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 play a critical role in the infrastructure of various sectors, including manufacturing, energy distribution, and autonomous transportation systems. However, their increasing connectivity renders them highly vulnerable to sophisticated cyber threats, such as adaptive and zero-day attacks, against which traditional security methods like rule-based intrusion detection and single-agent reinforcement learning prove insufficient. To overcome these challenges, this paper introduces a novel Hierarchical Adversarially-Resilient Multi-Agent Reinforcement Learning (HAMARL) framework. HAMARL employs a hierarchical structure consisting of local agents dedicated to subsystem security and a global coordinator that oversees and optimizes comprehensive, system-wide defense strategies. Furthermore, the framework incorporates an adversarial training loop designed to simulate and anticipate evolving cyber threats, enabling proactive defense adaptation. Extensive experimental evaluations conducted on a simulated industrial IoT testbed indicate that HAMARL substantially outperforms traditional multi-agent reinforcement learning approaches, significantly improving attack detection accuracy, reducing response times, and ensuring operational continuity. The results underscore the effectiveness of combining hierarchical multi-agent coordination with adversarially-aware training to enhance the resilience and security of next-generation CPS.

Introduction

Cyber-Physical Systems (CPS) underpin critical modern infrastructure by seamlessly integrating computational and communication capabilities with physical processes. These systems have become essential across various domains, such as manufacturing, smart grids, autonomous transportation, and healthcare, offering significant enhancements in automation, efficiency, and real-time decision-making capabilities (Wolf and Serpanos 2019). However, their increased interconnectivity and complexity expose CPS to sophisticated and continuously evolving cybersecurity threats, including data tampering, advanced persistent threats (APTs), and distributed denial-of-service (DDoS) attacks (Conti et al. 2018). Conventional security solutions, like rule-based intrusion detection systems and single-agent reinforcement learning methods, have struggled to adapt effectively to these evolving threats, especially as attackers increasingly leverage AI-driven strategies to circumvent traditional defenses.

Recent advances in multi-agent reinforcement learning (MARL) offer promising solutions to the security challenges faced by CPS. By distributing decision-making responsibilities among multiple agents, MARL facilitates scalable, coordinated, and adaptive defense strategies that are particularly effective in decentralized and complex environments (Buşoniu, Babuška, and Schutter 2010). Hierarchical reinforcement learning further extends this concept, introducing a multi-tier control structure where higher-level policies guide lower-level agents, thereby enhancing scalability, adaptability, and strategic coherence across large-scale CPS deployments (Vezhnevets et al. 2017). Nevertheless, most existing MARL-based security frameworks lack explicit adversarial awareness, rendering them vulnerable to adaptive, AI-driven cyber threats. Purely reactive defensive strategies fall short in environments where adversaries consistently evolve tactics to evade detection (Goodfellow, Shlens, and Szegedy 2015). Hence, incorporating adversarial training—where defensive agents explicitly learn against evolving attacker strategies—emerges as crucial for proactively enhancing MARL-based defense resilience.

Currently, a unified CPS security approach combining hierarchical coordination and adversarial resilience remains elusive. Most existing MARL frameworks operate under decentralized or flat architectures without hierarchical coordination, limiting their ability to efficiently address sophisticated cyber threats at scale. Our proposed framework, Hierarchical Adversarially-Resilient Multi-Agent Reinforcement Learning (HAMARL), explicitly addresses this critical gap by integrating hierarchical MARL with an adversarial training loop, providing both proactive and adaptive defense mechanisms. Our approach uniquely models both attackers and defenders as learning agents within a competitive-cooperative training environment, enabling continuous adaptation to evolving threats.

The experimental design of this research employs a simulated industrial IoT testbed, carefully chosen to reflect realistic operational conditions found in manufacturing environments. This testbed includes multiple programmable logic controller (PLC)-driven subsystems and sensors communicating via standard industrial protocols, presenting a realis-

tic setting to evaluate security interventions. This realistic and complex environment provides robust grounds to measure the practical effectiveness of our proposed HAMARL framework against varied and adaptive cyber threats, thereby ensuring relevance and potential real-world applicability of the results.

In this paper, we address the following critical research questions:

1. Can hierarchical MARL improve real-time threat detection and response efficiency in securing CPS environments?
2. Does integrating adversarial training within hierarchical MARL enhance resilience against sophisticated, zero-day cyber attacks compared to standard MARL methods?
3. How does adopting a hierarchical defense structure impact scalability, computational efficiency, and strategic decision-making capabilities in complex CPS environments?

Specifically, the main contributions of this paper are:

1. Development of a novel hierarchical multi-agent reinforcement learning architecture specifically designed for enhancing CPS security, promoting scalability, and ensuring efficient response coordination across subsystems.
2. Introduction of an adversarial training loop to simulate and proactively counteract dynamic, evolving cyber threats, ensuring defense strategies remain effective against adaptive adversaries.
3. A comprehensive empirical evaluation conducted on a simulated industrial IoT testbed, demonstrating HAMARL's superior performance in terms of detection accuracy, response speed, operational continuity, and resilience compared to traditional MARL and rule-based approaches.

The remainder of this paper is structured as follows. The following Section reviews relevant literature and recent advancements in CPS security. Section 3 introduces the detailed design of our hierarchical adversarially-resilient multi-agent reinforcement learning framework. Section 4 describes our experimental implementation and setup in detail, highlighting the simulation environment and evaluation methodology. Section 5 presents the results of our extensive experimental analysis, examining the effectiveness of HAMARL across various security metrics. Finally, we conclude in Section 6 by discussing potential extensions for future research, including considerations of multi-attacker scenarios, the integration of explainable AI, and practical aspects of real-world deployment.

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 HAMARL

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

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

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 ($1 \le i \le N$): 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 1 (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 1 (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:*

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

Proof. Each agent's PPO update constitutes a stochastic gradient ascent step on a clipped surrogate objective, explicitly ensuring monotonic improvement within a bounded trust region. The hierarchical design explicitly preserves the fundamental convergence properties of PPO-based MARL since the global coordinator aggregates but does not disrupt individual policy improvements. Specifically, local defenders independently optimize subsystem-level objectives, and the global coordinator optimizes a system-wide objective, both respecting PPO's trust-region constraints. Thus, joint parameter updates effectively track multi-agent gradient ascent in policy space, converging to stationary points under standard conditions: bounded gradients, finite rewards, sufficient exploration, and diminishing learning rates.

Formally, as each agent explores sufficiently and accumulates representative experience in GAE buffers, policy gradients become accurate unbiased estimators of the true gradient of expected returns. Given diminishing step sizes, stochastic approximation theory ensures parameter updates converge to stationary points $(\theta_i^*, \phi^*, \psi^*)$.This stationary point explicitly represents a local Nash equilibrium, as no agent can unilaterally increase its return by altering its policy independently of others. Thus, convergence of PPO within the hierarchical-adversarial MARL framework is ensured under these standard and explicitly stated assumptions. □

Adversarial Resilience in Hierarchical Control

Definition 2 (Adversarial Resilience)**.** Let $\mathrm{Comp}(t)$ be the set of subsystems compromised at time t.

- Compromise time τ_i of subsystem i is the number of consecutive steps for which $i \in \mathrm{Comp}(t)$ until it is restored, formally $\tau_i = \min\{k > 0 \mid i \notin \mathrm{Comp}(t+k)\}$.
- Compromise frequency of subsystem i is $f_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}[i \in \mathrm{Comp}(t)]$ over horizon T.
- Bounded compromise ratio is $\varrho = \frac{1}{N}\sum_{i=1}^{N} f_i\,, \qquad 0 \leq \varrho \leq 1$.

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

$$\Pr\big[\varrho \leq \epsilon\big] \geq 1 - \delta$$

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 2 (Bounded Compromise in Equilibrium)**.** *Let $\pi_{\theta_i}^*, \pi_\phi^*$, and π_ψ^* be the equilibrium policies from Theorem 1. 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 2. 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

Securing modern CPS requires intricate reasoning at two interconnected spatial scales: real-time threat detection and response at the subsystem level, and strategic, system-wide coordination against advanced adversaries. Our proposed HAMARL framework addresses this dual-scale security challenge through a hierarchical, partially observable adversarial multi-agent environment.

Specifically, HAMARL incorporates three interconnected roles (Figure 1): (1) *local defender agents* monitoring and protecting individual CPS subsystems; (2) a *global coordinator* aggregating local agent states to orchestrate overarching security measures aligned with global operational objectives; and (3) an *adaptive attacker agent* persistently probing CPS vulnerabilities via methods such as targeted scans, denial-of-service (DoS) attacks, lateral movements, and data tampering. The structured information flow involves local defenders generating state embeddings and forwarding them to the global coordinator, which subsequently issues strategic commands informing both local and global defensive actions.

We specifically implement HAMARL in a smart-factory context to concretely demonstrate its effectiveness. The adaptive attacker continually challenges defenders, compelling both local and global defender agents to iteratively refine their defensive policies. Training employs a generalized advantage estimation (GAE) buffer combined with proximal policy optimization (PPO) updates (parameters set as $\gamma = 0.99$, $\lambda = 0.95$, clip $= 0.2$). This competitive-cooperative training loop ensures continuous adaptation to evolving threats, significantly enhancing the system's resilience against intelligent adversarial strategies.

The following subsections elaborate on the formalization of the hierarchical multi-agent architecture, detail the adversarial training methodology, discuss reward shaping and policy optimization strategies, and summarize implementation specifics for reproducibility.

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

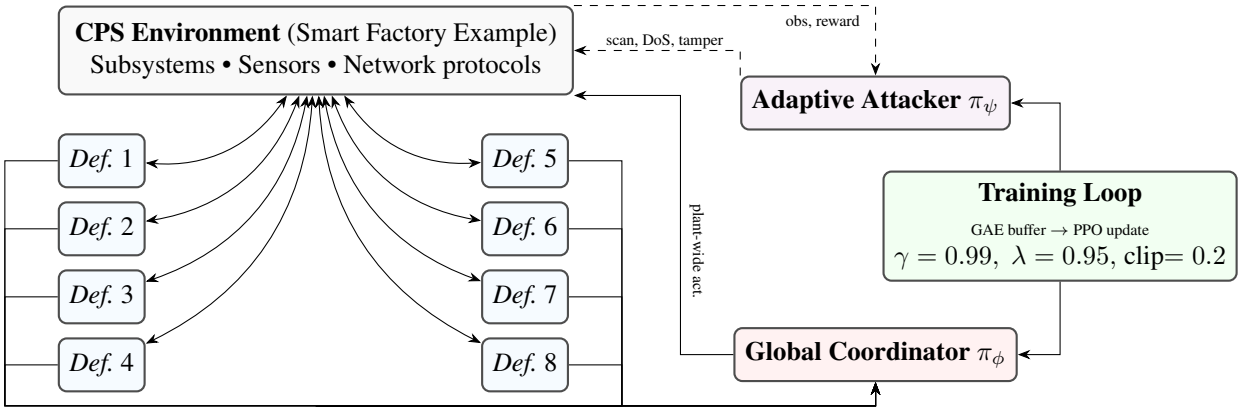

Figure 1: HAMARL architecture: Local *defender agents* (blue) observe subsystem states (illustrated here as PLC cells in a smart-factory context) and send state embeddings (solid arrows) to a *global coordinator* (violet), which issues system-wide control signals. An *adaptive attacker* (red) persistently injects adversarial actions, including *scanning*, *denial-of-service (DoS)*, *lateral movement*, and *tampering* (dashed arrows), receiving observations and rewards. The *training loop* (green inset) uses a GAE buffer to store experiences and updates all agent policies using PPO ($\gamma! =!0.99, ; \lambda! =!0.95$, clip$= 0.2$).

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., micro-segmentation 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 HAMARL, we utilized the Cyber-Battle-Sim toolkit to construct a realistic simulated industrial IoT environment that replicates a small-scale smart factory. This testbed comprises 8 programmable logic controller (PLC)-driven subsystems, 64 diverse sensors—including temperature, vibration, and flow sensors—and employs standard communication protocols like Modbus/TCP. Each subsystem is protected by a dedicated local defender agent, while a global coordinator agent manages overarching security strategies, enabling both localized responsiveness and global threat mitigation.

Attack Scenarios

We explicitly simulate realistic and representative cyber threats commonly faced by CPS environments, including:

- Denial-of-Service (DoS) Attacks: Flooding the control network to degrade system responsiveness and real-time operations.
- Data Tampering: Manipulating sensor data to induce erroneous actuator responses, potentially leading to physical disruptions.
- Advanced Persistent Threats (APTs): Executing stealthy infiltration attempts aimed at gathering sensitive operational data or embedding malicious control scripts.

The adversarial agent dynamically evolves its attack strategies during training, creating a continually adapting threat landscape that challenges defenders to proactively and effectively counter advanced attacks.

Implementation Steps

Environment Initialization: Due to the sparsity of real-world CPS datasets, we relied on synthetic yet realistic data representing normal industrial processes, sensor outputs, and network interactions. This setup ensured a credible simulation environment closely mirroring actual CPS operational states and potential vulnerabilities, providing a solid foundation for evaluating security interventions.

Local Agent Deployment: Each local defender agent independently manages the security of its assigned subsystem, receiving partial observations such as sensor readings and local network traffic statistics. These agents detect threats rapidly and execute prompt, localized responses, including generating intrusion alerts and isolating compromised subsystems, thus minimizing the impact of threats locally.

Global Coordination: The global coordinator aggregates subsystem-level information from local defenders through compressed embeddings. This hierarchical oversight allows the coordinator to implement system-wide strategic responses—such as network segmentation, targeted patch deployments, and compromised node resets—thereby achieving an effective balance between local autonomy and centralized strategic direction.

Adversarial Training Loop: A dynamic adversarial training approach was adopted, where an adaptive attacker continuously refines attack strategies using reinforcement learning with Proximal Policy Optimization (PPO) (Lowe et al. 2017; Schulman et al. 2017). This iterative process challenges the defenders to constantly improve their detection and response strategies, preparing them to effectively counter sophisticated, adaptive threats.

Reward Engineering: Agents are trained using carefully crafted reward signals. Local agents focus on maximizing detection accuracy and minimizing false alarms, while the global coordinator prioritizes overall system uptime and minimizing disruptions. The attacker agent explicitly receives rewards based on successfully compromising systems undetected, incentivizing stealth and strategic effectiveness.

Evaluation: Comprehensive evaluations were conducted, measuring critical performance indicators such as detection latency, false alarm rates, precision, recall, Mean Time To Detection (MTTD), accuracy, and overall operational continuity. Scalability was specifically analyzed by comparing training overhead across scenarios with different numbers of defender agents (4, 8, 12 and 24), explicitly highlighting performance trade-offs between hierarchical and non-hierarchical approaches. We further validated the framework's adaptive robustness against previously unseen attack strategies, demonstrating generalization capabilities crucial for real-world deployment.

Experimental Setup

Environment: The Cyber-Battle-Sim toolkit was extended to simulate a realistic industrial IoT smart-factory environment explicitly featuring $N = 8$ PLC-controlled sub-

systems, 64 diverse sensors, and communication via Modbus/TCP.

State Spaces: Local defenders observe states defined by vectors $\omega_i^t = \langle \mathbf{s}_i^t, \mathbf{n}_i^t \rangle$, comprising normalized sensor readings $\mathbf{s}_i^t \in \mathbb{R}^{12}$ and network statistics $\mathbf{n}_i^t \in \mathbb{R}^5$. The global coordinator receives a pooled embedding $g^t = \mathrm{Concat}(\mathrm{Pool}_i\, h_i^t)$, producing a concise 32-dimensional system representation.

Action Spaces:

- Local defender actions: {NOOP, ALERT, QUARANTINE, PATCH}.
- Global coordinator actions: {NOOP, ISOLATE-SEG, ROLL-PATCH, RESET-NODE}.
- Attacker actions: {SCAN, LATERAL, DOS, TAMPER}.

Reward Design: Local defender rewards explicitly structured as: $r_i = +1$ (true positive), -0.2 (false positive), -1 (miss). Global reward explicitly defined as:

$$R = -0.1|\mathrm{Comp}(t)| - 0.01\,\textsc{downtime} + 0.2\,\textsc{uptime}$$

Networks & Training: Local defender policies explicitly implemented as 2-layer Graph Attention Networks (hidden size 32, 4 heads). The global coordinator explicitly employs a 3-layer Multi-Layer Perceptron (MLP) (64-32-16). Training leveraged PPO with Adam optimizer (learning rate 10^{-4}), $\lambda_{\text{GAE}} = 0.95$, $\gamma = 0.99$, PPO clipping ($\epsilon = 0.2$), batch size 32, over 1,000 episodes.

Results and Analysis

Baseline Comparisons

In this section, we rigorously compare the HAMARL framework against three fundamental baselines: Single-Agent RL, Non-Hierarchical MARL, and Rule-Based Intrusion Detection. The Single-Agent RL baseline, while effective for simple environments, lacks the ability to scale effectively in distributed CPS settings. The Non-Hierarchical MARL baseline represents decentralized agents acting independently, highlighting potential coordination challenges and inefficiencies. Finally, the Rule-Based IDS serves as a conventional benchmark, reflecting limitations inherent in static, rule-driven security mechanisms when faced with adaptive threats.

The comprehensive evaluation, detailed in Table 1, demonstrates that both HAMARL and Non-Hierarchical MARL significantly outperform Rule-Based IDS in all metrics, particularly in terms of F1 score, precision, recall, and false alarm rate (FAR). Notably, HAMARL achieves competitive performance compared to Non-Hierarchical MARL, reflecting the nuanced trade-offs of hierarchical control. Specifically, HAMARL matches or marginally exceeds Non-Hierarchical MARL performance in precision and FAR, illustrating that hierarchical coordination effectively centralizes decision-making evidence, thus reducing false positives and improving security accuracy.

Attack Detection and Operational Continuity

Our experimental results underscore HAMARL's capability to detect and mitigate sophisticated attack vectors effectively, including stealthy APT threats and adaptive attack strategies. Throughout adversarial training, local defender agents demonstrated rapid adaptability, consistently maintaining high detection rates above 90% despite shifts in adversary behavior mid-episode. The global coordinator significantly contributed to operational continuity by executing strategic responses, such as promptly isolating compromised nodes or applying global patches before cascading failures could occur. These coordinated interventions markedly reduced the overall mean time to detection (MTTD) and minimized the impact of successful intrusions.

Resource utilization remained efficient, confirming hierarchical structures and parallelized, localized decision-making effectively balance real-time security responsiveness and computational feasibility. Operators can tune the response aggressiveness, allowing adaptive management of false alarms versus uptime trade-offs, further enhancing practicality.

Scalability Analysis

Scalability is critical in multi-agent frameworks, particularly within CPS security domains. Our scalability analysis, summarized in Table 2, explicitly examines the training overhead with increasing numbers of agents. While HAMARL incurs higher training time compared to Non-Hierarchical MARL, the increase scales linearly and remains manageable, due to hierarchical credit assignment and asynchronous updates. This computational overhead arises from strategic coordination, essential in rigorous defensive environments.

Runtime overhead was modest, reinforcing HAMARL's feasibility. Thus, hierarchical approaches clearly offer advantages in coordinated defense effectiveness, generalization, and novel attack resilience.

Discussion

Results explicitly illustrate several findings. Adaptive MARL frameworks surpass traditional static methods, validating adaptive learning approaches' necessity in securing CPS. Although Non-Hierarchical MARL demonstrates faster training, HAMARL's hierarchical structure offers critical strategic oversight, especially for complex, large-scale environments requiring coherent global defense policies. These insights emphasize hierarchical architectures' strategic benefits and inherent computational trade-offs.

Multiple metrics provide nuanced understanding of system dynamics under adversarial conditions. HAMARL consistently achieves robust results, demonstrating practical deployment potential. Future work could optimize hierarchical coordination, explore sophisticated structures, or incorporate transfer learning to enhance efficiency and scalability.

Ultimately, integrated experimental analysis highlights hierarchical adversarial resilience's value in MARL frameworks, guiding future research and practical CPS security implementations.

Method	Seed	Return ↑	F1 ↑	Precision ↑	Recall ↑	FAR ↓ (%)	MTTD ↓	Accuracy ↑ (%)
Rule-Based IDS	42	278.4	0.436	0.482	0.398	50.06	99.18	47.0
	100	299.2	0.522	0.546	0.500	49.98	99.20	51.5
	2025	263.9	0.526	0.515	0.537	50.56	99.00	54.0
PPO Non-Hier. MARL	42	**1423.54**	**0.802**	**0.935**	**0.702**	**6.48**	**496.35**	**82.72**
	100	**1464.66**	0.805	0.932	**0.708**	6.84	501.07	82.89
	2025	1380.74	0.799	0.934	0.698	6.65	502.27	82.41
HAMARL (ours)	42	1354.82	0.797	0.932	0.696	6.78	500.21	82.29
	100	1462.94	**0.805**	**0.934**	0.707	**6.63**	**499.06**	**82.93**
	2025	**1397.28**	**0.799**	**0.934**	**0.698**	**6.56**	**500.05**	**82.38**

Table 1: Detailed comparative evaluation across all metrics, seeds, and methods. Arrows indicate if higher (↑) or lower (↓) values are explicitly preferred. Best-performing metrics for each seed are highlighted in bold.

# Agents	4	8	12	24
Non-Hier. MARL (h)↓	**0.025**	**0.024**	**0.024**	**0.028**
HAMARL (ours) (h)↓	0.036	0.069	0.100	0.204

Table 2: Scalability comparison (wall-clock training time in hours) between Non-Hierarchical MARL and HAMARL across varying numbers of defender agents. Training explicitly conducted over 500 episodes on an Apple MacBook Pro (M4 Max, 36 GB RAM). Best (lowest) times highlighted in bold.

Conclusion and Future Work

This work introduced HAMARL, a Hierarchical, Adversarially-Resilient Multi-Agent Reinforcement Learning framework designed to secure Cyber-Physical Systems (CPS) against sophisticated, adaptive cyber threats. By integrating decentralized local anomaly detection with centralized global coordination, HAMARL effectively balances rapid response capabilities and comprehensive strategic oversight. This hierarchical approach demonstrates significant advantages over conventional flat multi-agent reinforcement learning (MARL) and traditional static rule-based intrusion detection systems, achieving notably higher detection accuracy, shorter mean time-to-detect, and reduced false alarms, while maintaining operational continuity under previously unseen attack vectors.

The incorporation of an adversarial training loop was critical for enhancing the adaptability of HAMARL. By continuously training against a dynamic and adaptive red-team attacker agent, the defender agents developed robust generalization capabilities, ensuring effectiveness even when confronted with novel and evolving threats. From an industrial perspective, such adaptability is crucial, as HAMARL eliminates the need for manual retuning common in static defenses, offering a proactive and continuously improving cybersecurity solution suitable for modern CPS environments characterized by rapidly changing threat landscapes.

Despite these advances, several challenges remain. Foremost among these is the substantial computational cost associated with training hierarchical MARL systems, presenting practical constraints for deployment in resource-constrained operational technology networks typical of industrial settings. Future research efforts should focus on reducing these computational demands through techniques such as lightweight policy distillation, transfer-learning-based initialization, and federated or distributed training approaches. Additionally, optimizing reward shaping and hierarchical credit assignment currently requires careful domain-specific tuning, presenting another critical area for improvement to facilitate broader applicability. Real-world adoption also necessitates demonstrable compliance with industrial standards (e.g., IEC 62443), rigorous fail-safe validations, and comprehensive field trials under realistic production conditions.

Several promising research directions emerge for future exploration. Transfer and meta-learning methods could significantly reduce the data and computational overhead associated with deploying HAMARL across diverse CPS domains, such as adapting policies from smart manufacturing environments to smart grids or medical IoT systems. Enhancing HAMARL with explainability features or integrating formal verification techniques could further increase its trustworthiness and auditability, crucial for regulatory compliance and operator confidence. Finally, extending adversarial training scenarios to include multiple or colluding attackers could expose critical vulnerabilities and facilitate the development of even more robust defensive coordination strategies among defender agents.

As CPS deployments continue to scale and become increasingly autonomous, the importance of actively adaptive cybersecurity frameworks becomes more pronounced. HAMARL represents a significant advancement toward achieving resilient, scalable, and proactive security solutions. Continued research along these outlined pathways is essential for transitioning HAMARL from an academic prototype to dependable, industry-grade protection mechanisms, ultimately safeguarding the next generation of critical infrastructure against emerging and adaptive cyber threats.

References

Baheti, R.; and Gill, H. 2011. Cyber-physical systems. *The impact of control technology*, 12(1): 161–166.

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*.

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.