Title

xACT: Explanation-informed ACceptance Testing of deep adaptive programs

Team Composition

Oregon State University (OSU), Alan Fern, Principal Investigator (PI), alan.fern@oregonstate.edu
OSU co-PIs: Margaret Burnett, Tom Dietterich, Martin Erwig, Liang Huang, Fuxin Li, Prasad Tadepalli, and Weng-Keen Wong

Problem Statement

The overall objective of the OSU xACT effort is to develop tools for explaining learned agents for sequential decision making and to discover best principles for designing explanation user interfaces. OSU's explainable agent model is enabled by Explainable Deep Adaptive Programs (xDAPs) which are a combination of their prior work on adaptive programs, deep Reinforcement Learning (RL), and explainability. xDAPs are programs where certain "choice points" are attached to deep networks and trained via Deep Reinforcement. After training, explanation-informed acceptance testing is conducted via an explanation user interface (xUI), which allows a human to observe the agent behaving in selected simulation scenarios. Importantly, the xUI also allows the user to pause the agent at any decision point in order to explore explanations for that decision. The current xUI supports several forms of explanations, with the Phase 1 user study focusing on visual explanations including reward-decomposition bars and visual saliency maps linked to those bars. The Phase 1 user study involved non-experts in machine learning using the interface to understand an agent trained for a simple real-time strategy game.

Approach and Technique

The overall technical approach illustrated in Figure 1 begins with the AI System developer and ends with the test pilot (acceptance tester). The system developer creates an agent as an xDAP and then trains the agent offline using RL. An explanation learning process is then conducted, which aims to learn explanation structures to be used for explaining the agent's decisions. This explanation learning process may or may not be integrated with the original RL process depending on the type of explanations desired. The final step on the developer side is an optional process of interactive naming, where a subject matter expert interacts with a specialized interface to attach semantically meaningful concepts to the explanation structures. The result of the offline training and interactive naming is an agent with learned explanation structures that can be used by the explanation generator to interactively produce explanations for the xUI.

The Phase 1 user study focused on a subset of the proposed framework as shown in Figure 2. The left part of the figure shows a traditional acceptance testing user interface for a simple real-time strategy game. The user is able to navigate through the decisions of the agent (labeled D1-D4) to judge whether the agent is doing the right thing. The right side of the figure shows the xACT xUI, which adds two main types of explanations to the traditional interface. First, at each decision point, the user is shown a visualization of the reward decomposition for each action. The key idea is that there are multiple sources of reward in the game (e.g., negative reward for taking damage, positive reward for giving damage, etc.) and the bars show how much of each type of reward the agent predicts for each action. This gives insight into the trade-offs being considered by the agent. In addition, the user can display a saliency map (displayed at the bottom), which shows which parts of the game information were most important to assigning the value to the selected bar. Together these visualizations can be explored to help determine if the agent is doing the right thing for the right reasons.

Approved for Public Release, Distribution Unlimited

Figure 1. xACT system architecture

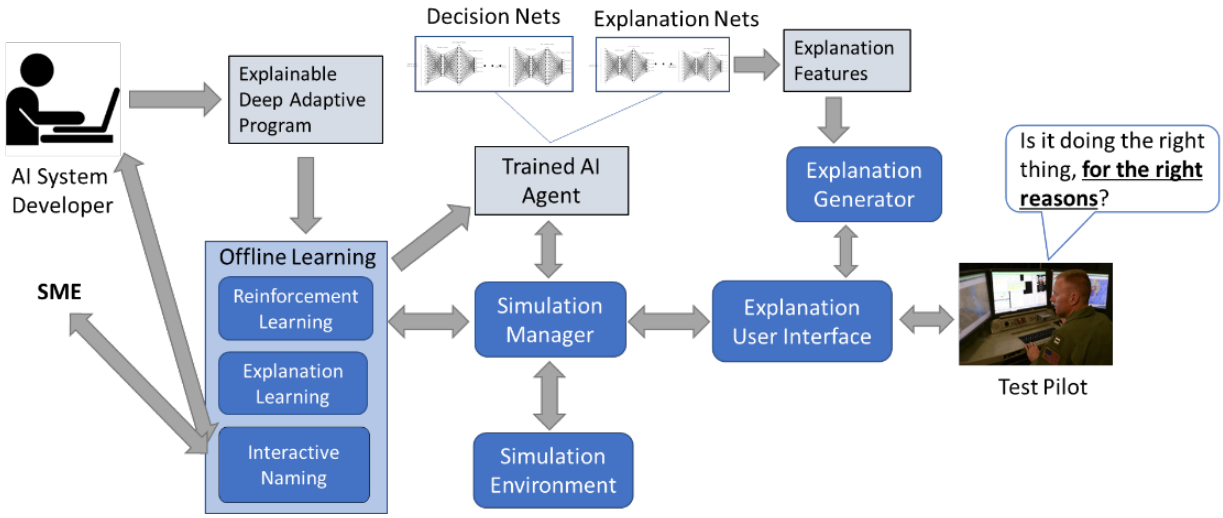
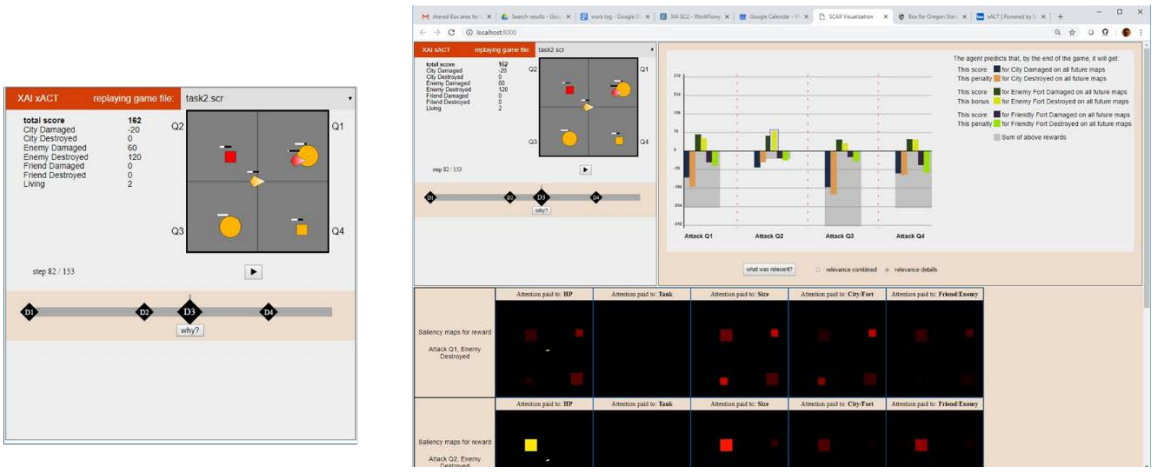


Figure 2. (Left, without explanation) Interface for standard acceptance testing, where a user is able to observe the behavior of a trained agent. (Right, with explanation) xUI where the user is able to explore visual explanations of decisions of the agent along the timeline



Key Objectives and Current Status

The primary user study for Phase 1 aimed at determining which explanation components if any from the xUI are able to improve users' mental models of the agent. In particular, the study focused on non-experts in machine learning. To do this, OSU performed a randomized experiment with 124 participants, explaining an AI agent via saliency maps (the eyeballs of an AI), reward bars (the crystal ball the AI uses to predict future score), neither, or both. These experiments were designed to answer the following research question: *Which treatment is better (and how) at improving a participant's mental model of an intelligent agent's behavior?*

To assess mental models, users in the study took part in a two-hour acceptance testing exercise, where they would interactively observe the agent's behavior at game decision points and answer questions about the decision points posed by the user study system. Users in treatments that included explanation

Approved for Public Release, Distribution Unlimited

components were able to explore the explanations as well as observe the behavior when answering questions. In order to assess the mental models of users, two researchers qualitatively coded responses from participants' final responses to describe the decision making process of the AI. Once they agreed, a score was assigned based on positive/negative codes assigned to each response. Quantitative analysis techniques were then applied to understand which of the treatments (Control, Saliency, Rewards, or Everything) performed best for building mental models in the users.

Table 1. Measures, Objectives, and Actuals to Date

Measures	Objectives	Actuals to Date
Mental Model Score – based on qualitative codes assigned to elements of user responses	Improve the user's understanding of the agent to develop an appropriate level of trust	1) Everything participants had significantly higher mental model scores than the Control participants 2) Treatments that saw reward bars (Everything and Rewards) led to significantly higher mental model scores than the Control group
Prediction Score – based on a user's ability to predict the next action of the agent based on previous observations of behavior and explanations (for treatments that have explanations)	Improve the user's understanding of what the agent is likely to do in a particular situation	There were no significantly significant differences between the control treatment and treatments with explanations

	Numerical Value		Hypothesis Results	Mental Model Task
	With Explanation	**Without Explanation**		
Mental Model Understanding	48 (mental model score)	70 (mental model score)	Significant, p = 0.00534	Describing agent behavior model

	Size	Length	Type
User Study	124 subjects	4 Treatments with approximately 30 subjects per treatment Each treatment was 2 hours	Between subject design with approximately 30 randomized participants per treatment

Test and Evaluation

In addition to the above user study, OSU has developed and evaluated other explanation components. In one study, OSU developed an interactive naming interface, which allowed users to assign semantics to internal nodes of neural networks for use in explanation generation. In a preliminary evaluation OSU studied the consistency and differences among five different users who assigned semantics to the same network nodes based on the same data sets. Two main results were established: 1) [Naming Coverage] after naming, 80% of the images in the data set activated at least one network node that had been named by a user, and 2) there was significant overall consistency between the names assigned to nodes by different users.

The second study evaluated whether explanations could help users better understand fine-grained image classifiers. In particular, explanation neural networks were learned as a way of compressing the distributed feature representation of deep networks into a small number of meaningful features. The features can be visualized by humans and used to help understand the key features considered by the network when making classifications. A user study was conducted using two datasets, one a seagull

Approved for Public Release, Distribution Unlimited

classification dataset and the other a cell microscopy classification dataset. It was found that users that used the explanations achieve significantly higher mental model scores compared to those that did not use explanations (seagulls p = 0.0086 and microscopy p = 0.0011).

Zoe Juozapaitis, Anurag Koul, Alan Fern, Martin Erwig, Finale Doshi-Velez. (under review). E*xplainable Reinforcement Learning via Reward Decomposition.* Conference on Uncertainty in Artificial Intelligence (UAI 2019).

Zhongang Qi, Saeed Khorram, and Fuxin Li. (under review). *Embedding Deep Networks into Visual Explanations.* International Joint Conference on Artificial Intelligence (IJCAI 2019).

Rey Pocius, Lawrence Neal, and Alan Fern. (2019). *Strategic Tasks for Explainable Reinforcement Learning.* AAAI-2019 Student Abstract.

Mandana Hamidi-Haines, Zhongang Qi, Alan Fern, Fuxin Li, and Prasad Tadepalli. (2019). *Interactive Naming for Explaining Deep Neural Networks: A Formative Study.* IUI Workshop on EXplainable Smart Systems (EXSS).

Zhongang Qi and Fuxin Li. *Learning Explainable Embeddings for Deep Networks.* (2017). NIPS Workshop on Interpreting, Explaining and Visualizing Deep Learning.

S. Penney, J. Dodge, C. Hilderbrand, A. Anderson, L. Simpson, and M. Burnett. (2017). *Toward Foraging for Understanding of StarCraft Agents: An Empirical Study.* IUI 2018.

J. Dodge, S. Penney, C. Hilderbrand, A. Anderson, and M. Burnett. (2017). *How the Experts Do It: Assessing and Explaining Agent Behaviors in Real-Time Strategy Games.* CHI 2018.

Martin Erwig, Alan Fern, Magesh Murali, and Anurag Koul. (2018). *Explaining Deep Adaptive Programs via Reward Decomposition.* IJCAI Workshop on Explainable Artificial Intelligence.

Sam Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. (2018). *Visualizing and Understanding Atari Agents.* International Conference on Machine Learning.

Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Alan Fern, and Margaret Burnett. (2019). *Explaining Reinforcement Learning to Mere Mortals: An Empirical Study.* International Joint Conference on Artificial Intelligence (IJCAI 2019).

Approved for Public Release, Distribution Unlimited