PPO-Based Cyber Defense Agent Training with Adaptive Decoy Deployment

Zhenhong (Kevin) Zhong*, Jie Gao*, and Thomas Kunz†
*School of Information Technology, Carleton University, Ottawa, ON, Canada, K1S 5B6
†Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada, K1S 5B6
Emails: {kevinzhong@cmail.carleton.ca, jie.gao6@carleton.ca, tkunz@sce.carleton.ca}

***Abstract*—This poster explores the use of PPO to train an autonomous cyber defense agent within the CybORG framework. Our agent features a novel decoy deployment strategy that dynamically deploys and reallocates decoys based on adversary behavior, leveraging scan-state tracking and adversary recognition to adjust the defense against targeted and exploratory attacks. Experimental results demonstrate that PPO-based cyber defense agent training enhances adaptivity, highlighting its potential for real-world AI-empowered security applications.**

I. Introduction

As cyber threats evolve, advanced defense mechanisms become essential. Traditional cyber defense struggles against adversaries that can adapt their tactics, while Deep Reinforcement Learning (DRL) is a promising approach due to its ability to dynamically learn, adapt, and optimize defense strategies [1]. Among DRL methods, Proximal Policy Optimization (PPO) offers a balance between sample efficiency and stability, making it well-suited for cyber defense applications [2].

This poster presents a PPO-based cyber defense agent training approach. A key innovation is adaptive decoy deployment that dynamically positions decoys based on observed attack patterns. The trained agent learns to deploy decoys strategically, prioritizing critical assets against targeted attacks while distributing defense resources against broader threats.

Our approach integrates adversary recognition and scan-state tracking to enhance decision-making. Experimental results show that our agent can outperform benchmarks including Hierarchical PPO (HPPO) and Ensembled Dueling Double Deep Q Network (DDDQN) baselines, demonstrating effectiveness and adaptability in various attack scenarios.

II. Network Scenario

A general network with interconnected hosts, servers, and security infrastructure is illustrated in Fig. 1. A red agent (attacker) and a blue agent (defender) interact, with each action altering the network state. The red agent can perform reconnaissance, exploitation, and lateral movement to compromise systems, either executing targeted attacks (B_line) or exploring broadly (Meander), depending on the selected red agent type. The blue agent counters with actions such as analysis, malware removal, system restoration, and decoy deployment [3].

Multiple types of decoys can be deployed based on host vulnerabilities and attacker behavior. The blue agent should optimize decoy placement to maximize their effectiveness in disrupting the strategy of the red agent.

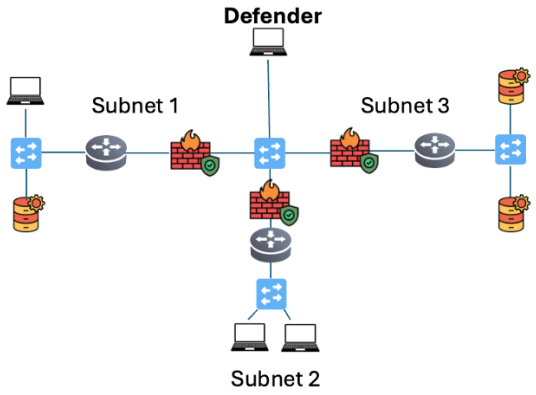

Fig. 1: An Example Network Scenario for Cyber Defense

The reward, which is always negative, penalizes system compromises (e.g., when the red agent gains access to hosts or servers) depending on the level of the corresponding impact.

III. Key Components in our PPO-Based Approach

Fig. 2 shows the workflow of our PPO-based agent training, with the following four key components.

1) *Enhanced State Tracking*: Our agent maintains a 10-element state vector that tracks network scanning history across all hosts. This vector separately records recent and older scans, using a mechanism which updates the status of scan records as new scans occur. Such temporal state tracking allows the agent to recognize scan patterns that indicate specific attack strategies and maintain situational awareness across the network.

2) *Adversary Behavior Recognition*: Our training approach efficiently identifies attacker types by using sum-based heuristics on our 10-element state vector. It distinguishes between targeted attacks by a B_line red agent and broader exploratory compromises by a Meander red agent via examining the scan record. Upon identification, the agent selects appropriate defensive strategies tailored to the specific red agent type, improving defense effectiveness against different threats without incurring a high complexity.

3) *Strategic Decoy Deployment*: We train the blue agent to implement a structured decoy deployment strategy using a host-specific priority design. A greedy_decoy dictionary associates network hosts with tailored decoy actions, allowing the agent to mislead attackers by deploying decoys where they would be effective [4]. It also tracks deployed decoys to

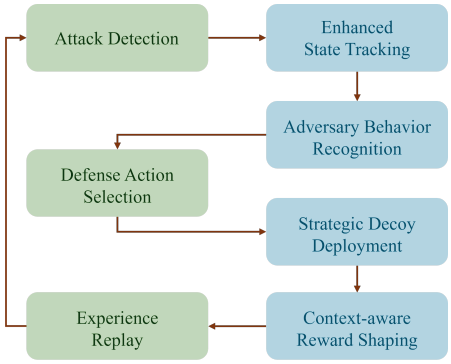

Fig. 2: Workflow of the PPO-based Agent Training. The 3 blocks on the left represent standard training steps, while the 4 blocks on the right highlight our design.

prevent redundancy while ensuring comprehensive coverage for the nodes across the network.

4) *Context-aware Reward Shaping*: We design a reward mechanism, used only for the training phase, that considers the temporal context, the appropriateness of the action, and the progression of the attack. It provides incentives for early decoy deployment, bonuses for successful analysis actions when anomalies are detected, penalties for unnecessary interventions and premature system restoration, and substantial rewards for complete attack mitigation. This helps shape the blue agent behavior toward the optimal defense against evolving threats.

IV. Experiments and Results

Our PPO-based approach was tested within the CybORG environment, a simulation framework designed for evaluating cyber defense strategies. CybORG provides a controlled and reproducible setting for training and testing autonomous defense agents against adversaries [5], [6].

A. Agent and Environment Setup

PPO Architecture: The PPO-based agent employs a three-layer actor-critic network with a 64-64 hidden structure. It processes 52 environmental features and a 10-dimensional state vector, merging real-time state observations with scan records for improved situational awareness. The actor outputs a probability distribution over 36 actions, including 27 defense actions unrelated to decoys (such as analysis, restore, etc.) and 9 actions dedicated to decoy deployment, while the critic estimates the state value.

Experimental Methodology: The agent was trained for 100,000 episodes using experience batches of 512 steps for policy updates, with a discount factor of 0.99 to balance immediate responses with long-term rewards. Two types of blue agents were trained to defend against the B_line red agent and the Meander red agent, respectively.

B. Training Results and Discussions

The proposed agent was evaluated against two benchmarks, an HPPO agent (ranked 11 in the global leader board of TTCP Cage 2 submissions [5]) and a DDDQN agent (ranked 17). As shown in Table I, our agent achieved competitive reward scores against both types of red agents, outperforming both benchmarks in half of the cases, demonstrating effective defense.

TABLE I: Performance Comparison of Different Agents

Steps	B_line Agent			Red Meander Agent		
	Proposed Agent	HPPO	DDDQN	Proposed Agent	HPPO	DDDQN
30	-5.35	-4.44	-5.87	-5.54	-6.57	-10.93
50	-12.96	-7.70	-10.82	-8.82	-11.78	-26.29
100	-36.89	-15.68	-23.96	-16.47	-29.52	-67.89

Competitive results were achieved due to the following:

- By tracking network scan records over time, the PPO-based agent recognized recurring attack patterns and adjusted its defenses accordingly. This improved the adaptivity of the agent.
- The PPO-based agent demonstrated efficient learning, with rapid early improvements that stabilized after approximately 30,000 episodes. The enhanced state tracking mechanism in Fig. 2 played a crucial role in guiding policy development toward effective defense strategies.
- To counter targeted attacks by a B_line agent, the PPO-based agent learned to strategically deploy decoys on a few hosts to effectively thwart the red agent's access to the critical server. Against exploratory attacks by a Meander agent, our agent learned to deploy decoys more broadly, i.e., across multiple subnets while covering critical infrastructure.

V. Conclusion

This poster shows that our PPO-based agent provides effective and adaptive defense against two red agent types by combining state tracking, behavior recognition, strategic decoy deployment, and context-aware reward shaping. Our approach achieves satisfactory performance, beating approaches from the leaderboard in some cases. Designed for early-stage detection and decoy deployment, our approach offers timely disruption of reconnaissance activities. By prioritizing rapid scan pattern recognition and decoy placement, we intentionally de-emphasize resource intensive containment and forensic actions, favoring early-stage interception over deep remediation of established intrusions. This trade-off aligns with our reconnaissance-focused threat model. For future work, active post-compromise responses, live system isolation, or automated patching may be considered.

References

[1] A. Molina-Markham, C. Miniter, B. Powell, and A. Ridley, "Network Environment Design for Autonomous Cyberdefense," arXiv:2103.07583, 2021.

[2] M. Wolk, et. al., "Beyond CAGE: Investigating Generalization of Learned Autonomous Network Defense Policies," *NeurIPS RL4RealLife Workshop*, New Orleans, U.S., 2022.

[3] M. Kiely, D. Bowman, M. Standen, and C. Moir, "On Autonomous Agents in a Cyber Defence Environment," in *Proc. 2nd Int. Workshop Adaptive Cyber Defence* Melbourne, USA, 2023, pp. 1-8.

[4] E. Bates, V. Mavroudis, and C. Hicks, "Reward Shaping for Happier Autonomous Cyber Security Agents," in *Proc. 16th ACM Workshop Artif. Intell. Secur.*, Copenhagen, Denmark, 2023, pp. 221-232.

[5] "TTCP CAGE Challenge 2" [Online]. Available at: https://github.com/cage-challenge/cage-challenge-2

[6] M. O. Farooq and T. Kunz, "A Generic Blue Agent Training Framework for Autonomous Cyber Operations," in *Proc. IFIP Networking*, Thessaloniki, Greece, 2024, pp. 515-521.