CPSC 490: Senior Project Proposal
Student: Aadit Vyas
Advisor: Dragomir Radev, Ph.D.
Lab: Language, Information, and Learning at Yale (LILY)
Field: Explainable AI

Figure 1: Three examples of PHYRE tasks (left) and one example solution (right). Black objects are static; objects with any other color are dynamic and subject to gravity. The tasks describe a terminal goal state that can be achieved by placing additional object(s) in the world and running the simulator. The task in the left-most pane requires placement of two balls to be solved, whereas the others can be solved with one ball. The right-most pane illustrates a solution (red ball) and the solution dynamics.

I. Introduction

On August 15, 2019, Facebook published PHYRE: A New Benchmark for Physical Reasoning\(^1\) (Figure 1). PHYRE was developed to be a benchmark for physical reasoning to encourage the development of learning algorithms that are sample-efficient and generalize well across puzzles because, as tested in the PHYRE paper, several modern learning algorithms failed to solve these problems efficiently (Figure 4).

This project proposal seeks to leverage the PHYRE platform to explore the field of Explainable AI.

II. Background Research

A. What is Explainable AI and why is it important?

As AI becomes increasingly a part of various industries and processes, the need to trust the decisions of these AI-agents has increased as well. Especially true for AI systems in healthcare, autonomous driving, and military, ambiguity with respect to the decisions reached by AI systems is a growing area of concern. In order for AI systems to be trusted, humans must be able to fully understand how decisions are being made.

\(^1\) PHYRE: A New Benchmark for Physical Reasoning
Explainable AI (XAI) is an emerging field in machine learning that aims to shed light on the mythical AI black box of how decisions are made by artificial intelligence (AI) systems. XAI aims to understand the steps and models involved in making decisions.

**B. PHYRE and XAI**

PHYRE presents a great playground for XAI research: each stage is a two-dimensional world that simulated simple deterministic Newtonian physics. Each task has an initial world state and a goal—the agent aims to achieve the goal by taking a single action, placing one or more new dynamic bodies into the world.

After a move is made and simulated, the logic behind an agent’s move can be discerned on a high level by a human observer. However, would it not be valuable if the agent provided an explanation for its move before it was simulated?

**C. NLP and RL**

![Figure 2: The LEARN framework consists of a standard RL module containing the agent-environment loop, augmented with a LEARN module.](image)

Reinforcement learning (RL) is a fundamental area of machine learning surrounding how agents ought to take actions to maximize some encoded reward. LanguagE-Action Reward Network (LEARN) is a framework that maps free-form natural language instructions to intermediate rewards, which can be used by an RL-agent (Figure 2). This idea of giving an agent intermediate rewards using instructions in natural language.

---

2 Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models
3 Using Natural Language for Reward Shaping in Reinforcement Learning
First, to use a LEARN framework, a mapping between language and objects/actions must be implicitly or explicitly learned, which is known as symbol grounding. Accomplishing this can be quite nuanced because natural language instructions are often incomplete and natural language can be ambiguous.

A University of Texas at Austin team leveraged LEARN to aid RL to help an AI agent play Montezuma’s Revenge from the Atari Learning Environment, showing that agents that used language-based rewards lead to successful completion of the given task 60% more often on average, compared to learning without language.

**D. NLP and the Physical World**

A team at the University of Washington explored how NLP can demonstrate physical commonsense reasoning. Their study across a dataset of over 200k annotations suggested that modern NLP representations can still only learn associations that are explicitly written down.

For this project, the demonstration of an agent to be able to learn explicit associations is promising, given the deterministic Neutonian physics underlying the PHYRE environment and its consistent map environments and objects.

**E. NLP and Inference**

![Diagram](image)

**Figure 3:** This is an overview of the Verb Physics paper approach for an NLP-system to infer physical knowledge. A verb’s usage in the language (top) implies physical relations between objects it takes as arguments. This allows people to reason about properties of specific objects (middle), as well as the knowledge implied by the verb itself (bottom).

---


5 Maxwell Forbes, Ari Holtzmann, Yejin Choi. Do Neural Language Representations Learn Physical Commonsense?
Another paper from the University of Washington explored how a program can infer relative physical knowledge of actions from unstructured natural language text\textsuperscript{6}. The team used modeled changes in physical attributes entailed by verbs together with objects that exhibit these properties and were able to infer knowledge using a large, crowd-sourced data set.

Salesforce published a paper studying commonsense reasoning in deep neural networks (DNNs) using both human and auto-generated explanations including transfer to out-of-domain tasks\textsuperscript{7}. Their research found that language model could be effectively leveraged for common sense reasoning. They proposed Common Sense Auto-Generated Explanations (CAGE) as a framework for generating explanations for Common sense Question Answerings (CQA) using a new Common Sense Explanations (CoS-E) dataset, achieving a state-of-the-art 65\% accuracy on CQA.

This research about language models being used to answer common sense questions is promising to the task this project is tackling. For XAI to be applied to PHYRE, the agent needs to be able to infer the results of actions in a well-defined environment with well-defined rules.

\textbf{F. General Artificial Intelligence}

Related to the core of this project is the idea of creating a general artificial intelligence system (GAI) that can solve a range of PHYRE problems while also providing explanations.

GAI seeks to develop AI agents that can develop a wide range of competencies in a variety of challenging tasks\textsuperscript{8}. Deep Q-network (DQN), which combined reinforcement learning with deep neural networks, has shown promise in creating GAI-agents. DQN combined with XAI could be a powerful implementation for achieving the goals of this project.

\section*{III. Project Overview}

Explainable Artificial Intelligence is a growing field within computer science seeking to demystify the artificial intelligence black box of AI decision making. As AI continues shaping society, XAI will help solidify trust in novel algorithms and applications of AI-agents.

Facebook’s PHYRE playground provides a great stage for XAI research.

\textit{A. Models discussed in the PHYRE paper}

\textsuperscript{6} VERB PHYSICS: Relative Physical Knowledge of Actions and Objects
\textsuperscript{7} Explain Yourself! Leveraging Language Models for Commonsense Reasoning
\textsuperscript{8} Human-level control through deep reinforcement learning
To better understand baseline agents, the PHYRE paper discussed a number of agents that rank agents given an observation of the initial state and reported their efficacies in solving PHYRE-tasks (Figure 4).

a. Random Agent (RAND): This agent does not perform any training and instead samples actions uniformly at random.

b. Non-parametric (MEM): Generates and trains on a set of random actions and uses the simulator to check if the action solves the state. This agent uses a list of “memorized” actions at test time.

c. Non-parametric (MEM-O): Same training phase as MEM, but continues to learn online at test time. Upon finishing each test task, the agent updates parameters based on the reward it received.

d. Deep Q-network (DQN): DQN collects a set of observation-action-reward triplets by randomly sampling actions and simulating them. The agent trains a deep network to predict the reward for an observation-action pair.

e. Deep Q-network with online learning (DQN-O): In the same sense as MEM-O, DQN-O uses rewards from test tasks to perform online updates. Upon completing a task, the agent performs a number of gradient descent updates for the model for the next task.

B. Applying Reinforcement Learning

Reinforcement learning (RL) is a branch of artificial intelligence which lets machines learn on their own in a way different from traditional machine learning. RL is particularly useful where decision making is sequential, and the goal is long-term such as game playing.9

The general flow of RL is an agent takes actions in an environment and gets observations and rewards from it10. To implement RL, it is important to define the main components of an RL algorithm.

C. Modeling RL for PHYRE

a. Agents: The agent would be the AI that plays the game.

---

9 An introduction to Q-Learning: Reinforcement Learning, Floydhub.com

10 Schooling Flappy Bird: A Reinforcement Learning Tutorial, toptal.com
b. Environment: The PHYRE playground along with its deterministic Newtonian Physics.

c. Actions: The AI-agent can introduce one or more balls into the environment anywhere in the environment as long as it doesn’t collide with an object (as long as it doesn’t break a law of physics).

d. Rewards: This is where a lot of thinking needs to be done. Considering an XAI-agent is the goal, we can look to some of the aforementioned XAI research for inspiration for how to relate NLP to reward propagation in RL so that a selected action also outputs a rationale.

e. States: There are a lot of potential states with the PHYRE. Depending on the game being played, the AI-agent can introduce more balls into an environment, and the balls themselves can be in different positions on the environment and modify the environment according to the laws of physics.

**IV. Summary**

Applying XAI to the PHYRE environment presents a promising stage for research. This project seeks to demonstrate how XAI can be coupled with an AI-agent (like an agent that uses RL) to elucidate the mystical “black box” of AI.

This project is exciting because XAI is still a burgeoning field in AI, but one of extreme importance. A proof-of-concept of an XAI agent in PHYRE could open the door to the creation of other XAI agents in other fields.
References