Title

Deeply Explainable Artificial Intelligence (DExAI): Informative Rollouts and Critical States

Team Composition

University of California, Berkeley (UCB), Trevor Darrell, Principal Investigator, trevor@eecs.berkeley.edu
UCB: Anca Dragan, Sandy Huang

Problem Statement

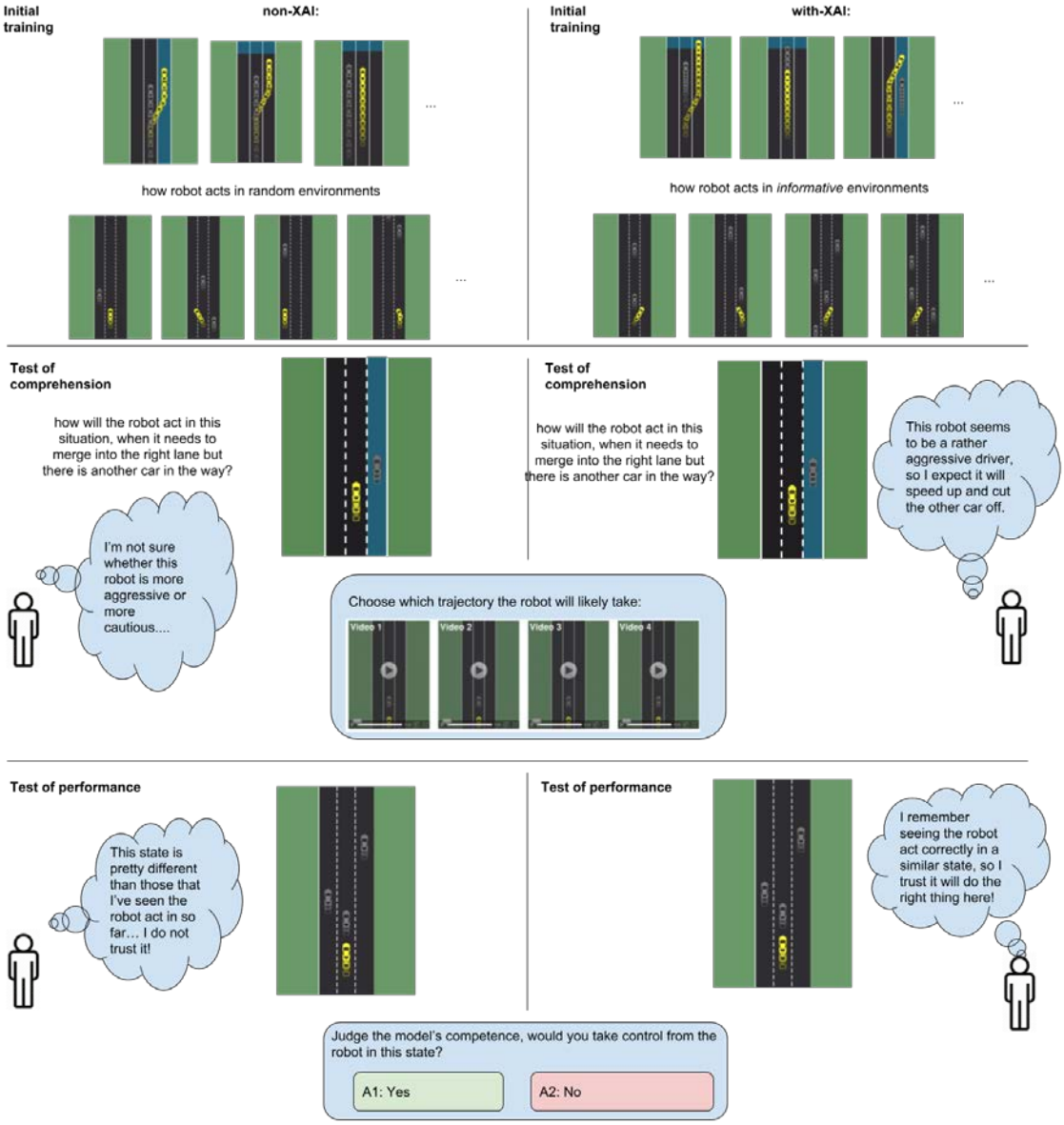
Safe and comfortable human-robot interaction depends on the human having a reasonably accurate mental model of the robot's policy – they need to understand how a robot will act, and which situations it can and cannot handle. Humans naturally improve their mental model of a robot over time, simply as they observe the robot and interact with it in the real world. But this process of passive familiarization can take a while.

Approach and Technique

UCB's approach aims to explain a robot's policy by showing informative examples of how that robot acts in different scenarios, that directly optimize for improving the human's mental model of the robot (see Figure 1). In other words, the informativeness of examples depends on how humans update their mental model of the robot, after seeing examples of robot behavior. The approach assumes the person has a particular way of updating their mental model, and selects examples that best improves the human's understanding of either the robot's objective function (which improves prediction), or which critical situations it can handle correctly (which establishes appropriate trust/reliance).

Approved for Public Release, Distribution Unlimited

Figure 1. Overview of the approach to explain a robot's policy. Left side: without explanation; right side: with explanation



Key Objectives and Current Status

Explanations should improve human prediction of how a robot will act in novel situations, including whether it will act correctly or not. In turn, this will help humans establish appropriate reliance on the robot. How these explanations impact user satisfaction is also measured.

Approved for Public Release, Distribution Unlimited

Table 1. Measures, Objectives, and Actuals to Date

Measures	Objectives	Actuals to Date
Human mental model: prediction	Improve user prediction of robot behavior	Informative example explanations led to significantly more accurate user predictions vs. no explanation
Trust/reliance	Establish appropriate trust/reliance	Informative example explanations led to more appropriate user trust vs. no explanation
User satisfaction	Increase user satisfaction	Informative example explanations led to higher user satisfaction vs. no explanation

	Size		**Length**	**Type**
User Study	161 subjects: average age 37.0 (SD = 11), 46% female		6 conditions; 6 test states	AMT, between-subjects
	Numerical Value		**Hypothesis Results**	**Mental Model Task**
	With Explanation	**Without Explanation**		
Mental Model Understanding: Prediction	-0.02 ± 0.48	-1.40 ± 0.48	Significant, p = 0.025	Prediction score = confidence level (5-point Likert) x accuracy (+1 if correct, -1 if incorrect)

	Size		**Length**	**Type**
User Study	60 subjects: average age 32.5 (SD = 6.7), 27% female		3 conditions; 10 test states; 2 policies (policy A is better than policy B)	AMT, between-subjects for conditions, within-subjects for policies
	Numerical Value		**Hypothesis Results**	**Trust Measure**
	With Explanation	**Without Explanation**		
Appropriate Trust / Reliance	Policy A: 1.35 ± 0.09 Policy B: 1.77 ± 0.1	Policy A: 2.29 ± 0.07 Policy B: 2.32 ± 0.07	Significantly more trust for policy A over policy B (p = 0.001) when there are explanations, but not without explanations; Significantly more trust for policies (both A and B) when there are explanations (p = 0.006)	Whether to take control (5-point Likert, lower score = more trust)
	Numerical Value		**Hypothesis Results**	**Satisfaction Measure**
	With Explanation	**Without Explanation**		
Explanation Satisfaction	Policy A: 3.00 ± 0.21 Policy B: 2.00 ± 0.27	N/A	Significantly more agreement for policy A over policy B	Whether user agrees with policy's critical states

Approved for Public Release, Distribution Unlimited

Test and Evaluation

See the following references for additional information:

Sandy H. Huang, David Held, Pieter Abbeel, and Anca D. Dragan. *Enabling robots to communicate their objectives.* RSS 2017. https://arxiv.org/abs/1702.03465.

Sandy H. Huang, Kush Bhatia, Pieter Abbeel, and Anca D. Dragan. *Establishing appropriate trust via critical states.* IROS 2018. https://arxiv.org/abs/1810.08174.

Approved for Public Release, Distribution Unlimited