WAKE FOREST UNIVERSITY
Department of Computer Science

Explainability for Safety in Reinforcement Learning

Tongtong Liu, Md Asifur Rahman, Joe McCalmon, **Sarra Alqahtani**
Department of Computer Science, Wake Forest University

The work is supported by NSF awards #2105007

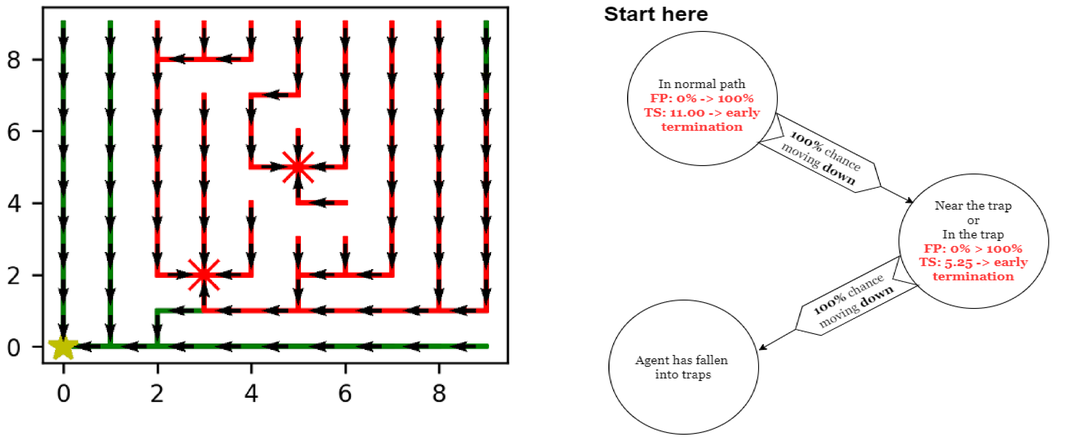
Introduction

Reinforcement learning (RL), a subfield of machine learning, can train autonomous agents to perform difficult tasks at a superhuman level. However, due to the nature of exploratory learning involved in RL agents, they are vulnerable to the adversarial threat which affects their usability in safety-critical applications. Moreover, these agents often learn behaviors that are inexplainable and unpredictable to humans.

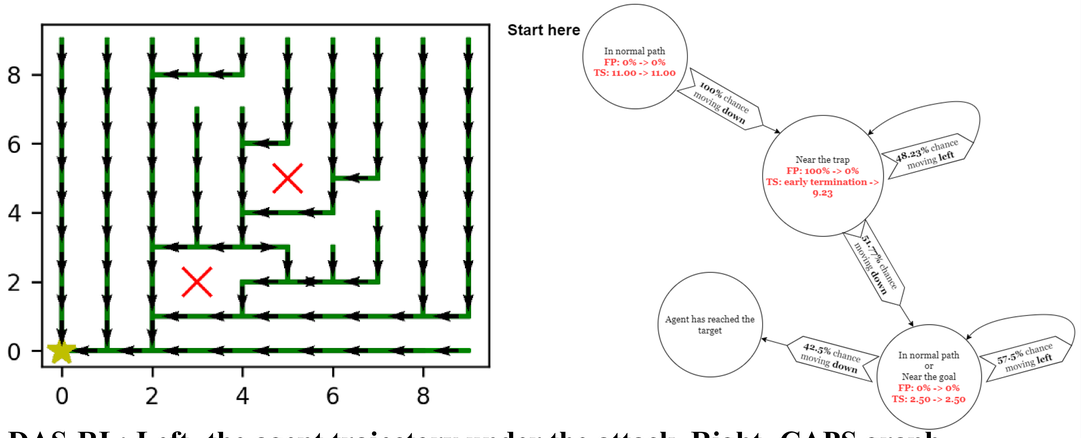
In this work, we demonstrate the usage of our explainable RL method, CAPS that can be integrated with existing RL algorithms to improve their agents' explainability with and without our proposed safety method.

Our method, CAPS, is a tool that can be used to analyze agent behavior and safety in critical situations. It can also be used in manual inspection, model checkers, or statistical analysis to elucidate the behavior of the underlying RL systems.

Results

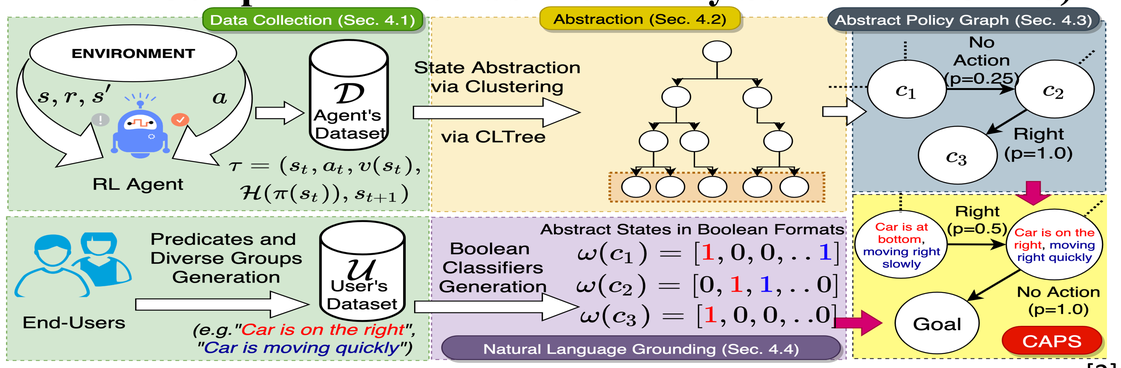


Unsafe RL: Left- the agent trajectory under the attack. Right- CAPS graph for the agent's behavior under the attack

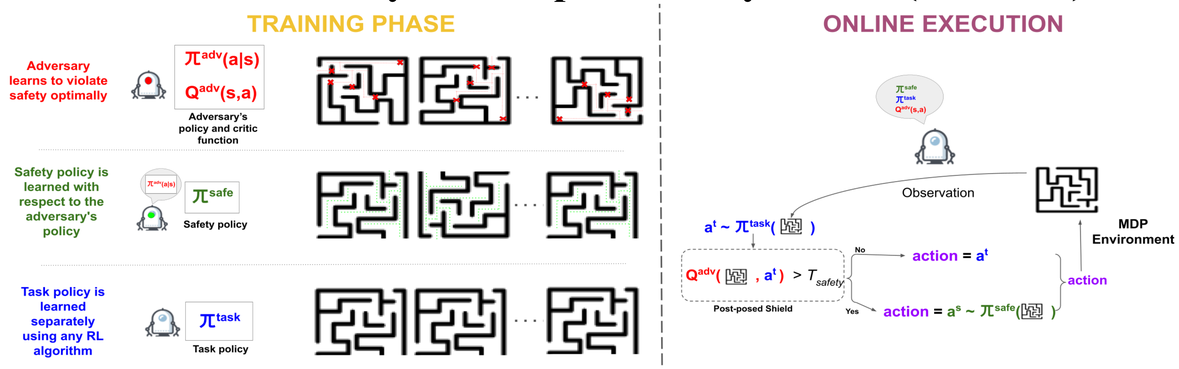


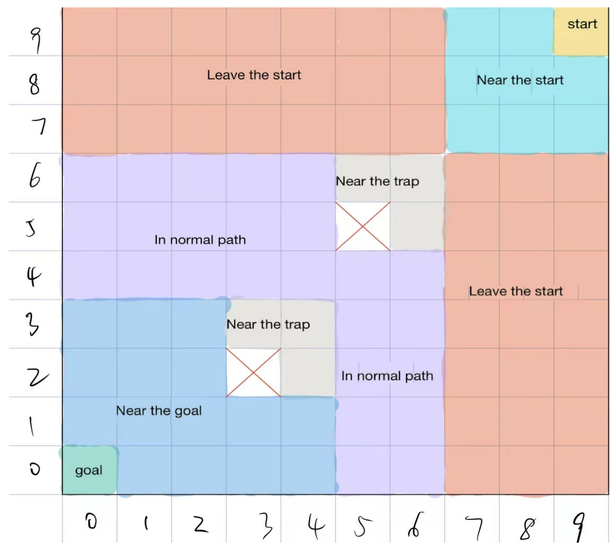
DAS-RL: Left- the agent trajectory under the attack. Right- CAPS graph for the agent's behavior under the attack

Comprehensible Abstract Policy Summaries CAPS)[1]



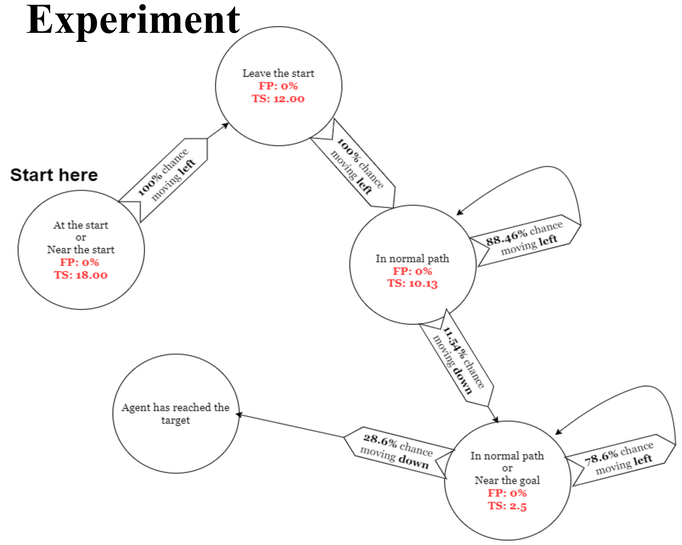
Diversity for Adaptive Safety for RL (DAS-RL)[2]





The environment with human-interpretable predicates

Experiment



CAPS graph of optimal RL policy without any attack

Citations

[1] Joe McCalmon, Thai Le, Sarra Alqahtani, and Dongwon Lee. 2022. CAPS: Comprehensible Abstract Policy Summaries for Explaining Reinforcement Learning Agents. In Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS '22). International Foundation for Autonomous Agents and Multiagent Systems.
[2] Md Asifur Rahman and Sarra Alqahtani.2023. Adversarial Behavioral Exploration for Safe Reinforcement Learning. Thirty-Seven AAAI Conference on Artificial Intelligence (AAAI-23). Under review.