Center for Human-Compatible Artificial Intelligence

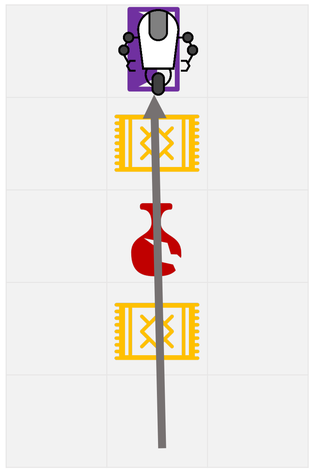
Berkeley UNIVERSITY OF CALIFORNIA

Preferences Implicit in the State of the World

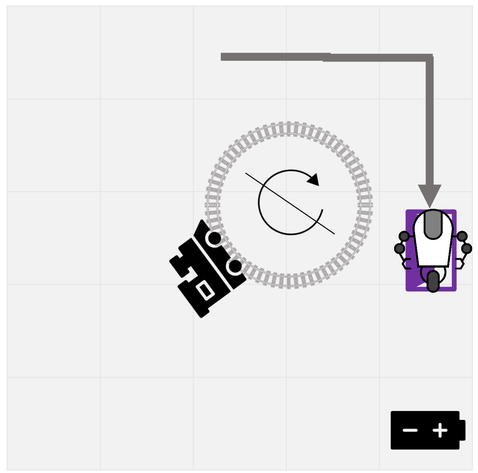
Rohin Shah*, Dmitrii Krasheninnikov*, Jordan Alexander, Pieter Abbeel, Anca Dragan

Problem setting: Specifying everything is hard

We care about many things in this room:
- Don't break stuff
- Keep floor clean
- Don't spill
- Don't hit humans



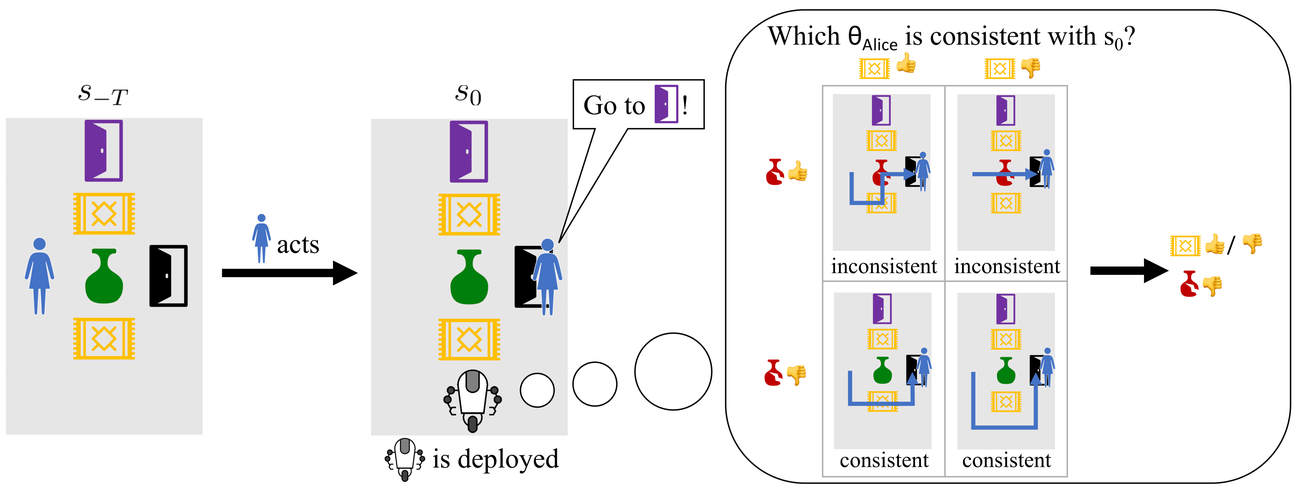
Assume known dynamics
Assume known features
But reward is misspecified



Naïve solution: Preserve the state
Problem: Does not distinguish between good and bad effects

Can we infer parts of the robot's specification, *without* having a human think of them?

The state contains preference information



Infer rewards by simulating past trajectories and checking consistency with the current state

By assuming that the human was Boltzmann-rational, we create a probabilistic model of past trajectories.

$$p(\tau \mid \theta) = p(s_{-T}) \prod_{t=-T}^{-1} \pi_t(a_t \mid s_t, \theta) \mathcal{T}(s_{t+1} \mid s_t, a_t)$$

But we only know the current state! So, we marginalize:

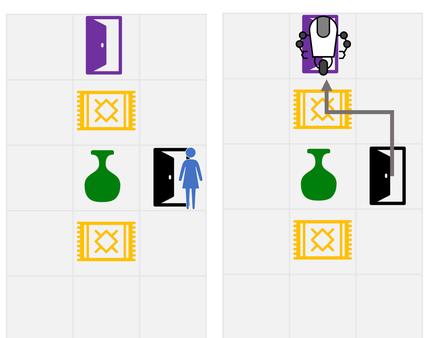
$$p(s_0 \mid \theta) = \sum_{s_{-T}, a_{-T}, \ldots, s_{-1}, a_{-1}} p(\tau \mid \theta)$$

Then we can find the MAP reward via gradient descent. We calculate gradients using dynamic programming.

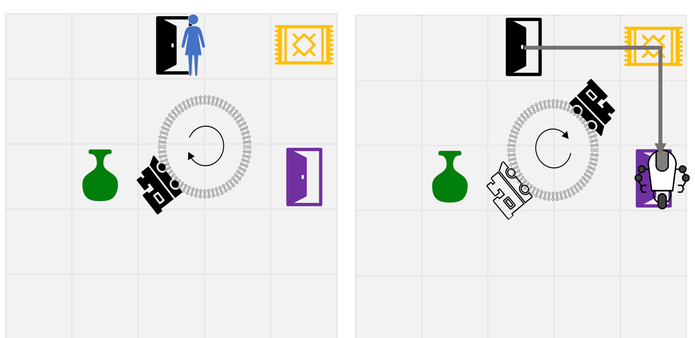
$$\theta^* = \arg\max_\theta \ln p(s_0 \mid \theta)$$

When a robot is deployed in an environment that humans have been acting in, the state of the environment is already optimized for human preferences and thus informative.

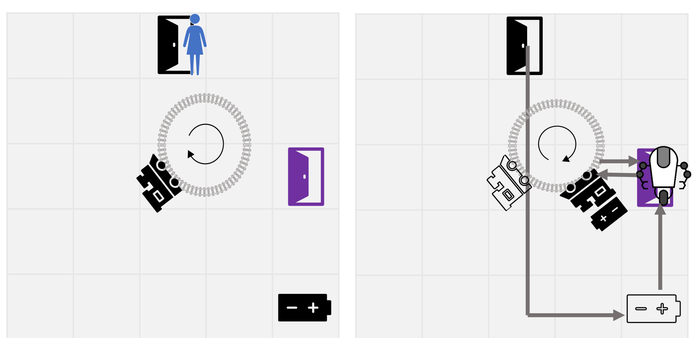
Proof of Concept



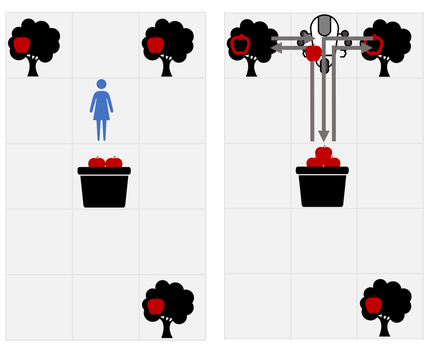
Negative side effects:
Don't break the vase



Environment effects:
Don't break the moving train



Desirable side effects:
Deliver the battery to the train



Implicit preferences:
Collect apples