HC-MARL: A General Hierarchical Cascaded Multi-Agent Collaborative Architecture for Cyber Wargaming Applications

Zhiqiang Qu , Jun He * , Bo Wu , Zhitao Long , Tao Xia

Posted Date: 21 May 2026

doi: 10.20944/preprints202605.1354.v1

Keywords: HC-MARL; multi-agent reinforcement learning; hierarchical reinforcement learning; command and control; cyber wargaming

 doi:10.20944/preprints202605.1354.v1

Article

HC-MARL: A General Hierarchical Cascaded Multi-Agent Collaborative Architecture for Cyber Wargaming Applications

Zhiqiang Qu [1,2], Jun He [2,*], Bo Wu [2], Zhitao Long [2] and Tao Xia [2]

[1] National University of Defense Technology, Changsha 410073, China

[2] Information Support Force Engineering University, Wuhan 430000, China

* Correspondence: hejun17c@nudt.edu.cn

Abstract

Cyber wargaming serves as a core tool for simulating cyber confrontations and supporting operational decision verification. Existing multi-agent methods face issues such as the absence of cross-level mechanisms, poor adaptability to dynamic environments, and inefficient collaboration when applied to cyber wargaming. To address these challenges, we innovatively propose HC-MARL, a general hierarchical cascaded multi-agent reinforcement learning architecture tailored for cyber wargaming. Agents are modeled as hierarchical cascaded units to achieve structural decoupling, while a cross-level bidirectional information transfer and threat-sharing mechanism enables command propagation and adaptation to dynamic node changes. Specifically, a Transformer-based message transformation function is designed to resolve bottom-up information fusion and alignment; a policy function integrating neural networks with empirical knowledge is constructed to enhance threat response efficiency while ensuring smooth transmission of up-level commands; and a reward mechanism combining global and local rewards as well as outcome-based and staged rewards is introduced to improve the stability of multi-agent policy learning. To the best of our knowledge, the proposed HC-MARL framework is a novel general hierarchical collaborative multi-agent architecture for cyber wargaming. Experimental results demonstrate that the architecture effectively addresses the challenges associated with cross-level information transfer and dynamic agent changes in cyber wargaming. Compared with methods such as Singh et al., the 10-episode average reward is improved approximately by 30%, and the policy converges faster and more smoothly.

Keywords: HC-MARL; multi-agent reinforcement learning; hierarchical reinforcement learning; command and control; cyber wargaming

1. Introduction

Cyberspace has emerged as the fifth operational domain, following land, sea, air, and space, and cyber confrontation has become a critical component of modern warfare. As a simulation tool that integrates offensive and defensive technologies, game theory, and operational coordination, cyber wargaming constructs virtual cyber confrontation scenarios to effectively simulate the attack and defense actions, decision-making processes, and confrontation outcomes of red and blue teams, thereby providing crucial support for the formulation of cyber operation plans, operational capability assessment, and personnel training. In particular, cyber wargaming exercises can construct a set of hypothesized capabilities, tools, and vulnerabilities using open-source information. These assumptions allow participants to gain more direct insight into how an adversary would act if equipped with such postulated resources and capabilities, and how the defending side should respond accordingly.

Against the backdrop of artificial intelligence empowering cyber offense and defense, it is foreseeable that cyber confrontation will increasingly become automated and intelligent in the near

 doi:10.20944/preprints202605.1354.v1

future. AI-in-the-loop constitutes a fundamental requirement for cyber wargaming, with agent-versus-agent confrontation and human–agent hybrid confrontation serving as the essential modes of such exercises. Currently, no publicly available intelligent cyber wargaming system has been reported. While other domains, such as land, naval, and air wargaming, have been effectively integrated with intelligent adversarial gaming to meet the demands of intelligent warfare, we argue that intelligence has already become a defining attribute of contemporary cyber offense and defense. Therefore, exploring and investigating intelligent cyber wargaming is of considerable significance.

In the process of cyber defense, operational actions often exhibit multi-task, multi-objective, and cross-level characteristics. These actions—such as network surveillance, threat analysis, incident response, and security hardening—may be executed by different operational units (or agents) and are bound by tight logical interdependencies and resource constraints. In essence, the intelligent offensive–defensive adversarial game in cyber wargaming constitutes a multi-agent cluster task planning and collaborative optimization problem.

In recent years, multi-agent reinforcement learning (MARL) has achieved a series of breakthroughs in autonomous cyber defense[1–8]. In particular, hierarchical multi-agent reinforcement learning (HMARL) has demonstrated considerable potential in addressing complex network environments, multi-task autonomous coordination, algorithmic stability, and convergence speed[9–13]. Currently, two main approaches to HMARL exist in autonomous cyber defense. The first approach stratifies the agents' policies; for instance, Singh et al.[9,14] categorized policies into a master policy and sub-policies. The second approach constructs a two-layer agent architecture, where the upper-layer agent is responsible for selection or planning and the lower-layer agent is responsible for concrete execution [10,11]. In the hierarchical method proposed by Tang et al., agents in both layers share the same global observations, and the rewards are shared between the two layers. These rewards consist of the reward for upper-layer selection, the reward for lower-layer actions, and the cost of actions.

However, these hierarchical architectures are not suitable for cyber wargaming for several reasons. First, they fail to capture the characteristics of hierarchical command, making effective bidirectional information transfer between levels impossible. Second, the rigid structure of current multi-agent frameworks cannot accommodate normal combat attrition and dynamic adjustments that occur during cyber wargaming exercises. Third, the inherent long-horizon and sparsity of shared rewards lead to instability in learning.

To design a general hierarchical collaborative framework that accommodates the cross-level command characteristics of cyber wargaming and enables flexible and efficient multi-agent coordination, the following issues must be further addressed. First, at the hierarchical level, the upper layer should be capable of controlling (or influencing) the actions of the lower layer through commands, while the lower layer should provide feedback on the status and outcomes of its actions to the upper layer; therefore, a cross-level bidirectional information exchange mechanism needs to be devised. Second, the layers in the architecture represent distinct hierarchical levels, and the decision-making scope and observed information vary according to the level at which an agent resides. Third, the architecture should allow flexible horizontal and vertical expansion to accommodate diverse operational scenarios and personnel configurations.

We argue that both biological systems and human cyber defense activities address complex task allocation and coordination problems through flexible, multi-scale hierarchical organization, and that future intelligent defense systems should more closely resemble an organized society of agents. In light of this, we draw on the TAG architecture [15] and design a general-purpose hierarchical collaborative multi-agent architecture tailored for intelligent cyber wargaming scenarios, as illustrated in Figure 1. The agent at layer $l+1$ can be likened to the commander-in-chief of cyber defense; the two agents at layer l are regional commanders, each responsible for defense tasks in different network regions. They receive commands from the superior layer, feed back the task results to the superior, and, when necessary, mutually communicate alert information. The agents at layer $l-1$ are responsible for the actual execution of actions; they likewise receive commands from the

 doi:10.20944/preprints202605.1354.v1

regional commanders, report the action outcomes back to them, and communicate alert information to one another as needed. Moreover, the number of agents at each layer is variable, enabling the construction of agent collectives with variable depth and variable width based on this architecture. Notably, this structure aligns well with the command system of cyber defense: each agent needs only to attend to its own responsibilities and optimize its individual reward, while simultaneously coordinating with sibling agents through the hierarchical layout to contribute to the overall reward.

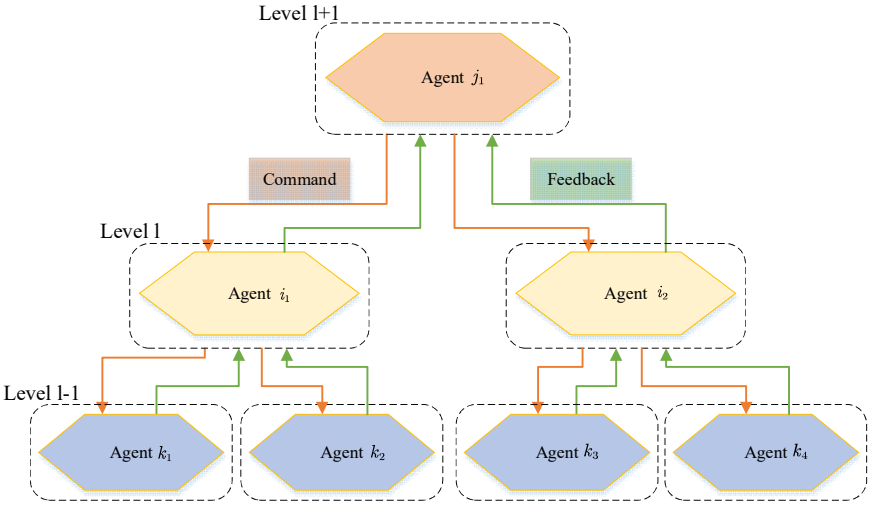


Figure 1. Diagram of the general hierarchical cascaded multi-agent cooperative architecture.

In the architecture illustrated in Figure 1, all agents share an identical structure. An agent influences its lower-level agents through commands, while feedback from the lower-level agents is transformed into its own observation vector. Regardless of the total number of layers in the system, each agent interacts solely with the immediately adjacent upper and lower layers, treating them as if they were part of the environment. Each layer can consist of one or multiple agents, and the entire architecture can be arbitrarily scaled through horizontal and vertical cascading of agent units. This constitutes one of the key innovations of this work.

This approach ensures effective information transfer across multiple levels while preserving scalability through loose coupling between levels. To validate the effectiveness of the proposed method, we instantiated hierarchical structures with two-layer and three-layer agents, where each layer is responsible for making decisions at different levels and scopes. Experimental results demonstrate that the HC-MARL framework can be effectively applied to intelligent cyber wargaming and that our method outperforms those such as Singh et al. (2024). The main contributions are summarized as follows:

1. We propose HC-MARL (Hierarchical Cascaded Multi-Agent Reinforcement Learning), a general-purpose hierarchical collaborative multi-agent architecture for intelligent cyber wargaming, which effectively handles cross-level information transfer and alert information sharing. The architecture is inherently consistent with the organizational form of cyber defense;
2. We design a message transformation function that converts variable-length messages fed back from lower layers into fixed-length observation vectors, enabling adaptation to different network scenarios and personnel configurations. An attention mechanism is also employed to optimize the efficiency of information transfer between cross-level agents;
3. We adopt an approach combining global rewards with local rewards, as well as long-horizon rewards with instantaneous rewards, to achieve more stable multi-agent policy learning;
4. We design a policy function that integrates neural networks with empirical knowledge to enhance threat response efficiency while ensuring smooth transmission of up-level commands.

The remainder of this paper is structured as follows. Section 2 reviews the theoretical background and contributions of prior studies. Section 3 elaborates on the design philosophy and

 doi:10.20944/preprints202605.1354.v1

methodology of the HC-MARL architecture, including the agent observation space, message transformation function, policy function, and reward mechanism. Section 4 validates our method through experiments and analyzes the experimental results. Section 5 discusses its limitations, and outlines future research directions.

2. Related Work

In recent years, cyber wargaming has played an increasingly prominent role in addressing cyber conflicts and cultivating professionals, with scholars worldwide conducting extensive research on its conceptual foundations, system architectures, and application scenarios. Fox et al. [16] focused on the adaptability of cyber wargaming in commercial settings, addressing the pain point that traditional wargames are predominantly oriented toward military and national security scenarios and are thus disconnected from actual business needs, and proposed an optimized wargaming framework for commercial scenarios. He et al. [17] targeted large-scale cyber confrontation scenarios, proposing a system architecture for wargaming based on large-scale cyber confrontation and conducting preliminary explorations of key technologies such as scenario reconstruction, offensive and defensive behavior modeling, and AI decision-making. Haggman [18] concentrated on the application of cyber wargaming in cybersecurity education, delving deeply into the design logic, implementation pathways, and effectiveness evaluation of educational wargames, and is a representative work in this niche area. Curry and Drage [19] analyzed the characteristics of different types of manual cyber wargames to help readers understand the categories of cyber wargames and meet their needs. Chen [20] through literature review, case analysis, and expert interviews, examined the practicality of cyber wargaming in education and analysis, constructed the first standardized evaluation framework covering multiple types of cyber wargaming—analytical, educational, and commercial—and provided a critical analysis of several wargames. Roche [21] established the first complete theoretical system for cyber wargaming, clearly delineating the two core application scenarios of “research wargames” and “educational wargames,” and proposed the “wargaming trilemma” framework, which explicitly defines the trade-offs among analytical utility, situational realism, and player engagement, serving as a core evaluation criterion for subsequent wargaming design. Yang et al. [22] constructed a wargaming framework based on “composability + minimal operational units” and proposed a wargaming method tailored to the security evolution of power networks. Overall, existing research emphasizes theoretical system and application scenario innovation, striving for technological breakthroughs and practical implementation, and has laid a certain foundation. However, there remains substantial room for improvement in adapting to future intelligent warfare and AI-empowered wargaming.

In essence, cyber wargaming is a game of adversarial confrontation between attackers and defenders within a given scenario, with the goal of enabling the wargaming party to take the most advantageous actions based on the offensive and defensive situation to maximize gains. Intelligent cyber wargaming, in which artificial intelligence (agents) assumes the role of the wargaming party, centers on solving core problems such as agent situational awareness, autonomous decision-making, and efficient collaboration. In recent years, deep reinforcement learning (DRL) and multi-agent reinforcement learning (MARL) have shown great potential in automated cyber defense. Papers [2,5,23,24] have successively modeled cyber confrontations between blue and red agents, developing DRL techniques to train agents and producing highly capable autonomous defense agents. However, such single-agent paradigms cannot meet the practical demands of hierarchical, region-specific, and multi-user wargaming. MARL leverages interactions and collaboration among multiple agents to detect, mitigate, and respond to cyber threats, offering a new avenue for intelligent cyber wargaming. Wiebe et al. [25] investigated the applicability of cooperative MARL in tactical-level cyber defense decision-making, comparing value-based independent learning and centralized training with decentralized execution (CTDE), and showed that both methods outperform heuristic defense methods. Oesch et al. [6] argued that agents should specialize in specific domains within the cyber defense process, and conducted research on the role positioning, game mechanisms, environmental

 doi:10.20944/preprints202605.1354.v1

adaptability, and training environments of agents in cyber defense. In particular, hierarchical multi-agent reinforcement learning (HMARL) decomposes and stratifies defense tasks, with each sub-task assigned to a specialized agent. Singh et al. [14] proposed a hierarchical proximal policy optimization (PPO) architecture that decomposes cyber defense tasks into specific sub-tasks and divides the policy into a master policy and sub-policies, with sub-policies responsible for handling a category of sub-tasks and the master policy selecting which sub-task to execute. Hürten et al. [11] designed a hierarchical agent structure where a high-level agent coordinates the overall defense strategy and low-level agents focus on tasks in specific regions. Tang et al. [10] modeled cyber offense and defense as a Stackelberg hyper-game and used hierarchical MARL as the driving force for game evolution, proposing a two-layer agent architecture in which a first-level defense agent serves as a selector that activates a type of second-level defense agent, which then executes the actions. To address the issue that existing methods commonly use communication mechanisms or hybrid action spaces to cope with environmental non-stationarity and partial observability, but rely on complex policy networks to process limited-bandwidth information. Wang et al. [26]propose a simplified cooperative multi-agent reinforcement learning framework that directly generates coordinated actions through customized communication protocols. Inspired by football games, Liu et al. [13] proposed a hierarchical commander-based policy optimization (HCPO) algorithm that deploys local commanders to retain the advantages of centralized training while circumventing communication constraints among agents.

Furthermore, Palmer et al. [27] provided a review and critique of deep reinforcement learning (DRL) methods in autonomous cyber defense. Landolt et al. [4] reviewed recent advances in the application of multi-agent reinforcement learning (MARL) to automated cyber defense, with a particular focus on autonomous intelligent cyber defense agents and training environments, and outlined existing challenges and future research directions. Lazer et al. [28] examined the impact of artificial intelligence agents on cybersecurity, highlighting emerging threat models, security frameworks, and evaluation processes for agent systems, and analyzed systemic risks including agent collusion, cascading failures, regulatory evasion, and memory poisoning.

Current research still exhibits the following deficiencies when applied to intelligent cyber wargaming: first, an inability to match the characteristics of hierarchical command, making effective bidirectional information transfer between levels difficult; second, the rigid structure of existing multi-agent frameworks cannot accommodate normal combat attrition and dynamic adjustments during cyber wargaming; and third, the inherent long-horizon and sparse nature of shared rewards leads to unstable policy learning.

3. Methods

3.1. Mathematical Representation of the Defense in Intelligent Cyber Wargaming

Multi-agent-based cyber wargaming exhibits the following core characteristics. First, it is inherently hierarchical, with clearly defined command and coordination relationships among agents operating at different levels. Second, it is highly dynamic: during adversarial engagement, the operational postures, resource states, and action strategies of both sides are in constant flux, requiring real-time adaptation by the agents. Third, the environment is partially observable: agents can only access local situational information and are unable to perceive the global picture, thereby necessitating information exchange to bridge the perceptual gap. Fourth, it entails multi-objective conflicts, in which the local objectives of individual agents may be misaligned, requiring collaborative mechanisms to achieve globally optimal outcomes.

Therefore, the coordination problem of defensive agents can be modeled as a decentralized partially observable Markov decision process (Dec-POMDP) [29]. A Dec-POMDP is a particular class of MDP in which multiple independent and decentralized agents with incomplete observations interact to optimize a shared reward signal. Formally, a Dec-POMDP is defined as a tuple $M = (N, S, A, T, O, \Phi, R, b_o)$, where $N = \{1, \cdots, n\}$ is the set of n agents. S is a (finite) set of states, A

 doi:10.20944/preprints202605.1354.v1

is the set of joint actions, T is the transition probability function that specifies the probabilities $Pr(s'|s,a)$, O is the set of joint observations, Φ is the observation probability function that specifies the probabilities $Pr(o|a,s')$, R is the immediate reward function, b_0 is the initial state distribution at time t = 0. In particular, $A = \times_{i \in N} A_i$ is the set of joint actions. Here, A_i is the set of actions available to agent i. Similar to the set of joint actions, $O = \times_{i \in N} O_i$ is the set of joint observations, where O_i is a set of observations available to agent i.

At each time step t, each agent follows its local policy π_i to select an action $a_i \in A$ based on its observations and received information. which causes the environment to transition to a new state $s' \sim Pr(s'|s,a)$ and generates a global reward $r = R(s,a)$. The main goal for the blue agents is to maximize the expected sum of rewards, $TR = \Sigma_t r_t$.

3.2. HC-MARL Framework Design

To address the demands for cross-level command and multi-agent collaboration in cyber wargaming, this work proposes a general hierarchical cascaded multi-agent collaborative architecture. Through hierarchical decomposition, the responsibility boundaries of agents at each level are clearly delineated. Meanwhile, a top-down guidance information flow and a bottom-up feedback information flow are designed between levels to ensure effective bidirectional interaction. A global threat alerting mechanism is adopted, together with a dynamic weight allocation strategy, to facilitate information exchange and collaborative decision-making among agents at the same level.

In this architecture, agents are organized in a tree-structured hierarchy according to command relationships. For a given agent, it may have multiple child nodes (i.e., command multiple operational units) but at most one parent node (the superior commander). Higher-level agents direct the lower-level ones, while the subordinates report execution results back to their parent node. Each agent makes decisions at the level that corresponds to its designated role, as shown in Figure 2.

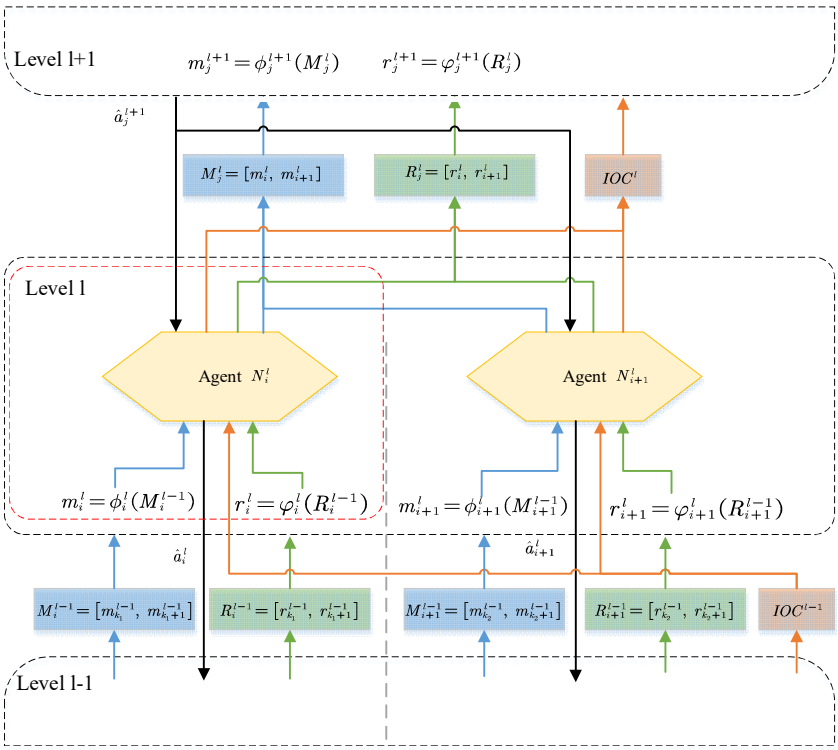


Figure 2. Diagram of the HC-MARL architecture and information flow.

Generally, for an *L*-layer multi-agent defense collective, the set of agents at layer l is denoted as $N^l = \{N_1^l,\ N_2^l,\ \cdots,\ N_n^l\}$. In the hierarchical structure, each agent N_i^l is connected to agents in the

 doi:10.20944/preprints202605.1354.v1

immediately adjacent upper and lower layers. We denote the sets of agents at layer $l+1$ and $l-1$ as N_i^{l+1} and N_i^{l-1} respectively.

Similarly, the observation space of agent N_i^l is denoted as O_i^l, its action space as A_i^l, and its reward set as R_i^l. For notational convenience, we denote the parent agent of agent N_i^l as $\hat{N}_i^l$—in cyber wargaming, each agent has at most one parent at the upper level—and the set of child agents of N_i^l as $\underset{\sim}{N}_i^l$.

In the HC-MARL architecture, four information flows exist between hierarchical levels. The parent node at the upper layer influences and shapes the observations of its child nodes through its chosen actions, forming a top-down guidance information flow. Conversely, the lower-layer agents transmit messages, rewards, and Indicators of Compromise (IOCs) to their superior agents, constituting a bottom-up feedback information flow. This design closely aligns with the operational process in which superior commanders issue orders to subordinate operators, while the latter report situational updates, emergent events, and action outcomes back to their superiors.

In cyber wargaming, operational scenarios and combat units can be highly dynamic, meaning that the number of agents at each layer may change over time. Existing MARL architectures are unable to accommodate this requirement. To address this challenge, we model agents as modular and reusable units, which we term Hierarchical Cascaded Units (HCUs), as illustrated by the red dashed box in Figure 2. This implies that the observation spaces, action spaces, and the forms of policy and value functions across all hierarchical levels must remain consistent. A primary issue is thus the observation space: the number of subordinate child nodes is variable, which causes the lengths of the messages, rewards, and IOCs to vary with the number of child nodes. Additionally, the message formats between different levels need to be aligned.

To address this issue, we define a message transformation function $m_i^l = \phi_i^l(M_i^{l-1})$ and a reward transformation function $r_i^l = \varphi_i^l(R_i^{l-1})$ that convert the feedback from lower-level agents into fixed-length vectors, which then serve as part of the observation vector at the current level, where $M_i^{l-1} = [m_k^{l-1}],\ k \in \underset{\sim}{N}_i^l$ and $R_i^{l-1} = [r_k^{l-1}],\ k \in \underset{\sim}{N}_i^l$. The policy function is defined to determine the action to be taken, which can be expressed as $a_i^l = \pi_i^l(\hat{a}^{l+1},\ m_i^l,\ IOC_i^l)$. Here, $\hat{a}^{l+1}$ incorporates the instruction action of the parent node, m_i^l is the transformed feedback messages from the child nodes, and the global Indicator of Compromise IOC_i^l. In other words, agent N_i^l determines its optimal action based on the parent's instruction action, the feedback from its child nodes, and the global threat alert message. This design not only highlights hierarchy and modularity, ensuring efficient communication among agents, but also enhances vertical cascadability and horizontal flexible scalability. From the perspective of agent modeling, only a single HCU needs to be modeled, which greatly simplifies the modeling complexity.

To improve the efficiency of inter-level information exchange, we introduce an attention mechanism that enables upper-layer agents to focus on the critical state information of lower-layer agents, thereby reducing the amount of information transmitted.

3.3. Observation Space Design

To accommodate the requirements of cyber wargaming, the basic observation of the blue agent we design comprises four components. The first component indicates the current mission phase, which can take values in $\{0, 1, 2\}$. The second component represents the instruction from the superior commander, with a length of 8, encoding the priority levels of different actions. The third component captures the feedback information received from subordinate nodes (except for the lowest-layer agents, generated by the message transformation function ϕ). The fourth component is the indicator of compromise, signaling whether a global intrusion alert is present. Illustrated in Figure 3.

 doi:10.20944/preprints202605.1354.v1

Mission	$\hat{a}^{l+1}$	m_i^l	IOC_i^l

Figure 3. Composition of the agent observation space.

Note that the various components of the observation are presented to the agent in heterogeneous data formats, and these fields may vary across different agents, thus necessitating normalization of the observation.

3.4. Transformation Function Design

In the HC-MARL architecture, the message transformation function ϕ and the reward transformation function φ serve to receive feedback information from lower-level agents and fuse it into fixed-length vectors. To accommodate the interface between the lowest-layer agents and the network environment, the message vector is designed with a length of 76 in this paper. For the lowest layer, messages originate from the environment, whereas for all other layers, they are generated by the message transformation function ϕ. Similarly, for the lowest layer, rewards come directly from the environment, while for other layers, they are produced by the reward transformation function φ. In this work, the reward transformation function employs a weighted average, while the message transformation function adopts a transformer model, the structure of which is illustrated in Figure 4.

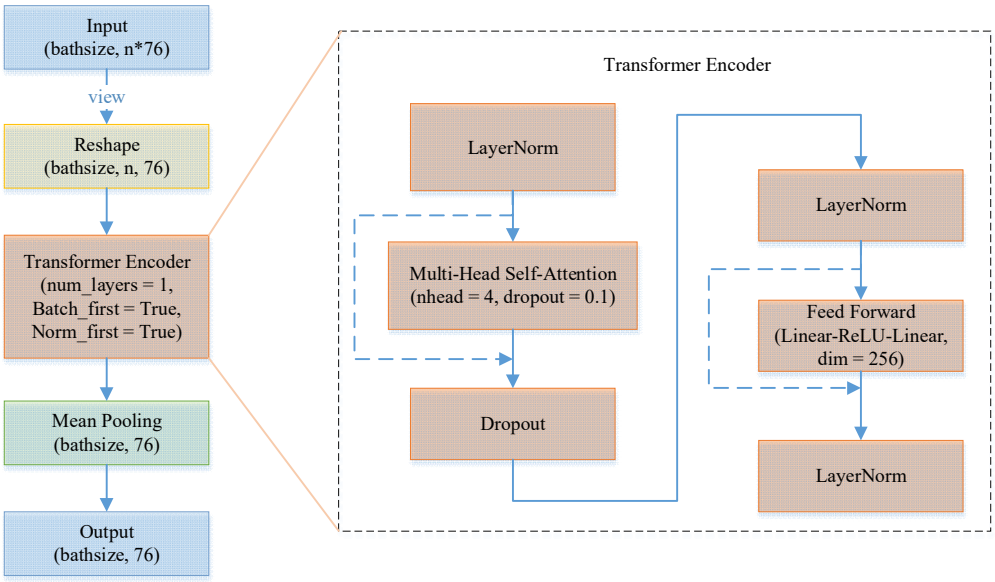


Figure 4. Network architecture of the message transformation function.

This model takes variable-length sequences as input and adopts the Transformer Encoder as its core feature extractor. Through an aggregation operation, it produces a fixed-length output. The advantages of designing the transformation function in this manner are twofold. First, it accommodates an arbitrary number of agents at each layer without requiring modifications to the model structure. Second, the introduction of the attention mechanism enables the model to focus on critical information, thereby reducing interference from redundant information and enhancing the efficiency of information exchange. The hyperparameters are listed in Table 1.

Table 1. Hyperparameter settings of the transformer model.

Hyperparameter	Values
d_model	76

 doi:10.20944/preprints202605.1354.v1

nhead	4
num_encoder_layers	1
dim_feedforward	256
dropout	0.1
activation	ReLU

3.5. Policy Function Design

The policy function π takes the observation vector as input and outputs the selected action, i.e., $a_i^l = \pi_i^l(\hat{a}^{l+1}, m_i^l, IOC_i^l)$, where $\hat{a}^{l+1}$ incorporates the guiding information from the upper-layer parent node, m_i^l is the messages received by the agent, and the global alert information IOC_i^l. In this work, we design a policy function that integrates neural networks with empirical knowledge. The information processing pipeline is illustrated in Figure 5.

The neural network used here has 76 input nodes, employs the ReLU activation function, contains two hidden layers of 256 nodes each, and produces 82 output nodes. The message vector m_i^l serves as the input to the neural network, which outputs an initial action weight vector. This initial weight vector is then element-wise multiplied with the guidance information $\hat{a}^{l+1}$ which expanded to 82 dimensions according to the action categories. The mapping to the mask also relies on fundamental knowledge in cyber defense. For example, once it is detected that an attacker has gained host privileges, remove or restore operations should be immediately executed to prevent further threat propagation. Finally, the result of the neural network and the result of the empirical knowledge are combined through a weighted sum to produce the final decision.

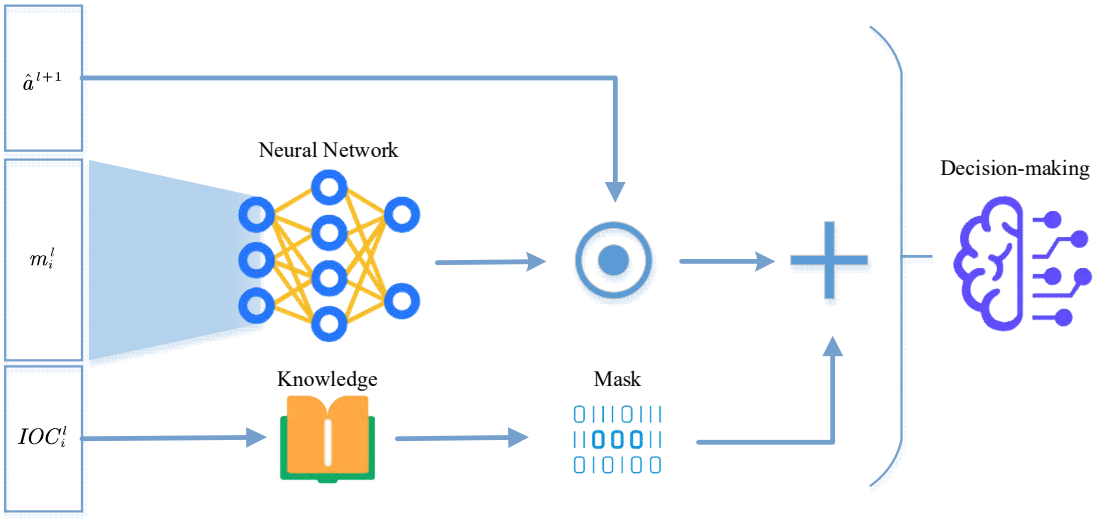


Figure 5. Information flow diagram of the policy function.

3.6. Reward Design

In our experiments, we observed that multi-agent policy learning tends to be unstable in such complex cyber offense–defense scenarios. To address this issue, we introduce local rewards to encourage agents to detect threats at an early stage. Specifically, a positive local reward (e.g., 0.2) is granted to an agent when it successfully executes tasks such as analysis, honeypot deployment, or removal. Furthermore, we incorporate staged rewards: a positive reward (e.g., 0.5) is given to an agent when it first discovers an anomalous process, an anomalous connection, or a honeypot-triggered alert within its responsible region. The composite reward for each agent is obtained by weighting the global reward—the original reward, which is typically negative—with these local rewards. During MAPPO training, we found that the value function loss (vf_loss) was very high, whereas the policy loss(policy_loss) was very small; we therefore scaled the reward by a factor λ.

 doi:10.20944/preprints202605.1354.v1

3.7. Learning and Optimization

In the HC-MARL architecture, each agent is required to learn two key functions: the policy function π for generating actions, and the message transformation function ϕ.

$$a_i^l = \pi_i^l(\hat{a}^{l+1},\ m_i^l,\ IOC_i^{l-1}) \tag{1}$$

$$m_i^l = \phi_i^l(M_i^{l-1}) \tag{2}$$

The policy function is responsible for mapping the combination of received guidance information and observations to the action of the current agent, while the message transformation function converts the feedback messages from lower-level agents into a fixed-length vector. The modular design of the framework allows agents at each level to learn independently using algorithms suited to their specific roles. This flexibility accommodates a wide range of learning methods. During training, each agent stores its experiences and updates its policy based on the received rewards, as outlined in Table 2.

Table 2. Learning algorithm.

Algorithm 1 Env.step()

Input: $\hat{a}^{l+1}$

$a_i^l = \pi_i^l(\hat{a}^{l+1},\ m_i^l,\ IOC_i^l)$

for agent $N_i^l \in$ Level l do

$m_k^{l-1},\ r_k^{l-1},\ IOC_k^{l-1} \leftarrow \text{step}(a_i^l)$

end for

$M_i^{l-1} = [m_k^{l-1}],\ k \in \underset{\sim}{N}_i^l,\ R_i^{l-1} = [r_k^{l-1}],\ k \in \underset{\sim}{N}_i^l$

$m_i^l = \phi_i^l(M_i^{l-1}),\ r_i^l = \varphi_i^l(R_i^{l-1}),\ IOC_i^l = \cup\, IOC_k^{l-1},\ k \in \underset{\sim}{N}_i^l$

if training **then**

store$(\hat{a}^{l+1},\ m_i^l,\ IOC_i^l,\ a_i^l,\ r_i^l)$

update()

end if

Return: $m_i^l,\ r_i^l,\ IOC_i^l$

4. Experiments

To validate the effectiveness of the proposed method, we conducted experimental evaluations in the CybORG environment of CAGE Challenge 4, which is the latest and most prominent open-source platform for autonomous cyber defense. This foundational platform provides simulation of the cyber behavior of attackers and defenders within realistic network topologies, enabling researchers to repeatedly construct, initialize, and test multi-agent models across a variety of cybersecurity scenarios. The original observation spaces, action spaces, and reward structures of the various agents are documented in detail, along with the source code, in the corresponding GitHub repository[30].

In our experiments, we employed the MAPPO algorithm to update the policies (using the Ray RLlib library). Both the actor and critic networks were implemented as feedforward neural networks with two hidden layers, each containing 256 neurons. The hyperparameters used during training were set as follows: learning rate of 1×10^{-3}, discount factor of 0.99, train batch size of 32,768, SGD minibatch size of 8,192, num_sgd_iter of 15, and clip_param of 0.3. The training environment was set with a maximum time step of 500, and a total of 200 training iterations were conducted. After each training iteration, the learned policy was evaluated under the same maximum time step of 500. The evaluation consisted of 10 episodes, and the average reward over these 10 episodes was recorded as the test result.

 doi:10.20944/preprints202605.1354.v1

In previous studies, a shared reward mechanism was adopted for all blue agents, while each agent maintained an independent policy. Experiments revealed that this approach resulted in policy instability and significant reward fluctuations across agents. In this work, we first apply reward scaling and then introduce local rewards and staged rewards. The benefit of this design is that different agents receive distinct rewards when selecting different actions within their respective areas of responsibility, which facilitates more effective policy learning. The comparative experimental results are presented in Figure 6.

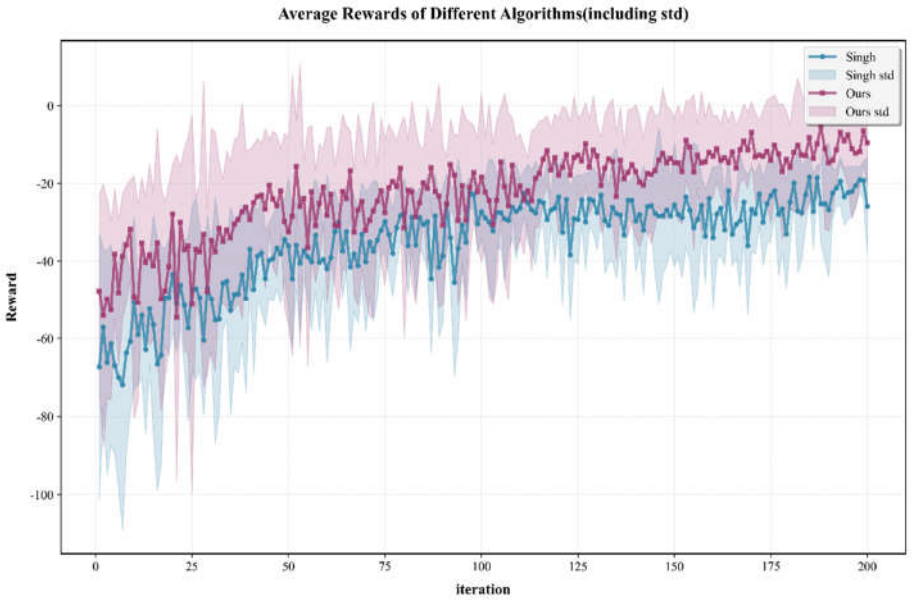
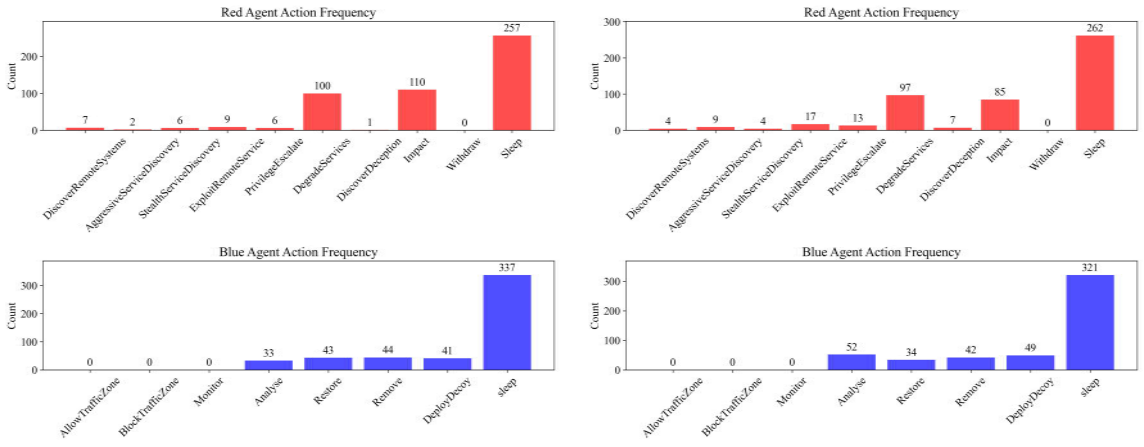


Figure 6. Comparison of various reward schemes.

We adopted the method of Singh as the baseline. The experimental results demonstrate that the introduction of local rewards and staged rewards yields an improvement of approximately 30% in average reward. As indicated by the variance, our method also leads to more stable reward fluctuations.

We further conducted a statistical analysis of the actions taken by blue agents before and after the incorporation of local rewards, illustrated in Figure 7. The results reveal that, with local rewards, the agents exhibit a stronger preference for early-stage preventive actions such as “Analyse” and “DeployDecoy”, which is consistent with our expectations.



(a) No Local and Immediate Rewards

(b) Integration of Local and Immediate Rewards

Figure 7. Influence of local and instantaneous rewards on agent action selection.

 doi:10.20944/preprints202605.1354.v1

To evaluate the proposed method in this network scenario, we instantiated multi-agent defense collectives with both two-layer and three-layer hierarchical structures. In the two-layer architecture, the top layer consists of a single agent that serves as the overall commander-in-chief of the defense, while the second layer comprises seven agents, each responsible for defending a distinct subnet. In the three-layer architecture, the top layer again consists of a single agent acting as the overall commander; the second layer contains three agents, which partition the seven subnets into three groups according to their functionality, with each agent responsible only for its assigned group; and the bottom layer is composed of seven agents, each tasked with executing specific operations—such as monitoring, analysis, deception, and removal—within its designated subnet. To verify the generality and flexibility of the architecture, the number of agents within each subnet is configured heterogeneously. The structure is illustrated in Figure 8.

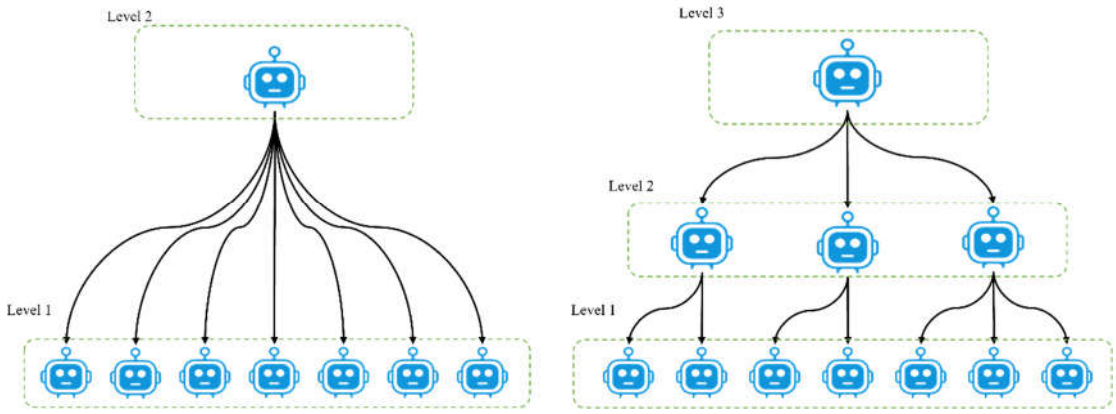


Figure 8. Hierarchical architecture of defensive agents.

We adopted the same algorithm and hyperparameters for training and testing, with the results presented in Figure 9. In the figure, the blue curve represents the results obtained by the Singh method, the purple curve corresponds to the two-layer agent cooperative defense constructed under our HC-MARL framework, and the orange curve denotes the three-layer agent cooperative defense developed by our proposed method.

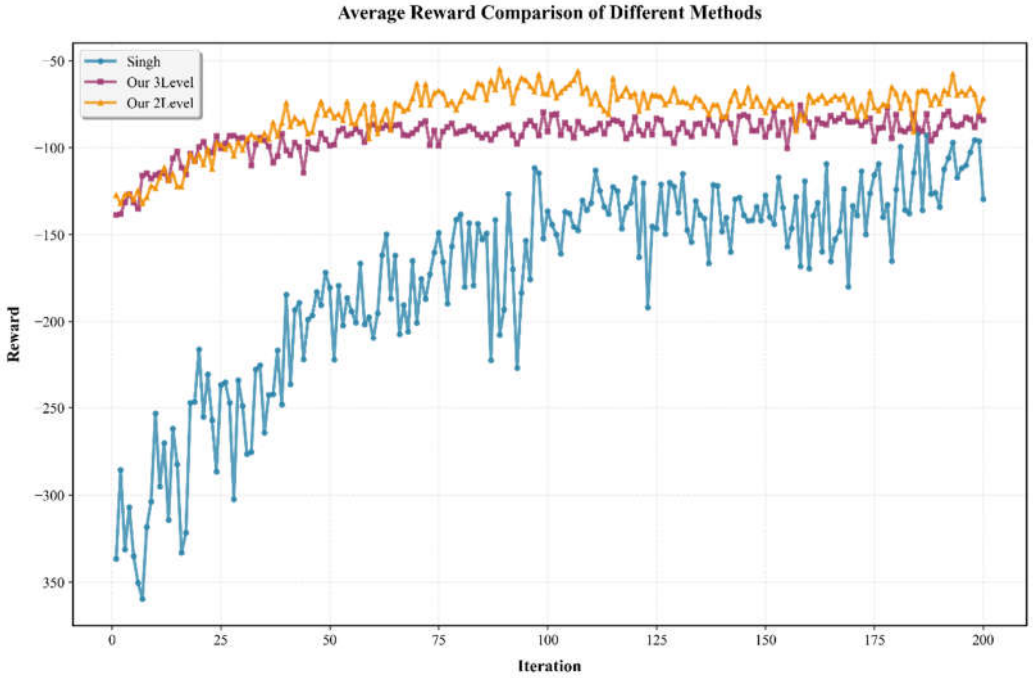


Figure 9. Comparison of various methods.

 doi:10.20944/preprints202605.1354.v1

As illustrated in the results, both our two-layer and three-layer structures outperform the baseline method, achieving an approximate 30% improvement in rewards. Additionally, the curve trends demonstrate that our method converges faster and exhibits more stable performance. Furthermore, the three-layer structure is more stable than the two-layer counterpart; however, a deeper hierarchy increases the complexity of agent collaboration, rendering policy learning and optimization more challenging.

5. Discussion

In this work, we proposed HC-MARL, a general hierarchical cascaded multi-agent framework that integrates cross-level command and dynamic node adjustment to overcome the limitations of existing MARL methods in intelligent cyber wargaming. The main conclusions are as follows:

(1) The bidirectional information flow design of the HC-MARL framework not only ensures effective information transmission within the multi-agent system but also clarifies the core responsibilities and interaction boundaries among agents, thus effectively characterizing the information activities between superior and subordinate levels in intelligent cyber wargaming.

(2) Modeling all agents as unified hierarchical cascaded units allows the defensive collective to scale both horizontally and vertically, which greatly simplifies the modeling process. Moreover, the message transformation function—which converts variable-length feedback from child nodes into fixed-length observation vectors through an attention mechanism—enables dynamic addition or removal of child nodes without retraining the whole model, thereby adapting naturally to evolving network topologies and personnel configurations.

(3) The reward mechanism, which combines global and local rewards together with outcome-based and staged (instantaneous) rewards, substantially stabilizes multi-agent policy learning. Instantaneous rewards can be intentionally set to guide agents toward proactive actions such as “Analyse” and “DeployDecoy”, shifting the defensive emphasis from post-attack restoration to early warning. Furthermore, the dynamic weight allocation strategy balances intra-region action conflicts by prioritizing the current agent’s reward, further smoothing collaborative decision-making.

(4) By integrating neural networks and empirical knowledge into the policy function—where the neural network’s initial action weights are first modulated by parent node guidance via element-wise multiplication and then combined with an empirical knowledge mask through a weighted sum—HC-MARL enhances threat response efficiency while ensuring smooth transmission of up-level commands.

As a result, the framework achieves approximately a 30% improvement in 10-episode average optimal reward compared with the baseline methods, while exhibiting faster and more stable policy convergence. These findings offer a new approach and perspective on multi-agent cyber wargaming; however, several directions warrant further investigation:

(1) The current method does not fully consider agent heterogeneity (e.g., capability differences among tactical agents of different types). Future work may incorporate heterogeneous multi-agent systems to optimize cooperative strategies for agents with diverse capabilities.

(2) This work involves multiple neural network components. While some hyperparameter settings were borrowed from prior research, systematic comparative experiments on network architectures (e.g., the message transformation function) and hyperparameter configurations have not been performed, leaving room for more thorough tuning.

(3) Although the HC-MARL framework supports horizontal and vertical scaling, optimization becomes increasingly difficult as the structure grows more complex. The training methodology for multi-level cascaded policies thus warrants further in-depth investigation.

Author Contributions: Conceptualization, Z.Q. and J.H.; methodology, Z.Q. and J.H.; software, Z.Q.; validation, Z.Q., Z.L. and T.X.; formal analysis, B.W.; investigation, Z.Q.; resources, Z.Q.; data curation, Z.L. and T.X.; writing—original draft preparation, Z.Q.; writing—review and editing, Z.Q. and J.H.; visualization, Z.Q. and

 doi:10.20944/preprints202605.1354.v1

Z.L.; supervision, J.H.; project administration, B.W.; funding acquisition, J.H. and B.W. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: The code presented in this study is available on request from the corresponding author.

Acknowledgments: During the preparation of this manuscript, the authors used DeepSeek for linguistic polishing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kiely, M., Ahiskali, M., Borde, E., Bowman, B., Bowman, D., Van Bruggen, D., Cowan, K., Dasgupta, P., Devendorf, E., Edwards, B., Fitts, A., Fugate, S., Gabrys, R., Gould, W., Huang, H.H., Jacobs, J., Kerr, R., King, I.J., Li, L., Martinez, L., Moir, C., Murphy, C., Naish, O., Owens, C., Purchase, M., Ridley, A., Taylor, A., Farmer, S., Valentine, W.J., Zhang, Y.: Exploring the efficacy of multi-agent reinforcement learning for autonomous cyber defence: A CAGE challenge 4 perspective. AAAI. 39, 28907–28913 (2025). https://doi.org/10.1609/aaai.v39i28.35158.
2. Kiely, M., Bowman, D., Standen, M., Moir, C.: On Autonomous Agents in a Cyber Defence Environment, http://arxiv.org/abs/2309.07388, (2023). https://doi.org/10.48550/arXiv.2309.07388.
3. Kunz, T., Fisher, C., Novara-Gsell, J.L., Nguyen, C., Li, L.: A Multiagent CyberBattleSim for RL Cyber Operation Agents, http://arxiv.org/abs/2304.11052, (2023). https://doi.org/10.48550/arXiv.2304.11052.
4. Landolt, C.R., Würsch, C., Meier, R., Mermoud, A., Jang-Jaccard, J.: Multi-Agent Reinforcement Learning in Cybersecurity: From Fundamentals to Applications, http://arxiv.org/abs/2505.19837, (2025). https://doi.org/10.48550/arXiv.2505.19837.
5. Nguyen, T.T., Reddi, V.J.: Deep Reinforcement Learning for Cyber Security. IEEE Trans. Neural Netw. Learning Syst. 34, 3779–3795 (2023). https://doi.org/10.1109/TNNLS.2021.3121870.
6. Oesch, S., Austria, P., Chaulagain, A., Weber, B., Watson, C., Dixson, M., Sadovnik, A.: The Path To Autonomous Cyber Defense, http://arxiv.org/abs/2404.10788, (2024). https://doi.org/10.48550/arXiv.2404.10788.
7. Wang, M., Dechene, R.: Multi-Agent Actor-Critics in Autonomous Cyber Defense, http://arxiv.org/abs/2410.09134, (2024). https://doi.org/10.48550/arXiv.2410.09134.
8. Standen, M., Lucas, M., Bowman, D., Richer, T.J., Kim, J., Marriott, D.: CybORG: A Gym for the Development of Autonomous Cyber Agents, http://arxiv.org/abs/2108.09118, (2021). https://doi.org/10.48550/arXiv.2108.09118.
9. Singh, A.V., Rathbun, E., Graham, E., Oakley, L., Boboila, S., Chin, P., Oprea, A.: Hierarchical multi-agent reinforcement learning for cyber network defense. (2025).
10. Tang, Y., Sun, J., Wang, H., Deng, J., Tong, L., Xu, W.: A method of network attack-defense game and collaborative defense decision-making based on hierarchical multi-agent reinforcement learning. Computers & Security. 142, 103871 (2024). https://doi.org/10.1016/j.cose.2024.103871.
11. Hürten, T., Loevenich, J.F., Spelter, F., Adler, E., Braun, J., Moxon, L., Gourlet, Y., Lefeuvre, T., Lopes, R.R.F.: Hierarchical multi-agent reinforcement learning for autonomous cyber defense in coalition networks. In: MILCOM 2024 - 2024 IEEE Military Communications Conference (MILCOM). pp. 176–181. IEEE, Washington, DC, USA (2024). https://doi.org/10.1109/MILCOM61039.2024.10773689.
12. Alshamrani, A.: Federated hierarchical MARL for zero-shot cyber defense. PLoS One. 20, e0329969 (2025). https://doi.org/10.1371/journal.pone.0329969.
13. Liu, Z., Tu, J., Hong, Y., Xiong, L., Jin, Y., Tang, Y., Li, F.: HCPO: Hierarchical Conductor-Based Policy Optimization in Multi-Agent Reinforcement Learning, http://arxiv.org/abs/2511.12123, (2025). https://doi.org/10.48550/arXiv.2511.12123.
14. Singh, A.V., Rathbun, E., Graham, E., Oakley, L., Boboila, S., Oprea, A., Chin, P.: Hierarchical Multi-agent Reinforcement Learning for Cyber Network Defense, http://arxiv.org/abs/2410.17351, (2024). https://doi.org/10.48550/arXiv.2410.17351.

 doi:10.20944/preprints202605.1354.v1

15. Paolo, G., Benechehab, A., Cherkaoui, H., Thomas, A., Kégl, B.: TAG: A Decentralized Framework for Multi-Agent Hierarchical Reinforcement Learning, http://arxiv.org/abs/2502.15425, (2025). https://doi.org/10.48550/arXiv.2502.15425.
16. Fox, D., McCollum, C., Arnoth, E., Mak, D.: Cyber Wargaming: Framework for Enhancing Cyber Wargaming with Realistic Business Context. The Homeland Security Systems Engineering and Development Institute (HSSEDI)TM (2018).
17. He, J.; Lian, X.; Qi, Q.; Zhang, H. Discussion on Key Technologies of Large-Scale Network Confrontation Wargaming System. Communications Technology 2018, 51, 450–456.
18. Haggman, A.: Cyber Wargaming: Finding, Designing, and Playing Wargames for Cyber Security Education, https://pure.royalholloway.ac.uk/en/publications/cyber-wargaming-finding-designing-and-playing-wargames-for-cyber-, (2019).
19. Curry, J., Drage, N.: The handbook of cyber wargames: wargaming the 21st century. History of Wargaming Project, London (2020).
20. Chen, S.: An analysis of cyber wargaming: Current games, limitations, and recommendations. CMC Senior Theses. (2022).
21. Roche, E.: Cyber wargaming: Research and education for security in a dangerous digital world. Journal of Strategic Security. 18, (2025).
22. Yang, Y.; Wu, J.; Gao, Z.; Liang, Z.; Hong, C.; Li, P.; Zhang, Y. Wargaming Technology Towards Cybersecurity Threat Evolution in Power Information Network. Southern Power System Technology 2025, 19(6), 52–62, doi:10.13648/j.cnki.issn1674-0629.2025.06.005.
23. Foley, M., Hicks, C., Highnam, K., Mavroudis, V.: Autonomous Network Defence using Reinforcement Learning. In: Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security. pp. 1252–1254. ACM, Nagasaki Japan (2022). https://doi.org/10.1145/3488932.3527286.
24. Hannay, J.: john-cardiff/-cyborg-cage-2, https://github.com/john-cardiff/-cyborg-cage-2, (2025).
25. Wiebe, J., Mallah, R.A., Li, L.: Learning Cyber Defence Tactics from Scratch with Multi-Agent Reinforcement Learning, http://arxiv.org/abs/2310.05939, (2023). https://doi.org/10.48550/arXiv.2310.05939.
26. Wang, Q., He, Z., Shi, H.: Simplifying communication control: A cooperative multi-agent reinforcement learning framework based on group decision-making. J. King Saud Univ. Comput. Inf. Sci. 37, 317 (2025). https://doi.org/10.1007/s44443-025-00326-6.
27. Palmer, G., Parry, C., Harrold, D.J.B., Willis, C.: Deep Reinforcement Learning for Autonomous Cyber Defence: A Survey, http://arxiv.org/abs/2310.07745, (2024). https://doi.org/10.48550/arXiv.2310.07745.
28. Lazer, S.J., Aryal, K., Gupta, M., Bertino, E.: A survey of agentic AI and cybersecurity: Challenges, opportunities and use-case prototypes, http://arxiv.org/abs/2601.05293, (2026). https://doi.org/10.48550/arXiv.2601.05293.
29. Oliehoek, F.A., Amato, C.: A concise introduction to decentralized POMDPs. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-28929-8.
30. TTCP CAGE Working Group: TTCP CAGE challenge 4, https://github.com/cage-challenge/cage-challenge-4, (2023).