AlphaSOC: Reinforcement Learning-based Cybersecurity Automation for Cyber-Physical Systems

Ryan Silva, Cameron Hickert, Nicolas Sarfaraz, Jeff Brush, Josh Silbermann, Tamim Sookoor
Johns Hopkins University Applied Physics Laboratory

ABSTRACT

Achieving agile and resilient autonomous capabilities for cyber defense requires moving past indicators and situational awareness into automated response and recovery capabilities. The objective of the AlphaSOC project is to use state of the art sequential decision-making methods to automatically investigate and mitigate attacks on cyber physical systems (CPS). To demonstrate this, we developed a simulation environment that models the distributed navigation control system and physics of a large ship with two rudders and thrusters for propulsion. Defending this control network requires processing large volumes of cyber and physical signals to coordinate defensive actions over many devices with minimal disruption to nominal operation. We are developing a Reinforcement Learning (RL)-based approach to solve the resulting sequential decision-making problem that has large observation and action spaces.

KEYWORDS

Reinforcement Learning, Cybersecurity, Cyber-Physical System

1 INTRODUCTION

Aging Critical Infrastructure (CI), composed of Cyber-Physical Systems (CPS), will be unable to support modern demands without increased inter-connectivity and intelligence to optimize performance. But, networking these systems and making them smart, through efforts such as Industrial Internet of Things (IIoT), exposes them to vulnerabilities and threats. There have already been a number of such attacks on systems such as power grids, water treatment plants, and oil pipelines. Attacks traditionally focused on IT systems, are now being deployed against CPS. In order to ensure critical infrastructure can be modernized to address the increasing demands while mitigating threats against them, cyber defenders need assistance, which approaches such as Active Cyber Defense (ACD) aim to provide by automating defensive capabilities.

AI enables novel advanced persistent threats (APTs) to overwhelm current generations of defenses by being fast, flexible, and adaptive [1]. The current industry standard is to manually predefine courses of action in response to alerts [2], but this approach often leaves complex decision-making to a human analyst, which increases the mean time to response (MTTR). In order to defend against these attacks, it will be necessary to create a corresponding class of active cyber defense measures that can intelligently command and control security mechanisms to respond to autonomous adversarial tactics, techniques, and procedures (TTPs) [3]. Through the AlphaSOC project, we are developing an RL-based approach to enhance ACD automation and tip the balance back in favor of defenders.

We trained an RL-based security operator to defend a ship's navigation control network from an agile attacker with the goal of disrupting the availability of critical devices. Previous work relies on a model of the attack and does not directly use a dynamic physical model [5]. We choose not to focus on the specifics of the threat model, but rather design a system that responds to the consequences of an attack. The simulation environment, while modeling a cyber-physical system within a ship, is similar to other environments used to train RL-based cyber defenders such as FARLAND [4]. We propose an agent trained using reinforcement learning due to its strengths as a general sequential decision making optimization algorithm that can potentially be applied to a wide range of cyber-physical systems. We have a video demonstration of this simulation and a trained RL defender in action.

2 ENVIRONMENT DESCRIPTION

Our ship simulation environment models a distributed control system with sensors (PVs), Proportional Integral Derivative (PID) controllers, and actuators (CVs) on a control network. This system is designed to minimize the error between the desired speed and heading and the actual speed and heading of the simulated ship. The details of the control system can be seen in Fig. 1.

The attacker has direct access to the network and performs a Denial of Service (DOS) attack on discovered devices. If successful, this prevents that device from sending and receiving data on the control network, resulting in loss of control and potentially causing significant deviation from the desired speed and heading.

The defender observes the desired heading and speed, the actual heading and speed, and anomalous network traffic alerts from an intrusion detection system (IDS). We assume for now that sensor readings are not influenced by the attacker, however we do model attacks that may go undetected by an IDS. We also model a small chance for an IDS to alert when there is no attack occurring (false positive). Based on the cyber and physical indicators, the defender can decide to investigate a device to determine if it is under attack or not, which lasts for a short amount of time during which the defender cannot take other actions. The defender can mitigate attacker actions by reconfiguring a device to a known safe state and changing its network address. This takes a device offline for a brief period of time while that device is being reconfigured.

3 METHODS

We represent the state of a device with a categorical distribution: "online," "offline, and "reconfiguring". The "reconfiguring" state means that the defender agent has chosen to take a mitigation action and that device is unresponsive while it reconfigures. The "reconfiguring" state is fully observable while the other states are hidden unless the defender agent takes an investigation action. We also calculate the current error between desired and actual heading and speed. We normalize these quantities as part of the defender agent observation. Any devices that are currently being investigated are also included in the observation.

978-1-6654-0967-4/22/$31.00 ©2022 IEEE
DOI 10.1109/ICCPS54341.2022.00036

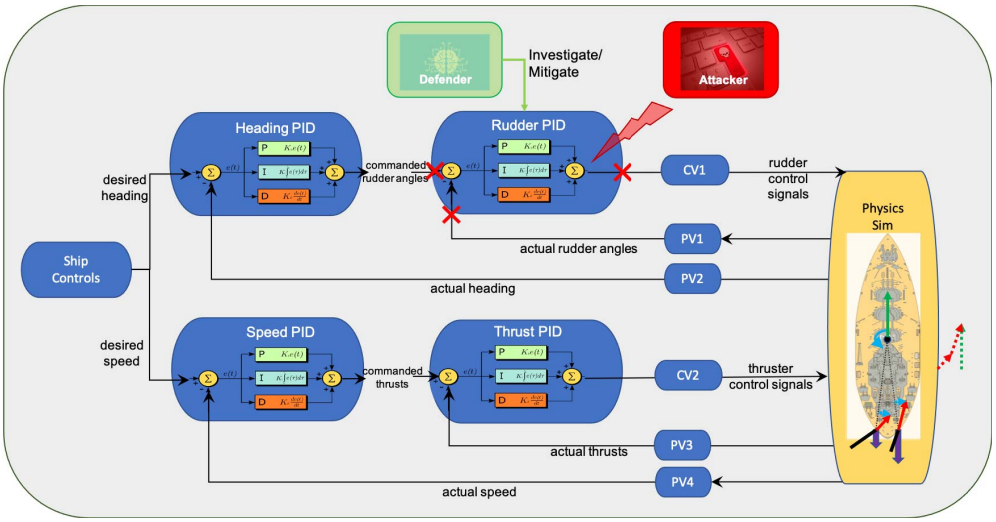

Figure 1: Simulation environment. Devices in blue form the navigation control system and are part of the attackable network. Arrows show the logical flow of information between devices. A physics simulator (yellow) receives rudder and thrust control signals and outputs sensor readings. The attacker (red) can disrupt all incoming traffic to a device, jamming the control loop and influencing the trajectory of the ship. The defender (green) can investigate and reconfigure a device to mitigate an attack.

The goal is to minimize disruption to the control network by minimizing the time devices are offline due to attack or reconfiguring due to defensive action. We formulate the reward function as shown in equation 1. Given the state s and action a, a reward of r (set to 1.0) is given for a correct mitigation a_m on a device d that is disrupted, while a penalty of w_d is given for each device that is disrupted by either being attacked or shuffled. The penalty w_d captures each devices' importance in the control network, which was calculated from experimental simulation data.

$$R(s,a) = \sum_{d \in devices} r\{(d\ disrupted)\&(a_m = d)\} - w_d\{d\ disrupted\} \tag{1}$$

We trained a reinforcement learning agent with the Proximal Policy Optimization algorithm, using the Stable Baselines 3 implementation. We split training up into episodes of 1000 time steps with each time step being the equivalent of 1 second of real time.

4 RESULTS

We observed that in 1500 training episodes, the defender agent learned a policy in the partially observable environment that approaches an optimal policy under fully observable conditions. This means that the defender learned to reduce the total time that devices on the network were disrupted compared to the random policy that it was initialized with. We evaluated the trained agent over 100 episodes and found the defender agent learned to take investigation actions a majority of the time, and on average attacks were discovered and mitigated within 25 timesteps.

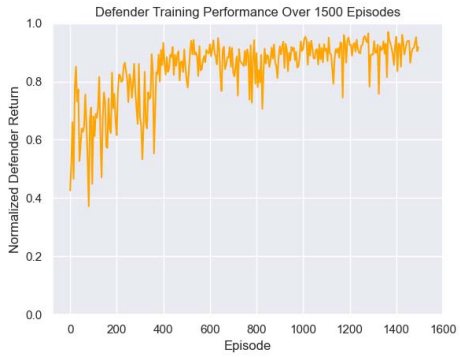

Figure 2: Training results. Episode returns are normalized such that a defender taking no mitigation actions would score 0, while an optimal defender that is immediately notified when the system is under attack would score 1

REFERENCES

[1] DarkTrace 2018. *The Next Paradigm Shift.* DarkTrace. Retrieved Jan 19, 2022 from https://www.oixio.ee/sites/default/files/the_next_paradigm_shift_-_ai_driven_cyber_attacks.pdf

[2] Andy Applebaum, Shawn Johnson, Michael Limiero, and Michael Smith. 2018. Playbook Oriented Cyber Response. In *2018 National Cyber Summit (NCS).* 8–15. https://doi.org/10.1109/NCS.2018.00007

[3] Neil Dhir, Henrique Hoeltgebaum, Niall Adams, Mark Briers, Anthony Burke, and Paul Jones. 2021. Prospective Artificial Intelligence Approaches for Active Cyber Defence. *arXiv preprint arXiv:2104.09981* (2021).

[4] Andres Molina-Markham, Cory Miniter, Becky Powell, and Ahmad Ridley. 2021. Network Environment Design for Autonomous Cyberdefense. *arXiv preprint arXiv:2103.07583* (2021).

[5] Saman A. Zonouz, Himanshu Khurana, William H. Sanders, and Timothy M. Yardley. 2014. RRE: A game-theoretic intrusion response and recovery engine. *IEEE Transactions on Parallel and Distributed Systems* 25, 2 (Feb. 2014), 395–406. https://doi.org/10.1109/TPDS.2013.211