost to the person. People's decision in this problem is how to take the object such that it is ergonomically low cost. The robot's decision is how to hold the object at the handover time to enable that. A perfect human optimizer would minimize cost at the handover *and* at the goal. Our myopic model minimized cost the handover time, which then could in some cases resulted in high cost at the goal, such as needing to twist their arm in an uncomfortable way. or regrasp. Fig.3 shows an example.

When the robot optimizes for its actions, it chooses ways to hold the object that *incentivize* good global human plans. The robot chooses grasps such that even when the person chooses their grasp greedily for the handover, that greedy grasp is also as close as possible to the global optimum, resulting in low cost (high reward) at the goal as well (Fig.3).

Overall, planning with a myopic collaborative model of the human behavior results in the robot taking actions that can guide the person towards plans that are globally optimal, helping them overcome the limitations of greedy action selection.

Non-Collaborative but Computing a Best Response to the Robot. Not every situations is collaborative. Take driving for example. The car has the same objective as its passenger, but a different objective from other human driven vehicles on the road – these are trying to reach their own destinations as efficiently as possible, and that sometimes competes with the car's objective.

Breaking the collaboration assumption that the human is optimizing the same reward function as the robot (or viceversa), puts making prediction of human behavior back to solving a 2-player game even if we assume known reward parameters. In [20], we introduced a model that alleviates this difficulty by assuming that the person is not computing to a Nash equilibrium, but instead computing a *best response* to the robot's plan using a different, yet known reward function $r_{\mathcal{H}}$. That is, rather than trying to influence the robot's behavior, the person is taking the robot's behavior as fixed, and optimizing their own reward function within that constraint:

$$\mathbf{u}^*_{\mathcal{H}}(\mathbf{u}_{\mathcal{R}}) = \arg\max_{\mathbf{u}_{\mathcal{H}}} R_{\mathcal{H}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}_{\mathcal{H}})$$

The robot can then compute the action sequence that, when combined with the human's response to that sequence, leads to the highest value for its *own* reward:

$$\mathbf{u}^*_{\mathcal{R}} = \arg\max_{\mathbf{u}_{\mathcal{R}}} R_{\mathcal{R}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}^*_{\mathcal{H}}(\mathbf{u}_{\mathcal{R}}))$$

It can then take the first action in $\mathbf{u}_{\mathcal{R}}$, observe the change in the world, and replan. This is what we did in the

collaborative replanning case with the Reactive robot, except now the robot has a model for how the person will respond to its actions as opposed to computing a joint global plan.

We applied this to autonomous driving, namely the interaction between an autonomous car and a human-driven vehicle. Both the robot and the person want to achieve their goal efficiently, which made their reward functions non-identical. We gave the robot access to $R_{\mathcal{H}}$ by learning it offline using IRL.

Typically in autonomous driving, cars treat people like obstacles, planning to stay out of their way. This leads to overly conservative behavior, like cars never getting to merge on a highway, or getting stuck at 4-way stops. In contrast, our car coordinates with people (Fig.4). It sometimes plans to merge in front of them knowing that they can slow down to accommodate the merge. Or at an intersection, for $R_{\mathcal{R}}$ being higher if the person goes first through the intersection, the robot does not just sit there, but coordinates by *starting to back up*, which makes it safer for the person to go (effectively *signaling* to the human driver). We ran a user study with a driving simulator, and the results suggested that people's behavior when the robot is planning with this model leads to significantly higher reward for the robot than in the baseline of treating the person as on obstacle.

Overall, treating interaction as an underactuated system where the person is not playing a game, but acting rationally as a best response to the robot's actions, leads to coordination behavior that naturally emerges out of the optimization over robot actions.

IV. Using Human Behavior to Infer Human Internal State

Thus far we've oversimplified the game by assuming that the robot knows the persons's reward parameters. In reality, the robot does have direct access to these. Even further, we argue that in reality, the robot does not really have access to its own reward parameters either – collaborative robots, meant to help a person, should optimize for whatever that person wants, not for some a-priori determined reward function.

Robots today take their reward function as given. But where does this reward function come from? Typically, it is designed by some person who does their best at writing down what they think the robot should optimize for. Unfortunately, we, people, are terrible at specifying what we want. From King Midas to The Sorcerer's Apprentice, we have countless stories that warn us about unintended, negative consequences of misspecified wishes or objectives. We propose that robots should have uncertainty over their objectives, and that they should try to optimize for what people want internally, but can't necessarily explicate. This is key to alleviating the negative consequences of a misspecified objective.

To achieve this, we use human actions as observations about their internal or desired reward function. as they would in the full dynamic game. But the robot will no longer assume that it knows $r_{\mathcal{H}}$. However, if we assume a rational model of human behavior, then the human actions become observations about this hidden internal human state. To estimate $\theta_{\mathcal{H}}$ from human actions, the robot needs an observation model – the probability of observed actions given $\theta_{\mathcal{H}}$. We assume the person is approximately rational [2, 25] with a model that comes from a maximum entropy distribution in which trajectories are more likely when their total reward is higher:

$$P(\mathbf{u}_{\mathcal{H}}|x^0, \mathbf{u}_{\mathcal{R}}, \theta_{\mathcal{H}}) \propto \exp(R_{\mathcal{H}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}_{\mathcal{H}}; \theta_{\mathcal{H}}))$$

Then the robot can update its belief over $\theta_{\mathcal{H}}$:

$$b'(\theta_{\mathcal{H}}) \propto b(\theta_{\mathcal{H}}) P(\mathbf{u}_{\mathcal{H}}|x^0, \mathbf{u}_{\mathcal{R}}, \theta_{\mathcal{H}})$$

If the robot observes a trajectory $\mathbf{u}_{\mathcal{H}}$ for the full time horizon, then the belief update equates to (Bayesian) Inverse Reinforcement Learning [17, 25].

But robots sometimes need to infer the human reward online, as the human trajectory is unfolding. Think back to the driving application: it is not the case that every person optimizes the same reward function: some drivers are more aggressive than others, some are not paying attention, and so on. It is therefore helpful to be able to update θ as the robot is interacting with the person. In such cases, the robot has only observed $\mathbf{u}_{\mathcal{H}}^{0:t}$ and must update its belief based just that, rather than a full trajectory.

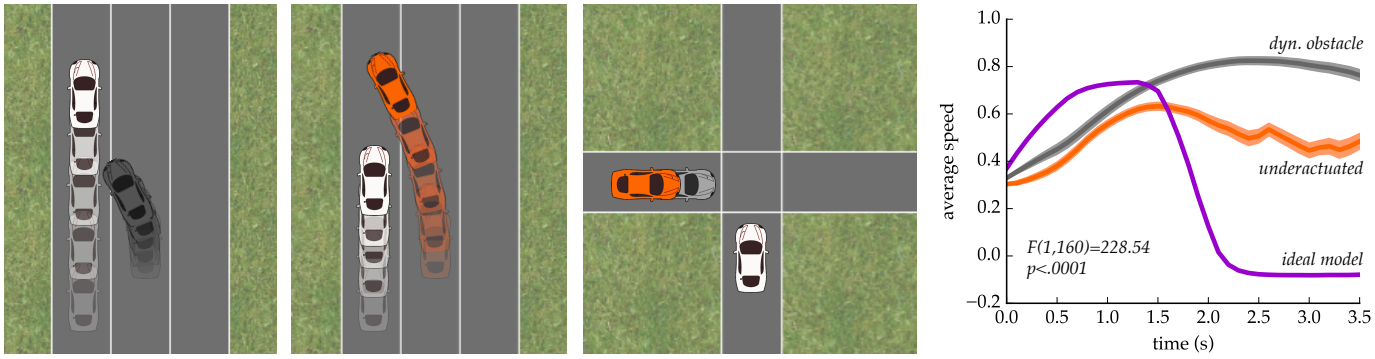
Further, robots have an opportunity to go beyond passive inference, and use their actions for *active* estimation, triggering informative human reactions. Instead of passively observing what people do, they can leverage their actions to gather information.

Finally, it's not just human physical actions that are informative. Human behavior in general, like physical corrections, comparisons, orders given to the robot, and even a reward function that a designer tries to write down – all of these are useful sources of information about what the true robot objective should be.

Online Inference by Integrating over Futures. In [5, 13], we integrated over the possible future human trajectories in order to compute the belief update:

$$P(\mathbf{u}_{\mathcal{H}}^{0:t}|x^0, \mathbf{u}_{\mathcal{R}}, \theta_{\mathcal{H}}) = \int P(\mathbf{u}_{\mathcal{H}}|x^0, \mathbf{u}_{\mathcal{R}}, \theta_{\mathcal{H}}) d\mathbf{u}_{\mathcal{H}}^{t+1:T}$$

We showed that for the case of the reward $r_{\mathcal{H}}$ being parametrized by which goal $\theta_{\mathcal{H}}$ the person is reaching for, the integral can be approximated via Laplace's method, assuming reward 0 for trajectories that do not reach the goal.



(a) Treating the Human as a Dynamic Obstacle

(b) Planning in the Underactuated System

(c) The Robot Inches Backwards to Get the Human to Cross First

(d) User Study: The Robot Does Affect the Human's Speed as Predicted

Fig. 4: Coordination behavior emerges out of not treating people like obstacles that are moving, but modeling how they will rationally react to robot actions.

Fig.2 shows a Predictive condition from [13], in which the robot infers the person's goal and uses it to proactively change its plan. This condition outperforms the Reactive condition. In [5], we used goal inference to adapt to an operator's goal during teleoperation of a robot arm.

Overall, human actions as observations about the underlying human goals, and inferring these enables robots to proactively adapt to what people want.

Active Online Inference using Robot Physical Actions. Inference does not have to be passive. In [21], we explored active inference, where the robot makes queries that the person responds to. But really, having a robot whose actions influence human behavior presents an opportunity: to leverage robot actions and trigger informative human reactions.

In [19], we ran online inference by modeling the human as optimizing with a shorter time horizon. We then took the human (short time horizon) trajectory as evidence about the underlying $\theta_{\mathcal{H}}$. There, $\theta_{\mathcal{H}}$ parametrized the reward function by representing weights on different important features of the human state and action. Unlike goals, this is a continuous and high-dimensional space. So rather than maintaining a belief over all possible $\theta_{\mathcal{H}}$s, we clustered users into styles and only maintained a belief over a discrete set of driving styles.

Further, we sped up the inference by leveraging the robot's actions: since the person will choose actions that depend on what the robot does, $\mathbf{u}_{\mathcal{R}}$, the robot has an opportunity to select actions that *maximize information gain* (trading off with maximizing reward using the current estimate $\hat{\theta}_{\mathcal{H}}$:

$$\mathbf{u}^*_{\mathcal{R}} = \arg\max_{\mathbf{u}_{\mathcal{R}}} R_{\mathcal{R}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}^*_{\mathcal{H}}(\mathbf{u}_{\mathcal{R}}; \hat{\theta_{\mathcal{H}}})) + \lambda(H(b) - \mathbb{E}_{\theta_{\mathcal{H}}} H(b'))$$

Note that if we were able to treat $\theta_{\mathcal{H}}$ as the hidden state in a POMDP with very complicated dynamics (that require planning for the person to solve for how the state will update given the robot's action), then the robot's policy would achieve an optimal trade-off between exploiting current information and gathering information. However, since even POMDPs with less complex dynamics are still intractable in continuous state and action, we resorted to an explicit trade-off.

We found that the robot planning in this formulation exhibited interesting behavior that could be seen as information-gathering. For instance, it would nudge closer to someone's lane, because the anticipated reaction from the person would be different depending on their driving style: attentive drivers break, distracted drivers continue. Or, at a 4-way stop, it would inch forward into the intersection, again anticipating different reactions for different styles.

The robot can leverage its actions' influence on the person to actively gather information about their internal reward parameters.

What If the Human Knows the Robot is Learning?. One issue with estimation arises when the observation model is wrong. People might act approximately optimal with respect to the reward function, except when they know that the robot is trying to learn something from them. This is why coaches are different from experts: when we teach, we simplify, we exaggerate, we showcase. A gymnastics coach does not demonstrate the same action they would perform if they were in the olympics.

In [9], we analyzed the difference between maximizing the reward function for the true $\theta^*_{\mathcal{H}}$:

$$\mathbf{u}^{expert}_{\mathcal{H}} = \arg\max_{\mathbf{u}_{\mathcal{H}}} R_{\mathcal{H}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}_{\mathcal{H}}; \theta^*_{\mathcal{H}})$$

and maximizing the probability that the robot will infer the true $\theta^*_{\mathcal{H}}$:

$$\mathbf{u}^{teacher}_{\mathcal{H}} = \arg\max_{\mathbf{u}_{\mathcal{H}}} b'(\theta^*_{\mathcal{H}}) = \arg\max_{\mathbf{u}_{\mathcal{H}}} P(\theta^*_{\mathcal{H}} | x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}_{\mathcal{H}})$$

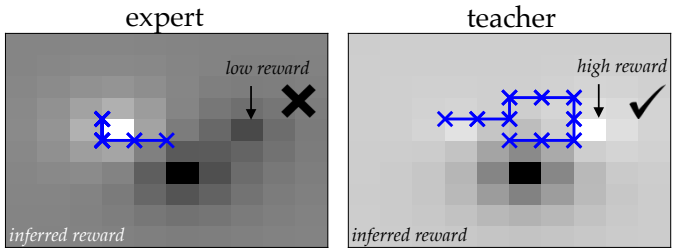


Fig. 5: We model teaching demonstrations as being informative about the underlying reward function. An expert demonstration (left) might lead to inferring the wrong reward function, e.g. because the expert goes straight for the high reward peak nearby. A teaching demonstration will deviate from optimality to showcase the underlying reward function, e.g. the teacher goes to both high reward peaks to clarify.

Fig.5 compares $\mathbf{u}_{\mathcal{H}}^{expert}$ to $\mathbf{u}_{\mathcal{H}}^{teacher}$ in a simple MDP where the reward function consists of high and low reward peaks. The expert demonstration heads straight to the closest high reward peak, but the robot has trouble inferring that there is another peak with also high reward. In contrast, the teaching demonstration visits both, leading to the robot inferring the correct $\theta_{\mathcal{H}}$.

Overall, we should expect and account for how people will act differently when they are trying to teach the robot about their internal reward parameters.

Learning Objectives from Rich Human Guidance. It is not just human physical actions as part of the task that should inform the robot about the internal human objective. We explored physical corrections [1] (Fig.6), comparisons [21], orders (human oversight) [8, 14], and even attempts at specifying an objective [10], all as sources of information for the robot. Each required its own observation model, and its own approximations for running the inference.

In [10], we proposed to mode the reward design process: the probability that a reward designer would choose $\theta_{\mathcal{R}}$ as the specified reward, given the true reward θ^* *and* the training environments they are considering. We then showed how the robot can invert this model to get a posterior distribution over what the true reward is, and that this alleviates consequences like reward hacking and negative side-effects. Surprisingly, we found that this works even when the important features affecting the true reward, like the presence of a dangerous kind of terrain, are latent and not directly observed.

Even more surprising is our finding from [8], where we focused on shut down orders to the robot. Intuitively, robots should just follow such orders as opposed to try to infer the underlying reward function that triggered them. Unfortunately though, designing a reward function that incentivizes accepting orders is challenging, and so is writing down the right hard constraints that the robot should follow. Instead, our work has proved that when the robot treats orders as a useful source of information about its objective, the incentive to accept them is positive.

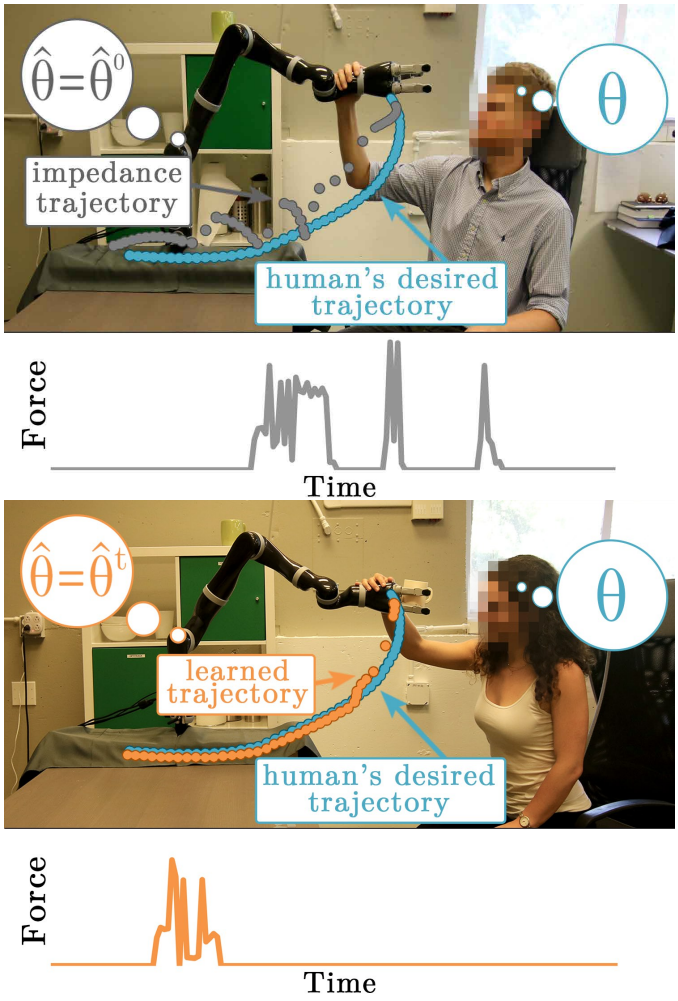


Fig. 6: Learning the desired objective function from physical corrections, in real-time, leads to completing the task in the desired way with less human intervention.

Overall, we find that treating human guidance and oversight as a useful source of information about the robot's true reward can alleviate the unintended consequences of misspecified robot reward functions.

V. Human Inferences about the Robot

The previous sections made approximations in which the person knew everything they needed about the robot – they were computing a best response to the robot and got access to the robot's planned trajectory $\mathbf{u}_{\mathcal{R}}$. Here, we relax this. Much like robots do not know everything about people and make inferences about their reward function or goals $\theta_{\mathcal{H}}$, people too will not know everything about robots and will try to make similar inferences when deciding on these actions.

Humans interacting with robots will have some belief about $\theta_{\mathcal{R}}$. This section focuses on how robot actions affect not just human actions, but also these human beliefs. This means the robot can specifically choose

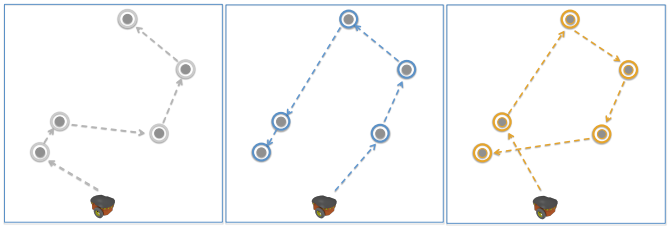


Fig. 7: After seeing the first t actions, the person should be able to infer confidently the remaining ones. Imagine seeing the first step of the most efficient plan, on the left. It is clear what the robot's 2nd action will be, but after that there are two courses that are relatively close in efficiency. On the other hand, with the middle plan, the first action only leaves one remaining plan sensible. It sacrifices efficiency to make the final T-1 actions clear.

actions to guide these beliefs towards being as accurate as possible, so that the human actions that follow are well informed. These robot actions end up *communicating* to the human, expressing the robot's internal state.

As of now, we modify the robot's objective to explicitly incentivize communication. We are actively working on making this communication emerge out of the robot optimizing for its own reward function, but now with this more sophisticated model of the person – one in which robot actions affect human beliefs, and human beliefs are what affect human actions.

A. Humans Expecting Robot Behavior to Be Approximately Rational

A simple but important inference that people make when observing other agents is what they expect the agent's actions to be. The principle of rational action [22] suggests that people expect rational agents, such as robots, to be rational – to maximize their own reward:

$$P_{\mathcal{H}}(\hat{\mathbf{u}_{\mathcal{R}}}|x^0) \propto e^{\hat{R_{\mathcal{R}}}(x^0,\mathbf{u}_{\mathcal{R}})}$$

Here, $\hat{\mathbf{u}_{\mathcal{R}}}$ is the person's estimate of what the robot's action sequence will be, and $\hat{R_{\mathcal{R}}}$ is the person's estimate of what the robot's reward function is.

Our work leveraged this to generate plans that match what people expect. Namely, at a time t, after the person has observed $\mathbf{u}_{\mathcal{R}}^{0:t}$, we can model what they expect the robot to do next as

$$P_{\mathcal{H}}(\hat{\mathbf{u}_{\mathcal{R}}}^{t+1:T}|x^0,\mathbf{u}_{\mathcal{R}}^{0:t}) \propto \exp(\hat{R_{\mathcal{R}}}(x^0,\mathbf{u}_{\mathcal{R}}^{0:t}) + \hat{R_{\mathcal{R}}}(x^{t+1},\mathbf{u}_{\mathcal{R}}^{t+1:T}))$$

The robot can use this model to choose a full time horizon plan or trajectory such that the beginning of the plan is informative of the remaining plan, i.e. makes the remaining plan have high probability:

$$t\text{-predictability}(\mathbf{u}_{\mathcal{R}}) = P_{\mathcal{H}}(\mathbf{u}_{\mathcal{R}}{}^{t+1:T}|x^0,\mathbf{u}_{\mathcal{R}}^{0:t})$$

$$\mathbf{u}_{\mathcal{R}}^{*} = \arg\max_{\mathbf{u}_{\mathcal{R}}} t\text{-predictability}(\mathbf{u}_{\mathcal{R}})$$

Note that the robot, when interested in its plan being t-predictable, might purpusefully deviate from the optimum with respect to $\hat{R_{\mathcal{R}}}$ in order to make sure that the remainder of the plan is what the person would predict after observing t time steps.

Fig.7 shows plans optimized for different t: 0, 1, and 2. The $t = 0$ one is the most efficient. The problem is that after seeing the first step, there are two possible plans that are relatively efficient, so this plan does not do a great job collapsing the person's belief over what will happen, even after they have seen some of the trajectory. In contrast, for $t = 1$, this is no longer the most efficient, but makes it very clear what the remainder of the plan will be. [7] details our user studies, both online and in person, with results suggesting that people have an easier time coordinating with robots that are more t-predictable.

Overall, the robot can leverage the person's expectations about its actions to make its plans more predictable.

B. Humans Using Robot Behavior to Infer Robot Internal State

Once people have a model of how the robot will behave, they can also start using that model to perform inference about hidden states, like the robot's goals or objectives.

Building on [2, 4], we have been exploring Bayesian Inference as a model of how people infer robot internal state $\theta_{\mathcal{R}}$ from observed robot actions. This model is analogous to the algorithms we used in the human behavior section to enable robots to infer human internal state from observer human actions:

$$P_{\mathcal{H}}(\mathbf{u}_{\mathcal{R}}|x^0,\theta_{\mathcal{R}}) \propto \exp(\hat{R_{\mathcal{R}}}(x^0,\mathbf{u}_{\mathcal{R}};\theta_{\mathcal{R}}))$$

$$b'_{\mathcal{H}}(\theta_{\mathcal{R}}) \propto b_{\mathcal{H}}(\theta_{\mathcal{R}}) P_{\mathcal{H}}(\mathbf{u}_{\mathcal{R}}|x^0,\theta_{\mathcal{R}})$$

The robot can now communicate a $\theta_{\mathcal{R}}^{*}$:

$$\mathbf{u}_{\mathcal{R}}^{*} = \arg\max_{\mathbf{u}_{\mathcal{R}}} b'_{\mathcal{H}}(\theta_{\mathcal{R}}^{*})$$

This is analogous to pragmatics, but the communication happens through physical behavior and not through language.

Communicating Robot Goals. In earlier work [6], we studied a version of this formulation where $\theta_{\mathcal{R}}$ is the robot's goal. A manipulator arm decides to exaggerate its trajectory to the right to convey that the correct goal is the one on the right and not the one on the left (Fig.8), and this does lead to participants' inferring the robot's goal faster.

The human inference model when $\theta_{\mathcal{R}}$ is a goal is one that has been heavily explored in cognitive science (e.g. [2]), and this work showed what happens when a robot uses it for communication. The result is analogous to findings in human-human collaborations about how people exaggerate their motion to disambiguate their goal or intention [16].

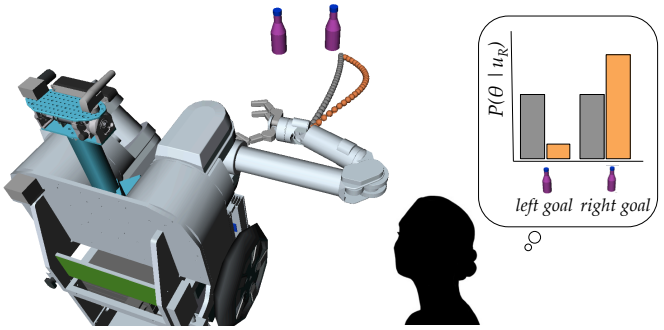


Fig. 8: The robot models the person as running Bayesian Inference to infer its goal. It chooses to exaggerate its motion to the right to convey that its goal is the bottle on the right.

Our ability to coordinate with each other relies heavily on predicting each others' intentions [16]. Modeling human inferences about intentions enables robots to purposefully deviate from efficiency in order to maximally clarify their intentions.

Communicating Robot Reward Parameters. More recently, we've been exploring how $\theta_{\mathcal{R}}$ does not have to be restricted to a goal. Much like how in Sec. IV we inferred not just human goals, but also more generally human reward function parameters, here too the robot can express not just goals.

In [12], we studied how an autonomous car can plan behavior that is informative of its reward function. The car decides to show an environment in which the optimal trajectory merges closely in front of another car (Fig.9, left), as opposed to an environment where merging into a lane away from the other car is optimal (Fig.9, right) – it finds behavior that is informative about the fact that its reward is rather aggressive as a driving style, prioritizing efficiency over safety.

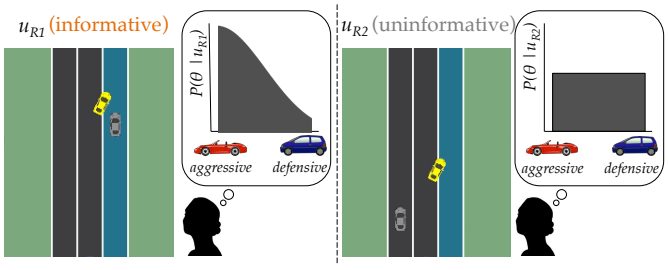


Fig. 9: The robot models the person as running Bayesian Inference to infer its objective parameters θ. It chooses its actions such that they maximally convey information about $\theta_{\mathcal{R}}$ – in this case that it prioritize aggressive driving/efficiency over safety.

As robots get more complex, understanding and verifying their reward functions is going to become more and more important to end-users. Modeling human inferences about reward parameters enables robots to choose actions sequences that are communicative of the true reward parameters.

Communicating Confidence. Even more recently, we have been exploring spaces of θs beyond even reward parameters. Robot actions implicitly communicate about many different aspects of robot internal state. We have found that people observe robot actions and make attributions about its confidence (Fig.10), or about the weight of the object that the robot is carrying [24].

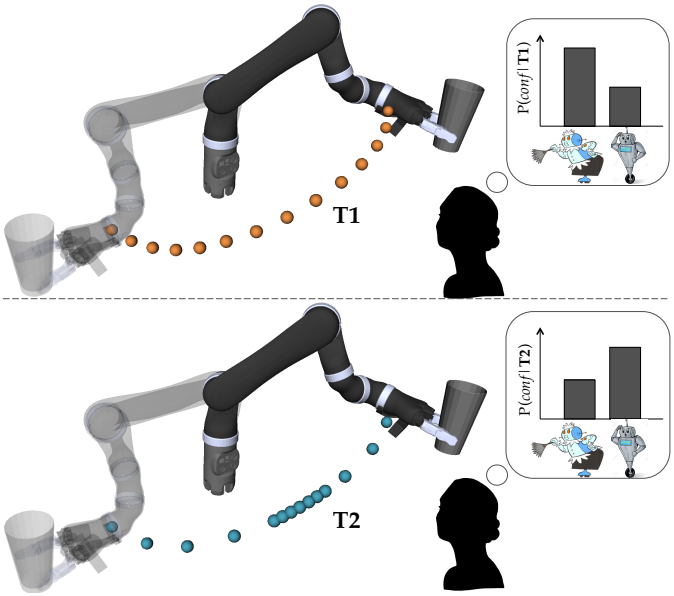


Fig. 10: Different motion timings communicate different levels of robot confidence.

VI. Discussion

This writeup synthesized our findings in integrating mathematical models of human state and action into robot planning of physical behaviors for interactive tasks. We focused on rational models of human behavior in a two-player game, showing how different approximations to the solution lead to different robot behaviors.

A first set of approximations assume that the person has access to what the robot will do, and the robot has to the person's overall reward or utility function. Still, we found that the robot generates behaviors that adapt to the person, that guide the person towards better performance in the task, or that account for the influence the robot will have on what the person ends up doing. We saw robots handing over objects to compensate for people's tendencies to just grasp them in the most comfortable way, and cars being more effective on the road by triggering responses from other drivers.

More sophisticated approximations accounted for the fact that different people have different reward functions, and showed that the robot can actively estimate relevant parameters online, leading to interesting coordination behaviors, like cars deciding on trajectories that look like inching forward at intersections or nudging into lanes to probe whether another driver will let them through.

Finally, even further approximations acknowledge that people will need to make predictions about the robot, in

the same way that the robot makes predictions about people. This leads to robots that are more transparent, communicating their reward function (e.g. their driving style) through the way they act.

This work is limited in many ways, including the fact that as models of people get more complex, it becomes harder to generate robot behavior in real time (especially behavior that escapes poor local optima). However, it is exciting to see the kinds of coordination behaviors that we typically need to hand-craft starting to emerge out of low-level planning *directly in the robot's control space*. This requires breaking outside of the typical AI paradigm, and formally reasoning about people's internal states and behavior.

References

[1] Andrea Bajcsy, Dylan Losey, Martia O'Malley, and Anca Dragan. Learning robot objectives from physical human interaction. In *in review*, 2017.

[2] Chris L Baker, Joshua B Tenenbaum, and Rebecca R Saxe. Goal inference as inverse planning. In *Proceedings of the Cognitive Science Society*, volume 29, 2007.

[3] A. Bestick, R. Bajcsy, and A.D. Dragan. Implicitly assisting humans to choose good grasps in robot to human handovers. In *International Symposium on Experimental Robotics (ISER)*, 2016.

[4] Gergely Csibra and György Gergely. ?obsessed with goals?: Functions and mechanisms of teleological interpretation of actions in humans. *Acta psychologica*, 124(1): 60–78, 2007.

[5] A.D. Dragan and S.S. Srinivasa. Formalizing assistive teleoperation. In *Robotics: Science and Systems (R:SS)*, 2012.

[6] A.D. Dragan and S.S. Srinivasa. Generating legible motion. In *Robotics: Science and Systems (R:SS)*, 2013.

[7] J. Fisac, C. Liu, J. Harick, K. Hedrick, S. Sastry, T. Griffiths, and A.D. Dragan. Generating plans that predict themselves. In *Workshop on the Algorithmic Foundations of Robotics (WAFR)*, 2016.

[8] D. Hadfield-Menell, A.D. Dragan, P. Abbeell, and S. Russell. The off-switch game. In *International Joint Confernece on Artificial Intelligence (IJCAI)*, 2017.

[9] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart J Russell. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 3909–3917, 2016.

[10] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Inverse reward design. In *in review*, 2017.

[11] Trey Hedden and Jun Zhang. What do you think i think you think?: Strategic reasoning in matrix games. *Cognition*, 85(1):1–36, 2002.

[12] S. Huang, P. Abbeel, and A.D. Dragan. Enabling robots to communicate their objectives. In *Robotics: Science and Systems (RSS)*, 2017.

[13] C. Liu, J. Harick, J. Fisac, A.D. Dragan, K. Hedrick, S. Sastry, and T. Griffiths. Goal inference improves objective and perceived performance in human-robot collaboration. In *Autonomous Agents and Multiagent Systems (AAMAS)*, 2016.

[14] S. Milli, D. Hadfield-Menell, A.D. Dragan, P. Abbeell, and S. Russell. Should robots be obedient? In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[15] Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, pages 663–670, 2000.

[16] Giovanni Pezzulo, Francesco Donnarumma, and Haris Dindo. Human sensorimotor communication: A theory of signaling in online social interactions. *PLoS One*, 8(11): e79876, 2013.

[17] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. *Urbana*, 51(61801):1–4, 2007.

[18] Ariel Rubinstein. *Modeling bounded rationality*. MIT press, 1998.

[19] D. Sadigh, S. Sastry, S. Seshia, and A.D. Dragan. Information gathering actions over human internal state. In *International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[20] D. Sadigh, S. Sastry, S. Seshia, and A.D. Dragan. Planning for autonomous cars that leverages effects on human drivers. In *Robotics: Science and Systems (R:SS)*, 2016.

[21] D. Sadigh, A.D. Dragan, S. Sastry, and S. Seshia. Active preference-based learning of reward functions. In *Robotics: Science and Systems (RSS)*, 2017.

[22] Beate Sodian, Barbara Schoeppner, and Ulrike Metz. Do infants apply the principle of rational action to human agents? *Infant Behavior and Development*, 27(1):31–41, 2004.

[23] Andrea Thomaz, Guy Hoffman, Maya Cakmak, et al. Computational human-robot interaction. *Foundations and Trends® in Robotics*, 4(2-3):105–223, 2016.

[24] A. Zhang, D. Hatfield-Menell, A. Nagabadi, and A.D. Dragan. Expressive robot motion timing. In *International Conference on Human-Robot Interaction (HRI)*, 2017.

[25] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *AAAI*. Chicago, IL, USA, 2008.