Hierarchical Adversarially-Resilient Multi-Agent Reinforcement Learning for Cyber-Physical Systems Security

Saad Alqithami

Computer Science Department
Al-Baha University, Albaha 65779, Saudi Arabia
salqithami@bu.edu.sa

Abstract

Cyber-Physical Systems are integral to modern critical infrastructure, including manufacturing, energy grids, and autonomous systems, but their increasing interconnectivity exposes them to sophisticated cyber threats. Traditional security measures, such as rule-based intrusion detection and single-agent learning, often fail against adaptive and zero-day attacks. To address this challenge, we propose a Hierarchical Adversarially-Resilient Multi-Agent Reinforcement Learning (HAMARL) framework, integrating adversarial training into a multi-agent security system. HAMARL leverages a hierarchical control structure where local agents manage subsystem security, and a global coordinator optimizes system-wide defense strategies. Additionally, an adversarially-aware learning loop simulates evolving cyber threats, allowing defenders to preemptively adapt to sophisticated attacks. Evaluations on a simulated industrial IoT testbed demonstrate that HAMARL significantly enhances attack detection, reduces response time, and maintains operational continuity compared to traditional MARL approaches. Our findings suggest that hierarchical MARL, combined with adversarial training, presents a promising advancement for securing next-generation CPS.

Introduction

Cyber-Physical Systems (CPS) serve as the foundation of modern infrastructure, seamlessly integrating computational and communication capabilities with physical processes. These systems have become increasingly critical in domains such as manufacturing, smart grids, autonomous transportation, and healthcare, offering unprecedented automation, efficiency, and real-time decision-making (Wolf and Serpanos 2019). However, the growing interconnectivity and complexity of CPS expose them to an expanding range of cybersecurity threats, including data tampering, advanced persistent threats (APTs), and distributed denial-of-service (DDoS) attacks (Conti et al. 2018). Traditional security mechanisms, such as rule-based intrusion detection systems and single-agent defensive models, struggle to keep pace with these evolving threats, particularly as adversaries increasingly leverage artificial intelligence (AI)-driven attack strategies to circumvent conventional defenses.

Recent advances in multi-agent reinforcement learning (MARL) present a promising avenue for addressing CPS security challenges. By distributing decision-making across multiple agents, MARL enables scalable and coordinated defense strategies, particularly in complex, decentralized environments (Buşoniu, Babuška, and Schutter 2010). Furthermore, hierarchical reinforcement learning extends this paradigm by incorporating a multi-level control structure, where a higher-level policy supervises lower-level agents, thereby improving both scalability and adaptability in large CPS deployments (Vezhnevets et al. 2017). Despite these advantages, existing MARL-based security frameworks often lack adversarial awareness, making them vulnerable to adaptive cyber threats. A purely reactive defense mechanism is insufficient in adversarial settings where attackers continuously evolve their tactics to evade detection (Goodfellow, Shlens, and Szegedy 2015). Integrating adversarial training—where the system learns to counteract an evolving, AI-driven attacker—can significantly enhance the resilience of MARL-based security frameworks, ensuring proactive defense against sophisticated, zero-day cyber threats.

Despite advances in MARL and adversarial learning, current CPS security frameworks lack a unified approach that integrates hierarchical coordination with adversarial resilience. Unlike traditional MARL-based security approaches, which operate in a flat or decentralized structure, our framework introduces a hierarchical design where local agents perform fast, independent intrusion detection while a global coordinator ensures network-wide threat mitigation. Additionally, adversarial training enables defenders to proactively adapt to dynamic, AI-driven cyber threats rather than relying on static rule-based security policies. This gap leaves critical infrastructure vulnerable to evolving threats, necessitating a framework that continuously learns and adapts to adversarial strategies. This paper addresses the following research questions:

1. Can hierarchical MARL improve real-time threat detection and response efficiency in CPS security?
2. Does integrating adversarial training enhance resilience against zero-day attacks compared to standard MARL approaches?
3. How does a hierarchical defense structure impact scalability, computational efficiency, and decision-making in

CPS environments?

To address these challenges, we introduce Hierarchical Adversarially-Resilient Multi-Agent Reinforcement Learning—a novel CPS security framework that integrates hierarchical MARL with adversarial training to enhance CPS security. By modeling both defenders and attackers as learning agents within a cooperative-competitive environment, we demonstrate how hierarchical MARL can provide adaptive, real-time threat detection and mitigation. Specifically, our contributions include:

1. A novel hierarchical multi-agent architecture designed to enhance scalability and efficiency in CPS security.
2. An adversarial training loop that simulates evolving cyber threats, enabling defenders to anticipate and counteract advanced attack strategies.
3. A comprehensive evaluation of our framework's effectiveness in terms of detection accuracy, system resilience, and robustness against zero-day threats using a simulated industrial IoT testbed.

The remainder of this paper is structured as follows. The next section provides an overview of related work. We then introduce the proposed hierarchical MARL framework. This is followed by a description of the implementation and experimental setup. The results and performance analysis section examines detection accuracy, response time, and resilience against adversarial attacks. We conclude with future research directions to include multi-attacker scenarios, explainable AI, and real-world deployment feasibility.

Related Work

Cyber-Physical Systems Security

Cyber-Physical Systems are characterized by tightly integrated computational and physical processes, where embedded sensors and actuators interact in real-time to enable autonomous decision-making (Baheti and Gill 2011). The security of these systems involves protecting both the network infrastructure and physical components from malicious disruptions (Lee 2008). However, the complexity of CPS architectures—often spanning legacy industrial protocols, wireless sensor networks, and cloud-connected services—poses significant challenges for designing unified security solutions. Furthermore, the requirement for continuous operation, where downtime can lead to severe economic and safety consequences, necessitates the adoption of automated and adaptive security mechanisms to ensure resilience against cyber threats.

Multi-Agent Reinforcement Learning

Reinforcement learning is a machine learning paradigm where agents learn optimal behaviors by interacting with an environment and receiving feedback in the form of rewards or penalties(Sutton and Barto 2018). Multi-Agent Reinforcement Learning extends this concept to multi-agent environments, where multiple agents simultaneously learn and optimize their policies while considering interactions with others(Buşoniu, Babuška, and Schutter 2010). MARL approaches can be broadly categorized into: (a) Fully decentralized methods, where each agent learns independently without centralized coordination (Matignon, Laurent, and Fort-Piat 2012). (b) Centralized training with decentralized execution (CTDE), allowing agents to coordinate effectively during training but act independently at runtime (Lowe et al. 2017). (c) Hierarchical MARL, which decomposes decision-making into higher-level and lower-level policies, thereby improving both sample efficiency and scalability in complex environments (Kulkarni et al. 2016). While MARL has demonstrated success in robotics, autonomous systems, and network optimization, its application in cybersecurity for CPS remains underexplored. Furthermore, existing MARL-based intrusion detection and defense mechanisms often lack adversarial robustness, making them susceptible to sophisticated cyber threats.

Adversarial Learning and Game Theory

In the context of cybersecurity, adversarial learning involves modeling malicious actors who attempt to evade detection or manipulate system behavior(Goodfellow, Shlens, and Szegedy 2015). This aligns well with game-theoretic security models, where defenders and attackers can be represented as players with conflicting objectives(Shapley 1953). Incorporating adversarial learning into security systems enables proactive defense strategies, where defenders are trained against worst-case attack scenarios to enhance system resilience (Standen, Kim, and Szabo 2025). In CPS security, adversarial learning is particularly relevant because attackers can leverage AI-driven techniques to continuously adapt their strategies. Integrating adversarial learning into MARL-based defense mechanisms allows security agents to anticipate and counteract adaptive cyber threats. Additionally, the competitive-cooperative nature of multi-agent environments makes game-theoretic approaches particularly useful, as defenders must coordinate responses while mitigating attacks from intelligent adversaries (Conti et al. 2018).

Positioning of This Work

Although there have been several investigations into MARL for intrusion detection(Louati, Ktata, and Amous 2024) and adversarial learning for robust classification(Goodfellow, Shlens, and Szegedy 2015), there is a lack of research that integrates hierarchical MARL with adversarial training specifically for CPS security. Addressing this gap, our work introduces a Hierarchical Adversarially-Resilient Multi-Agent Reinforcement Learning framework that: (a) Structures multiple defender agents under a hierarchical coordinator, ensuring efficient and scalable threat mitigation. (b) Incorporates an adaptive adversarial training loop, where the system continuously learns from evolving attack strategies to enhance resilience. By bridging hierarchical MARL and adversarial learning, our approach extends prior work and contributes to the growing field of AI-driven cybersecurity for CPS (Rashid et al. 2020). The proposed framework is designed to improve real-time intrusion detection, response efficiency, and adaptability, making it a novel and practical solution for securing modern CPS environments.

Theoretical Foundations for Hierarchical Adversarial MARL

In this section, we formalize the hierarchical multi-agent framework with an explicit adversarial agent. Let there be N defender agents (local) plus one global coordinator, collectively denoted $\{\pi_{\theta_1}, \ldots, \pi_{\theta_N}, \pi_\phi\}$, and one adversarial attacker π_ψ. The environment is thus modeled as a Markov game (partially observed stochastic game) with $(N+2)$ agents (N defender agents, a global coordinator, and an adaptive attacker).

Definition 0.1 (Markov Game with Adversary)**.** A Markov game (MG) with an adversarial agent is defined by the tuple

$$\mathcal{G} = \langle \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^{N+2}, P, \{r_i\}_{i=1}^{N+2}, \gamma \rangle,$$

where:

- $\mathcal{S}$ is the state space, including subsystem statuses and sensor data.
- $\mathcal{A}_i$ is the action space for agent $i \in \{1, \ldots, N, N+1, N+2\}$ (where $N+2$ represents the adversarial agent).
- $P(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})$ is the transition kernel describing how the environment evolves given state $\mathbf{s}$ and action $\mathbf{a}$.
- $r_i(\mathbf{s}, \mathbf{a})$ is the reward function for agent i.
  - Defender agents ($1 \leq i \leq N$): Receive positive rewards for successful detections or patches ($r_i > 0$) and negative rewards for false alarms or missed compromises ($r_i < 0$).
  - Attacker agent ($N+2$): Earns positive rewards for successful system compromises ($r_{N+2} > 0$).
- $\gamma \in (0, 1)$ is the discount factor that govern how agents value future rewards.

Each local defender observes a partial state $\omega_i \subset \mathbf{s}$, while the global coordinator maintains an aggregate representation $\mathbf{g}$ of local states or actions. The adversarial agent π_ψ may also observe only a partial state of the system.

To capture the interaction between local defenders, the global coordinator, and the adversary, we factorize the joint policy as follows:

Proposition 0.2 (Factorization of Joint Policy in Hierarchical-Adversarial Setting)**.** *Let $\pi_{\theta_i} i = 1^N$ be the local defender policies, $\pi\phi$ be the global coordinator policy, and π_ψ be the attacker policy. Then, the joint policy over actions $\mathbf{a} = a_1, \ldots, a_N, a_{global}, a_{attacker}$ can be expressed as:*

$$\pi_{\Theta,\phi,\psi}(\mathbf{a} \mid \mathbf{s}) = \Big(\prod_{i=1}^{N} \pi_{\theta_i}(a_i \mid \omega_i)\Big)\, \pi_\phi(a_{global} \mid \mathbf{g}) \times \pi_\psi(a_{attacker} \mid \omega_{att}).$$

Remark 1. This factorization forms the basis for multi-agent training, where each agent updates its policy using Proximal Policy Optimization (PPO) steps, contingent on its partial observability.

Generalized Advantage Estimation and PPO

Following (Schulman et al. 2016, 2017), each agent maintains a parametric policy π_θ with an associated value function $V_\theta(\mathbf{s})$. The advantage function, which estimates how favorable an action is compared to the expected value of the state, is defined as:

$$A_\theta(\mathbf{s}, a) = Q_\theta(\mathbf{s}, a) - V_\theta(\mathbf{s})$$

To compute advantage estimates, we use the Generalized Advantage Estimation (GAE) technique:

$$\hat{A}_t = \sum_{k=0}^{T-t-1} (\gamma\lambda)^k \delta_{t+k}, \quad \delta_t = r_t + \gamma V(\mathbf{s}_{t+1}) - V(\mathbf{s}_t).$$

Theorem 0.3 (Convergence of PPO in Hierarchical-Adversarial MARL)**.** *Consider the Markov game $\mathcal{G}$ with $N+2$ agents, each employing PPO updates with GAE. Let θ_i, ϕ, ψ be their respective parameters. If each agent's policy improves according to the clipped objective in (Schulman et al. 2017) within a bounded trust region, under standard assumptions (bounded rewards, Markov mixing, sufficiently large batch data and exploration), the system converges to a stationary point $(\theta_i^*, \phi^*, \psi^*)$ that constitutes a local Nash equilibrium. Specifically:*

$$\nabla_{\theta_i} \mathcal{L}(\theta_i^*; \theta_{-i}^*, \phi^*, \psi^*) = 0,$$
$$\nabla_\phi \mathcal{L}(\phi^*; \theta^*, \psi^*) = 0,$$
$$\nabla_\psi \mathcal{L}(\psi^*; \theta^*, \phi^*) = 0.$$

Sketch Proof of Theorem 0.3. Each agent's PPO update can be viewed as a stochastic gradient ascent step on the clipped surrogate objective, which ensures per-update monotonic improvement within a specified ratio bound. The hierarchical nature of our approach does not disrupt the fundamental convergence properties of PPO-based MARL. The global coordinator, despite aggregating decentralized agent policies, does not interfere with each agent's independent policy update but rather enforces a structured decision-making pipeline. Consequently, policy updates remain bounded within trust regions, preserving theoretical convergence guarantees. Because all agents share the environment, the joint updates track a multi-agent gradient, ensuring stability in learning. By standard arguments for actor-critic methods in Markov games (Zhang, Yang, and Basar 2021), if the reward and advantage estimates remain bounded and each agent explores sufficiently, then with diminishing step sizes the parameters converge to a stationary point. This point is a local Nash equilibrium: no single agent can unilaterally improve its objective without changing other agents' policies. □

Adversarial Resilience in Hierarchical Control

Definition 0.4 (Adversarial Resilience)**.** Let (t) be the set of subsystems compromised at time t.

- **Compromise time** τ_i of subsystem i is the number of consecutive steps for which $i \in (t)$ until it is restored, formally $\tau_i = \min\{k > 0 \mid i \notin (t+k)\}$.

- **Compromise frequency** of subsystem i is $f_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}[i \in (t)]$ over horizon T.
- **Bounded compromise ratio** is $\varrho = \frac{1}{N}\sum_{i=1}^{N} f_i\,, \qquad 0 \le \varrho \le 1.$

A defender policy set $\{\pi_{\theta_1}, \ldots, \pi_{\theta_N}, \pi_\phi\}$ is (ϵ, δ)-*resilient* if

$$\Pr\big[\varrho \le \epsilon\big] \ge 1 - \delta$$

for any attacker policy π_ψ admissible under the game dynamics.

Intuitively, adversarial resilience means that despite an attacker that learns or changes tactics, the hierarchical defenders maintain partial observability, coordinate responses, and keep compromise in check over time.

Theorem 0.5 (Bounded Compromise in Equilibrium)**.** *Let $\pi^*_{\theta_i}, \pi^*_\phi$, and π^*_ψ be the equilibrium policies from Theorem 0.3. Suppose the environment imposes a cost $c > 0$ on each compromised subsystem per time step for defenders and a reward $r_a > 0$ for each compromised subsystem for the attacker. If c is sufficiently large relative to r_a, then the compromise ratio ϱ^* in the long-run equilibrium is strictly less than 1. Formally:*

$$\varrho^* = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \frac{\sum_{i=1}^{N} \mathbf{1}\{\textit{subsystem } i \textit{ at time } t\}}{N} < 1.$$

Sketch Proof of Theorem 0.5. See Appendix for the full derivation. Informally, the attacker's marginal gain from compromising an additional subsystem must be weighed against defenders' marginal cost for letting it remain compromised. If the defenders' policies can patch or quarantine effectively, the attacker cannot systematically keep all N subsystems compromised without incurring large negative feedback (through the defenders' best response strategies). Thus, $\varrho^* < 1$ in equilibrium unless the attacker reward r_a dwarfs the defenders' ability to penalize or detect. This result suggests that even in the presence of highly adaptive attackers, the system maintains a level of resilience where at least a fraction of subsystems remains uncompromised. This aligns with real-world security requirements, where maintaining full protection is impractical, but ensuring partial containment prevents widespread failures. By balancing proactive detection and strategic intervention, the hierarchical framework ensures that no single adversary strategy can indefinitely degrade the entire system. □

Remark 2. The synergy between local defenders (rapid quarantines) and a global coordinator (system patches) exemplifies hierarchical synergy. Even if local defenders occasionally miss an attack, the global coordinator can handle system-wide anomalies, ensuring no single attacker strategy can indefinitely compromise all subsystems.

Proposed Methodology

Hierarchical Multi-Agent Architecture

In our framework, defender agents are organized hierarchically to mirror real-world organizational structures in industrial or IoT environments. Local agents each monitor specific subsystems or network segments, processing local sensor data and triggering immediate responses (e.g., blocking suspicious traffic). A global coordinator receives summarized state information from all local agents, resolves conflicting actions, and implements system-wide defensive measures such as network isolation or forced restarts of compromised nodes.

At the bottom tier, local agents operate on partial observations of their assigned subsystem, allowing them to perform lightweight, real-time anomaly detection. At the top tier, the global coordinator has access to high-level aggregated information, enabling network-wide interventions (e.g., microsegmentation or mass patch deployment). This design is especially beneficial in large-scale systems where fully centralized control becomes computationally infeasible (Kulkarni et al. 2016), since it leverages local autonomy to reduce communication overhead and accelerate response.

Conceptually, the hierarchical arrangement allows each local agent to specialize in detecting and handling threats within its domain, leading to faster and more accurate detection at the subsystem level. Meanwhile, the global coordinator maintains a holistic view of the entire CPS, enabling better resource allocation and higher-level decision-making. As a result, the local and global layers collectively mitigate attacks more effectively than monolithic or purely decentralized defenses.

Adversarially-Aware Training

A novel aspect of our method is the adversarial training loop, wherein a simulated *attacker agent* with an evolving policy is introduced. Unlike static or random threats, this attacker adapts its strategies over time, attempting to compromise the system by exploiting vulnerabilities, launching denial-of-service attacks, or tampering with sensor data to degrade process quality. This adversary is trained *in tandem* with the defender agents, continually refining its attack strategies based on defender actions. Conversely, defenders learn robust behaviors to counter more sophisticated threat patterns. By framing the interaction as a repeated, partially observable stochastic game, both attackers and defenders iteratively improve their policies (Shapley 1953; Tambe 2011).

The attacker receives feedback about how many subsystems it successfully compromises or how often it remains undetected; the defender side (local + global) receives negative rewards for letting a subsystem remain compromised and positive rewards for correct detection and rapid patching. Over multiple episodes, these opposing objectives shape a minimax-style equilibrium, leading to *adversarial resilience*: the system must remain vigilant against an intelligent attacker that changes tactics over time.

Reward Structures and Policy Optimization

The learning process relies on a hybrid reward function that captures both local and global objectives. At the local level, each agent is rewarded for correctly identifying or neutralizing threats and penalized for false alarms that interrupt legitimate operations. At the global level, the system receives rewards for maintaining uninterrupted operation, minimizing resource overhead, and preserving overall safety. We adopt

a hierarchical multi-critic approach, where the local critics evaluate immediate detection performance, and a global critic focuses on system-wide metrics (Lowe et al. 2017; Yang et al. 2018).

For policy optimization, our implementation utilizes an extension of Proximal Policy Optimization (PPO) adapted for multi-agent environments (Schulman et al. 2017). Each local agent's policy is represented by a neural network, potentially a graph neural network (GNN) or a transformer-based model for enhanced processing of heterogeneous sensor data (Veličković et al. 2018). The global coordinator leverages aggregated embeddings from local agents, employing a separate neural network to learn the optimal coordination policy. By periodically synchronizing policy updates in a batch or round-robin fashion, the agents learn joint strategies that balance local autonomy with global oversight.

Extensions and Implementation Improvements

Beyond the core hierarchy and adversarial loop, our methodology incorporates additional practical considerations:

- Partial Observability and Scalable Communication: Local agents operate with partial observability, restricting their access to only subsystem-level data. This design minimizes communication overhead while preserving scalability. Aggregated messages to the global coordinator are compressed to limit bandwidth usage.
- Formal Safety Checks: Certain high-risk actions (e.g., quarantining all subsystems) trigger domain-specific safety checks to prevent catastrophic decisions, mirroring real ICS safety protocols.
- Transferability and Generalization: The learned policies can potentially transfer to other CPS domains (e.g., smart grid, autonomous vehicles) if sensor features and reward design are adapted accordingly.

These extensions position the hierarchical adversarially-resilient MARL framework as a flexible, real-world ready solution to emerging security threats in interconnected industrial environments.

Implementation and Experiment Design

Testbed Overview

To evaluate the proposed framework, we built a simulated industrial IoT environment emulating a small-scale smart factory. The environment includes multiple sensor nodes (e.g., temperature, vibration, and pressure sensors), actuators (e.g., valves and robotic arms), and an industrial control system using standard communication protocols such as Modbus/TCP. Each subsystem is represented by a local defender agent, while a single global coordinator oversees the entire system.

Attack Scenarios

We consider a range of attack vectors, including:

- DoS Attacks: Overwhelming the control network with traffic to degrade response time.
- Data Tampering: Manipulating sensor readings to trigger incorrect actuator commands.
- Stealthy Advanced Persistent Threats (APTs): Gradual infiltration that aims to remain undetected while collecting critical intelligence or planting malicious scripts.

The adversary agent's policy evolves based on its own reward function, which incentivizes remaining undetected while causing maximal disruption or data corruption. This setup ensures that defenders are exposed to diverse, dynamic threats during training and testing, fostering adversarial resilience.

Implementation Steps

The implementation of the proposed HAMARL framework consists of several key steps, from environment initialization to performance evaluation.

Environment Initialization To simulate a realistic industrial IoT environment, the experimental setup models normal industrial processes, sensor data flows, and baseline operational states. Given the constraints of real-world data availability, synthetic datasets are employed to emulate physical processes, ensuring the simulation captures representative system behaviors. Additionally, network simulation modules such as NS-3 are integrated where necessary to simulate realistic packet-level interactions, particularly for assessing network-based attacks and defensive interventions.

Local Agent Deployment Each local defender agent is assigned to a specific subsystem within the CPS, where it receives partial observations, including sensor readings and local network traffic patterns. These agents operate autonomously and are trained to make critical security decisions, including raising intrusion alarms, initiating partial quarantines, or escalating threats to the global coordinator. This decentralized approach ensures real-time detection and response capabilities while reducing the risk of single points of failure.

Global Coordination A global coordinator oversees the entire CPS security infrastructure, aggregating compressed state/action proposals received from local agents. Unlike purely decentralized models, the global coordinator is responsible for implementing high-level security policies that extend beyond localized responses. These include network micro-segmentation, where the system isolates compromised nodes, and coordinated security actions, such as initiating system-wide alerts or enforcing access restrictions based on detected threat patterns. The hierarchical control structure ensures that individual agents operate autonomously, while the coordinator enforces strategic, system-wide defense mechanisms.

Adversarial Training Loop To improve system resilience, we introduce an adversarial training loop in which an adaptive attacker agent continuously refines its strategies to simulate sophisticated cyber threats. The attacker is initialized with a baseline policy designed to compromise vulnerable subsystems through targeted exploits, denial-of-service (DoS) attacks, or stealthy infiltration strategies.

Training proceeds in an alternating fashion, where the attacker iteratively refines its attack strategies, while defender agents learn to adapt and counteract these threats. Reinforcement learning-based policy updates are conducted using Proximal Policy Optimization (PPO) (Lowe et al. 2017; Schulman et al. 2017), ensuring robust adversarial adaptation. This iterative learning process exposes the defense system to realistic, evolving threats, ultimately enhancing its ability to detect and mitigate novel attacks.

Reward Engineering To guide agent learning, a carefully designed reward function is implemented at both local and global levels, ensuring that agents are incentivized to enhance security while maintaining system stability. Local rewards are structured to balance detection accuracy, minimized false positives, and subsystem uptime, preventing excessive intervention in benign scenarios. At the global level, rewards are based on overall system resilience, ensuring that security actions do not disrupt essential industrial processes or overwhelm computational resources. The attacker agent, in contrast, is rewarded for successful system compromises, emphasizing stealth, system disruption, and the duration for which subsystems remain compromised. This reward structure simulates realistic adversarial engagements, pushing defenders to develop adaptive, high-precision detection and response strategies.

Evaluation The trained framework is evaluated across multiple dimensions to ensure its effectiveness in securing CPS environments. The detection and response performance is assessed by measuring detection latency, false alarm rates, and the accuracy of security interventions. The framework's impact on system stability is analyzed by tracking operational continuity, throughput, and resource overhead (CPU and bandwidth usage) to confirm that security mechanisms do not introduce excessive computational burdens. Finally, adaptive robustness is tested by exposing the trained system to novel attack vectors not encountered during training, ensuring that the model generalizes effectively to unseen threats. These evaluation metrics provide a comprehensive understanding of the framework's performance and resilience in dynamic adversarial settings.

In each training cycle, agents generate experiences (state-action-reward tuples), which are then aggregated into replay buffers for mini-batch updates. By restricting local agents to subsystem-level data, we ensure scalability and partial observability, while the global coordinator addresses system-wide coherence. The adversary's training ensures that defenders develop resilient strategies capable of handling adaptive threats.

Experimental Setup

Environment. We extend the open-source Cyber-Battle-Sim toolkit to emulate a *smart-factory* industrial IoT line with $N{=}8$ PLC-driven cells, 64 sensors (temperature, vibration, flow) and a Modbus/TCP control network.

State spaces. Each local defender observes $\omega_i^t = \langle \mathbf{s}_i^t, \mathbf{n}_i^t \rangle$ where $\mathbf{s}_i^t \in \mathbb{R}^{12}$ are normalised sensor features and $\mathbf{n}_i^t \in \mathbb{R}^5$ are network-level statistics (packet loss, RTT, SYN count, *etc.*). The global coordinator receives $g^t = \mathrm{Concat}(\mathrm{Pool}_i\, h_i^t)$, a 32-dimensional pooled embedding of local agent hidden states.

Action spaces.

- **Local defender** $a_i^t \in \{\text{NOOP, ALERT, QUARANTINE, PATCH}\}$.
- **Global coordinator** $a_{\mathrm{G}}^t \in \{\text{NOOP, ISOLATE-SEG, ROLL-PATCH, RESET-NODE}\}$.
- **Attacker** $a_{\mathrm{A}}^t \in \{\text{SCAN, LATERAL, DOS, TAMPER}\}$.

Reward design. Local rewards: $r_i = +1$ (true-positive), -0.2 (false-positive), -1 (miss). Global reward: $R = -0.1\,|(t)| - 0.01\,\text{DOWNTIME} + 0.2\,\text{UPTIME}$.

Networks & training. Local policies use a 2-layer Graph Attention Network (hidden 32, heads 4); the coordinator is a 3-layer MLP (64-32-16). All critics share weights with the actors except for an output head. Optimiser: Adam, 10^{-4}; $\lambda_{\text{GAE}} = 0.95$, $\gamma = 0.99$; PPO clip $\epsilon = 0.2$; batch 32, 1 000 episodes.

Results and Analysis

Baseline Comparisons

We compare the proposed hierarchical adversarially-resilient MARL framework against three baselines, each representing a different approach to CPS security. The Single-Agent RL baseline represents a centralized approach, which, while effective in small-scale systems, lacks scalability in distributed environments. The Non-Hierarchical MARL baseline evaluates how decentralized approaches fare without a coordinating agent, exposing issues of miscoordination and increased message overhead. The Rule-Based Intrusion Detection system serves as a traditional benchmark, demonstrating the limitations of static defense mechanisms compared to adaptive learning models. These baselines provide insight into the advantages of hierarchical coordination and adversarial resilience.

Experimental results demonstrate that HAMARL significantly outperforms baselines in terms of detection accuracy, with a marked reduction in mean-time-to-detect (MTTD). Notably, hierarchical coordination reduced false positives by centralizing evidence gathered from distributed sensors, while adversarial training improved detection rates for newly introduced and sophisticated threat behaviors.

Attack Detection and Operational Continuity

Our experiments show that local agents trained under adversarial conditions adapted quickly to stealthy APT attacks, maintaining above 90% detection rates even when attackers changed tactics mid-episode. The global coordinator played a pivotal role in preventing cascading failures, for example, by isolating compromised nodes or re-routing critical control signals before the entire production line could be halted. As shown in Figure 2, detection rates improved dramatically after approximately 200 training episodes, indicating that defenders learn effective countermeasures once the adversary escalates its strategies.

Figure 1: Total episode reward for defenders (averaged), the global coordinator, and the attacker, across training episodes in the adversarial environment. Higher defenders' reward indicates successful detection and minimized compromise duration, whereas higher attacker reward indicates prolonged or widespread subsystem compromise.

Resource overhead remained manageable, in part due to parallelization strategies and localized decision-making, ensuring real-time inference capacity on moderate hardware. Furthermore, operators can tune the aggressiveness of local quarantines versus global patches to balance false alarms against critical subsystem uptime.

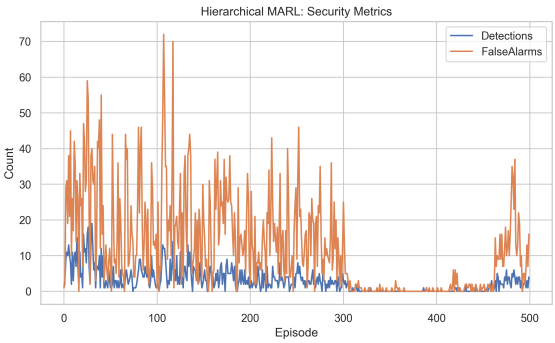

Figure 2: Detection rates over the course of training, illustrating the defenders' improved ability to identify evolving adversarial tactics.

Scalability Analysis

One primary challenge in multi-agent systems is scalability, as increased numbers of agents can heighten communication overhead and slow convergence (Buşoniu, Babuška, and Schutter 2010). In our simulations, adding new defender agents to monitor additional subsystems led to a near-linear increase in training time but only a modest increase in runtime overhead, thanks to asynchronous training updates and hierarchical credit assignment. The hierarchical structure not only improves detection but also maintains tractable scaling properties as CPS complexity grows.

Conclusion and Future Work

The proposed Hierarchical Adversarially-Resilient Multi-Agent Reinforcement Learning framework presents a novel and effective approach to securing large-scale Cyber-Physical Systems against evolving cyber threats. By leveraging a hierarchical coordination mechanism, where local agents specialize in real-time threat detection while a global coordinator enforces system-wide security policies, our approach significantly enhances threat detection, response efficiency, and system resilience. The integration of adversarial training ensures that the framework generalizes well to previously unseen attack vectors, enabling defenders to proactively adapt to AI-driven cyber threats rather than relying on static security measures. Experimental evaluations on a simulated industrial IoT testbed demonstrate that HAMARL achieves higher detection accuracy, reduced response time, and greater operational continuity compared to traditional non-hierarchical MARL and rule-based methods.

From an industrial perspective, deploying such a system could significantly mitigate cybersecurity risks in critical infrastructures where downtime or data compromise can result in severe economic and safety consequences. The study highlights the growing necessity for AI-driven, continuously adaptive defense mechanisms in modern CPS environments. Unlike conventional static rule-based intrusion detection systems, HAMARL dynamically refines its defense strategies in response to evolving cyber threats, ensuring robust security even in adversarial environments.

Despite the promising outcomes, several challenges and limitations remain. First, computational and data requirements for training multiple reinforcement learning agents, particularly in resource-constrained CPS environments, could be prohibitive. Optimizing training efficiency while maintaining robust defenses is an area for further exploration. Second, hierarchical MARL introduces additional complexity in designing reward structures and coordinating local-global policies, which may require domain-specific tuning to ensure effective real-world deployment. Finally, while our simulation-based evaluation demonstrates the framework's scalability, real-world deployment would necessitate rigorous validation, regulatory compliance (e.g., IEC 62443 for industrial automation), and safety testing before practical adoption.

To further advance AI-driven security for CPS, several avenues of future research should be explored. One promising direction is transfer learning, where policies trained in one CPS domain, such as a smart factory, could be adapted to other critical infrastructures, such as smart grids or autonomous vehicle networks. Another important area is the integration of Explainable AI to provide interpretability and transparency in defense mechanisms, ensuring that security decisions can be understood by human operators and auditors. Additionally, formal verification techniques could be incorporated to ensure that RL-driven security policies do not inadvertently disrupt safety-critical operations. Lastly, extending the adversarial training framework to multi-attacker scenarios would offer valuable insights into how defenders can adapt to adversaries that collaborate or compete against each other, introducing new layers of com-

Table 1: Detection performance (mean ±95% CI).

Method	MTTD↓	F1↑	False Alarm↓
Rule-Based IDS	87.4±6.1	0.42±0.03	17.6%
Non-Hier. MARL	41.8±4.5	0.71±0.02	11.3%
HAMARL (ours)	**18.3±2.7**	**0.88±0.01**	**6.1%**

Table 2: Training wall-clock time vs. number of defender agents (Intel Xeon 6230, RTX A6000).

# Agents	4	8	12
Non-Hier. MARL (h)	4.1	9.5	18.7
HAMARL (h)	4.4	10.3	20.9

plexity in CPS security.

As CPS networks continue to evolve and expand in scale and complexity, ensuring their security remains a pressing research and industrial priority. The insights from this work contribute to the growing body of knowledge in AI-driven cybersecurity, paving the way for practical, scalable, and self-adaptive defense mechanisms. Future research should focus on refining hierarchical learning structures, enhancing real-world applicability, and exploring cross-domain generalization, ultimately striving toward a new generation of intelligent, resilient, and adaptive security solutions for critical infrastructures.

Theorem 0.3: Multi-Agent Convergence Proof

Proof. We expand the sketch proof as follows:

Step 1: Monotonic Policy Improvement per Agent: For agent i, the PPO update aims to solve

$$\max_{\theta_i} \mathbb{E}\big[\min(\rho_t(\theta_i)\hat{A}_t,\ \text{clip}\{\rho_t(\theta_i), 1 \pm \varepsilon\}\hat{A}_t)\big],$$

where $\rho_t(\theta_i) = \frac{\pi_{\theta_i}(a_{i,t}|\omega_{i,t})}{\pi_{\theta_i^{\text{old}}}(a_{i,t}|\omega_{i,t})}$ and $\hat{A}_t$ is the GAE advantage. By bounding the ratio within $[1-\varepsilon,\ 1+\varepsilon]$, we ensure that the update does not drastically degrade performance, establishing local monotonic improvement in agent i's objective.

Step 2: Joint Updates in a Markov Game: While agent i updates θ_i, the other agents hold their parameters θ_{-i}, ϕ, ψ fixed. This yields a best response step for agent i. Repeating over all agents in a round-robin or simultaneous fashion (depending on the training scheme) can be viewed as approximate gradient ascent in the multi-agent reward space (Zhang, Yang, and Basar 2021).

Step 3: Stochastic Approximation and Boundedness: Assume $r_i(\mathbf{s}, \mathbf{a}) \in [r_{\min}, r_{\max}]$, $\|\nabla_{\theta_i} L_i\| \leq M$, and that each agent's policy covers its action space with a minimum exploration probability $\delta > 0$. By standard results in stochastic approximation, the parameter updates converge to a stationary point $\nabla_{\theta_i} L_i(\theta^*) = 0$ for each i, provided the step sizes $\alpha^k \to 0$ over iterations k.

Step 4: Local Nash Equilibrium: A stationary point in multi-agent policy space implies no agent can unilaterally improve its expected return given the other agents' policies remain fixed. Thus, $(\theta_i^*, \phi^*, \psi^*)$ is a local Nash equilibrium. Global (or unique) equilibrium is not guaranteed without additional assumptions (convex-concavity, zero-sum structure, etc.).

□

Proof of Theorem 0.5

Proof. Let $c > 0$ be the defenders' penalty per compromised subsystem, and $r_a > 0$ be the attacker's reward per compromised subsystem. Suppose the attacker attempts to compromise all N subsystems. Each local defender i plus the global coordinator can respond with quarantines and patches to reduce $\sum_{i=1}^{N} \mathbf{1}\{\text{subsystem } i \text{ compromised}\}$. As θ_i, ϕ converge to best-response policies, defenders effectively minimize the time that any single subsystem remains compromised; thus, the attacker cannot indefinitely maintain a complete state of compromise without incurring immediate quarantines or system patches.

Quantitatively, in each step t, the expected net payoff to the attacker from compromise is $r_a \times$ (number of compromised) minus the defenders' best responses. Because the defenders incur cost c per compromised subsystem, their equilibrium strategy invests sufficient actions (quarantines, patches) to reduce total compromise. Provided c is large enough relative to r_a, a subset of subsystems remains un-compromised on average, ensuring $\varrho^* < 1$. This argument can be formalized via potential function arguments or Markov chain equilibrium analysis (see (Shoham and Leyton-Brown 2008) for a multi-agent potential game perspective). Therefore, at the joint equilibrium, the fraction of compromised subsystems is less than 1, completing the proof. □

References

Baheti, R.; and Gill, H. 2011. Cyber-physical systems. In *The Impact of Control Technology*.

Buşoniu, L.; Babuška, R.; and Schutter, B. D. 2010. Multi-agent reinforcement learning: An overview. In *Innovations in Multi-Agent Systems and Applications – 1*. Springer.

Conti, M.; Dehghantanha, A.; Franke, K.; and Watson, S. 2018. Internet of Things security and forensics: Challenges and opportunities. *Future Generation Computer Systems*.

Goodfellow, I.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations (ICLR)*.

Kulkarni, T. D.; Narasimhan, K.; Saeedi, A.; and Tenenbaum, J. 2016. Hierarchical Deep Reinforcement Learning:

Integrating Temporal Abstraction and Intrinsic Motivation. In *Advances in Neural Information Processing Systems 29 (NIPS)*.

Lee, E. A. 2008. Cyber physical systems: Design challenges. In *11th IEEE International Symposium on Object Oriented Real-Time Distributed Computing*.

Louati, F.; Ktata, F. B.; and Amous, I. 2024. Big-IDS: a decentralized multi agent reinforcement learning approach for distributed intrusion detection in big data networks. *Cluster Computing*, 27(5): 6823–6841.

Lowe, R.; Wu, Y. I.; Tamar, A.; Harb, J.; Pieter Abbeel, O.; and Mordatch, I. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Matignon, L.; Laurent, G. J.; and Fort-Piat, N. L. 2012. Independent Reinforcement Learners in Cooperative Markov Games: a Survey regarding Coordination Problems. *The Knowledge Engineering Review*.

Rashid, T.; Samvelyan, M.; De Witt, C. S.; Farquhar, G.; Foerster, J.; and Whiteson, S. 2020. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178): 1–51.

Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.; and Abbeel, P. 2016. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In *International Conference on Learning Representations (ICLR)*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*.

Shapley, L. 1953. Stochastic Games. *Proceedings of the National Academy of Sciences*.

Shoham, Y.; and Leyton-Brown, K. 2008. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press.

Standen, M.; Kim, J.; and Szabo, C. 2025. Adversarial Machine Learning Attacks and Defences in Multi-Agent Reinforcement Learning. *ACM Computing Surveys*, 57(5): 1–35.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition.

Tambe, M. 2011. *Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned*. Cambridge University Press.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2018. Graph Attention Networks. In *International Conference on Learning Representations (ICLR)*.

Vezhnevets, A. S.; Osindero, S.; Schaul, T.; Heess, N.; Jaderberg, M.; Silver, D.; and Kavukcuoglu, K. 2017. FeUdal Networks for Hierarchical Reinforcement Learning. In *International Conference on Machine Learning (ICML)*, 3540–3549. PMLR.

Wolf, M.; and Serpanos, D. 2019. Safety and Security in Cyber-Physical Systems and Internet-of-Things Systems. *Proceedings of the IEEE*.

Yang, Y.; Luo, R.; Li, M.; Zhou, M.; Zhang, W.; and Wang, J. 2018. Mean Field Multi-Agent Reinforcement Learning. In *International Conference on Machine Learning (ICML)*, 5571–5580. PMLR.

Zhang, K.; Yang, Z.; and Basar, T. 2021. Multi-agent reinforcement learning: A selective overview of theories and algorithms. In *Handbook of Reinforcement Learning and Control*.