Closed-loop Reasoning about Counterfactuals to Improve Policy Transparency

Michael S. Lee Henny Admoni Reid Simmons

The Robotics Institute, Carnegie Mellon University

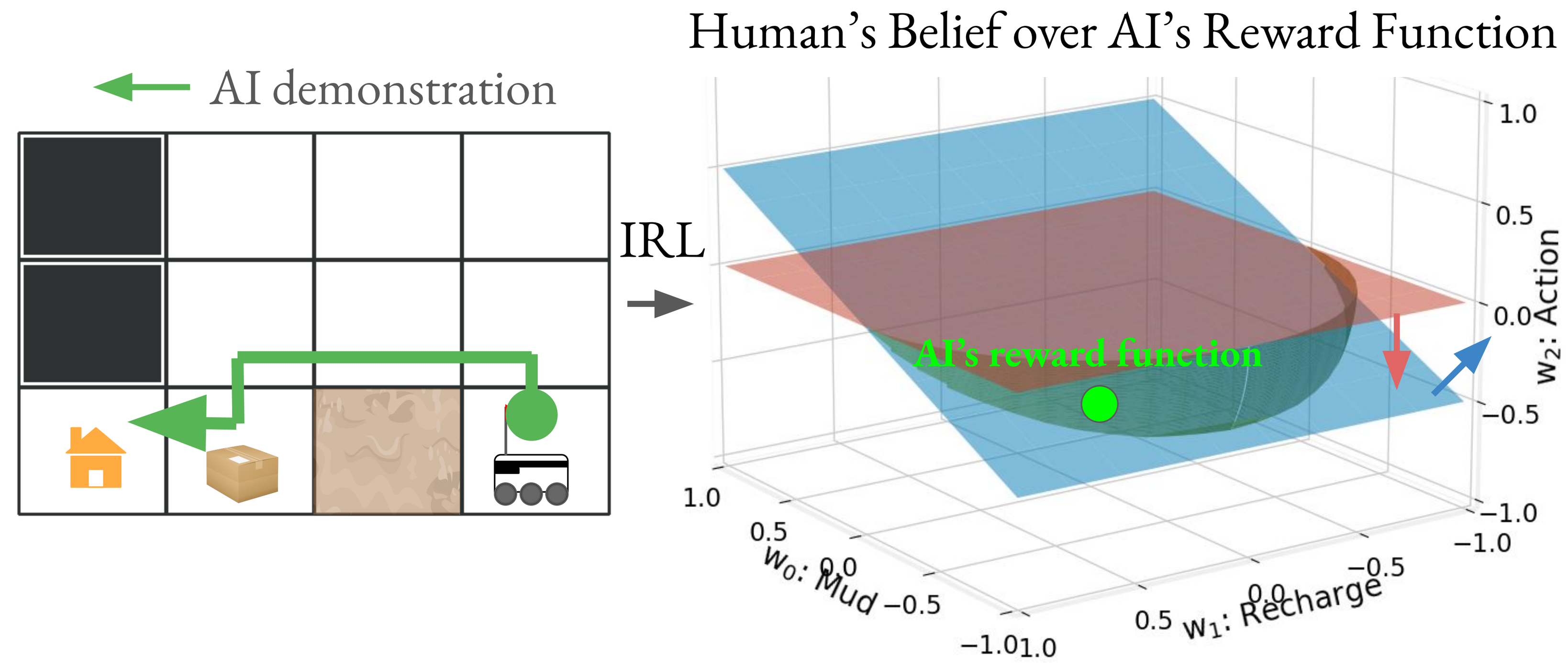
Overview

How can we **increase** the **transparency of AI policies** by providing demonstrations, tests, and feedback to humans?

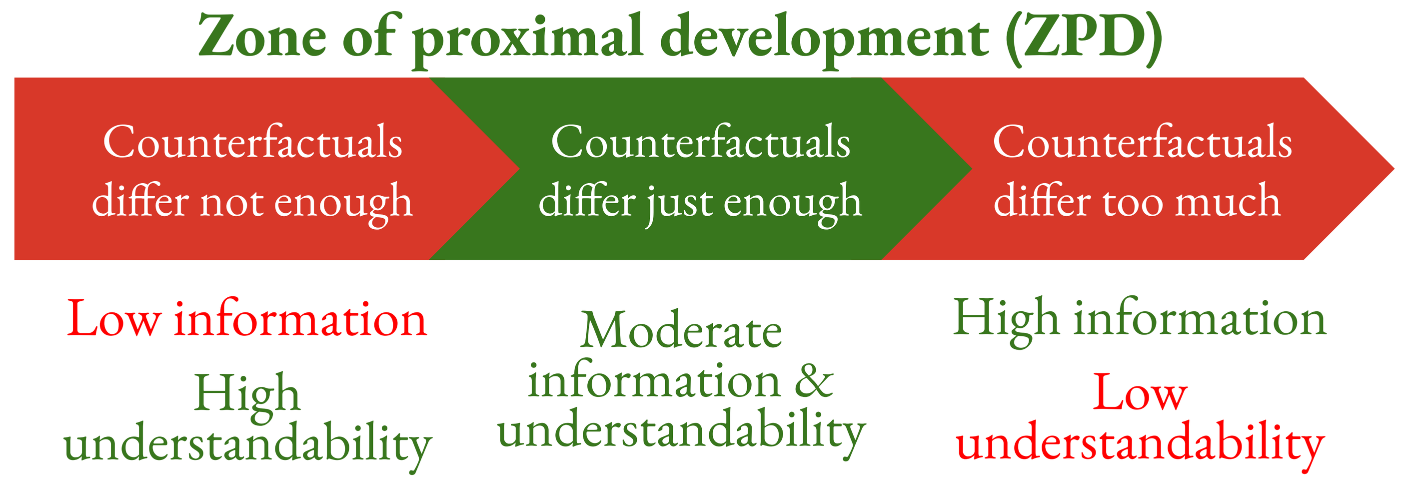
We design a **closed-loop teaching scheme** inspired by the human education literature, where the AI iteratively provides **informative and understandable demonstrations** given a **human's counterfactual expectations** of the AI's policy.

Background

Calculate **informativeness** of a **demonstration,** assuming humans infer others' policies and reward functions through **inverse reinforcement learning (IRL)**.

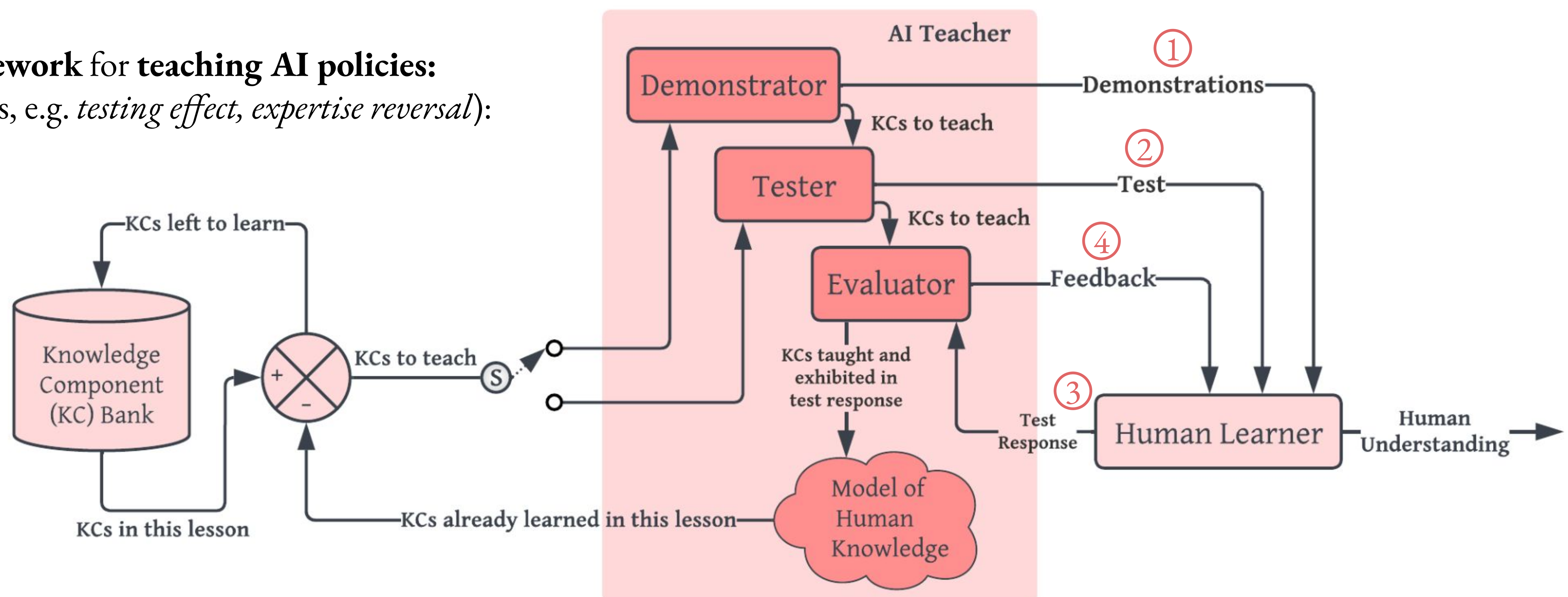


Increase **understandability** by teaching in the **zone of proximal development**, showing demonstrations that **differ just enough** from the **human's counterfactual expectations** of AI behavior.

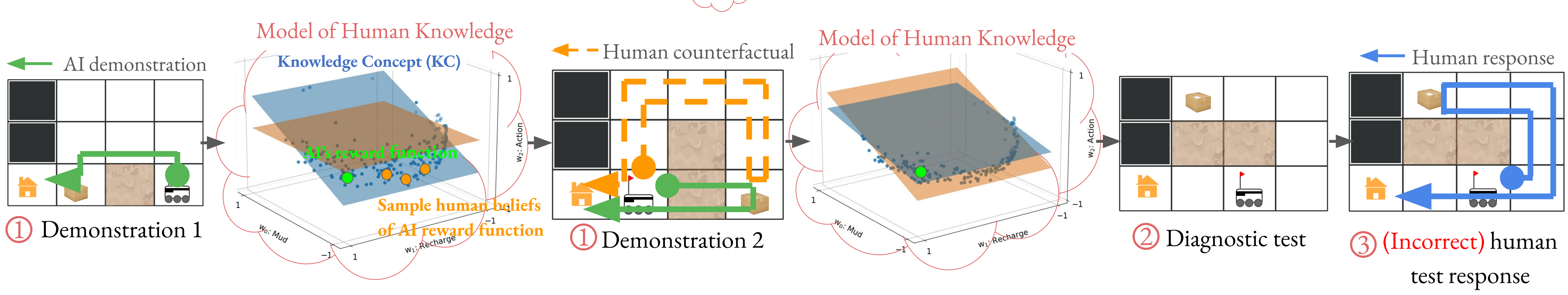


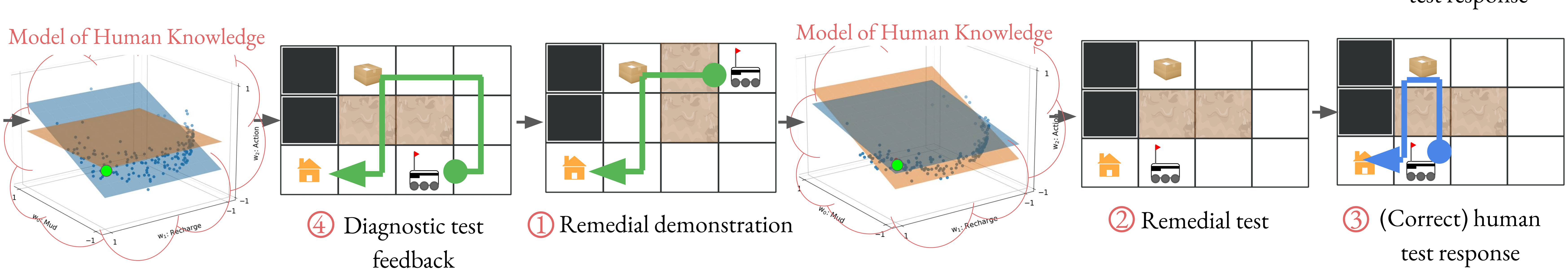
Approach

Proposed closed-loop framework for **teaching AI policies:**
(based on education principles, e.g. *testing effect, expertise reversal*):



An **example teaching progression** given a model of a human's beliefs and counterfactual expectations:





Evaluation & Future Work

User study will test whether the closed-loop teaching improves **learning outcomes** (via a held out set of tests), **learning efficiency** (via subjective reports of improved understanding), and **user engagement** (via User Engagement Scale [1]) over baseline of showing only demonstrations.

Future work: scale approach to policies that operate on **high dimensional states and reward functions** across different **contexts**; explore synergies between increasing policy transparency via demonstrations and **language.**

[1] O'Brien, Cairns, and Hall. "A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form." International Journal of Human Computer Studies (2018).
This work was supported by the ONR award N00014-18-1-2503.