Proactive Environment Design to Enhance Reward Alignment and Human Intent Prediction

Yi-Shiuan Tung

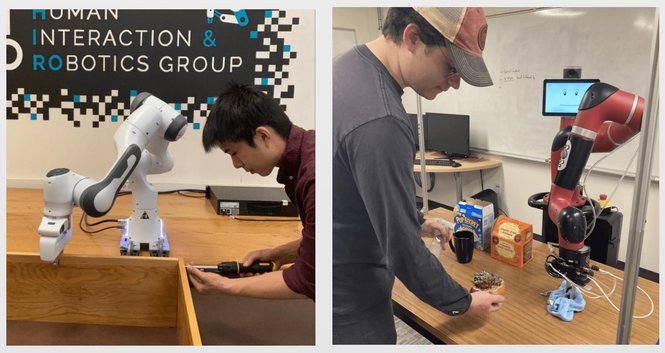
University of Colorado Boulder

HIRO Human Interaction & Robotics Group

CAIRO Collaborative AI and Robotics Lab

Why Environment Design (ED)?

- Human-robot collaboration requires the robot to **predict human intent** and **adapt to their preferences.**
- Current approaches focus on better inference algorithms but overlook an important lever for improving collaboration: ED.



Problem Formulation

Markov Decision Process (MDP) with **S** (state space), **A** (action space), **T** (transition function), **R** (reward function), and **γ** (discount factor)

We define ED as modifying the transition function T in a MDP.

$$\max_{T} \max_{\pi} \mathrm{E}[\sum_{t} \gamma^{t} R(s_t, a_t) \mid \pi]$$

Reward Alignment

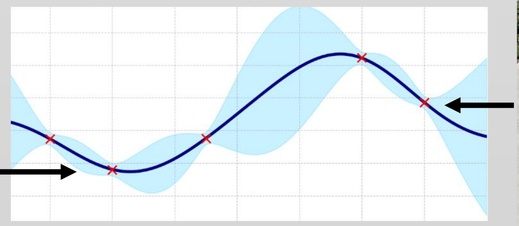
Key Insight

Environment design can generate informative scenarios to query the human, improving the efficiency of reward function learning.

1. Generate Environments

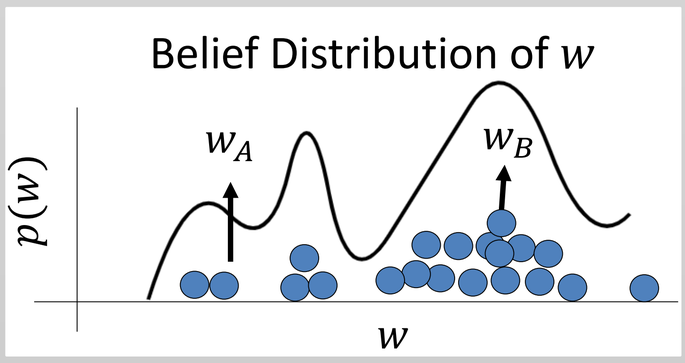
Bayesian Optimization

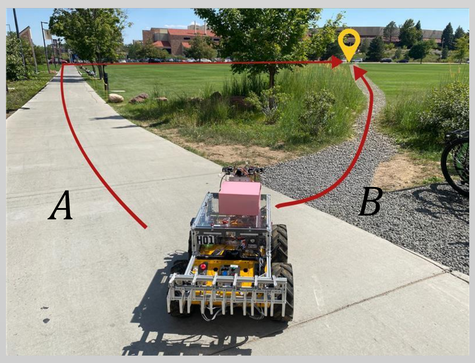
Information Gain



Environment Parameters

2. Create Counterfactual Trajectories





We use the Bradley-Terry Model of preferences where trajectories with higher rewards are exponentially more likely.

$$p(A > B) = \frac{\exp(R(A))}{\exp(R(A)) + \exp(R(B))}$$

$$R(A) = \phi(A)^T w$$

$R(A)$ – reward of trajectory A

$\phi(A)$ – features of trajectory A

We maintain a Bayesian belief of the human's reward functions by using Monte Carlo Markov Chain (MCMC).

Goal: Find trajectories to ask the human for preferences such that the uncertainty of our belief over the human's reward function is minimized.

$$max_{A,B}\ \mathrm{H(w)} - \mathrm{E[H(w|I)]}$$

$H(w)$ – entropy of our belief w

I – human's input

Research Questions

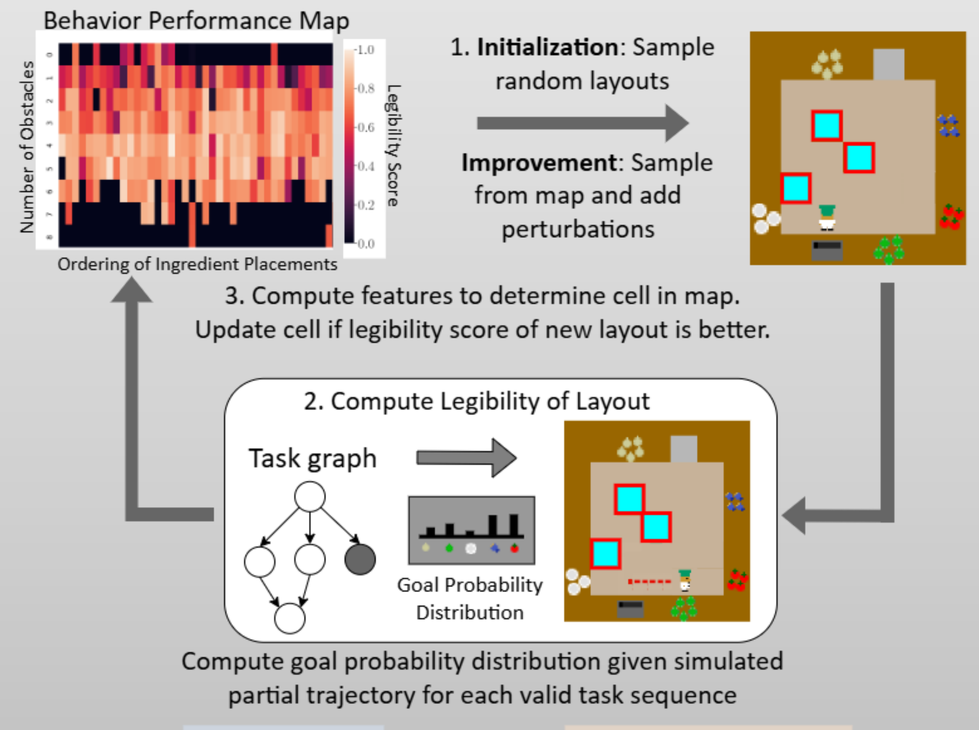
RQ 1: What aspects of the environment can we modify to improve the overall task performance and fluency of human-robot interaction?

RQ 2: How can we efficiently optimize over the high-dimensional space of possible environment layouts?

Human Goal Prediction

Key Insight

By shaping the workspace through object placement and AR obstacles, the robot can make human motion more legible and improve goal prediction.

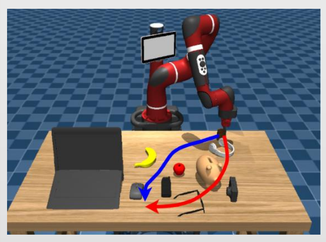


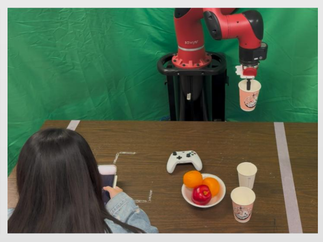
$$\sum_{\tau \in \mathrm{Perm}(T)} \mathbb{I}\{\mathrm{valid}(\tau)\} \sum_{G} \mathbb{E}_{\xi_{S \to Q}} [-\log p(G \mid \xi_{S \to Q})]$$

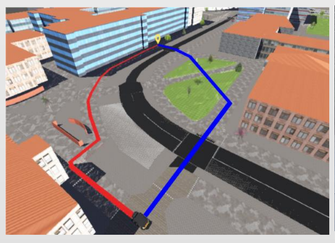
Ensure task sequence is valid — Sampled trajectories — Cross Entropy Loss

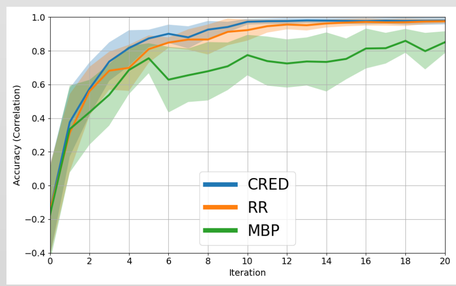
Experiments / Results

1. Reward Alignment

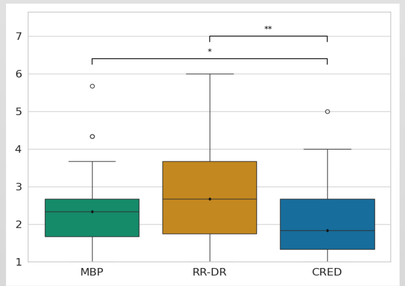




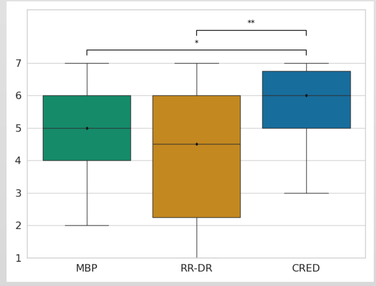




Reward Accuracy



NASA-TLX



Ease of Choice

2. Human Goal Prediction