Preference is a Tie-Breaker, A New Definition for Reinforcement Learning

Tuhina Tripathi, Xinyu Cao, Bradley Hayes

Collaborative AI and Robotics Lab

Motivation

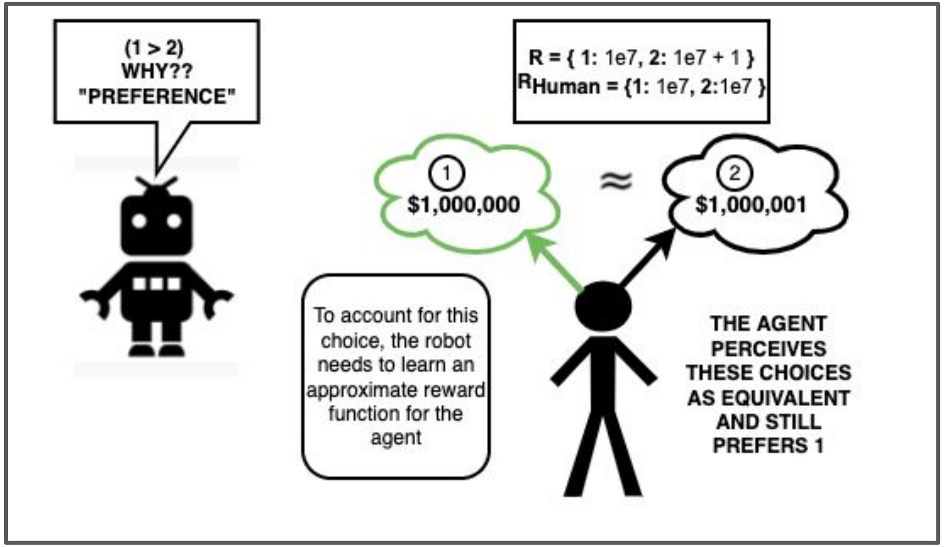
To improve the efficiency of shared operations between humans and robots, it is crucial that autonomous systems be capable of accounting for human preferences. Preference learning is an active area of research with decades of valuable work, but surprisingly there does not yet exist a rigorous specification of “preference” that disentangles it from what could be considered a base task specification. In this work, we provide a novel, formal definition of preference that differentiates it from a base task specification and enables a more precise treatment in the learning literature.

Main Idea

In many cases, optimal behavior is fully and uniquely determined by a task’s reward function and thus leaves no room for preference. **Preference can only exist for decisions where expected reward is equivalent across multiple options, otherwise the behavior is suboptimal (incorrect) or prescribed (determined solely by the reward function).** We also consider cases where an agent may view some choices as functionally equivalent, even if the reward function does not (i.e., choices where expected rewards are **”close enough”** but not equal, such as receiving $1,000,000 or receiving $1,000,001), constructing an *approximately correct reward function* from the agent’s perspective.

Contribution

- We introduce a new reward function (R_{Human}) that represents the human’s interpretation of reward, derived by observing their behavior given that “preference must break ties” (observed “suboptimal” choices in R must be equivalent to optimal ones in R_{Human}).
- With this decoupling of reward functions, we achieve two principal outcomes: 1) A lower bound of cumulative reward to define whether the preference is usable or not, and 2) A framework to characterize how an agent perceives a task, given a base model.



(1 > 2)
WHY??
"PREFERENCE"

R = { 1: 1e7, 2: 1e7 + 1 }
R_{Human} = {1: 1e7, 2:1e7 }

To account for this choice, the robot needs to learn an approximate reward function for the agent

THE AGENT PERCEIVES THESE CHOICES AS EQUIVALENT AND STILL PREFERS 1

Task Environment

In order to adapt to the preferences of the demonstrator, we analyse how the human reward compares to the true reward for the task. The different reward scenarios are:

- R_{Human} **Identical:** Human agrees with the true reward function and the human policy is understandably adaptable.
- R_{Human} **Admissible:** There is disagreement but the human policy satisfies baseline success requirements and hence, is admissible.
- R_{Human} **Inadmissible:** The human policy does not satisfy the baseline criteria and therefore cannot be adapted.