

Principles of Robot Autonomy II

Human-Robot Interaction



Stanford
University



Today's itinerary

- Recap (IL, IRL, pairwise comparisons)
- Game-Theoretic Views on Multi-Agent Interactions
- Partner Modeling: Active Info Gathering over Human's Intent
- Partner Modeling: Learning and Influencing Latent Intent
- Partner Modeling: Role Assignment

Today's itinerary

- Recap (IL, IRL, pairwise comparisons)
- Game-Theoretic Views on Multi-Agent Interactions
- Partner Modeling: Active Info Gathering over Human's Intent
- Partner Modeling: Learning and Influencing Latent Intent
- Partner Modeling: Role Assignment

Types of Imitation Learning

Behavioral Cloning

$$\arg \min_{\theta} \mathbb{E}_{(s, a^*) \sim P^*} L(a^*, \pi_{\theta}(s))$$

Works well when P^* is close to P_{θ}

Direct Policy Learning (via Interactive Demonstrator)

Requires Interactive Demonstrator (BC is a 1-step special case)

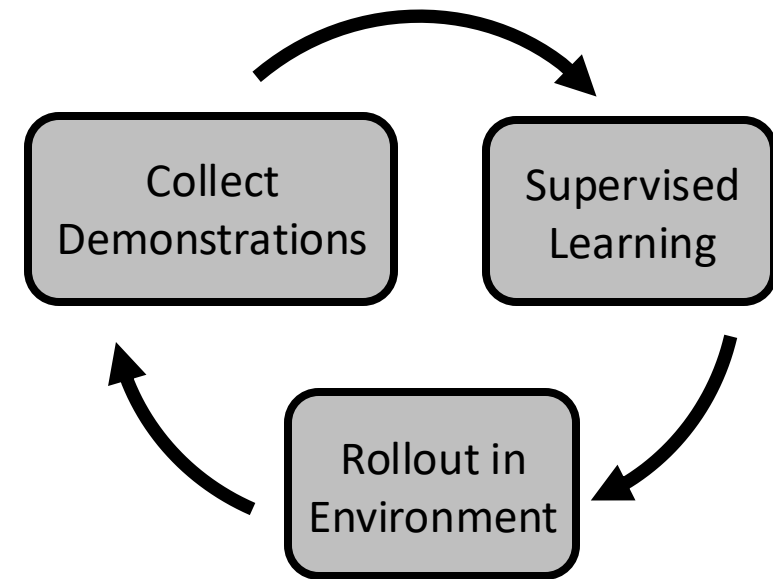
Inverse RL

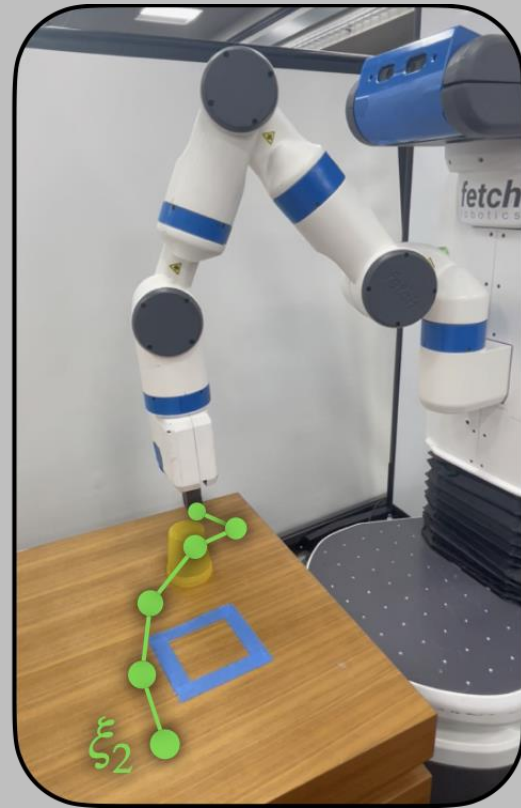
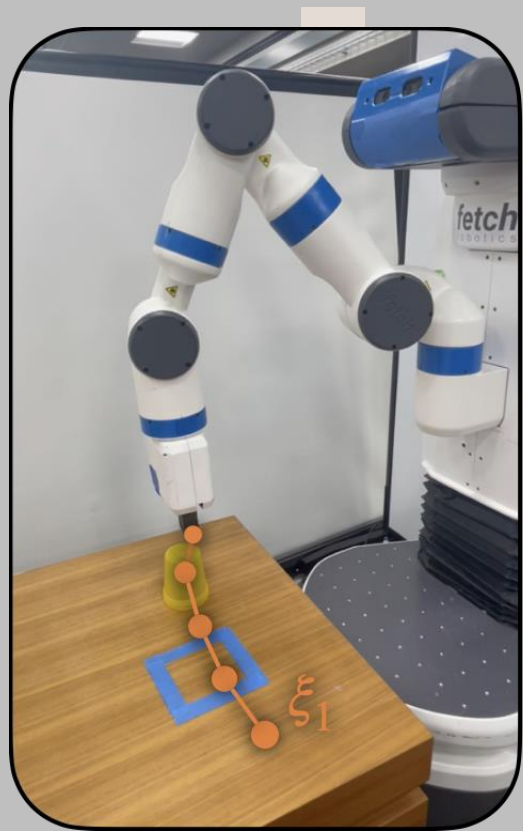
Learn r such that:

$$\pi^* = \arg \max_{\theta} \mathbb{E}_{s \sim P(s|\theta)} r(s, \pi_{\theta}(s))$$

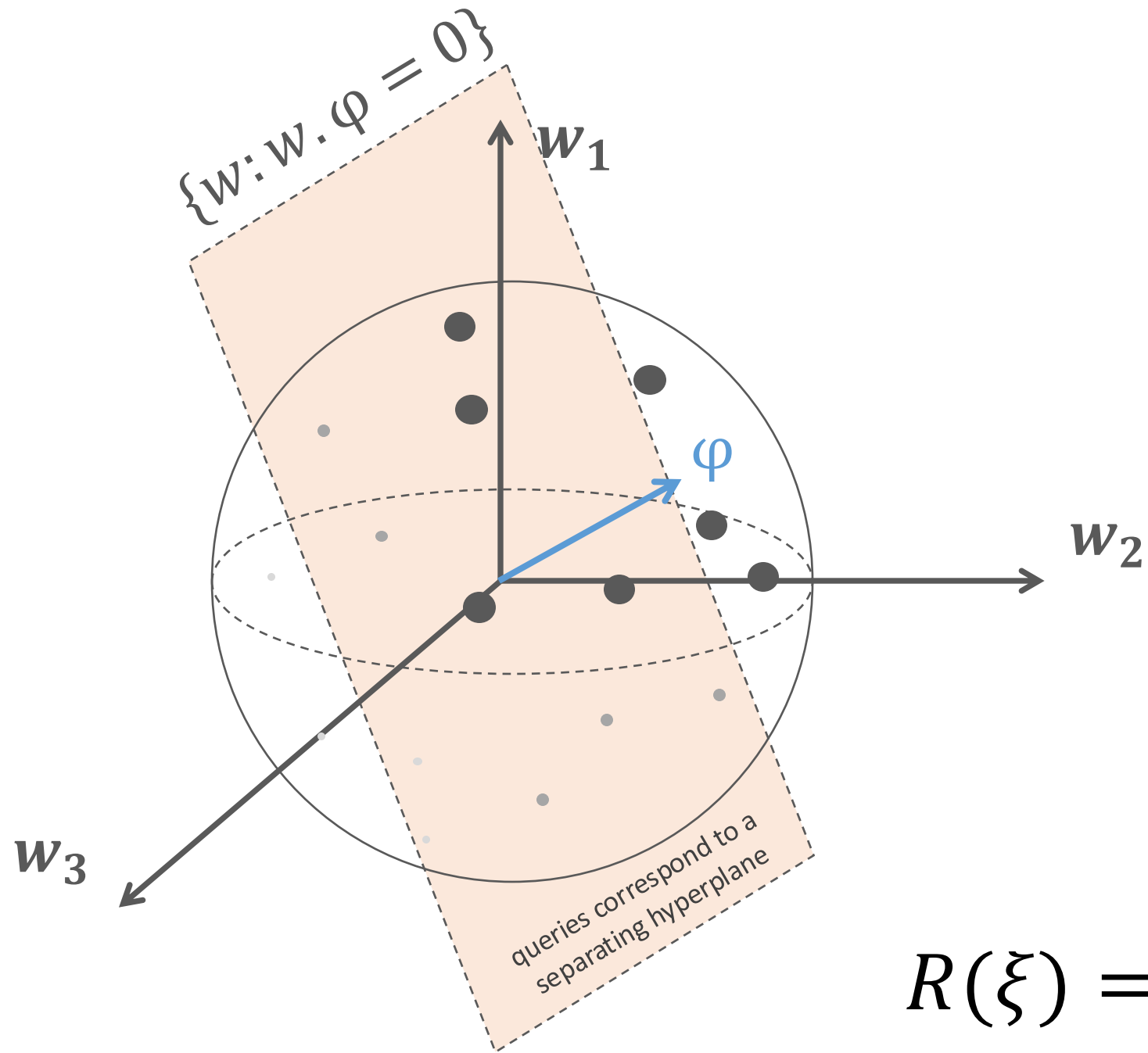
RL problem

Assume learning r is statistically easier than directly learning π^*



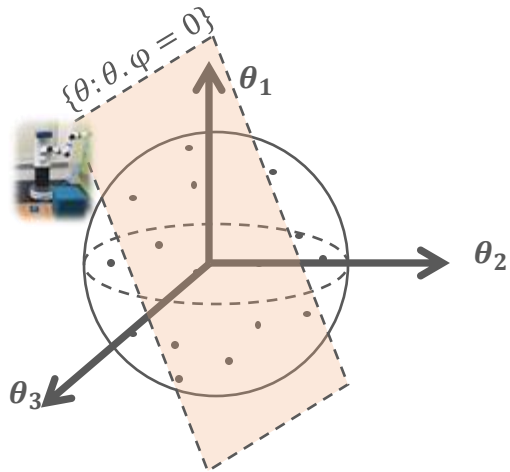


ξ_A or ξ_B ?



$$R(\xi) = w \cdot \phi(\xi)$$

Actively synthesizing queries



minimum volume removed

$$\max_{\varphi} \min\{\mathbb{E}[1 - f_{\varphi}(w)], \mathbb{E}[1 - f_{-\varphi}(w)]\}$$

Subject to $\varphi \in \mathbb{F}$

$$\mathbb{F} = \{\varphi: \varphi = \Phi(\xi_A) - \Phi(\xi_B), \xi_A, \xi_B \in \Xi\}$$

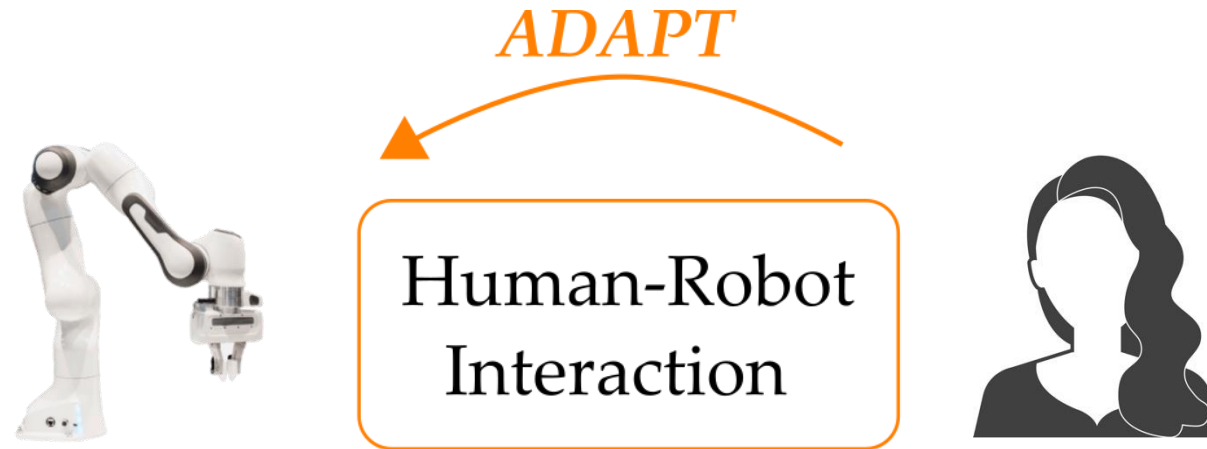
Human update function $f_{\varphi}(w) = \min(1, \exp(I_t w^T \varphi))$

- [Sadigh et al. RSS17]
- [Biyik et al. CoRL18]
- [Biyik et al. CDC19]
- [Palan et al. RSS19]
- [Biyik et al. CoRL19]
- [Basu et al. IROS19]
- [Biyik et al. RSS20]
- [Myers et al. CoRL21]
- [Myers et al. ICRA22]

Today's itinerary

- Recap (IL, IRL, pairwise comparisons)
- Game-Theoretic Views on Multi-Agent Interactions
- Partner Modeling: Active Info Gathering over Human's Intent
- Partner Modeling: Learning and Influencing Latent Intent
- Partner Modeling: Role Assignment

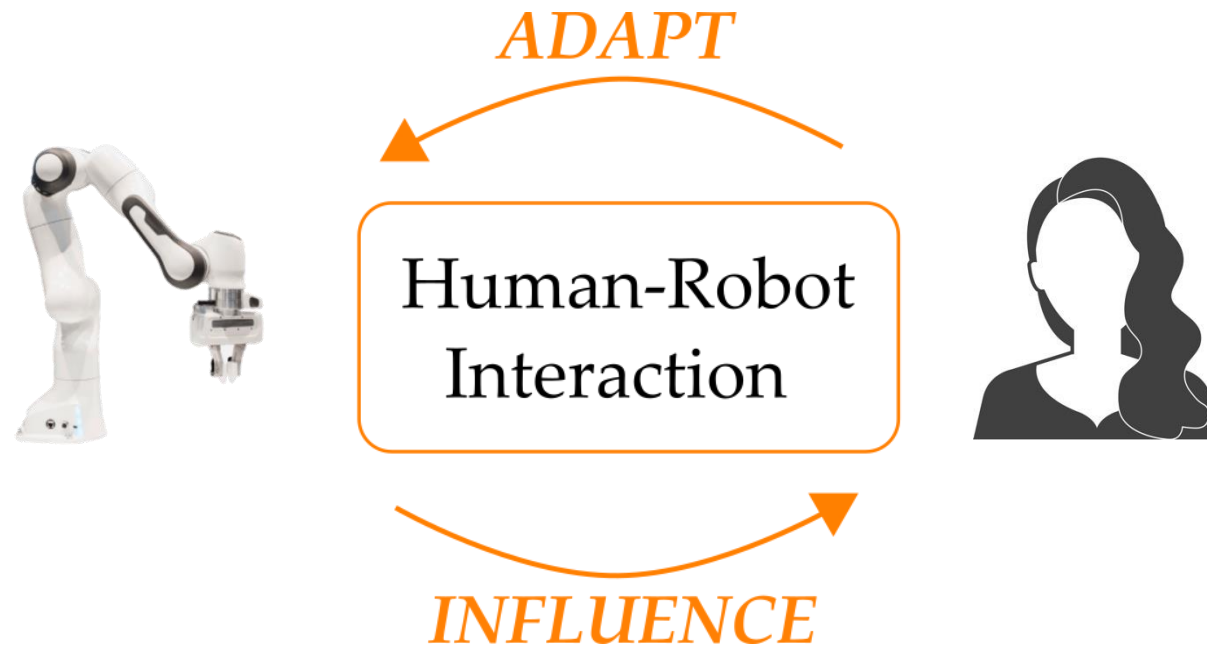
Learning from Humans



Existing research explores how robots *adapt* to humans

- Imitation learning
- Learning from demonstrations

Influencing Humans



Far less studies how robots *influence* humans

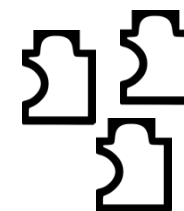
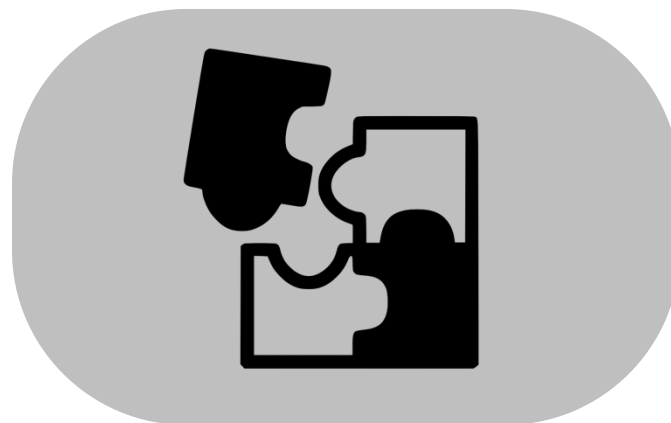
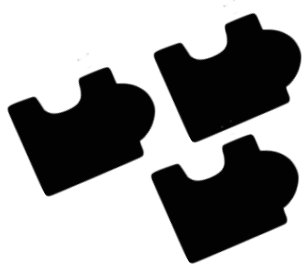




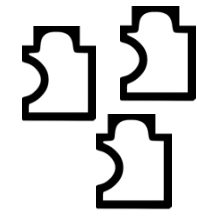
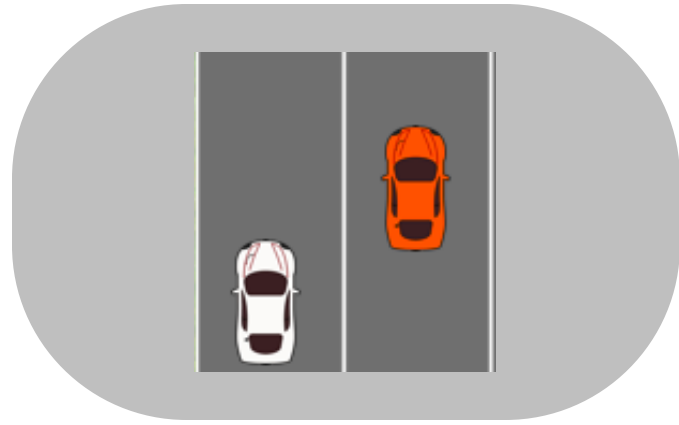
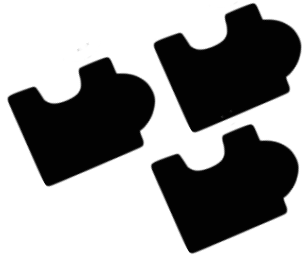
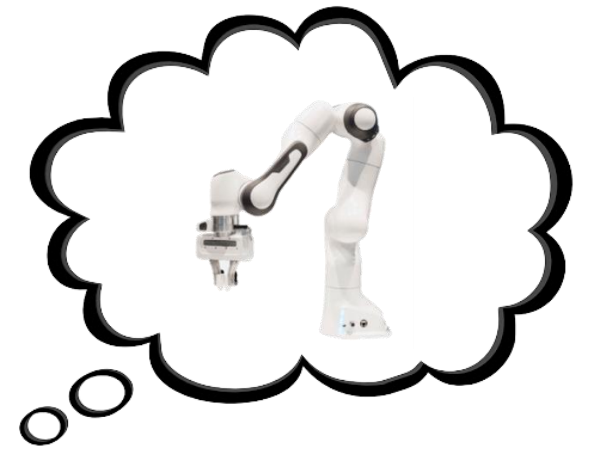
Nth order Theory of Mind



Nth order Theory of Mind

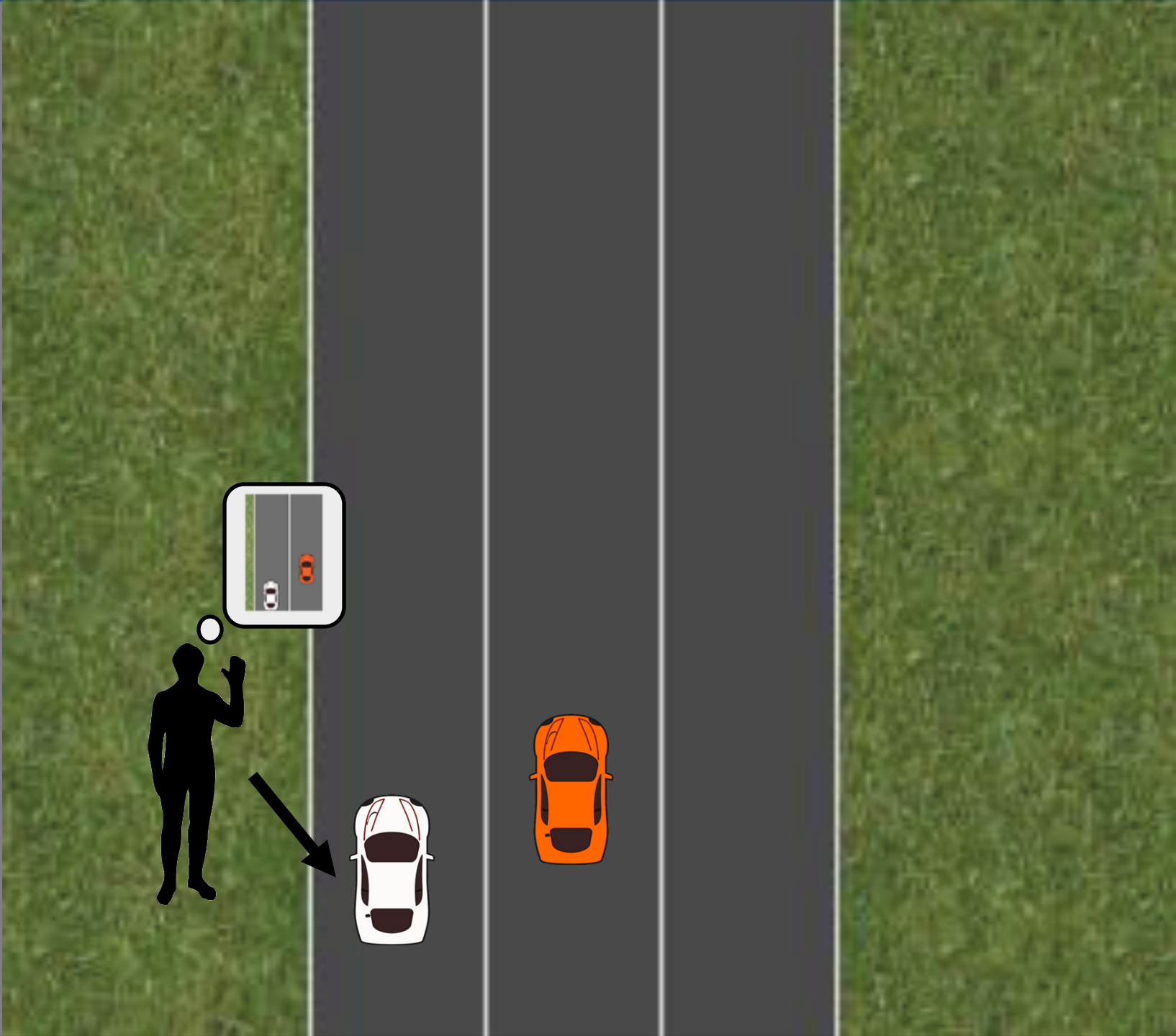


Nth order Theory of Mind







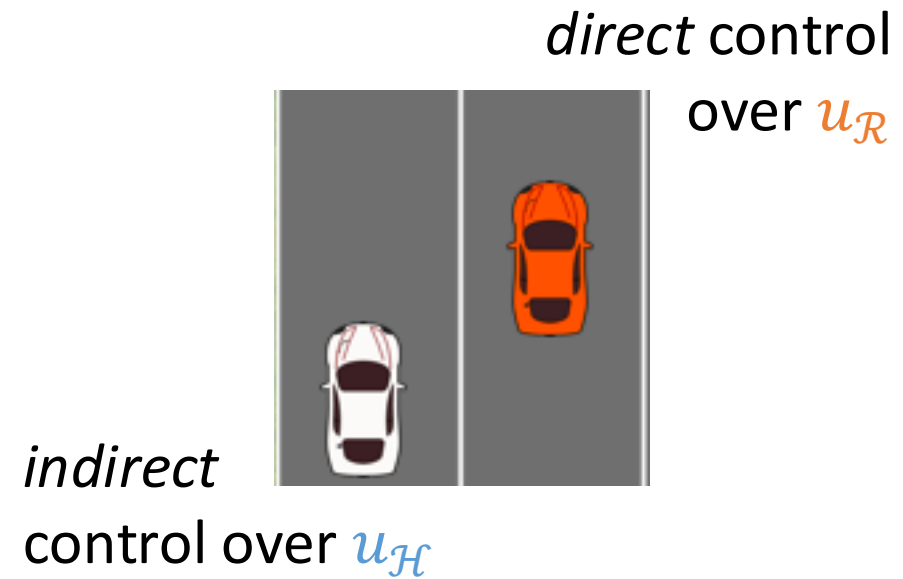


An autonomous car's actions will *affect* the actions of other drivers.



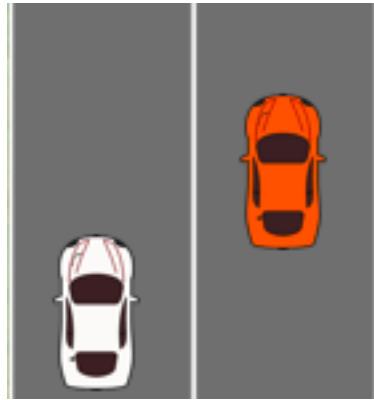


Interaction as a Dynamical System



Interaction as a Dynamical System

$$u_{\mathcal{R}}^* = \operatorname{argmax}_{u_{\mathcal{R}}} R_{\mathcal{R}}(x, u_{\mathcal{R}}, u_{\mathcal{H}}^*(x, u_{\mathcal{R}}))$$



Find optimal actions for the robot while accounting for the human response $u_{\mathcal{H}}^*$.

Model $u_{\mathcal{H}}^*$ as optimizing the human reward function $R_{\mathcal{H}}$.

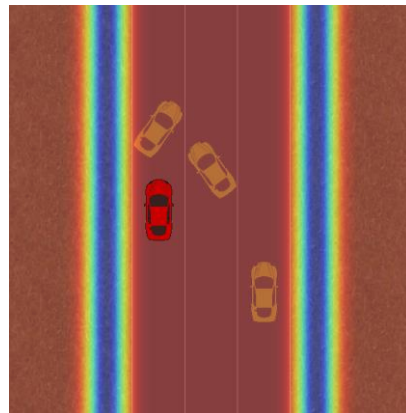
$$u_{\mathcal{H}}^*(x, u_{\mathcal{R}}) \approx \operatorname{argmax}_{u_{\mathcal{H}}} R_{\mathcal{H}}(x, u_{\mathcal{R}}, u_{\mathcal{H}})$$

Learning Driver Models

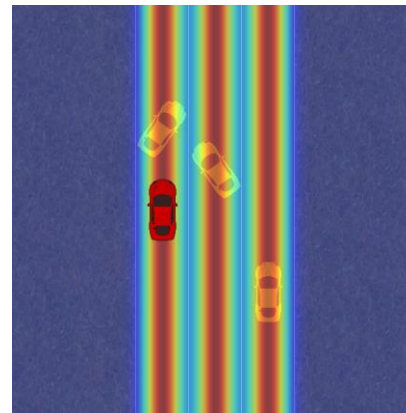
Learn Human's reward function based on Inverse Reinforcement Learning:

$$P(u_{\mathcal{H}}|x, w) = \frac{\exp(R_{\mathcal{H}}(x, u_{\mathcal{R}}, u_{\mathcal{H}}))}{\int \exp(R_{\mathcal{H}}(x, u_{\mathcal{R}}, \tilde{u}_{\mathcal{H}})) d \tilde{u}_{\mathcal{H}}}$$

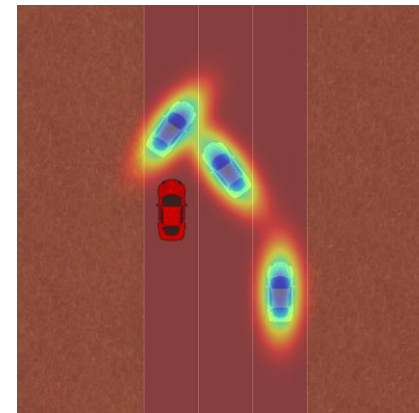
$$R_{\mathcal{H}}(x, u_{\mathcal{R}}, u_{\mathcal{H}}) = w^{\top} \phi(x, u_{\mathcal{R}}, u_{\mathcal{H}})$$



Features for the boundaries of the road.



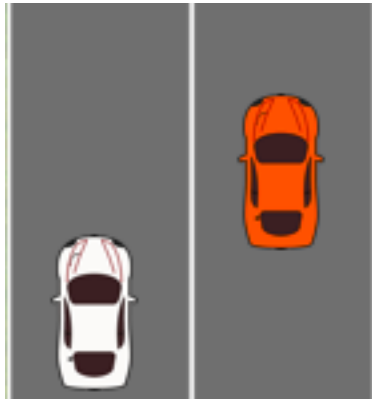
Features for staying inside the lanes.



Features for avoiding other vehicles.

Interaction as a Dynamical System

$$u_{\mathcal{R}}^* = \operatorname{argmax}_{u_{\mathcal{R}}} R_{\mathcal{R}}(x, u_{\mathcal{R}}, u_{\mathcal{H}}^*(x, u_{\mathcal{R}}))$$



Find optimal actions for the robot while accounting for the human response $u_{\mathcal{H}}^*$.

Model $u_{\mathcal{H}}^*$ as optimizing the human reward function $R_{\mathcal{H}}$.

$$u_{\mathcal{H}}^*(x, u_{\mathcal{R}}) \approx \operatorname{argmax}_{u_{\mathcal{H}}} R_{\mathcal{H}}(x, u_{\mathcal{R}}, u_{\mathcal{H}})$$

Approximations for Tractability

- Receding Horizon Control:

Plan for short time horizon, replan at every step.

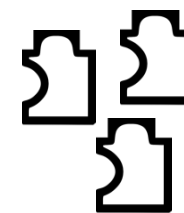
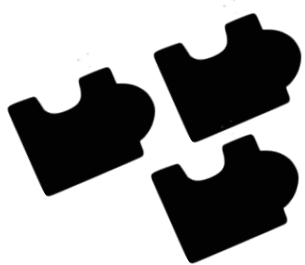
- Model the problem as a *Stackelberg game*.

Give the human full access to $u_{\mathcal{R}}$ for the short time horizon.

Nth order Theory of Mind



Nth order Theory of Mind



Approximations for Tractability

- Receding Horizon Control:

Plan for short time horizon, replan at every step.

- Model the problem as a *Stackelberg game*.

Give the human full access to $u_{\mathcal{R}}$ for the short time horizon.

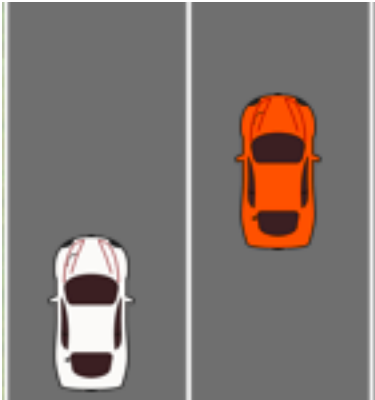
$$u_{\mathcal{H}}^*(x, u_{\mathcal{R}}) = \operatorname{argmax}_{u_{\mathcal{H}}} R_{\mathcal{H}}(x, u_{\mathcal{R}}, u_{\mathcal{H}})$$

- Assume deterministic human model.

Solution of Nested Optimization

$$u_{\mathcal{R}}^* = \operatorname{argmax}_{u_{\mathcal{R}}} R_{\mathcal{R}}(x, u_{\mathcal{R}}, u_{\mathcal{H}}^*(x, u_{\mathcal{R}}))$$

$$R_{\mathcal{R}}(x, u_{\mathcal{R}}, u_{\mathcal{H}}) = \sum_{t=1}^N r_{\mathcal{R}}(x^t, u_{\mathcal{R}}^t, u_{\mathcal{H}}^t)$$



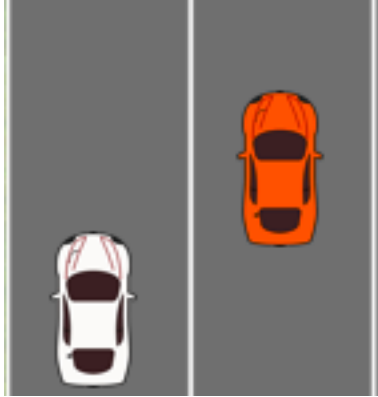
Gradient-Based Method (Quasi-Newton):

$$\left\{ \begin{array}{l} R_{\mathcal{R}}(x, u_{\mathcal{R}}, u_{\mathcal{H}}^*) \\ \frac{\partial R_{\mathcal{R}}}{\partial u_{\mathcal{R}}} = \frac{\partial R_{\mathcal{R}}}{\partial u_{\mathcal{H}}} \frac{\partial u_{\mathcal{H}}^*}{\partial u_{\mathcal{R}}} + \frac{\partial R_{\mathcal{R}}}{\partial u_{\mathcal{R}}} \end{array} \right.$$

$$u_{\mathcal{H}}^*(x, u_{\mathcal{R}}) \approx \operatorname{argmax}_{u_{\mathcal{H}}} R_{\mathcal{H}}(x, u_{\mathcal{R}}, u_{\mathcal{H}})$$

$$R_{\mathcal{H}}(x, u_{\mathcal{R}}, u_{\mathcal{H}}) = \sum_{t=1}^N r_{\mathcal{H}}(x^t, u_{\mathcal{R}}^t, u_{\mathcal{H}}^t)$$

Solution of Nested Optimization



Given $R_{\mathcal{H}}$ is:

- smooth,
 - its minimum is attained,
- for an *unconstrained optimization*, the partial $\frac{\partial R_{\mathcal{H}}}{\partial u_{\mathcal{H}}}$ at the optimum $u_{\mathcal{H}}^*$ evaluates to zero.

Quasi-Newton method:

$$\frac{\partial R_{\mathcal{R}}}{\partial u_{\mathcal{R}}} = \frac{\partial R_{\mathcal{R}}}{\partial u_{\mathcal{H}}} \cdot \frac{\partial u_{\mathcal{H}}^*}{\partial u_{\mathcal{R}}} + \frac{\partial R_{\mathcal{R}}}{\partial u_{\mathcal{R}}}$$

$$\frac{\partial R_{\mathcal{H}}}{\partial u_{\mathcal{H}}}(x, u_{\mathcal{R}}, u_{\mathcal{H}}^*(x, u_{\mathcal{R}})) = 0$$

$$\frac{\partial^2 R_{\mathcal{H}}}{\partial u_{\mathcal{H}}^2} \cdot \frac{\partial u_{\mathcal{H}}^*}{\partial u_{\mathcal{R}}} + \frac{\partial^2 R_{\mathcal{H}}}{\partial u_{\mathcal{H}} \partial u_{\mathcal{R}}} \cdot \frac{\partial u_{\mathcal{R}}}{\partial u_{\mathcal{R}}} = 0$$

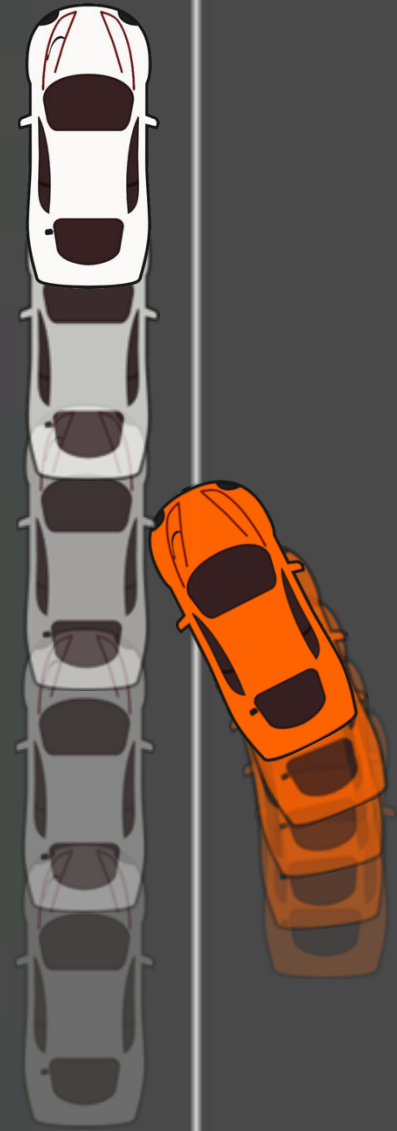
Implication: Efficiency

Human

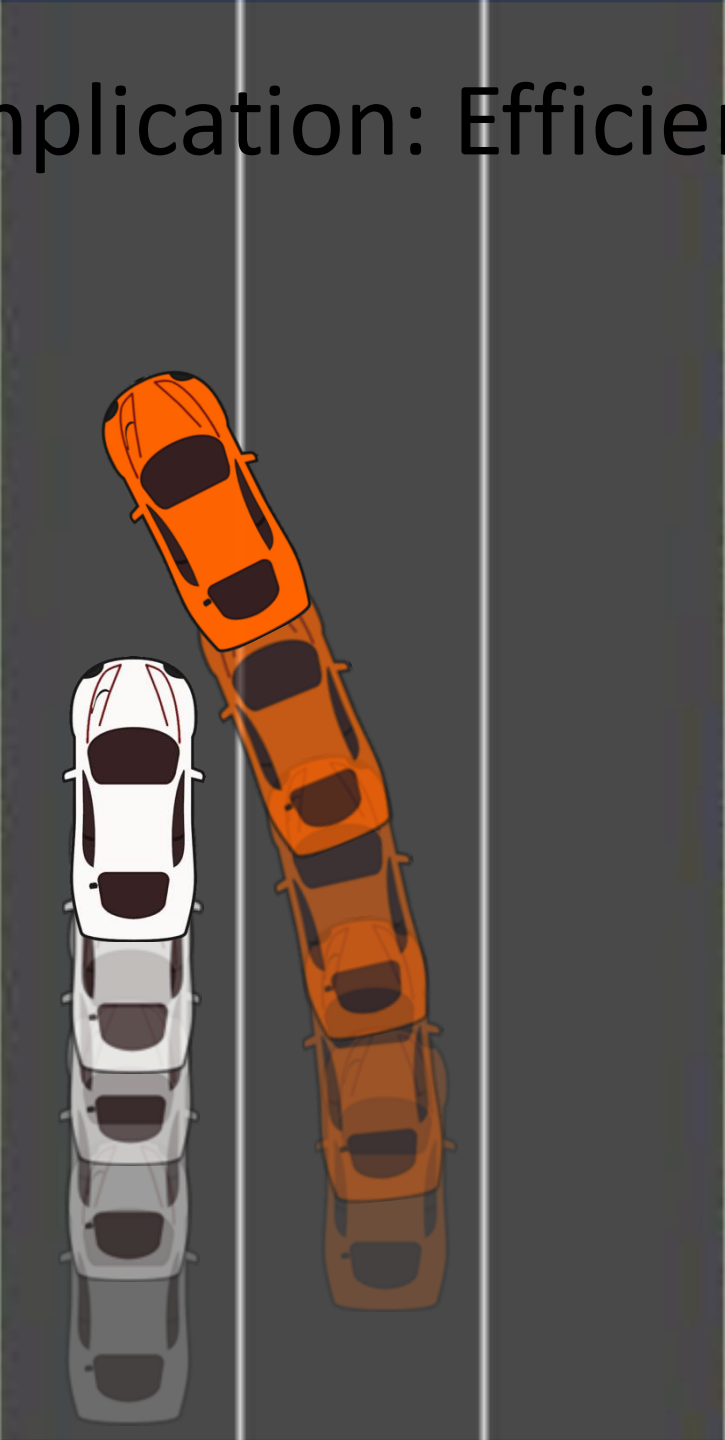


Robot

Implication: Efficiency



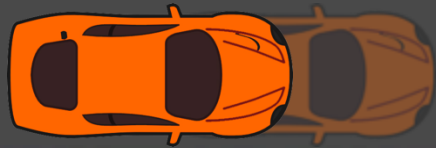
Implication: Efficiency



Implication: Coordination

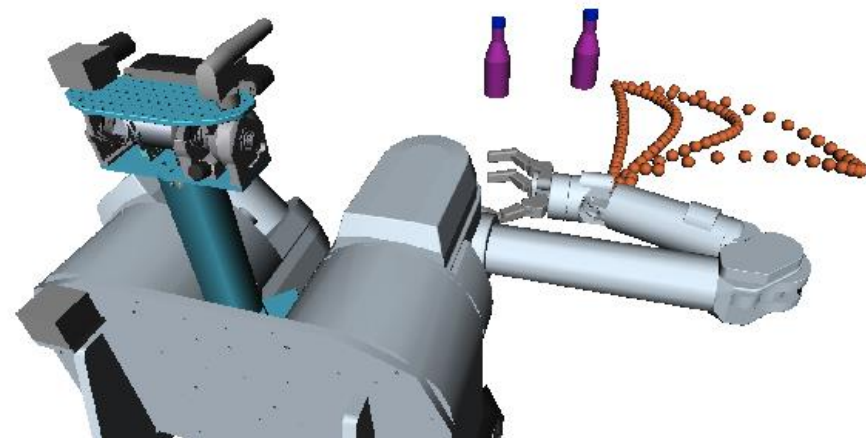
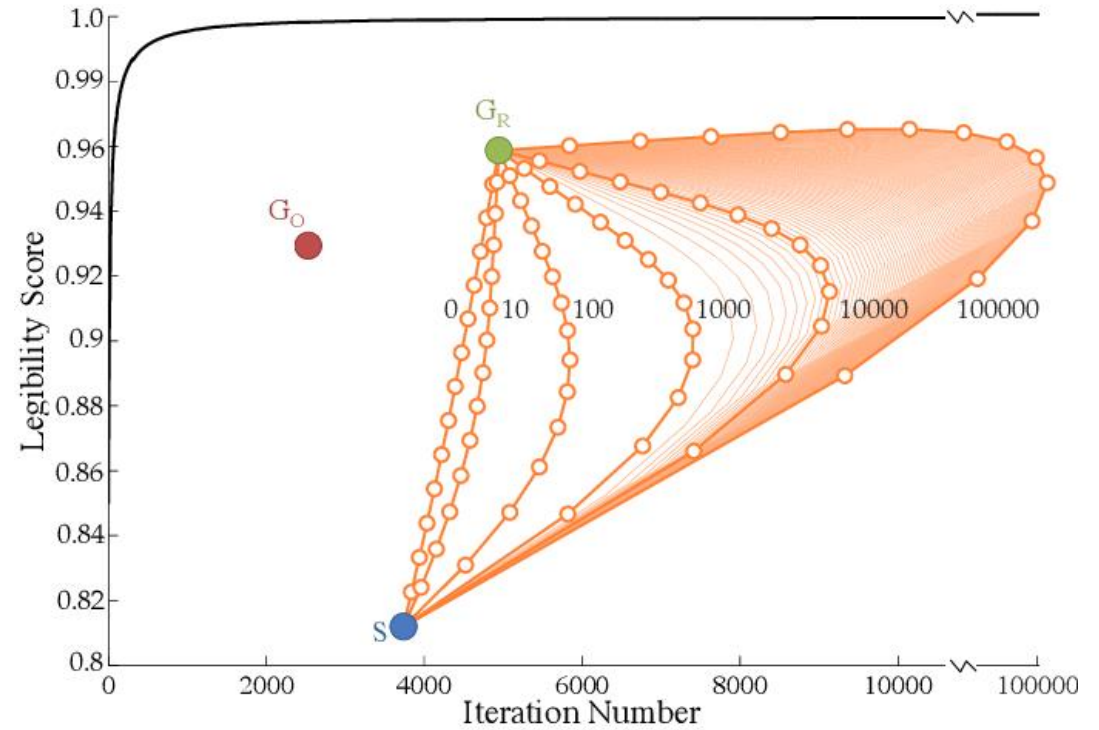


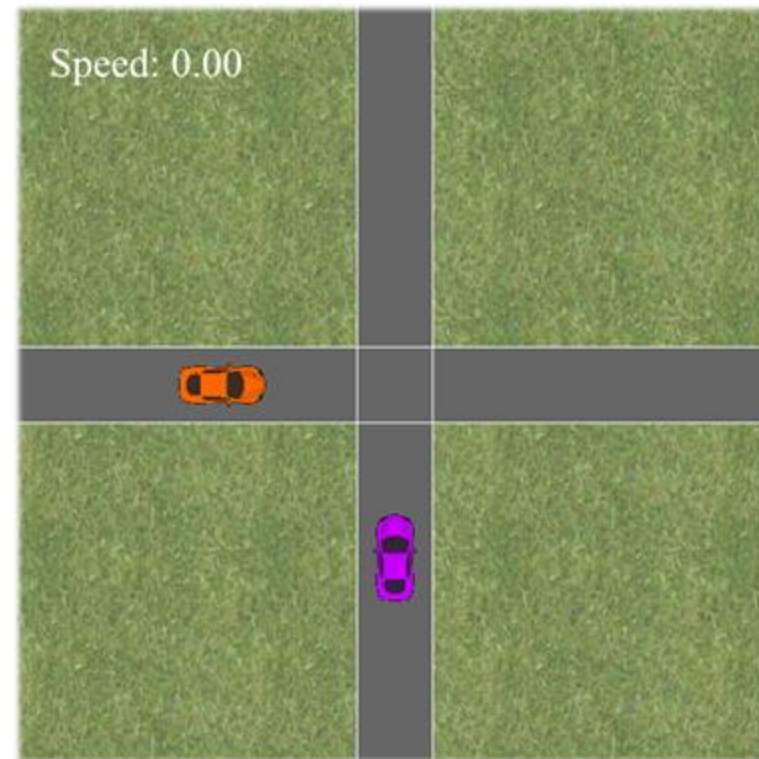
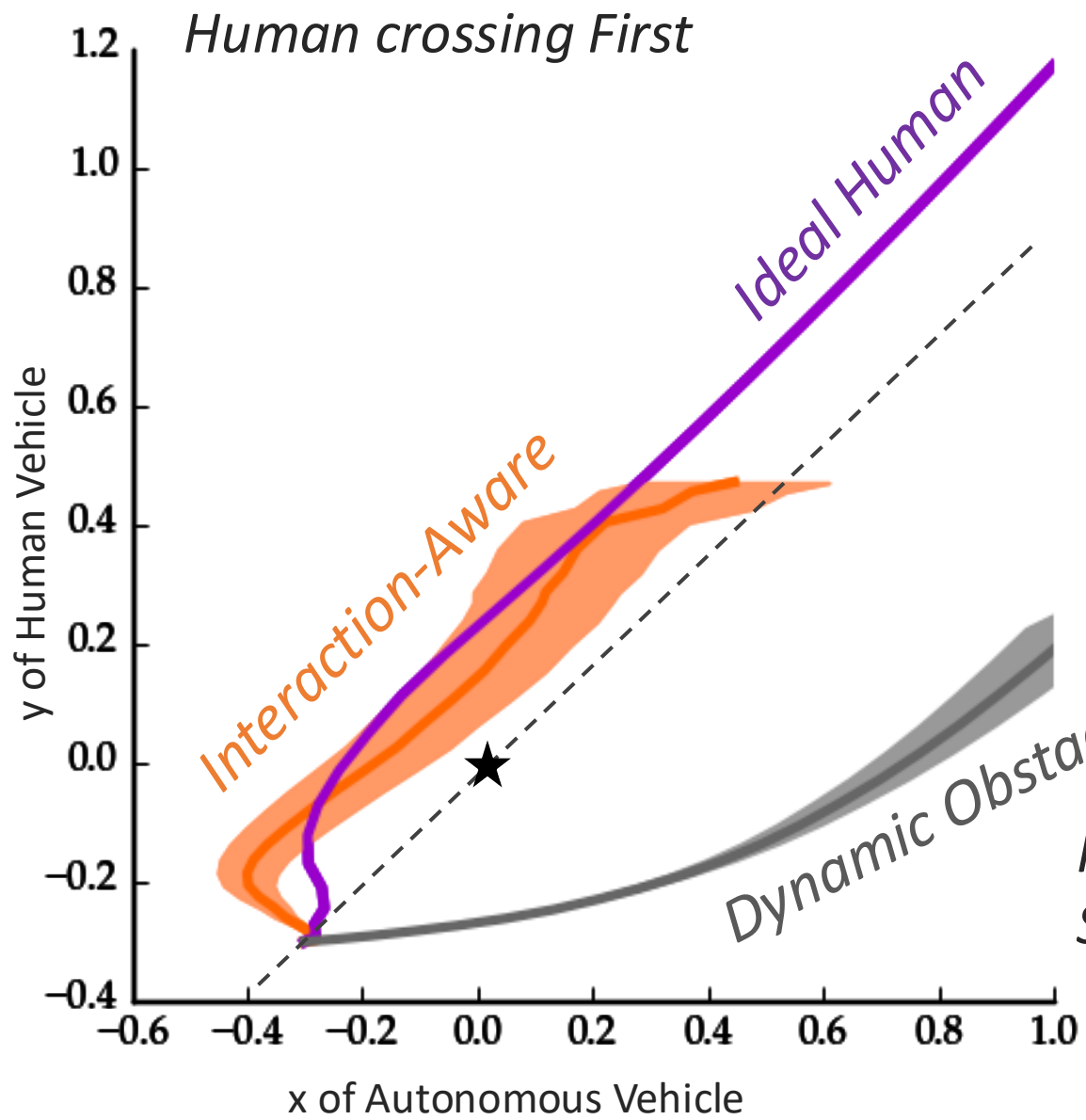
Implication: Coordination



Legible Motion

Using robot motion to coordinate with the human better about the robot's goal







$$p(u_{\mathcal{H}}|x) \propto \exp(R_{\mathcal{H}}(x, u_{\mathcal{H}}))$$





2015/02/06 23:10:05

We can't rely on a
single driver model.

We need to *differentiate*
between different drivers.



$$p(u_{\mathcal{H}} | x, \theta) \propto \exp(R_{\mathcal{H}}(x, u_{\mathcal{H}}, \theta))$$



$$b_{t+1}(\theta) \propto b_t(\theta) \cdot p(u_{\mathcal{H}} | x_t, \theta)$$



$$p(u_{\mathcal{H}} | x, \theta) \propto \exp(R_{\mathcal{H}}(x, u_{\mathcal{H}}, \theta))$$



$$b_{t+1}(\theta) \propto b_t(\theta) \cdot p(u_{\mathcal{H}} | x_t, \theta)$$

$$u_{\mathcal{R}} = \operatorname{argmax}_{u_{\mathcal{R}}} R_{\mathcal{R}}$$



Drivers *respond* to
actions of other cars.

...We have an opportunity to
actively gather information.



$$p(u_{\mathcal{H}} | x, \theta, u_{\mathcal{R}}) \propto \exp(R_{\mathcal{H}}(x, u_{\mathcal{H}}, \theta, u_{\mathcal{R}}))$$



$$b_{t+1}(\theta) \propto b_t(\theta) \cdot p(u_{\mathcal{H}} | x_t, \theta, u_{\mathcal{R}})$$

Info Gathering

$$R_{\mathcal{R}}(x, u_{\mathcal{H}}, \theta, u_{\mathcal{R}}) = \underbrace{\text{IH}(b_t) - \text{IH}(b_{t+1})}_{\text{Info Gathering}} + \underbrace{\lambda \cdot R_{\text{goal}}(x, u_{\mathcal{H}}, \theta, u_{\mathcal{R}})}_{\text{Goal}}$$

Goal





$$p(u_{\mathcal{H}} | x, \theta, u_{\mathcal{R}}) \propto \exp(R_{\mathcal{H}}(x, u_{\mathcal{H}}, \theta, u_{\mathcal{R}}))$$



$$b_{t+1}(\theta) \propto b_t(\theta) \cdot p(u_{\mathcal{H}} | x_t, \theta, u_{\mathcal{R}})$$

Info Gathering

$$R_{\mathcal{R}}(x, u_{\mathcal{H}}, \theta, u_{\mathcal{R}}) = \underbrace{\mathbb{H}(b_t) - \mathbb{H}(b_{t+1})}_{\text{Info Gathering}} + \underbrace{\lambda \cdot R_{goal}(x, u_{\mathcal{H}}, \theta, u_{\mathcal{R}})}_{\text{Goal}}$$

Goal

$$u_{\mathcal{R}} = \operatorname{argmax}_{u_{\mathcal{R}}} \mathbb{E}_{\theta} [R_{\mathcal{R}}]$$





$$p(u_{\mathcal{H}} | x, \theta, u_{\mathcal{R}}) \propto \exp(R_{\mathcal{H}}(x, u_{\mathcal{H}}, \theta, u_{\mathcal{R}}))$$



$$b_{t+1}(\theta) \propto b_t(\theta) \cdot p(u_{\mathcal{H}} | x_t, \theta, u_{\mathcal{R}})$$

Info Gathering

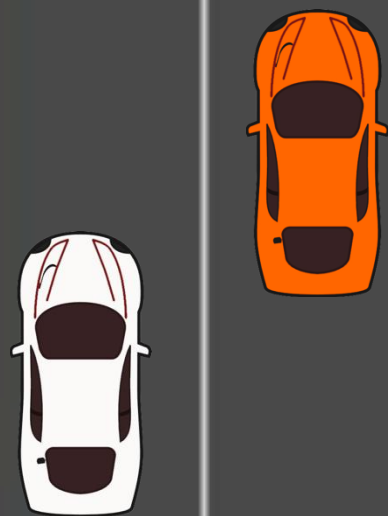
$$R_{\mathcal{R}}(x, u_{\mathcal{H}}, \theta, u_{\mathcal{R}}) = \underbrace{\mathbb{H}(b_t) - \mathbb{H}(b_{t+1})}_{\text{Info Gathering}} + \underbrace{\lambda \cdot R_{goal}(x, u_{\mathcal{H}}, \theta, u_{\mathcal{R}})}_{\text{Goal}}$$

Goal

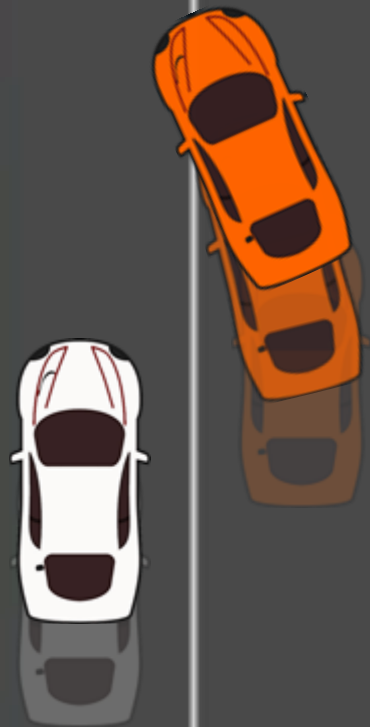
$$u_{\mathcal{R}} = \operatorname{argmax}_{u_{\mathcal{R}}} \mathbb{E}_{\theta} [R_{\mathcal{R}}]$$



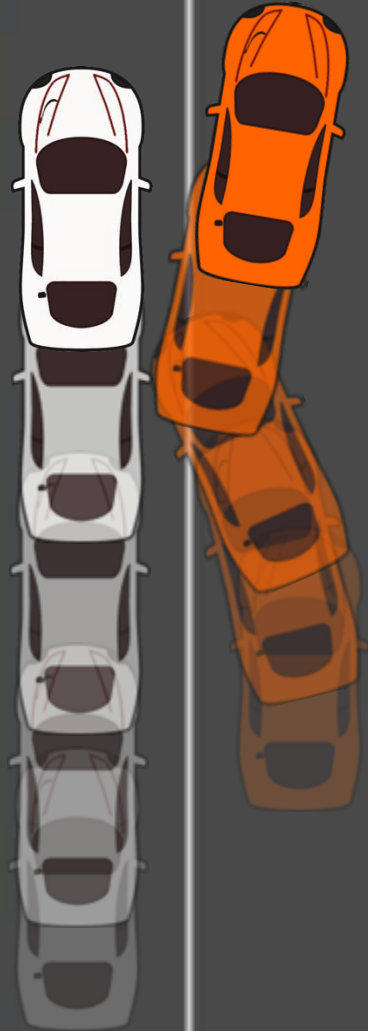
Nudging in for Active Info Gathering



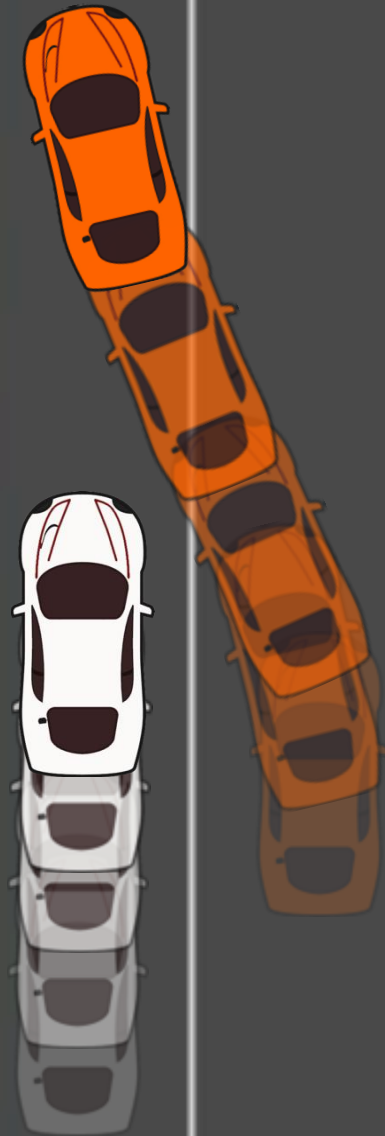
Nudging in for Active Info Gathering

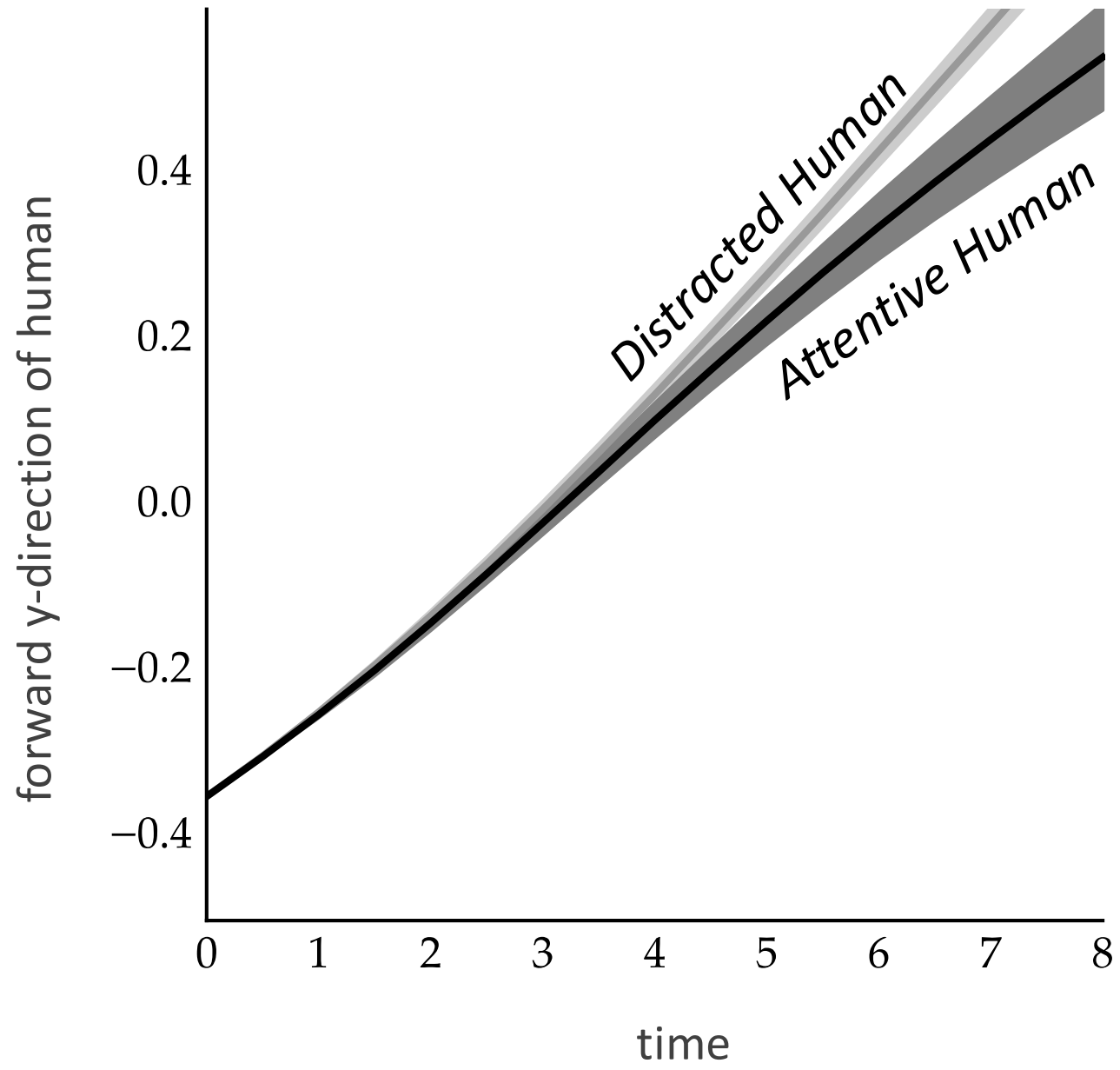


Distracted Human

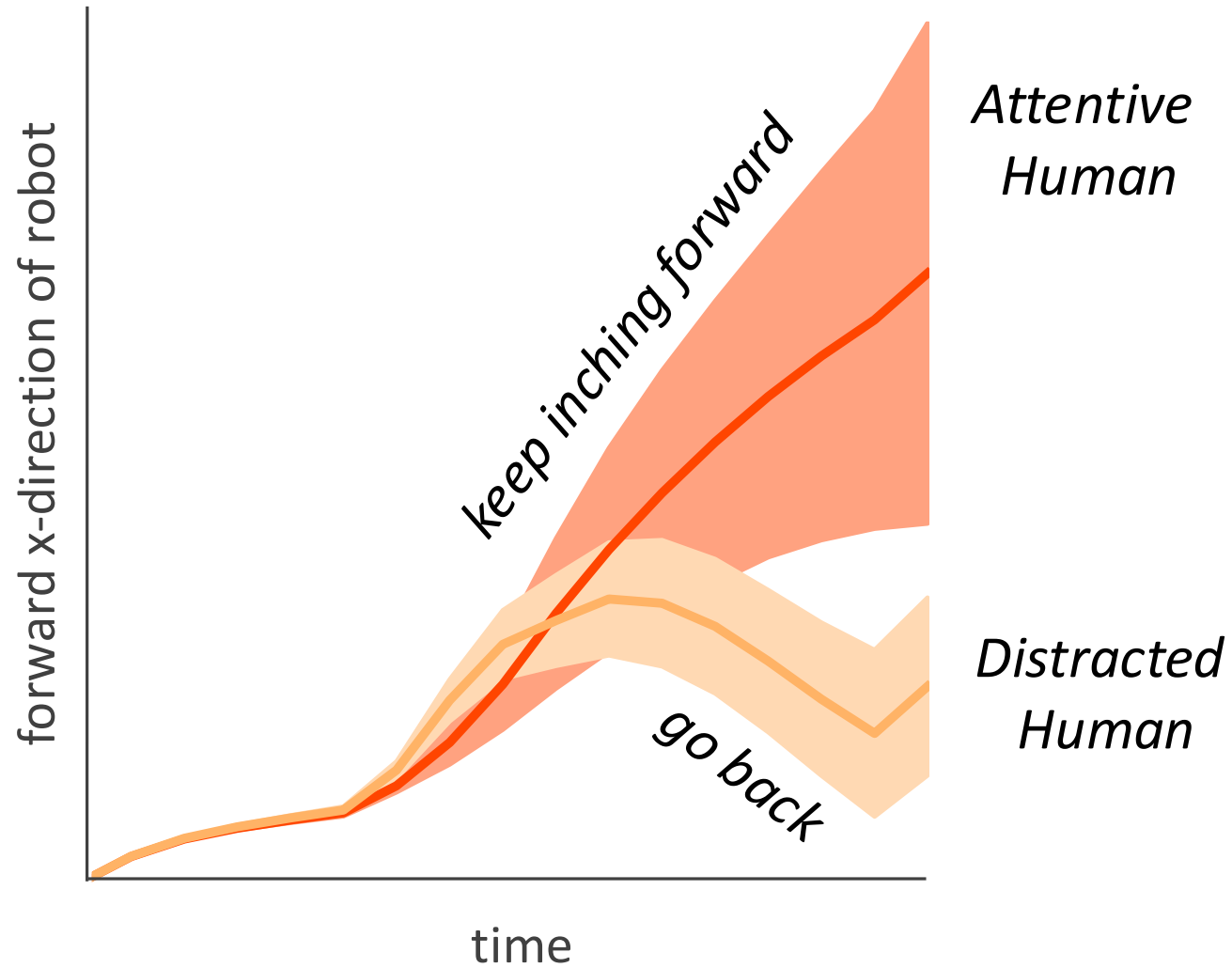


Attentive Human

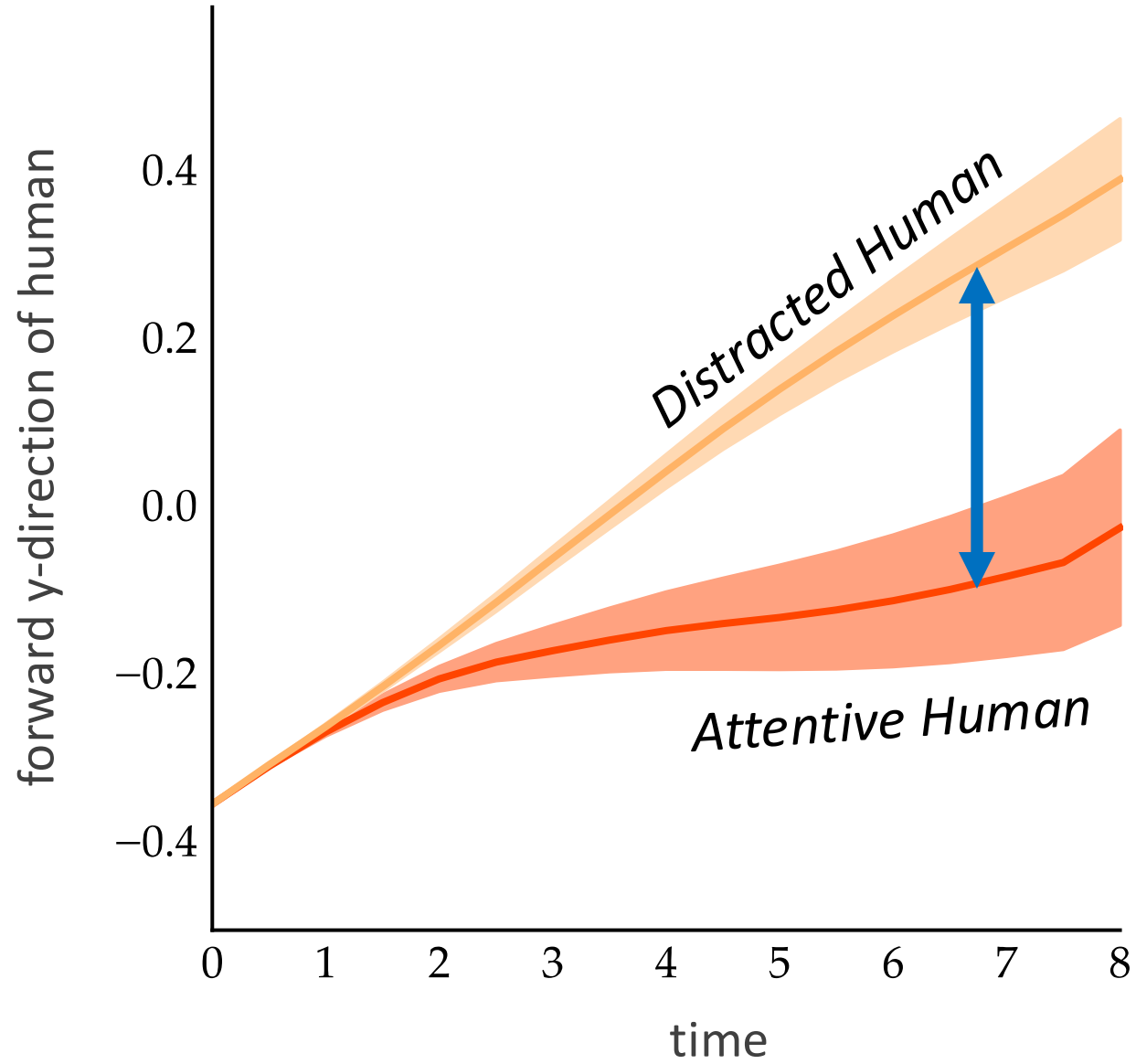




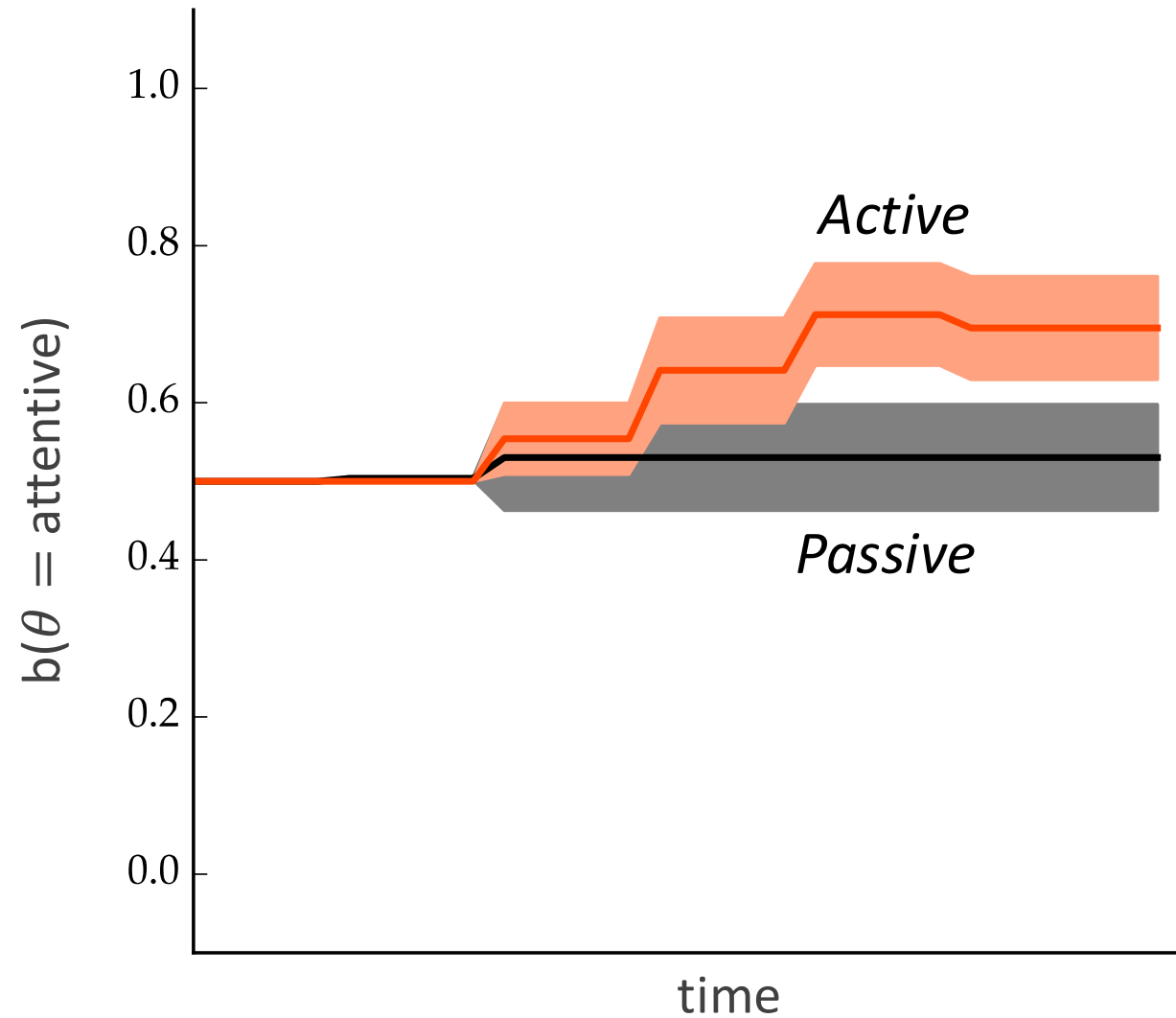
Robot Active Info Gathering



Human Responses



Belief over Driving Style: Active vs Passive



Key Idea:

Robot's actions *affect* human's actions. We want to *leverage* these effects for better safety and efficiency and better estimation.

Today's itinerary

- Recap (IL, IRL, pairwise comparisons)
- Game-Theoretic Views on Multi-Agent Interactions
- Partner Modeling: Active Info Gathering over Human's Intent
- Partner Modeling: Learning and Influencing Latent Intent
- Partner Modeling: Role Assignment



$$p(u_{\mathcal{H}} | x, \theta, u_{\mathcal{R}}) \propto \exp(R_{\mathcal{H}}(x, u_{\mathcal{H}}, \theta, u_{\mathcal{R}}))$$



$$b_{t+1}(\theta) \propto b_t(\theta) \cdot p(u_{\mathcal{H}} | x_t, \theta, u_{\mathcal{R}})$$

Info Gathering

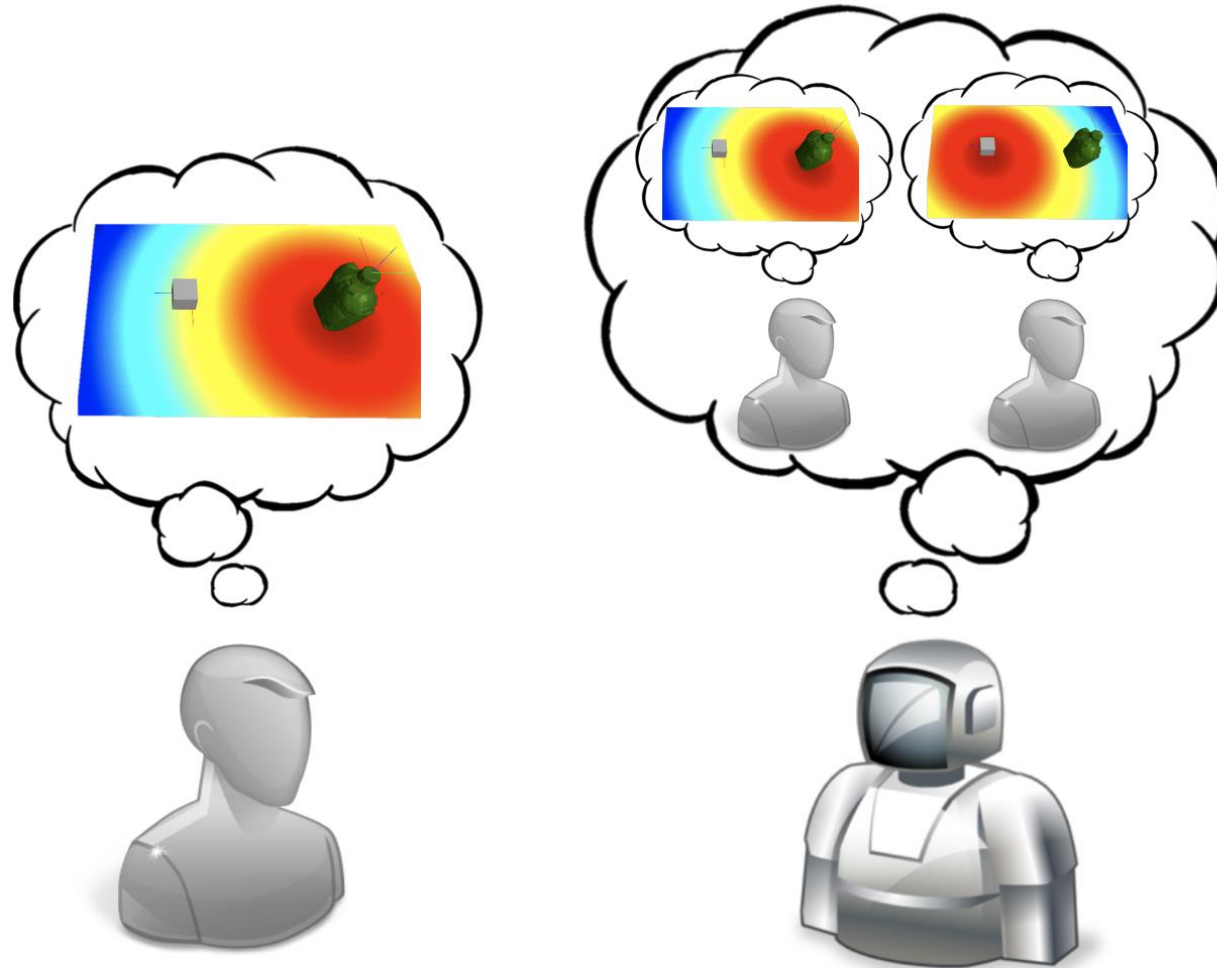
$$R_{\mathcal{R}}(x, u_{\mathcal{H}}, \theta, u_{\mathcal{R}}) = \underbrace{\mathbb{H}(b_t) - \mathbb{H}(b_{t+1})}_{\text{Info Gathering}} + \underbrace{\lambda \cdot R_{goal}(x, u_{\mathcal{H}}, \theta, u_{\mathcal{R}})}_{\text{Goal}}$$

Goal

$$u_{\mathcal{R}} = \operatorname{argmax}_{u_{\mathcal{R}}} \mathbb{E}_{\theta} [R_{\mathcal{R}}]$$



Modeling Intent Inference using POMDPs



POMDP Formulation

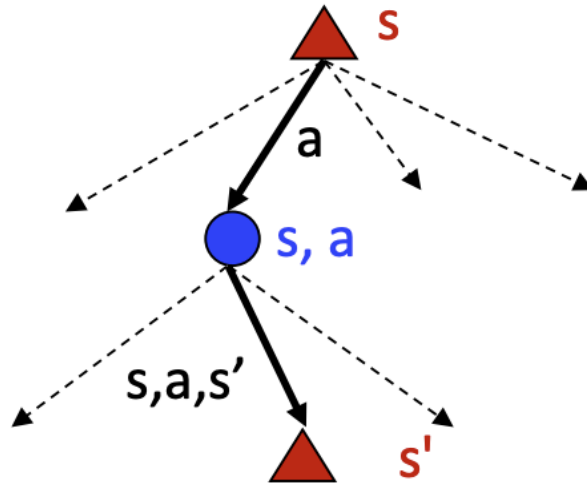
MDPs have:

States S

Actions A

Transition Function $P(s'|s, a)$

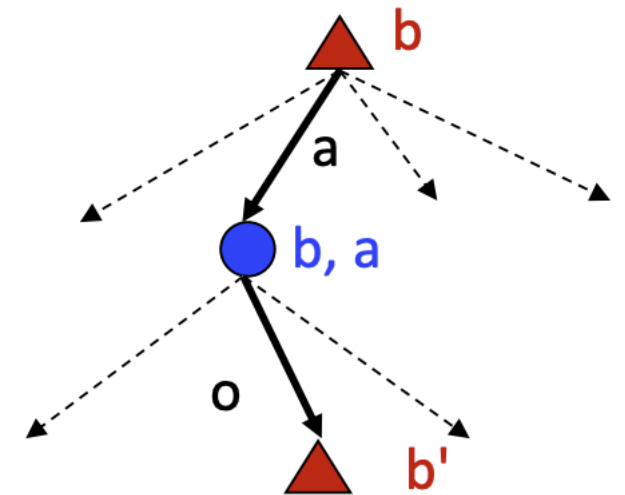
Reward $R(s, a, s')$



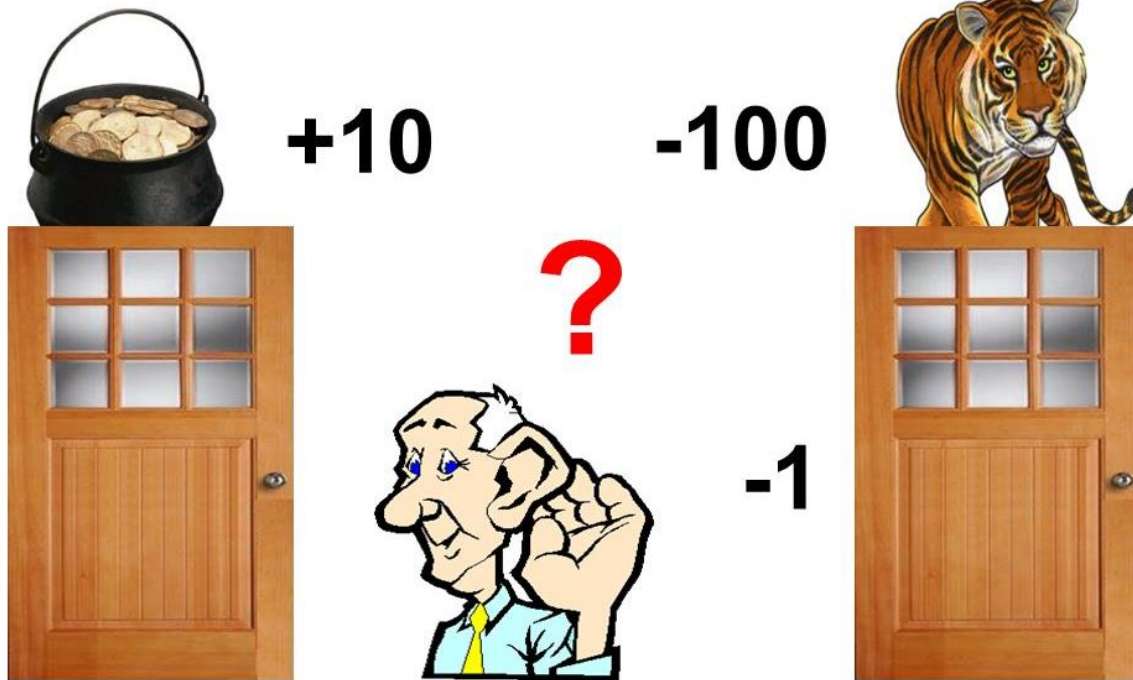
POMDPs add:

Observations O

Observation Function $P(o|s)$



Tiger Example



Actions $a = \{0, 1, 2\}$: 0: listen, 1: open left, 2: open right

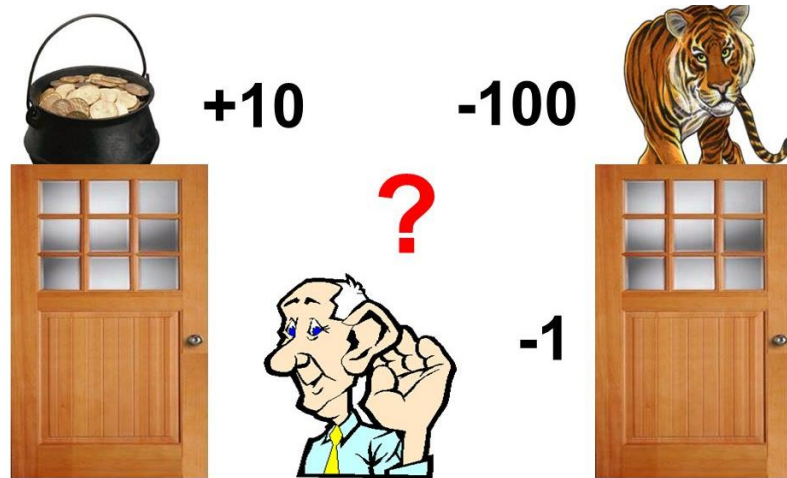
Reward Function:

- Penalty for wrong opening: -100
- Reward for correct opening: +10
- Cost of listening: -1

Observations:

- To hear the tiger on the left
- To hear the tiger on the right

Tiger Example



Belief update based on observations:

$$b_1(s_i) \propto p(o|s_i, a) \sum_{s_j \in \mathcal{S}} p(s_i|s_j, a) \cdot b_0(s_j)$$

Immediate return

Discounted future return

Value Iteration
over Beliefs

$$V^*(b) = \max_{a \in A} \left[\sum_{s \in \mathcal{S}} b(s) \cdot R(s, a) + \gamma \sum_{o \in \mathcal{O}} P(o|b, a) \cdot V^*(b_o^a) \right]$$

Hard to compute continuous space MDPs -> Approximation

Tiger Example

Value Iteration
over Beliefs

$$V^*(b) = \max_{a \in A} \left[\sum_{s \in S} b(s) \cdot R(s, a) + \gamma \sum_{o \in O} P(o|b, a) \cdot V^*(b_o^a) \right]$$

Immediate return *Discounted future return*

Hard to compute continuous space MDPs -> Approximation

Q-MDP
Approximation

$$V^*(b) = \mathbb{E}_s[V^*(s)] = \sum_s b(s) \cdot V^*(s)$$

Intent Inference

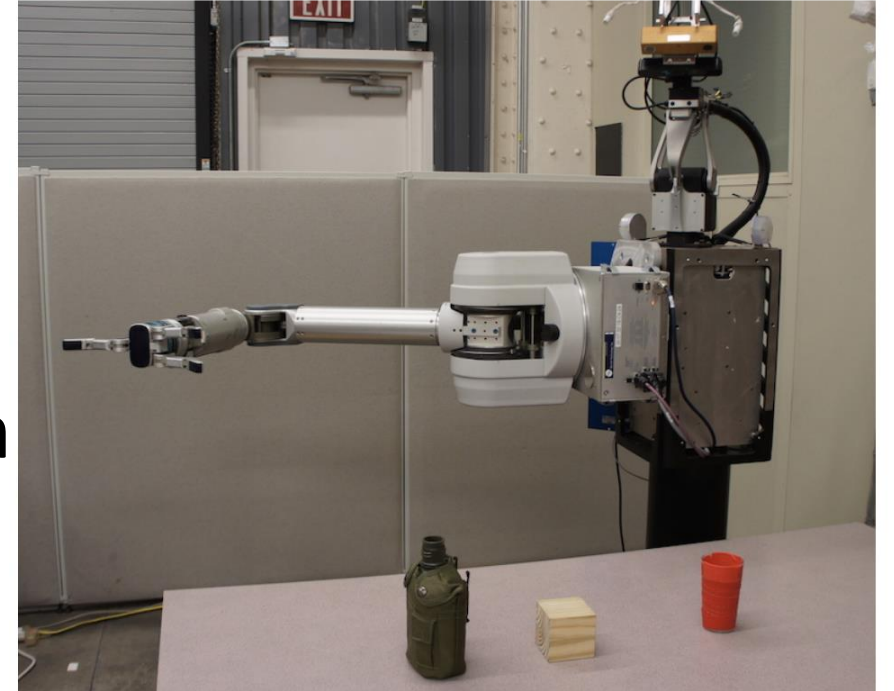
X Robot States

A Robot Actions

$T: X \times A \rightarrow X$ Transition function

$u \in U$ Human continuous input

$D: U \rightarrow A$ Mapping between human input and robot actions



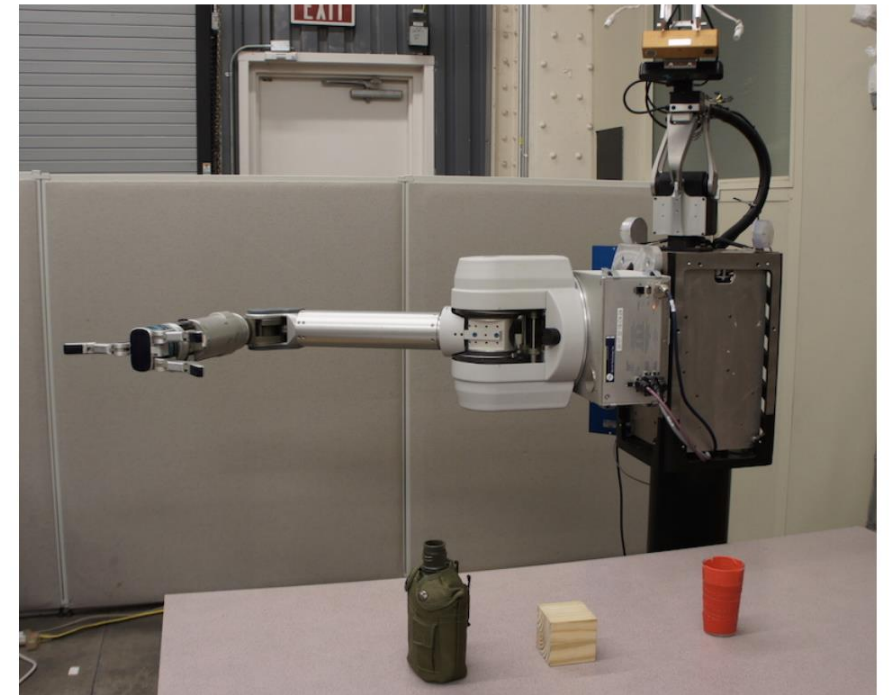
User's Policy is Learned from IRL

$\pi_g^{usr}(x) = p(u|x, g)$ We learn a policy for each goal

$$p(\xi|g) \propto \exp(-C_g^{usr}(\xi))$$

$$p(g|\xi) \propto p(\xi|g) \cdot p(g) \quad \text{Bayes Rule}$$

POMDP Observation Model



Hindsight Optimization (Q-MDP)

Estimate cost-to-go of the belief by assuming full observability will be obtained at the next time step.

You never gather information, but can plan efficiently in deterministic subproblems.

$$b(s) = b(g) = p(g|\xi) \quad \text{Uncertainty is only over goals}$$

$$Q(b, a, u) = \sum_g b(g) \cdot Q_g(x, a, u)$$

Action-Value function of the POMDP

Cost-to-Go of Acting optimally and going towards goal g

Shared Autonomy with Hindsight Optimization

