Title

Deeply Explainable Artificial Intelligence (DExAI): Heterogeneous Hierarchical Policies

Team Composition

University of California, Berkeley (UCB), Trevor Darrell, Principal Investigator, trevor@eecs.berkeley.edu
UCB: Haoran Tang, Pieter Abbeel

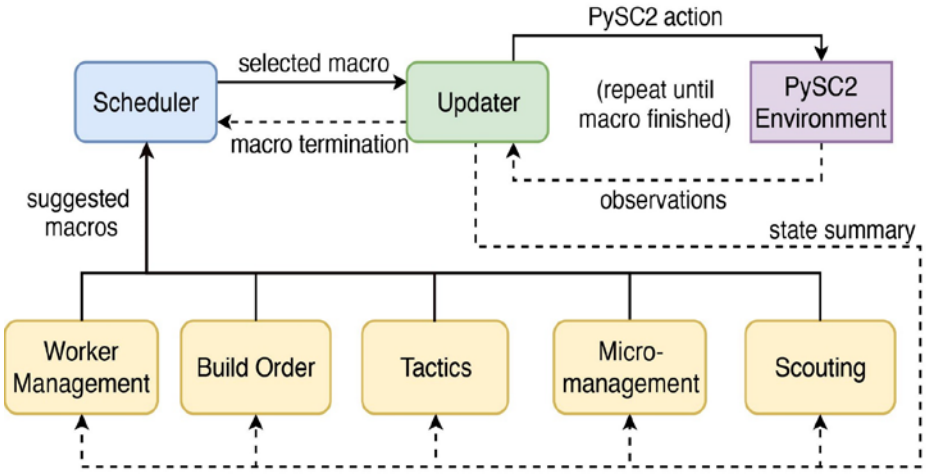
Problem Statement

The goal of the effort is to develop an AI to play the full game of StarCraft II and to explain its behavior automatically. Conventional techniques of deep reinforcement learning are data-inefficient and generate black-box behavior. Therefore UCB designed a novel modular architecture to improve both data efficiency and explainability.

Approach and Technique

Playing StarCraft II is a complex task that requires handling several aspects of the game simultaneously, including resource management and army control. Inspired by prior efforts, UCB's agent decomposes the task into five modules that have separate responsibilities, as detailed in Figure 1 below, and optimizes them independently with either scripting or deep reinforcement learning. The modular architecture is not only strong against the game's built-in bots, but also designed to be explainable. Every decision can be traced back to the responsible module and thus be understood and corrected effectively. The agent can automatically generate explanations of the form"(time)[module] action." For example, "(3:05) [Tactics] Attack enemy bases" means that the Tactics module decides to attack enemy bases at 3 minutes and 5 seconds in the game. By reading the explanations, AI experts can quickly identify problems in the agent and can correct them efficiently.

Figure 1. Overview of AI modular architecture



Key Objectives and Current Status

Sixteen experts in frontier development of AI for StarCraft II were invited to evaluate the explainability of the agent. The objectives are summarized below in Table 1.

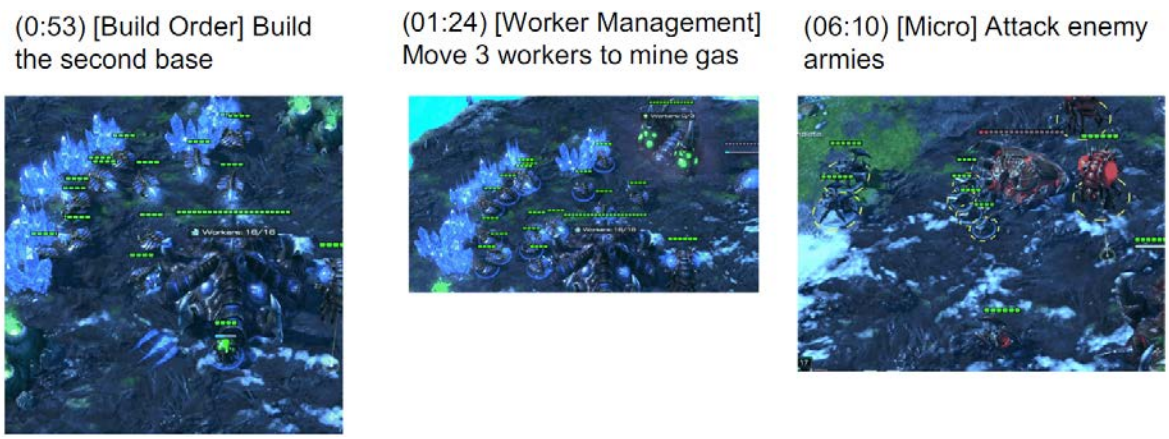
Approved for Public Release, Distribution Unlimited

Table 1. Measures, Objectives, and Actuals to Date

Measures	Objectives	Actuals to Date
Explanation Goodness	Explanations match expectations	Explanations are good
Explanation Satisfaction	Explanations are satisfactory	Explanations are satisfactory
User-Machine Task	Improve AI development efficiency	Experts debug better with explanations

In the first section, experts were presented with three video clips of the agent taking a certain action, along with the textual explanations (see example below in Figure 2).

Figure 2. Examples of explaining certain behavior



The experts rated on Explanation Goodness and Explanation Satisfaction on a Lickert scale from "I agree strongly" to "I disagree strongly" (5 to 1). The evaluation results, summarized below, suggest that most experts rated the explanations as good and satisfactory.

Table 2 (continued). Measures, Objectives, and Actuals to Date

	Numerical Value		Hypothesis Results	Goodness Measure
	With Explanation	**Without Explanation**		
Explanation Goodness	4.2	N/A	Significant, $p < 0.05$	Faithfulness

	Numerical Value		Hypothesis Results	Satisfaction Measure
	With Explanation	**Without Explanation**		
Explanation Satisfaction	3.8	N/A	Significant, $p < 0.05$	2, range of responses

In the second section, a user-machine task was performed to evaluate how explanations help AI developers to improve the system. The experts were evenly divided into groups A and B. They watched two video clips of two of the agents playing against each other and were tasked to find out a module that was malfunctioning in one of the players. Group A saw explanations in the first video, but not in the second, while Group B saw explanations in the second video, but not in the first. The evaluation interface is shown in Figure 3 below. As summarized below, the experts were more likely to complete the task with the help of explanations.

Approved for Public Release, Distribution Unlimited

Figure 3. Evaluation interface with explanations

Table 1 (continued). Measures, Objectives, and Actuals to Date

	Numerical Value		Hypothesis Results	User Machine Task
	With Explanation	**Without Explanation**		
User-Machine Task Performance	87.5%	25.0%	Significant, $p < 0.01$	Find the bad module

Test and Evaluation

For additional information see:

Lee, Dennis, et al. "Modular Architecture for StarCraft II with Deep Reinforcement Learning." *Fourteenth Artificial Intelligence and Interactive Digital Entertainment Conference.* 2018.