Information Gathering Actions over Human Internal State

Dorsa Sadigh S. Shankar Sastry Sanjit A. Seshia Anca Dragan

Abstract— **Much of estimation of human internal state (goal, intentions, activities, preferences, etc.) is *passive*: an algorithm observes human actions and updates its estimate of human state. In this work, we embrace the fact that robot actions affect what humans do, and leverage it to improve state estimation. We enable robots to do *active information gathering*, by planning actions that probe the user in order to clarify their internal state. For instance, an autonomous car will plan to nudge into a human driver's lane to test their driving style. Results in simulation and in a user study suggest that active information gathering significantly outperforms passive state estimation.**

I. Introduction

Imagine driving on the highway. Another driver is in the lane next to you, and you need to switch lanes. Some drivers are aggressive and they will never brake to let you in. Others are more defensive and would gladly make space for you. You don't know what kind of driver this is, so you decide to gently nudge in towards the other lane to *test* their reaction. At an intersection, you might *nudge in* to test if the other driver is distracted and they might just let you go through (Fig.1 bottom left). Our goal in this work is to give robots the capability to plan such actions as well.

In general, human behavior is affected by internal states that a robot would not have direct access to: intentions, goals, preferences, objectives, driving style, etc. Work in robotics and perception has focused thus far on estimating these internal states by providing algorithms with observations of humans acting, be it intent prediction [23], [13], [6], [3], [4], [16], Inverse Reinforcement Learning [1], [12], [15], [22], [18], driver style prediction [11], affective state prediction [10], or activity recognition [20].

Human state estimation has also been studied in the context of human-robot interaction tasks. Here, the robot's reward function depends (directly or indirectly) on the human internal state, e.g., on whether the robot is able to adapt to the human's plan or preferences. Work in assistive teleoperation or in human assistance has cast this problem as a Partially Observable Markov Decision Process, in which the robot does observe the physical state of the world but not the human internal state — that it has to estimate from human actions. Because POMDP solvers are not computationally efficient, the solutions proposed thus far use the current estimate of the internal state to plan (either using the most likely estimate, or the entire current belief), and adjust this estimate at every step [8], [7], [5]. Although efficient, these approximations sacrifice an important aspect of POMDPs: the ability to *actively gather information*.

Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley 94720 dsadigh,sastry,sseshia,dragan@eecs.berkeley.edu

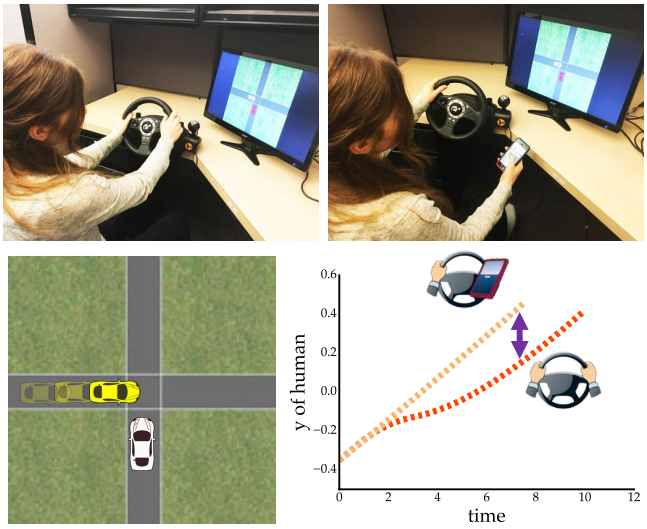


Fig. 1: We enable robots to generate actions that actively probe humans in order to find out their internal state. We apply this to autonomous driving. In this example, the robot car (yellow) decides to inch forward in order to test whether the human driver (white) is attentive. The robot expects drastically different reactions to this action (bottom right shows attentive driver reaction in light orange, and distracted driver reaction in dark orange). We conduct a user study in which we let drivers pay attention or distract them with cellphones in order to put this state estimation algorithm to the test.

> *Our key insight is that robots can leverage their own actions to help estimation of human internal state.*

Rather than relying on passive observations, robots can actually account for the fact that humans will react to their actions: they can use this knowledge to select actions that will trigger human reactions which in turn will clarify the internal state.

We make two contributions:

An Algorithm for Active Information Gathering over Human Internal State. We introduce an algorithm for planning robot actions that have high expected information gain. Our algorithm uses a reward-maximization model of how humans plan their actions in response to those of the robot's [19], and leverages the fact that different human internal states will lead to different human reactions to speed up estimation. Fig.1 shows an example of the anticipated difference in reaction between a distracted and an attentive driver.

Application to Driver Style Estimation. We apply our algorithm to estimating a human driver's style during the interaction of an autonomous vehicle with a human-driven vehicle. Results in simulation as well as from a user study suggest that our algorithm's ability to leverage robot actions for estimation leads to significantly higher accuracy in identifying the correct human internal state. The autonomous car plans actions like inching forward at an intersection (Fig.1), nudging into another car's lane, or braking slightly in front of a human-driven car, all to estimate whether the human driver is attentive.

Overall, we are excited to have taken a step towards giving robots the ability to actively probe end-users through their actions in order to better estimate their goals, preferences, styles, and so on. Even though we chose driving as our application domain in this paper, our algorithm is general across different domains and types of human internal state. We anticipate that applying it in the context of human goal inference during shared autonomy, for instance, will lead to the robot purposefully committing to a particular goal in order to trigger a reaction from the user, either positive or negative, in order to clarify the desired goal. Of course, further work is needed in order to evaluate how acceptant end-users are of different kinds of such probing actions.

II. Information Gathering Actions

We start with a general formulation of the problem of a robot needing to maximize its reward by acting in an environment where a human is also acting. The human is choosing its actions in a manner that is responsive to the robot's actions, and also influenced by some internal variables. While most methods that have addressed such problems have proposed approximations based on *passively* estimating the internal variable and exploiting that estimate, here we propose a method for *active* information gathering that enables the robot to purposefully take actions that probe the human. Finally, we discuss our implementation in practice, which trades off between exploration and exploitation.

A. General Formulation

We define a human-robot system in which the human's actions depend on some human internal state φ that the robot does not directly observe. In a driving scenario, φ might correspond to *driving style*: aggressive or timid, attentive or distracted. In collaborative manipulation scenarios, φ might correspond to the human's current *goal*, or their *preference* about the task.

We let $x \in X$ be a continuous physical state of our system. For our running example of autonomous cars, this includes position, velocity and heading of the autonomous and human driven vehicles. Let $\varphi \in \Phi$ be the hidden variable, e.g., human driver's driving style. We assume the robot observes the current physical state x^t, but not the human internal state φ.

The robot and human can apply continuous controls $u_\mathcal{R}$ and $u_\mathcal{H}$. The dynamics of the system evolves as robot's and human's control inputs arrive at each step:

$$x^{t+1} = f_\mathcal{H}(f_\mathcal{R}(x^t, u^t_\mathcal{R}), u^t_\mathcal{H}). \quad (1)$$

Here, $f_\mathcal{R}$ and $f_\mathcal{H}$ represent how the actions of the robot and of the human respectively affect the dynamics, and can be applied synchronously or asynchronously. We assume that while x changes via (1) based on the human and robot actions, φ does not. For instance, we assume that the human maintains their preferences or driving style throughout the interaction.

The robot's reward function in the task depends on the current state, the robot's action, as well as the action that the human takes at that step in response, $r_\mathcal{R}(x^t, u^t_\mathcal{R}, u^t_\mathcal{H})$.

If the robot has access to the human's policy $\pi_\mathcal{H}(x, u_\mathcal{R}, \varphi)$, maximizing robot reward can be modeled as a POMDP [9] with states (x, φ), actions $u_\mathcal{R}$, and reward $r_\mathcal{R}$. In this POMDP, the dynamics model can be computed directly from (1) with $u^t_\mathcal{H} = \pi_\mathcal{H}(x^t, u^t_\mathcal{R}, \varphi)$. The human's actions can serve as observations of φ via some $P(u_\mathcal{H}|x, u_\mathcal{R})$. In Sec. II-C we introduce a model for both $\pi_\mathcal{H}$ and $P(u_\mathcal{H}|x, u_\mathcal{R}, \varphi)$ based on the assumption that the human is maximizing their own reward function.

If we were able to solve the POMDP, the robot would estimate φ based on the human's actions, and optimally trade off between exploiting its current belief over φ, and *actively taking information gathering actions* meant to cause human reactions that give the robot a better estimate of the hidden variable φ.

Because POMDPs cannot be solved tractably, several approximations have been proposed for similar problem formulations [8], [11], [7]. These approximations are *passively* estimating the human internal state, and exploiting the belief to plan robot actions.[1]

In this work, we take the opposite approach: we focus explicitly on active information gathering. We enable the robot to decide to actively probe the person to get a better estimate of φ. Our method can be leveraged in conjunction with exploitation methods, or be used alone when human state estimation is robot's primary objective.

B. Reduction to Information Gathering

At every step, the robot can update its belief over φ via:

$$b^{t+1}(\varphi) \propto b^t(\varphi) \cdot P(u_\mathcal{H}|x^t, u_\mathcal{R}, \varphi). \quad (2)$$

To explicitly focus on taking actions to estimate φ, we redefine the robot's reward function to capture the

[1]One exception is Nikolaidis et al. [17], who propose to solve the full POMDP, albeit for discrete and not continuous state and action spaces.

information gain at every step:

$$r_{\mathcal{R}}(x^t, u_{\mathcal{R}}, u_{\mathcal{H}}) = H(b^t) - H(b^{t+1}) \tag{3}$$

with $H(b)$ being the entropy over the belief:

$$H(b) = -\frac{\sum_{\varphi} b(\varphi)\log(b(\varphi))}{\sum_{\varphi} b(\varphi)}. \tag{4}$$

Optimizing expected reward now entails reasoning about the effects that the robot actions will have on what observations the robot will get, i.e., the actions that the human will take in response, and how useful these observations will be in shattering ambiguity about φ.

C. Solution: Human Model & Model Predictive Control

We solve the information gathering planning problem via Model Predictive Control (MPC) [14]. At every time step, we find the optimal actions of the robot $\mathbf{u}^*_{\mathcal{R}}$ by maximizing the expected reward function over a finite horizon.

Notation. Let x^0 be the state at the current time step, i.e., at the beginning of the horizon. $\mathbf{u}_{\mathcal{R}} = (u^0_{\mathcal{R}}, \dots, u^{N-1}_{\mathcal{R}})$ be a finite sequence of the robot's continuous actions, and $\mathbf{u}_{\mathcal{H}} = (u^0_{\mathcal{H}}, \dots, u^{N-1}_{\mathcal{H}})$ be a finite sequence of human's continuous actions. Further, let $R_{\mathcal{R}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}_{\mathcal{H}})$ denote the reward over the finite horizon, if the agents started in x^0 and executed $\mathbf{u}_{\mathcal{R}}$ and $\mathbf{u}_{\mathcal{H}}$, which can be computed via the dynamics in (1).

MPC Maximization Objective. At every time step, the robot is computing the best actions for the horizon:

$$\mathbf{u}^*_{\mathcal{R}} = \arg\max_{\mathbf{u}_{\mathcal{R}}} \quad \mathbb{E}_{\varphi}\big[\, R_{\mathcal{R}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}^{*\varphi}_{\mathcal{H}}(x^0, \mathbf{u}_{\mathcal{R}}))\,\big] \tag{5}$$

where $\mathbf{u}^{*\varphi}_{\mathcal{H}}(x^0, \mathbf{u}_{\mathcal{R}})$ corresponds to the actions the human *would* take from state x^0 over the horizon of N steps if the robot executed actions $\mathbf{u}_{\mathcal{R}}$. Here, the expectation is taken over the current belief over φ, b^0.

Simplifying (5) using the definition of reward from (3), we get:

$$\mathbf{u}^*_{\mathcal{R}} = \arg\max_{\mathbf{u}_{\mathcal{R}}} \quad \mathbb{E}_{\varphi}\big[\, H(b^0) - H(b^N)\,\big], \tag{6}$$

$$\mathbf{u}^*_{\mathcal{R}} = \arg\max_{\mathbf{u}_{\mathcal{R}}} \quad \mathbb{E}_{\varphi}\big[\, -H(b^N)\,\big], \tag{7}$$

where the expectation remains with respect to b^0.

Human Model. We assume that the human maximizes their own reward function at every step. We let $r^{\varphi}_{\mathcal{H}}(x^t, u^t_{\mathcal{R}}, u^t_{\mathcal{H}})$ represent human's reward function at time t, which is parametrized by the human internal state φ. Then, the sum of human rewards over horizon N is:

$$R^{\varphi}_{\mathcal{H}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}_{\mathcal{H}}) = \sum_{t=0}^{N-1} r^{\varphi}_{\mathcal{H}}(x^t, u^t_{\mathcal{R}}, u^t_{\mathcal{H}}) \tag{8}$$

Building on our previous work [19], which showed how the robot can plan using such a reward function when there are no hidden variables, we compute $\mathbf{u}^{*\varphi}_{\mathcal{H}}(x^0, \mathbf{u}_{\mathcal{R}})$ through an approximation. We model the human as having access to $\mathbf{u}_{\mathcal{R}}$ a priori, and compute the finite horizon human actions that maximize the human's reward:

$$\mathbf{u}^{*\varphi}_{\mathcal{H}}(x^0, \mathbf{u}_{\mathcal{R}}) = \arg\max_{\mathbf{u}_{\mathcal{H}}} \quad R^{\varphi}_{\mathcal{H}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}_{\mathcal{H}}) \tag{9}$$

One can find $r^{\varphi}_{\mathcal{H}}$ through Inverse Reinforcement Learning (IRL) [1], [12], [15], [22], by getting demonstrations of human behavior associated with a direct measurement of φ.

The robot can use this reward in the dynamics model in order to compute human actions via (9). To update the belief b and compute expected reward in (7), we still need an observation model. We assume that actions with lower reward are exponentially less likely, building on the principle of maximum entropy [22]:

$$P(u_{\mathcal{H}} | x, u_{\mathcal{R}}, \varphi) \propto \exp(\varphi_{\mathcal{H}}(x, u_{\mathcal{R}}, u_{\mathcal{H}})) \tag{10}$$

Optimization Procedure. To solve (5) (or equivalently (7)), we use a gradient descent optimization method, L-BFGS, designed for unconstrained nonlinear problems [2]. Therefore, we would like to find the gradient of the objective in equation (5) with respect to $\mathbf{u}_{\mathcal{R}}$. Since the objective is the expectation of $R_{\mathcal{R}}$, we can reformulate this gradient as:

$$\begin{aligned}&\frac{\partial \mathbb{E}_{\varphi}\big[\, R_{\mathcal{R}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}^{*\varphi}_{\mathcal{H}}(x^0, \mathbf{u}_{\mathcal{R}}))\,\big]}{\partial \mathbf{u}_{\mathcal{R}}} \\ &= \sum_{\varphi} \frac{\partial R_{\mathcal{R}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}^{*\varphi}_{\mathcal{H}}(x^0, \mathbf{u}_{\mathcal{R}}))}{\partial \mathbf{u}_{\mathcal{R}}} \cdot b^0(\varphi)\end{aligned} \tag{11}$$

Then, we only need to find $\frac{\partial R_{\mathcal{R}}}{\partial \mathbf{u}_{\mathcal{R}}}$, which is equivalent to:

$$\begin{aligned}&\frac{\partial R_{\mathcal{R}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}^{*\varphi}_{\mathcal{H}}(x^0, \mathbf{u}_{\mathcal{R}}))}{\partial \mathbf{u}_{\mathcal{R}}} = \\ &\frac{\partial R_{\mathcal{R}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}_{\mathcal{H}})}{\partial \mathbf{u}_{\mathcal{H}}} \frac{\partial \mathbf{u}^{*\varphi}_{\mathcal{H}}}{\partial \mathbf{u}_{\mathcal{R}}} + \frac{\partial R_{\mathcal{R}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}_{\mathcal{H}})}{\partial \mathbf{u}_{\mathcal{R}}}\Big|_{\mathbf{u}_{\mathcal{H}} = \mathbf{u}^{*\varphi}_{\mathcal{H}}(x^0, \mathbf{u}_{\mathcal{R}})}\end{aligned} \tag{12}$$

Because $R_{\mathcal{R}}$, as indicated by (7), simplifies to the negative entropy of the updated belief, we can compute both $\frac{\partial R_{\mathcal{R}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}_{\mathcal{H}})}{\partial \mathbf{u}_{\mathcal{H}}}$ and $\frac{\partial R_{\mathcal{R}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}_{\mathcal{H}})}{\partial \mathbf{u}_{\mathcal{R}}}\Big|_{\mathbf{u}_{\mathcal{H}} = \mathbf{u}^{*\varphi}_{\mathcal{H}}(x^0, \mathbf{u}_{\mathcal{R}})}$ symbolically.

This leaves $\frac{\partial \mathbf{u}^{*\varphi}_{\mathcal{H}}}{\partial \mathbf{u}_{\mathcal{R}}}$. We use the fact that the gradient of $R_{\mathcal{H}}$ will evaluate to zero at $\mathbf{u}^{*\varphi}_{\mathcal{H}}$:

$$\frac{\partial R_{\mathcal{H}}}{\partial \mathbf{u}_{\mathcal{H}}}(x^0, \mathbf{u}_{\mathcal{R}}, \mathbf{u}^{*\varphi}_{\mathcal{H}}(x^0, \mathbf{u}_{\mathcal{R}})) = 0 \tag{13}$$

Now, differentiating this expression with respect to $\mathbf{u}_{\mathcal{R}}$ will result in:

$$\frac{\partial^2 R_{\mathcal{H}}}{\partial \mathbf{u}^2_{\mathcal{H}}} \frac{\partial \mathbf{u}^{*\varphi}_{\mathcal{H}}}{\partial \mathbf{u}_{\mathcal{R}}} + \frac{\partial^2 R_{\mathcal{H}}}{\partial \mathbf{u}_{\mathcal{H}} \partial \mathbf{u}_{\mathcal{R}}} \frac{\partial \mathbf{u}_{\mathcal{R}}}{\partial \mathbf{u}_{\mathcal{R}}} = 0 \tag{14}$$

Then, solving for $\frac{\partial \mathbf{u}_{\mathcal{H}}^{*\varphi}}{\partial \mathbf{u}_{\mathcal{R}}}$ enables us to find the following symbolic expression:

$$\frac{\partial \mathbf{u}_{\mathcal{H}}^{*\varphi}}{\partial \mathbf{u}_{\mathcal{R}}} = [-\frac{\partial^2 R_{\mathcal{H}}}{\partial \mathbf{u}_{\mathcal{H}} \partial \mathbf{u}_{\mathcal{R}}}][\frac{\partial^2 R_{\mathcal{H}}}{\partial \mathbf{u}_{\mathcal{H}}^2}]^{-1}. \tag{15}$$

This expression allows finding a symbolic expression for the gradient in equation (11).

D. Explore-Exploit Trade-Off

In practice, we use information gathering in conjunction with exploitation. We do not solely optimize the reward from Sec. II-B, but optimize it in conjunction with the robot's actual reward function *assuming the current estimate of* φ:

$$\begin{aligned} r_{\mathcal{R}}(x^t, \mathbf{u}_{\mathcal{R}}, \mathbf{u}_{\mathcal{H}}) = H(b^t) - H(b^{t+1}) \\ + \lambda \cdot r_{goal}(x^t, \mathbf{u}_{\mathcal{R}}, \mathbf{u}_{\mathcal{H}}, b^t) \end{aligned} \tag{16}$$

At the very least, we do this as a measure of safety, e.g., we want an autonomous car to keep avoiding collisions even when it is actively probing a human driver to test their reactions. We choose λ experimentally, though existing techniques that can better adapt λ over time [21].

Despite optimizing this trade-off, we do not claim that our method as-is can better solve the general POMDP formulation from Sec. II-A: only that it can be used to get better estimates of human internal state. The next sections test this in simulation and in practice, in a user study, and future work will look at how to leverage this ability to better solve human-robot interaction problems.

III. Simulation Results

In this section, we show simulation results that use the method from the previous section to estimate human driver type in the interaction between an autonomous vehicle and a human-driven vehicle.

In this section, we consider three different autonomous driving scenarios. In these scenarios, the human is either distracted or attentive during different driving experiments. The scenarios are shown in Fig.2, where the yellow car is the autonomous vehicle, and the white car is the human driven vehicle. Our goal is to plan to actively estimate the human's driving style in each one of these scenarios, by using the robot's actions.

A. Attentive vs. Distracted Human Driver Models

Our technique requires reward functions $r_{\mathcal{H}}^{\varphi}$ that model the human behavior for a particular internal state φ. We obtain a generic driver model via Continuous Inverse Optimal Control with Locally Optimal Examples [12] from demonstrated trajectories in a driving simulator in an environment with multiple autonomous cars, which followed precomputed routes.

We parametrize the human reward function as a linear combination of features, and learn weights on the features. We use various features including features for bounds on the control inputs, features that keep the vehicles within the road boundaries and close to the center of their lanes. Further, we use quadratic functions of speed to capture reaching the goal, and Gaussians around other vehicles on the road to enforce collision avoidance as part of the feature set.

We then adjust the learned weights to model attentive vs. distractive drivers. Specifically, we modify the weights of the collision avoidance features, so the distracted human model has less weight for these features. Therefore, the distracted driver is more likely to collide with the other cars while the attentive driver has high weights for the collision avoidance feature.

B. Manipulated Factors

We manipulate the *reward* function that the robot is optimizing. In the ***passive*** condition, the robot optimizes a simple reward function for collision avoidance based on the current belief estimate. It then updates this belief passively, by observing the outcomes of its actions at every time step. In the ***active*** condition, the robot trades off between this reward function and the Information Gain from (3) in order to explore the human's driving style.

We also manipulate the *human internal state* to be ***attentive*** or ***distracted***. The human is simulated to follow the ideal model of reward maximization for our two rewards.

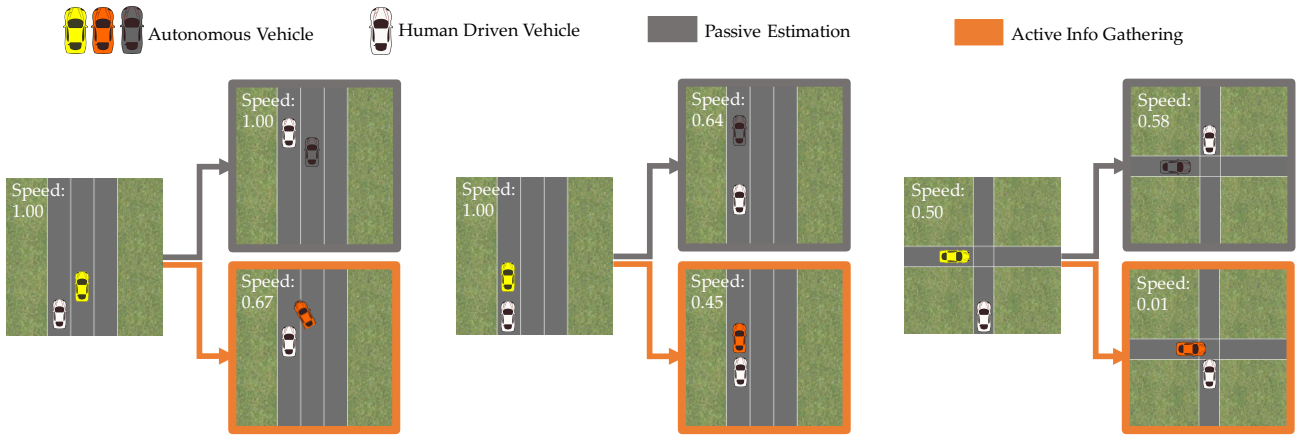
C. Driving Simulator

We use a simple point-mass model for the dynamics of the vehicle, where $\mathbf{x} = \begin{bmatrix} x & y & \theta & v \end{bmatrix}^\top$ is the state of the vehicle. Here, x and y are the coordinates of the vehicle, θ is the heading, and v is the speed. Each vehicle has two control inputs $\mathbf{u} = \begin{bmatrix} u_1 & u_2 \end{bmatrix}^\top$, where u_1 is the steering input, and u_2 is acceleration. Further, we let α be a friction coefficient. Then, the dynamics of each vehicle is formalized as:

$$[\dot{x} \quad \dot{y} \quad \dot{\theta} \quad \dot{v}] = [v \cdot \cos(\theta) \quad v \cdot \sin(\theta) \quad v \cdot u_1 \quad u_2 - \alpha \cdot v]. \tag{17}$$

D. Scenarios and Qualitative Results

Scenario 1: Nudging In to Explore on a Highway. In this scenario, we show an autonomous vehicle actively exploring the human's driving style in a highway driving setting. We contrast the two conditions in Fig.2(a). In the passive condition, the autonomous car drives on its own lane without interfering with the human throughout the experiment, and updates its belief based on passive observations gathered from the human car. *However, in the active condition, the autonomous car actively probes the human by nudging into her lane in order to infer her driving style. An attentive human significantly slows down (timid driver) or speeds up (aggressive driver) to avoid the vehicle, while a distracted*



(a) Scenario 1: Nudging in to Explore on a Highway (b) Scenario 2: Braking to Explore on a Highway (c) Scenario 3: Nudging in to Explore at an Intersection

Fig. 2: Our three scenarios, along with a comparison of robot plans for passive estimation (gray) vs active information gathering (orange). In the active condition, the robot is purposefully nudging in or braking to test human driver attentiveness. The color of the autonomous car in the initial state is yellow, but changes to either gray or orange in cases of passive and active information gathering respectively.

driver might not realize the autonomous actions and maintain their velocity, getting closer to the autonomous vehicle. It is this difference in reactions that enables the robot to better estimate φ.

Scenario 2: Braking to Explore on a Highway. In the second scenario, we show the driving style can be explored by the autonomous car probing the human driver behind it. The two vehicles start in the same lane as shown in Fig.2(b), where the autonomous car is in the front. In the passive condition, the autonomous car drives straight without exploring or enforcing any interactions with the human driven vehicle. *In the active condition, the robot slows down to actively probe the human and find out her driving style. An attentive human would slow down and avoid collisions while a distracted human will have a harder time to keep safe distance between the two cars.*

Scenario 3: Nudging In to Explore at an Intersection. In this scenario, we consider the two vehicles at an intersection, where the autonomous car actively tries to explore human's driving style by nudging into the intersection. The initial conditions of the vehicles are shown in Fig.2(c). In the passive condition, the autonomous car stays at its position without probing the human, and only optimizes for collision avoidance. This provides limited observations from the human car resulting in a low confidence belief distribution. *In the active condition, the autonomous car nudges into the intersection to probe the driving style of the human. An attentive human would slow down to stay safe at the intersection while a distracted human will not slow down.*

E. Quantitative Results

Throughout the remainder of the paper, we use a common color scheme to plot results for our experimental conditions. We show this common scheme in Fig.3: darker colors (black and red) correspond to attentive humans, and lighter colors (gray and orange) correspond to distracted humans. Further, the shades of orange correspond to active information gathering, while the shades of gray indicate passive information gathering. We also use solid lines for real users, and dotted lines for scenarios with an ideal user model learned through inverse reinforcement learning. This table is representative for the legends of Fig.4, Fig.**??**, Fig.5, and Fig.6.

	Attentive Human	Distracted Human
Active Robot	Real User (solid line) Ideal User Model (dotted line)	Real User (solid line) Ideal User Model (dotted line)
Passive Robot	Real User (solid line) Ideal User Model (dotted line)	Real User (solid line) Ideal User Model (dotted line)

Fig. 3: Legends indicating active/passive robots, attentive/distracted humans, and real user/ideal model used for Fig.4, Fig.**??**, Fig.5, and Fig.6.

Fig.4 plots, using dotted lines, the beliefs over time for the attentive (left) and distracted (right) conditions, comparing in each the passive (dotted black and gray respectively) with the active method (dotted dark orange and light orange respectively). In every situation, the active method achieves a more accurate belief (higher values for attentive on the left, when the true φ is attentive, and lower values on the right, when the true φ is distracted). In fact, passive estimation sometimes incorrectly classifies drivers as attentive when they are distracted and vice-versa.

The same figure also shows (in solid lines) results from our user study of what happens when the robot no longer interacts with an ideal model. We discuss these in the next section.

Fig.5 and Fig.6 plot the corresponding robot and

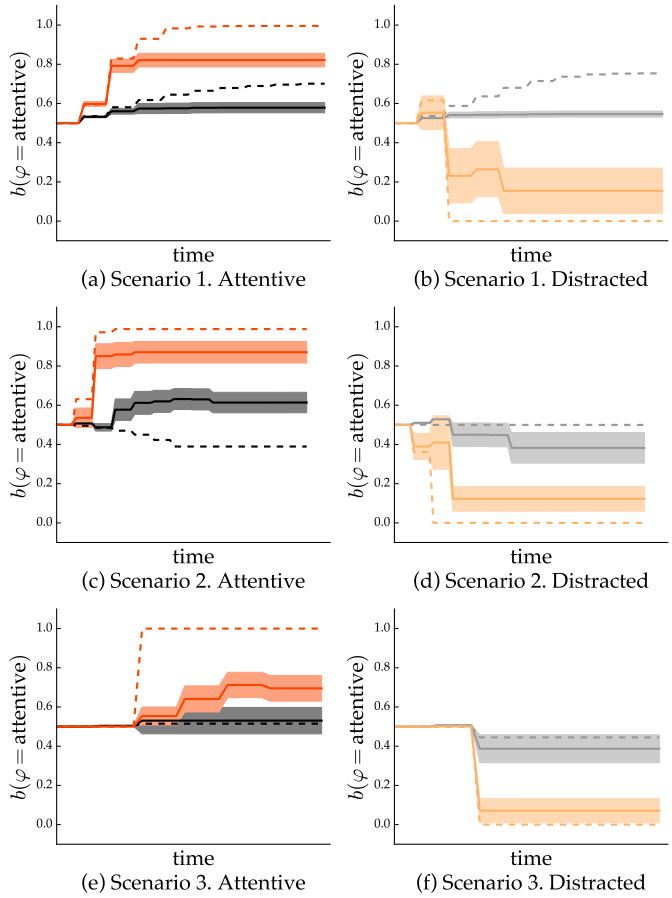


Fig. 4: The probability that the robot assigns to attentive as a function of time, for the attentive (left) and distracted (right). Each plot compares the active algorithm to passive estimation, showing that active information gathering leads to more accurate state estimation, in simulation and with real users.

human trajectories for each scenario. The important takeaway from these figures is that there tends to be a larger gap between attentive and distracted human trajectories in the active condition (orange shades) than in the passive condition (gray shades), especially in scenarios 2 and 3. It is this difference that helps the robot better estimate φ: *the robot in the active condition is purposefully choosing actions that will lead to large differences in human reactions*, in order to more easily determine the human driving style.

IV. User Study

In the previous section, we explored planning for an autonomous vehicle that actively probes a human's driving style, by braking or nudging in and expecting to cause reactions from the human driver that would be different depending on their style. We showed that active exploration does significantly better at distinguishing between attentive and distracted drivers using simulated (ideal) models of drivers. Here, we show the results of a user study that evaluates this active exploration for attentive and distracted human drivers.

A. Experimental Design

We use the same three scenarios discussed in the previous section.

Manipulated Factors. We manipulated the same two factors as in our simulation experiments: the *reward* function that the robot is optimizing (whether it is optimizing its reward through passive state estimation, or whether it is trading off with active information gathering), and the *human internal state* (whether the user is attentive or distracted). We asked our users to pay attention to the road and avoid collisions for the attentive case, and asked our users to play a game on a mobile phone during the distracted driving experiments.

Dependent Measure. We measured the probability that the robot assigned along the way to the human internal state.

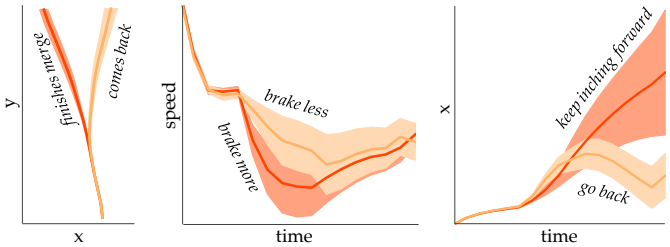
Hypothesis. *The active condition will lead to more accurate human internal state estimation, regardless of the true human internal state.*

Subject Allocation. We recruited 8 participants (2 female, 6 male) in the age range of 21-26 years old. All participants owned a valid driver license and had at least 2 years of driving experience. We ran the experiments using a 2D driving simulator with the steering input and acceleration input provided through a steering wheel and a pedals as shown in Fig.1. We used a within-subject experiment design with counterbalanced ordering of the four conditions.

B. Analysis

We ran a factorial repeated-measures ANOVA on the probability assigned to "attentive", using reward (active vs passive) and human internal state (attentive vs distracted) as factors, and time and scenario as covariates. As a manipulation check, attentive drivers had significantly higher estimated probability of "attentive" associated than distracted drivers (.66 vs .34, $F = 3080.3$, $p < .0001$). More importantly, there was a signifiant interaction effect between the factors ($F = 1444.8$, $p < .000$). We ran a post-hoc analysis with Tukey HSD corrections for multiple comparisons, which showed all four conditions to be significantly different from each other, all contrasts with $p < .0001$. In particular, the active information gathering did end up with higher probability mass on "attentive" than the passive estimation for the attentive users, and lower probability mass for the distracted user. This supports our hypothesis that our method works, and active information gathering is better at identifying the correct state.

Fig.4 compares passive (grays and blacks) and active (light and dark oranges) across scenarios and for attentive (left) and distracted (right) users. It plots the probability of attentive over time, and the shaded regions correspond to standard error. From the first column, we can see that our algorithm in all cases



(a) Scenario 1: R nudges in (b) Scenario 2: R brakes (c) Scenario 3: R inches forward

Fig. 5: Robot trajectories for each scenario in the active information gathering condition. The robot acts differently when the human is attentive (dark orange) vs. when the human is distracted (light orange) due to the trade-off with safety.

detects human's attentiveness with much higher probably than the passive information gathering technique shown in black. From the second column, we see that our algorithm places significantly lower probability on attentiveness, which is correct because those users were distracted users. These are in line with the statistical analysis, with active information gathering doing a better job estimating the true human internal state.

Fig.5 plots the robot trajectories for the active information gathering setting. Similar to Fig.4, the solid lines are the mean of robot trajectories and the shaded regions show the standard error. We plot a representative dimension of the robot trajectory (like position or speed) for attentive (dark orange) or distracted (light orange) cases. The active robot probed the user, but ended up taking different actions when the user was attentive vs. distracted in order to maintain safety. For example, in Scenario 1, the trajectories show the robot is nudging into the human's lane, but the robot decides to move back to its own lane when the human drivers are distracted (light orange) in order to stay safe. In Scenario 2, the robot brakes in front of the human, but it brakes less when the human is distracted. Finally, in Scenario 3, the robot inches forward, but again it stops when if the human is distracted, and even backs up to make space for her.

Fig.6 plots the user trajectories for both active information gathering (first row) and passive information gathering (second row) conditions. We compare the reactions of distracted (light shades) and attentive (dark shades) users. There are large differences directly observable, with user reactions tending to indeed cluster according to their internal state. These differences are much smaller in the passive case (second row, where distracted is light gray and attentive is black). For example, in Scenario 1 and 2, the attentive users (dark orange) keep a larger distance to the car that nudges in front of them or brakes in front of them, while the distracted drivers (light orange) tend to keep a smaller distance. In Scenario 3, the attentive drivers tend to slow down and do not cross the intersection, when the robot actively inches forward. None of these behaviors can be detected clearly in the passive information gathering case (second row). This is the core advantage of active information gathering: the actions are purposefully selected by the robot such that users would behave drastically differently depending on their internal state, clarifying to the robot what this state actually is.

Overall, these results support our simulation findings, that our algorithm performs better at estimating the true human internal state by leveraging purposeful information gathering actions.

V. Discussion

Summary. In this paper, we formalized the problem of active information gathering between robot and human agents, where the robot plans to actively explore and gather information about the human's internal state by leveraging the effects of its actions on the human actions. The generated strategy for the robot actively probes the human by taking actions that impact the human's action in such a way that they reveal her internal state. The robot generates strategies for interaction that we would normally need to hand-craft, like inching forward at a 4-way stop. We evaluated our method in simulation and through a user study for various autonomous driving scenarios. Our results suggest that robots are indeed able to construct a more accurate belief over the human's driving style with active exploration than with passive estimation.

Limitations and Future Work. Our work is limited in many ways. First, state estimation is not the end goal, and finding how to trade off exploration and exploitation is still a challenge. Second, our optimization is close to real-time, but higher computational efficiency is still needed. Further, our work relies on a model (reward function) of the human for each φ, which might be difficult to acquire, and might not be accurate.

Thus far, we have assumed a static φ, but in reality φ might change over time (e.g. the human adapts her preferences), or might even be influenced by the robot (e.g. a defensive driver becomes more aggressive when the robot probes her).

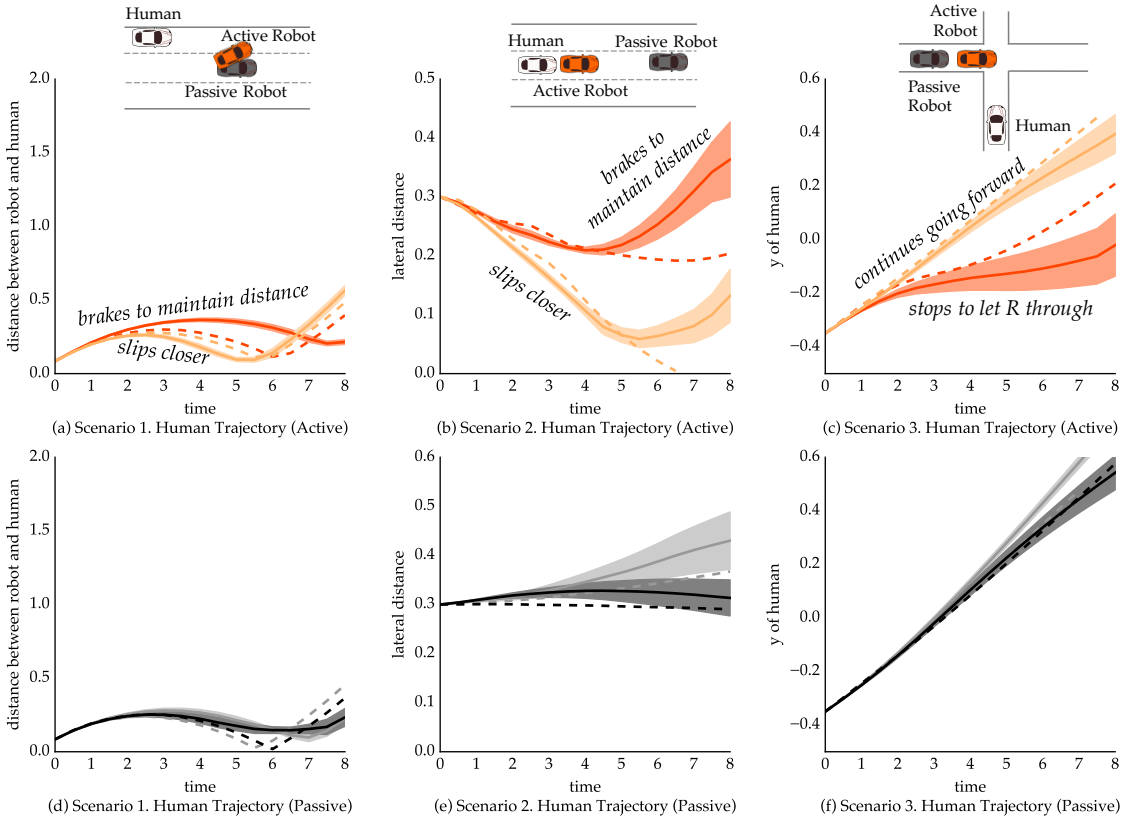
We also have not tested the users' acceptance of information gathering actions. Although these actions are useful, people might not always react positively to being probed.

Last but not least, exploring *safely* will be of crucial importance.

Conclusion. We are encouraged by the fact that robots can generate useful behavior for interaction autonomously, and are excited to explore information-gathering actions on human state further, including beyond autonomous driving scenarios.

Acknowledgements

This work was partially supported by Berkeley DeepDrive, NSF grants CCF-1139138 and CCF-1116993, ONR N00014-09-1-0230, and an NDSEG Fellowship.



(a) Scenario 1. Human Trajectory (Active)
(b) Scenario 2. Human Trajectory (Active)
(c) Scenario 3. Human Trajectory (Active)
(d) Scenario 1. Human Trajectory (Passive)
(e) Scenario 2. Human Trajectory (Passive)
(f) Scenario 3. Human Trajectory (Passive)

Fig. 6: The user trajectories for each scenario. The gap between attentive and distracted drivers' actions is clear in the active information gathering case (first row).

References

[1] Pieter Abbeel and Andrew Y Ng. Exploration and apprenticeship learning in reinforcement learning. In *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005. intelligent driver model. In *Intelligent Vehicles Symposium (IV), 2012 IEEE*, pages 1162–1167. IEEE, 2012.

[2] Galen Andrew and Jianfeng Gao. Scalable training of l 1-regularized log-linear models. In *Proceedings of the 24th international conference on Machine learning*, pages 33–40. ACM, 2007.

[3] Muhammad Awais and Dominik Henrich. Human-robot collaboration by intention recognition using probabilistic state machines. In *Robotics in Alpe-Adria-Danube Region (RAAD), 2010 IEEE 19th International Workshop on*, pages 75–80. IEEE, 2010.

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

[5] Tirthankar Bandyopadhyay, Kok Sung Won, Emilio Frazzoli, David Hsu, Wee Sun Lee, and Daniela Rus. Intention-aware motion planning. In *Algorithmic Foundations of Robotics X*, pages 475–491. Springer, 2013.

[6] Anca D Dragan and Siddhartha S Srinivasa. *Formalizing assistive teleoperation*. MIT Press, July, 2012.

[7] Alan Fern, Sriraam Natarajan, Kshitij Judah, and Prasad Tadepalli. A decision-theoretic model of assistance. In *IJCAI*, 2007.

[8] Shervin Javdani, J Andrew Bagnell, and Siddhartha Srinivasa. Shared autonomy via hindsight optimization. *arXiv preprint arXiv:1503.07619*, 2015.

[9] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1):99–134, 1998.

[10] Dana Kulic and Elizabeth A Croft. Affective state estimation for human–robot interaction. *IEEE Transactions on Robotics*, 2007.

[11] Chi-Pang Lam, Allen Y Yang, and S Shankar Sastry. An efficient algorithm for discrete-time hidden mode stochastic hybrid systems. In *Control Conference (ECC), 2015 European*. IEEE, 2015.

[12] Sergey Levine and Vladlen Koltun. Continuous inverse optimal control with locally optimal examples. *arXiv preprint arXiv:1206.4617*, 2012.

[13] Martin Liebner, Michael Baumann, Felix Klanner, and Christoph Stiller. Driver intent inference at urban intersections using the

[14] Manfred Morari, CE Garcia, JH Lee, and DM Prett. *Model predictive control*. Prentice Hall Englewood Cliffs, NJ, 1993.

[15] Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th international conference on Machine learning*, pages 663–670, 2000.

[16] Truong-Huy Dinh Nguyen, David Hsu, Wee-Sun Lee, Tze-Yun Leong, Leslie Pack Kaelbling, Tomas Lozano-Perez, and Andrew Haydn Grant. Capir: Collaborative action planning with intention recognition. *arXiv preprint arXiv:1206.5928*, 2012.

[17] Stefanos Nikolaidis, Anton Kuznetsov, David Hsu, and Siddhartha Srinivasa. Formalizing human-robot mutual adaptation via a bounded memory based model. In *Human-Robot Interaction*, March 2016.

[18] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. *Urbana*, 51:61801, 2007.

[19] Dorsa Sadigh, S. Shankar Sastry, Sanjit A. Seshia, and Anca Dragan. Planning for autonomous cars that leverages effects on human actions. In *Proceedings of the Robotics: Science and Systems Conference (RSS)*, June 2016.

[20] Tim Van Kasteren, Athanasios Noulas, Gwenn Englebienne, and Ben Kröse. Accurate activity recognition in a home setting. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 1–9. ACM, 2008.

[21] Hastagiri P Vanchinathan, Isidor Nikolic, Fabio De Bona, and Andreas Krause. Explore-exploit in top-n recommender systems via gaussian processes. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 225–232. ACM, 2014.

[22] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, pages 1433–1438, 2008.

[23] Brian D Ziebart, Nathan Ratliff, Garratt Gallagher, Christoph Mertz, Kevin Peterson, J Andrew Bagnell, Martial Hebert, Anind K Dey, and Siddhartha Srinivasa. Planning-based prediction for pedestrians. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 3931–3936. IEEE, 2009.