Proceedings of the Annual General Donald R. Keith Memorial Conference
West Point, New York, USA
May 4, 2023
A Regional Conference of the Society for Industrial and Systems Engineering

Reinforcement Learning in Cyber Wargaming Defense

Joseph Chedzoy, D'Andre Tobias, Joseph Miller, Max Hwang, and Trent Geisler

Department of Systems Engineering
United States Military Academy
West Point, New York 10996

Corresponding author's Email: dandretobias16@gmail.com

Author Note: Cadets Joseph Chedzoy, Max Hwang, Joseph Miller, and D'Andre Tobias are Firsties (Seniors) at the United States Military Academy (USMA). A special thanks goes to LTC Trent Geisler in the Department for System Engineering for his wisdom and guidance. Also, thanks to our client for teaching and guiding us this past year. By request, we will not be referencing them directly within this paper. "The views in this paper are the views of the team and do not reflect the views of the United States Military Academy, Department of System Engineering, or the United States Army."

Abstract: In recent decades the necessity for cyber security has grown for both private companies as well as government agencies. This growth is the result of increasing ability for organizations to mount cyber-attacks. As a response, organizations have been developing cyber defense artificial intelligence (AI), which greatly improves cyber-security capabilities. This necessitates not only the development of cyber-attack, defense, and vulnerability frameworks to simulate a realistic environment, but also methods with which to train the AI. Further, the number and variety of networks necessitates a framework with which AI can be quickly and cost-effectively trained. This paper will explore how our team has worked to develop an efficient and comprehensive framework under which a variety of AI can be trained to fulfill the need for cyber resiliency.

Keywords: Cyber Wargaming, Reinforcement Learning, Ontology

1. Introduction

Cyber security is among the greatest considerations of modern organizations and governments. In a study done by Check Point Research, a recognized leader in cyber threat intelligence, they found that in 2020 the number of cyber attacks attempted on businesses increased by 50% (Mello, 2022). Given the increase in cyber attacks and their sophistication, it is no surprise that cyber security expenditures are expected to be as high as $172 billion in 2022 for organizations in the United States alone (Pratt, 2021). Other nations, such as Taiwan and Ukraine, have experienced a similarly dramatic increase in the number of attacks and their sophistication. In a 2020 article, the author found that Taiwan has suffered from cyber attacks since 1999 and that these attacks are sophisticated enough to jeopardize power systems and cause large-scale power outages like those seen along the front in Ukraine (Huang, 2020). Attacks and defense costs are only projected to grow, with little expectation of cyber security defense capabilities surpassing attack capabilities; this is a result of the inherent disadvantages that defenders experience within the current environment of cyberspace.

Cyber security professionals currently do not have adequate defensive tools necessary to respond to cyber attacks. The relative ineffectiveness of cyber security is due to the disadvantages faced by cyber defense. Wendt credits the disadvantage of defense to the nature of cyber conflict which favors the attacker (Wendt, 2019). He continues by stating the attacker identifies and exploits one vulnerability, while the defender must identify and mitigate every vulnerability. The attacker can choose when to initiate, while the defender must first identify an attack before they can react. Attackers can reuse similar attacks to exploit similar vulnerabilities, while defenders must coordinate across many unique systems to identify common vulnerabilities and develop mitigations. The attacker does not incur a significant cost if they fail, while defenders' networks can be significantly damaged if they fail. If defenders do not identify attacks and respond to them quickly, the attacker will have more time to penetrate deeper into the system. This allows the attacker to disrupt the network and create back-doors that can be used to further harm the network. In efforts to increase the response time and effectiveness of the defender, cyber security professionals practice using the methodology of cyber wargaming.

Cyber wargaming is a simulation of a real attack wherein an attacking team attempts to disrupt a network and a defending team attempts to prevent this. Cyber wargaming allows cyber security practitioners to develop their abilities, improve their networks, and ultimately respond faster and more effectively. Nonetheless, historical increases in cyber security expenditures and cyber wargaming alone have not proved sufficient. Though it improves response time and effectiveness, cyber wargaming does not enable cyber security professionals to react to threats quickly enough. Furthermore, the current cyber security methodology is extremely costly. The White House reported that the President allocated "$10.9 billion of budget authority for civilian cybersecurity-related activities," for the 2023 fiscal year. This is an "11 percent increase," from 2022, and shows how there is a strong need for the development of cyber security (White House, 2022).

Cyber defense practitioners can use reinforcement learning (RL), a subset of artificial intelligence, to develop agents that meet the need for efficient and near-instantaneous responses. The goal of reinforcement learning is to have the agent choose the optimal actions to a variety of different stimuli, which are provided reinforcement learning environmental stimuli. RL functions by using a set of possible actions updated by a RL environment. In this set of possible actions, there are optimal actions, which are dictated by the environment. The agent learns which actions are optimal and sub-optimal through rewards from an objective function. The agent attempts to find the correct responses to environmental stimuli to maximize the rewards from the objective function. Supplementary information on AI and RL can be found in Sewak, Sahay, and Rathore (2022). To train a defense agent using RL, environmental stimulus is needed to prompt responses by a defense agent. By using a cyber wargame methodology, where an RL attack agent simulates a cyber attack, stimuli can be provided to train a defense agent in an efficient manner.

Responding to the overwhelming and growing number of cyber threats across many unique networks is becoming increasingly expensive and difficult to be achieved by human cyber security practitioners. This study focuses on developing a comprehensive framework by which many RL defense agents can be trained. This framework will leverage cyber wargame methodology to simulate the interactions between attack and defense agents, allowing for rapid and cost-effective training of cyber defense agents.

2. Related Works

Though many studies have been done on using AI to supplement cyber security, there is no significant literature that has approached this problem from the same angle. While our problem is focused on creating a generic RL training framework using a knowlege base, other studies focus on creating a specific defense or attack agents with pre-established training data. In Nguyen and Reddi (2021), the authors discussed the application of RL agents to security methods for cyber-physical systems and in game theory simulations for defense strategies against cyber attacks.

To pursue their study of RL agents in the security of cyber-physical systems, they used a series of real world examples and attempted to apply a robust model as a solution. In this study, it was found that although they could construct an agent capable of providing effective autonomous defense, there were certain attack techniques that caused the autonomous system to fail. So, to improve the robustness of their agent, they had to restructure and retrain their agent (Nguyen & Reddi, 2021). This application of RL to cyber security, is feasible in the defense of a single or very related networks against a known attacker. However, to construct and train many separate agents using different data compiled from many attackers presents a cost-benefit issue. To be able to train many agents using a generic training framework, informed by industry standard knowledge base could provide greater resiliency to systems.

In Nguyen and Reddi (2021), the authors introduce a cyber security application of RL using game theory, wherein multiple defense agents are trained in simulations to respond to different types of attacks. In this study, they discuss training agents to respond to *jamming attacks*, *spoofing attacks*, and *malware attacks*. As with their application of RL agents to cyber-physical systems, agents trained by this method saw reasonable success. However, in each case, the agent required significant fine-tuning to properly respond to the specific type of attack it defended against. Furthermore, the agents struggled when the attacker inputs were randomized or not well documented in the data sets that the agents were trained on (Nguyen & Reddi, 2021). In real-world scenarios, an attacker has many different potential attacks that can be used. Our teams proposed training methodology, addresses this weakness by allowing many agents to be trained quickly, and by using a vast knowledge base that is constantly updated by organizations within the industry. As such, unless the attacker uses novel attack methods, the agent should be capable of responding. The next section discusses how this training framework was developed.

3. Methodology

Defense is a critical aspect of cyber wargaming, as it involves protecting computer systems and networks from attacks launched by adversaries. In cyber wargaming, defense is typically achieved through a combination of preventive measures, such as firewalls and access controls, along with reactive measures, such as incident response and recovery. The goal of defense in cyber wargaming is to minimize the impact of attacks and to prevent adversaries from causing significant damage to the system or network. To achieve this, defenders must understand the tactics, techniques, and patterns used by attackers and develop effective countermeasures that can adapt to changing threat landscapes. By developing an attack AI, many attack scenarios can be simulated, thus allowing a defense agent to be trained.

3.1. Ontologies

To better understand the relationships between the attack and defense, an ontology was developed. An ontology is a graphical representation of the relationships, both direct and inferred, between concepts. Because the focus for this project was to develop a framework by which defense agents can be trained using stimuli from an attack agent, the rules on how the two agents interact must be established to build the cyber wargame simulation. The complex network of relationships is best understood in an ontology, which illustrates the relationship between the network, the attacker, and the defense. This ontology can be seen in Figure 1. The network vulnerability (also known as a Common Weakness Enumeration, or CWE) is the factor within the network that can be exploited. CWE's are exploited by an adversary's attack pattern. Each attack pattern is assigned a number or a Common Attack Pattern Enumeration & Classification (CAPEC). The CAPEC is the classification of the pathway that is utilized by an attack technique. The attack technique implements a more general and overarching attack tactic to implement the attack pattern and exploit the vulnerability.

The concept used for this framework and informed by the ontology in Figure 1 will remain constant. The attack agent chooses an attack technique that uses a CAPEC to exploit a CWE. The defense agent will then use a defend technique to defend against the attack technique. Each action and reaction update the environment and prompts the other agent to act in a turn-based simulation. However, the complexity of a cyber attack is such that in their final form an attack or defense agent could each potentially have thousands of possible actions. The goal of this project is to create a minimum viable product as a framework that can be appended to. So, for the time being, a small subset of attack techniques and corresponding defend techniques were chosen as a proof of concept. The selected subset of attack techniques are *Default Accounts*, *Web Session Cookies*, and *Software Deployment Tools*. The optimal defend techniques selected to defend against the attack techniques respectively are *Account Locking*, *Inbound Traffic Filtering*, and *Active Directory Configuration*. In addition to these techniques, several other techniques were selected randomly so that the agents would have to explore its options and choose the optimal techniques. Having established the possible techniques that the agent can use in the cyber wargame simulation, we progressed to develop the interactivity between agents.

From the ontology in Figure 1, defend techniques are used to defend against attack techniques. The defense agent responds to the actions taken by the attack agent by implementing a defense tactic that will combat the attack agents' corresponding action. There are several possible defend techniques that the defense agent can choose from. Along with providing the properties of each technique and tactic, the ontology provides the direct relationships between the attack and defend agents. Specifically, the ontology informs whether a defend technique is the proper choice to defend against each attack technique. The optimal answers of defend techniques are being *Account Locking* for *Default Accounts*, *Inbound Traffic Filtering* for *Web session Cookies*, and *Active Directory Configuration* for *Software Deployment Tools*.

In its final form, the defense agent would have more possible answers to choose from with varying rewards as some options are sub-optimal. However, to establish a functioning framework, the agent was developed in the least complex manner.

This forces the agent to learn that these options are the optimal answers for the specific attack agent. By rewarding the defense agent for choosing the optimal defend technique, the agent's behavior will be reinforced and it will train itself to correctly counter the attack agent. However, cyber wargaming plays out similarly to a turn-based game. The attack agent must act, then the defense agent must respond. Our team helped to develop the framework of interactions by which the agents could wargame in order to train the defense agent.

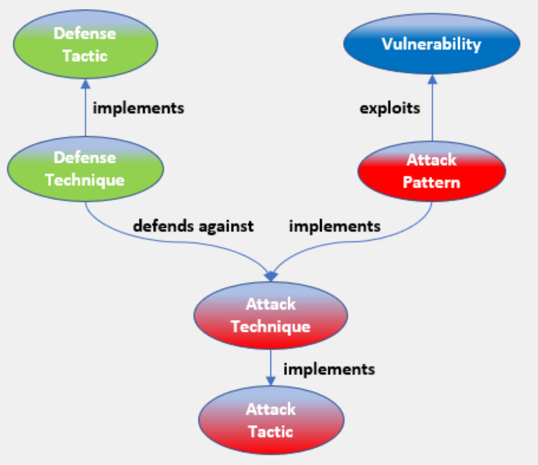


Figure 1: Cyber Wargame Ontology. The ontology is how the attack techniques and defense techniques interact with one another through attack patterns and vulnerabilities. This is a simplified version of the overall ontology.

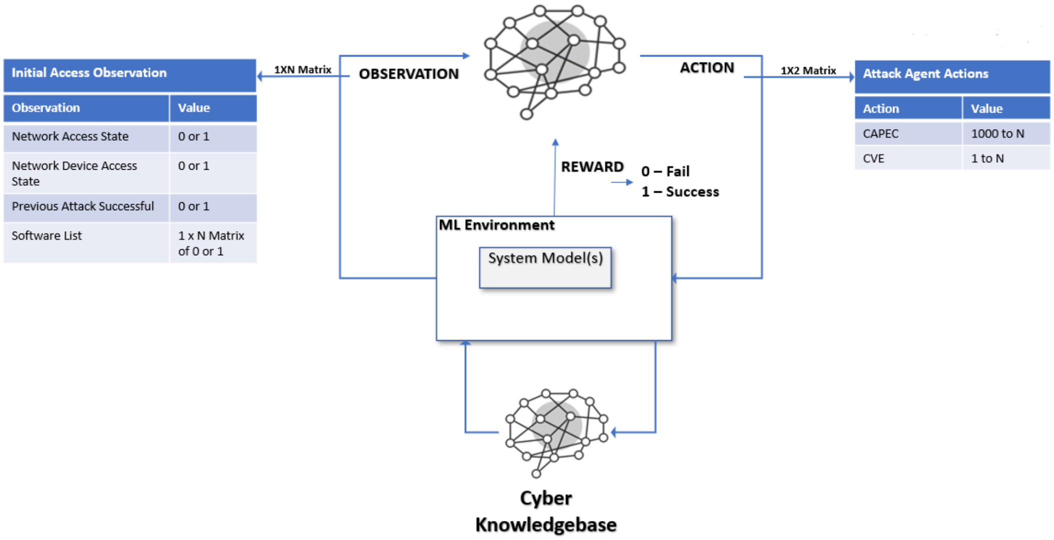
3.2. Creating the Scenario

In the ontology, Figure 1, the defense and attack agents interact through the defend techniques defending against attack techniques. Our team selected attack techniques to be used by the attack agent and defended by corresponding defend techniques. When building the defense agent, there was a constant balance between realistic complexity and necessary simplicity. There are hundreds of attack techniques that could have been, and eventually will be added to the attack agent. However, to keep troubleshooting manageable, we limited the number of attack techniques to three, each of which have a corresponding defend technique. Though only having three possible attack techniques is not realistic, it establishes a framework that can be expanded upon. Similarly, a defense agent would typically have over 150 possible defense techniques at any time, but only 10 possible responses were provided in the development of the prototype. It is also important to note that typically, when training an agent using RL, the optimal choices are unknown. However, to establish a proof of concept, our team selected known optimal actions as informed by industry standards (*MITRE ATT&CK Enterprise Matrix*, n.d.)

The three attack techniques selected each correspond to different phases of an attack. The first phase is initial access, wherein *Default Accounts* is the attack technique utilized by the attack agent. Initial access is an attack tactic that is composed of many attack techniques, the goal of these techniques is to allow an attacker to access a network. Specifically, *Default Accounts* is an attack technique that targets unmonitored accounts that exist on a network in order to gain access. The corresponding defend technique is *Account Locking*, which prevents a user from accessing the network on a specified account. The next phase of an attack is how the attacker moves through a network to reach a target node and is conducted by the lateral movement attack tactic. The attack technique selected for the agent to utilize in this phase is *Web Session Cookies*. The *Web Session Cookies* technique targets currently activated cookies on web browsers. These cookies allow the agent to bypass authentication as the authentication is already active on the web session. The corresponding defense technique is *Inbound Traffic Filtering*, which prevents web sessions from being exploited in this way. The final phase of the attack is to run an executable file on the target node in order to harm the network or gain access to sensitive information. The attack tactic for this phase is execution. The selected attack technique is *Software Deployment Tools* which are executable files run on a target node that can harm the network in many ways. The execution of this file can be prevented by the *Active Directory Configuration* defense technique. After having selected the attack and defend techniques, the logical relationship between both agents is established and the interaction between the two can be developed. (*MITRE D3FEND*, n.d.)

3.3. Machine Learning Environment

The interaction framework between the attack and defense agent is captured in a RL environment. From Figure 2, we see that the program runs a loop between the environment and the RL agent. The loop starts at the environment and provides the agent with the information of the network. The environment references the cyber knowledgebase to check the vulnerability of each node for later use when the agent creates an action. The information given compiled in a matrix which contains the network access state, network device access state, and the software list. For each part of the list, the network access state is defined as a 1 or 0 representing if the agent has access to the network (1) or the agent does not have access (0). This is similar to the network device as it is either 1 or 0. The software list represents the specific node that the agent is attacking, as for specific attacks it requires a specific software and if the node does not have it, the attacker needs another method to gain access. Once the agent takes its action it will choose an attack pattern and vulnerability to attack to gain access to the specific node. After the agent chooses an action, the environment will be notified and update the environment and give the agent a reward according to if the agent's choice was correct or not. The program will repeat this loop until the agent gains command and control over the target node. For initial training there was an assumption put into place, which is that there is only one optimal answer per node. Without this assumption the agent would learn to take an action we do not want it to learn, such as not acting or choosing a sub-optimal action.



Initial Access Observation	
Observation	Value
Network Access State	0 or 1
Network Device Access State	0 or 1
Previous Attack Successful	0 or 1
Software List	1 x N Matrix of 0 or 1

Attack Agent Actions	
Action	Value
CAPEC	1000 to N
CVE	1 to N

Figure 2: RL Environment. The figure above depicts an overview of the interaction between the agent, environment, and cyber knowledgebase.

Figure 2 only represents one agent acting within the environment. The goal of this loop is to have an attack and defense agent responding to one another in a manner that simulates a real network cyber attack and defense. The attack agent will attack the network, which in this simulation, alters the environment. The environment notifies the defense agent of which node was attacked, which leads to the defense agent responding. For example, when the attack agent attempts to access the network, it will use *Default Accounts* on the initial node. The defense agent will be notified that there is an anomaly within the system on the node that was attacked and use their response to attempt a defense technique. The defense agent will not know the specific attack, but will know through inferred pathways in the ontology what CWE's are present on that node and therefore what attack techniques could be used. If the defense agent uses the optimal defense technique, Account Locking, to evict the attack agent from the network, it will be successful. If it chooses a different defense technique to respond, the defense agent will fail. When the optimal defense technique is chosen, the defense agent is rewarded by the environment. This reward is the number that the objective function of the agent is attempting to maximize. Through many iterations of this loop, the defense agent explores the effect that each possible response has based on the information given by the environment. By rewarding the defense agent for choosing Account Locking as a response to the *Default Accounts* attack technique, this behavior is reinforced. The agent will understand that this choice increases its reward, and will learn the appropriate response. The feedback loop between the environment and agents functions similarly in the lateral movement and executing phases of the cyber wargame simulation. Through this feedback loop, the defense agent can be train itself to respond effectively to cyber attacks in an efficient manner.

4. Results

This project has established a proof of concept for a for a generic RL cyber agent training framework. Using this modular framework, a defensive RL agent can be quickly trained in a cost-effective manner to defend a variety of different networks. Though only a fraction of the possible attack and defense techniques are in the current knowledge base, this framework can be expanded upon, while keeping the same functionality. Once the defense agent is trained on every existing attack technique, the only remaining threat will be from the occasional error made by the agent and by newly developed attack techniques. However, with lateral development and an expanded knowledge base, this training framework may revolutionize the field of cyber security. In other similar studies, the focus is training a defense agent to defend a specific network. However, by focusing on a generic training method, many varying defense agents can be trained. Because the focus of this project was to develop a minimum viable product, there will be significant improvements to be made before this training framework is viable.

5. Conclusion & Future Works

Reinforcement learning, with additional development, could prove to be a valuable tool in the field of cyber wargaming. By allowing agents to learn and adapt based on their interactions with simulated environments, it provides a powerful framework for simulating and analyzing complex security scenarios. With its ability to handle uncertainty and learn from experience, reinforcement learning is well-suited for modeling the dynamic and evolving nature of cyber threats. The results of our study demonstrate that reinforcement learning algorithms can effectively learn to defend against cyber attacks, improving over time as they gain more experience. As the field of cyber security continues to grow and evolve, we believe that reinforcement learning will play an increasingly important role in helping organizations understand, anticipate, and respond to cyber threats. A major limitation put in place within this project was the limiting of choices for the attack and defend agents to select. This was done to scale down the problem. A limitation to this project was the defense agent. The detection aspect of the defense agent was hand-waved, allowing the agent to automatically detect the attacker rather than creating a detection agent for the simulation. This was incorporated due to the amount of time allotted on this project as well as to enable the focus of the simulation to be between the attack agent and defense agent rather than rely on a detection agent that occasionally may have not detected the attacker. Another limitation was the scope of the original project. After the defense was decided as the focus of the paper the team had to narrow down the project based off the original attack agents. The attack techniques were decided from the original attack tactics from when the team received the project. From there it was predetermined what attack tactics the techniques would be chosen from. Future work that could be possible is to work on a more complex environment as it requires both agents to learn the network before competing against each other. The attack agent then can learn different routes to the decision node rather than following a mission plan set by the operator of the code that is predetermined. It will cause for a greater learning curve for both agents creating more advance agents capable of traveling a network and being able to defend or attack it without needing a mission plan.

References

Huang, H. (2020). A collaborative battle in cybersecurity? threats and opportunities for taiwan. *Asia Policy*, *15*(2), 101 - 106. Retrieved from `https://search.ebscohost.com/login.aspx?direct=true&db=poh&AN=143225698&site=ehost-live&scope=site`

Lapan, M. (2018). *Deep reinforcement learning hands-on: Apply modern rl methods, with deep q-networks, value iteration, policy gradients, trpo, alphago zero and more*. Birmingham: Packt Publishing, Limited.

Mello, J. P. (2022). Software supply chain attacks hit three out of five companies in 2021. *CSO Online*.

Mitre att&ck enterprise matrix. (n.d.). `https://attack.mitre.org/matrices/enterprise/`. (Accessed: 2022-09-27)

Mitre d3fend. (n.d.). `https://d3fend.mitre.org/`. (Accessed: 2022-11-08)

Nguyen, T. T., & Reddi, V. J. (2021). Deep reinforcement learning for cyber security. *IEEE Transactions on Neural Networks and Learning Systems*.

Pratt, M. K. (2021). Cybersecurity spending trends for 2022: Investing in the future. *CSO Online*.

Sewak, M., Sahay, S. K., & Rathore, H. (2022). Deep reinforcement learning for cybersecurity threat detection and protection: A review. In *Secure knowledge management in the artificial intelligence era: 9th international conference, skm 2021, san antonio, tx, usa, october 8–9, 2021, proceedings* (pp. 51–72).

Smeets, M. (2018). The strategic promise of offensive cyber operations. *Strategic Studies Quarterly*, *12*(3), 90–113.

Wendt, D. (2019). Addressing both sides of the cybersecurity equation. *Journal of Cyber Security and Information Systems*, *7*(2).

White House. (2022). *Analytical perspectives: Budget of the u.s. government fiscal year 2023.* PDF. Retrieved from `https://www.whitehouse.gov/wp-content/uploads/2022/03/ap_16_it_fy2023.pdf`