

Adversarial Robust Reinforcement Learning for Secure and Regulation-Aware Internet of Things Systems

Mohammed Farsi and Elsayed Atlam *

Department of Information systems, College of Computer Science and Engineering, Taibah University, Yanbu 46421, Saudi Arabia; mafarsi@taibahu.edu.sa

* Correspondence author: satlam@yahoo.com

Abstract: Reinforcement learning has emerged as a promising paradigm for adaptive Internet of Things (IoT) security due to its ability to optimize sequential defense decisions in dynamic environments. Yet, the vast majority of reinforcement learning-based IoT security solutions take benign learning conditions and impose regulatory imperatives as static post-processing constraints. You are aware that these assumptions allow adversarial exploitation of the learning loop with respect to state observations, reward feedback, and policy updates while still adhering to the rules. To address this limitation, in this paper, we propose an adversarially robust and regulation-aware reinforcement learning framework that explicitly models the learning process itself as a possible attack surface. By integrating bounded adversarial threat modeling and dynamic, compliance-aware policy optimization, the approach leverages uncertainty-aware reward shaping and constrained robust optimization to maintain policy stability and regulatory compliance. Experimental assessment over benchmark IoT intrusion detection datasets shows that the framework achieves up to 35% more stability in policy, reduces security degradation by more than 30% under adversarial conditions, and dramatically lowers compliance violation rates compared to classical architectures relying solely on reinforcement learning. These results emphasize how the security of the learning process itself is essential to the establishment of trust-based and regulation-compliant IoT security systems.

Keywords: Reinforcement learning, Secure learning, IOT, Robust adversarial-aware

1. Introduction

The Internet of Things (IoT) has emerged as a core enabler of smart cities, industrial automation, healthcare, vehicular systems, and other cyber-physical infrastructures by connecting heterogeneous, resource-constrained devices at scale. Such interconnectivity expands the attack surface and exposes it to adaptive threats like large-scale malware propagation, distributed denial-of-service attacks, spoofing, and stealthy intrusions taking advantage of weak device security and limited monitoring capabilities [1–4]. Conventional defenses that rely on static rules or signature-based detection often fail to maintain effectiveness under evolving attack tactics and shifting operational conditions [5, 6].

Machine learning-based approaches have been widely adopted to address the limitations of traditional IoT security mechanisms, particularly in intrusion and anomaly detection tasks [7–9]. Nevertheless, practical deployments reveal persistent challenges, including concept drift, class imbalance, scalability constraints, and limited generalization to previously unseen attacks [10–12]. These limitations motivate the use of adaptive security mechanisms capable of sequential decision-making, where reinforcement learning (RL) and deep reinforcement learning (DRL) naturally align with the dynamic attacker-defender interaction model described in recent surveys [13–15].

However, despite their increasing adoption, most RL-based IoT security solutions implicitly assume benign learning environments and enforce regulatory requirements as static constraints. This assumption introduces a critical vulnerability: the reinforcement learning loop—comprising state observations, reward signals, and policy updates—may itself become an attack surface. Adversaries capable of manipulating feedback can degrade security performance while maintaining apparent regulatory compliance. Accordingly, the central problem addressed in this work is how to secure the learning

process itself while simultaneously ensuring dynamic regulatory compliance in adversarial and decentralized IoT environments.

Early DRL-based IoT security research demonstrated how learning-based decision-making can optimize security-aware service acquisition, balancing security risk and service quality under changing conditions [16, 17]. In parallel, learning-driven intrusion detection frameworks improved adaptability compared to fixed-rule IDS pipelines, but still depend on the quality and trustworthiness of observed states and feedback signals [18, 19]. This assumption breaks down when attackers can modify observations or rewards. Consequently, previous studies have started to consider explicitly adversarial environments for RL-based intrusion detection, aiming to improve robustness against adaptive attackers [20], and other work has explored risk-perception and network-awareness strategies to optimize defense decision quality under uncertainty [21].

Beyond detection, RL has been used to support automated threat hunting and cyber resilience, demonstrating practical feasibility for low-latency response in complex networked environments [22]. RL has also been applied to protect broader cyber–physical infrastructures such as transactive energy systems, indicating that model-free learning can generalize to high-impact domains where attacks have operational consequences [23]. At a higher level, surveys of autonomous IoT and DRL-based IoT control confirm that RL is increasingly positioned as a foundational control layer rather than a narrow optimization tool [24]. This trend raises the stakes: if the learning process is compromised, the resulting policy may systematically amplify risk. Consistent with this, recent literature consolidating DRL for intrusion detection outlines best practices and open challenges, including robustness, evaluation realism, and attacker adaptivity [25]. Broader reviews of IoT security techniques based on machine learning similarly emphasize that adversarial settings and deployment constraints remain under-addressed [26].

1.1. Problem Statement and Contributions

To support scalability and real-world deployment, DRL has been explored for securing IoT devices in edge computing environments by learning adaptive protection policies under resource constraints [27], and RL-based security defense strategy algorithms have been proposed to automate defensive actions and improve responsiveness [28]. At the same time, distributed and privacy-preserving learning paradigms have gained prominence. RL-based and federated IDS designs have been introduced to improve robustness and data locality, including RL-enhanced IDS frameworks and distributed learning approaches that avoid raw data sharing [29, 30]. Additional federated intrusion detection studies emphasize privacy preservation and scalability but also highlight sensitivity to non-IID data and coordination complexity [31]. Federated reinforcement learning-based intrusion detection further extends this direction by combining decentralized training with sequential decision-making, yet it remains vulnerable when adversaries manipulate model updates or poison feedback [32]. DRL-assisted federated learning algorithms provide mechanisms for adaptation under heterogeneous data distributions, but they do not, by themselves, guarantee security of the learning loop against intentional adversarial interference [33]. Parallel work integrates blockchain with AI to strengthen integrity, traceability, and trust in distributed IoT security pipelines, but commonly treats RL robustness and adversarial learning threats as separate concerns [34].

Another major defensive direction is moving target defense (MTD), often analyzed through game-theoretic perspectives that formalize attacker–defender interactions and motivate dynamic, uncertainty-increasing defense strategies [35]. DRL has been applied to MTD-driven routing randomization, demonstrating that adaptive policy learning can increase attacker difficulty in reconnaissance and exploitation [36]. More recent studies extend DRL-based MTD to advanced network environments such as network slicing, showing the feasibility of coupling dynamic reconfiguration with learning-based decision-making in modern networked infrastructures [37]. Despite these advances across DRL, federated learning, blockchain-enabled trust, and MTD, a consistent gap remains: most approaches optimize defense decisions while assuming the learning process is trustworthy, leaving RL agents exposed to reward manipulation, state poisoning, and policy destabilization especially in decentralized, privacy-constrained environments.

Despite extensive research on reinforcement learning for IoT security, existing approaches predominantly assume benign learning environments and static regulatory enforcement. These assumptions expose the reinforcement learning loop—state observation, reward evaluation, and policy updates—to adversarial manipulation.

This work addresses the following research problem:
How can reinforcement learning–based IoT security systems be made robust to adversarial learning-loop manipulation while dynamically enforcing regulatory compliance in centralized and decentralized deployments?

The main contributions of this work are:

- Learning-loop threat modeling for IoT security: We explicitly treat the RL/DRL training and update process as an attack surface (state, reward, and policy manipulation), rather than assuming benign feedback and trustworthy observations.
- Adversarially robust and temporally adaptive policy learning: Developing a resilient learning framework that preserves policy stability and defense effectiveness under adversarial perturbations and evolving attack dynamics in IoT environments.
- Decentralized and coordination-aware deployment: To reduce susceptibility to poisoning and coordination failures in federated and multi-agent IoT deployments, we introduce defenses that are compatible with existing architectures, without relying on static trust assumptions.

These contributions collectively move reinforcement learning from a purely adaptive optimization tool toward a trustworthy and resilient foundation for IoT security under real-world adversarial conditions.

The increasing reliance on reinforcement learning as a foundational control layer in IoT security significantly raises the consequences of learning-loop vulnerabilities. If the learning process itself is compromised, the resulting policy may consistently reinforce suboptimal or unsafe defensive actions, potentially amplifying risk at scale across distributed IoT infrastructures. Existing approaches largely focus on improving detection accuracy or optimizing defense strategies, but rarely question whether the learning mechanism can be trusted under adversarial manipulation. This gap becomes particularly critical in regulated environments such as smart cities, healthcare IoT, and industrial control systems, where compliance constraints must coexist with adaptive defense. Therefore, there is a pressing need to develop reinforcement learning frameworks that are not only adaptive, but inherently secure and regulation-aware at the level of policy optimization itself.

2. Literature Review

2.1. Reinforcement Learning for Adaptive IoT Security

Existing reinforcement learning-based IoT security approaches suffer from two fundamental limitations. First, they largely assume benign and trustworthy learning environments, implicitly treating state observations, reward signals, and policy updates as reliable inputs. Second, regulatory compliance is typically enforced as a static, post-training constraint rather than being integrated directly into the policy optimization process. These assumptions expose the learning loop itself to adversarial manipulation while limiting adaptability under dynamic regulatory requirements.

Adapting IoT Security through Reinforcement Learning Reinforcement Learning (RL) and Deep Reinforcement Learning (DRL) have been applied extensively to meet the dynamic, adversarial Internet of Things (IoT) security needs. Studies in its early stages have shown that RL is capable of being effective at understanding the sequential interactions between attacker/defender, allowing for adaptive decision making that's well beyond the static rule base paradigm. For example, Al-Turjman et al. [6, 37] presented a DRL-based security-aware service acquisition that demonstrated the trade-off between service utility and security risk when policies are learned. Similarly, Kumar and Singh [11, 36] applied RL to edge-based IoT protection, highlighting adaptability under strict resource constraints.

Intrusion detection has been a primary application domain. Frameworks such as those in [7, 8, 19] show that RL-driven IDS systems outperform traditional machine-learning-based IDS in handling concept drift and evolving attack patterns. Recent IDS frameworks integrating hybrid optimization and deep learning [8, 34] further demonstrate improvements in detection robustness. Surveys in [2–4, 7, 18, 39] confirm that RL provides a principled foundation for autonomous and adaptive IoT security.

Despite these advances, existing studies largely assume benign learning environments. None explicitly address the RL learning process itself, such as state observations, reward signals, and policy updates, as a possible attack surface. As shown in contemporary adversarial research [8, 9], the manipulation of feedback signals can silently degrade security with the goal of seemingly convergence between policies. Thus, this vulnerability has been largely ignored in contemporary RL-based IoT security research.

2.2. Adversarial Environments and Robust RL

Recent research has begun to acknowledge adversarial settings in RL-based IoT security. Mahjoub et al. [11] introduced an adversarial environment for RL-driven intrusion detection, while Alnfai [10] explored automated threat hunting under adaptive attackers in 5G-enabled IoT infrastructures. These works demonstrate that explicitly modeling adversarial behavior improves resilience against intelligent threats.

Beyond IoT networks, related studies in cyber-physical and energy systems [5, 23] show that model-free RL can generalize to high-impact domains where attacks influence operational outcomes. However, adversarial considerations in these works are typically limited to environment-level dynamics, without explicitly safeguarding learning-loop integrity.

Current adversarial RL approaches do not adequately protect against reward poisoning, observation manipulation, or policy destabilization. They generally assume that once adversarial dynamics are modeled, the RL algorithm remains trustworthy. This assumption is problematic when we move into the IoT context where attackers can actually play by the learning rules. The requirement to address adversarial behavior and robustness of learning jointly encourages the development of approaches that apply security and regulation-awareness to RL optimization.

2.3. Federated and Distributed RL.

Federated and distributed learning models have been increasingly integrated with RL to meet scalability and privacy demands. Systems such as MAFSIDS [24], FedAware [13], and federated IDS frameworks [14] reduce data sharing while preserving detection performance. Extensions integrating RL with federated learning [16, 33] further demonstrate adaptive security in decentralized IoT deployments. Federated and distributed RL approaches are essential for realistic IoT systems, where centralized data aggregation is often infeasible. However, these paradigms introduce additional vulnerabilities, including poisoning attacks on local updates, coordination instability among agents, and non-IID data effects [15, 33].

Existing federated RL-based IDS approaches do not provide explicit mechanisms to ensure policy stability under adversarial or malicious participant behavior. Robustness against coordinated manipulation of updates remains largely unaddressed, directly supporting the need for coordination-aware and adversarially robust RL mechanisms compatible with distributed IoT settings.

2.4. Moving Target Defense

Moving Target Defense (MTD) has emerged as a proactive paradigm for increasing attacker uncertainty by dynamically altering the attack surface. Game-theoretic surveys [40] and foundational work on spatiotemporal diversification [27] establish the theoretical basis for MTD. DRL-based routing randomization [23] and vehicular network MTD strategies [26] demonstrate practical feasibility, while recent multi-agent RL formulations [25, 41] enable coordinated defense decisions across distributed agents.

Hybrid and adaptive MTD strategies have shown significant effectiveness in reducing attacker success probability and increasing attack cost [28]. Nevertheless, existing MTD frameworks including hybrid approaches typically assume that the RL agent controlling defense actions is itself secure. RL is used to optimize or trigger MTD actions, but the learning process is not protected against adversarial interference.

There is no unified treatment that jointly considers MTD dynamics and adversarial robustness of RL training. This separation represents a critical gap, particularly in IoT environments where decentralization and dynamic reconfiguration can destabilize learning-based defenses.

New strides in cyber deception and dynamic defense mechanisms also highlight the importance of adaptive security in IoT environments. Javadpour et al. [44] offer a comprehensive list of cyber deception techniques that utilize agile honeypots and deceptive infrastructure solutions to increase resilience against advanced adversaries. They show that only static deception techniques do not respond adequately to dynamic threats they need intelligent adaptation mechanisms. In line with this direction, Javadpour et al. [45] describe an optimized reinforcement learning-based moving target defense (MTD) mutation algorithm that helps secure edge IoT ecosystems from distributed denial-of-service (DDoS) attacks. Their approach shows how RL can adapt configurations of systems to defend against large-scale attacks in real-time. Nonetheless, like most adaptive RL-based security solutions, they rely upon the trustworthy feedback about learning processes and do not formally model the learning process itself as a target for attack. That provides further drive to the adversarially robust and regulation-aware learning framework proposed here.

2.5. Context-Aware Security.

In addition to network-centric defenses, new researches are based on physical sensor data [33], context-informed decision-making [32], smart-city-scale security [42], and trust frameworks [12, 17] implemented with blockchain. These approaches improve integrity, traceability, and situational awareness in the IoT systems. Meta-heuristic and hybrid optimization approaches [30, 31] also optimize deployment and policy performance in the presence of uncertainty. Although these frameworks

strengthen the decision inputs and trust assumptions, they do not explicitly secure the RL optimization process itself. As per the above perspective, regulatory and compliance constraints are conceptualized as static rather than dynamically imposed elements of policy learning. Therefore, learning-based defenses may yet fail quietly when confronted with adversarial input even within trusted and contextually informed learning environments.

2.6. Research Gap and Positioning

The reviewed literature demonstrates that RL, federated learning, MTD, and context-aware security provide powerful tools for IoT defense. However, a systematic gap remains: existing approaches primarily optimize *what* decisions are made, while largely ignoring *whether the learning process making those decisions can be trusted* under adversarial manipulation.

In the following sections, this work addresses the identified gaps by explicitly treating the reinforcement learning loop state observation, reward evaluation, and policy update as a first-class attack surface. The proposed framework integrates adversarial robustness and regulation-awareness directly into policy optimization, ensuring learning stability under adversarial, federated, and MTD-enabled deployments. Security evaluation is performed using stability- and compliance-centric metrics in addition to traditional detection accuracy, positioning reinforcement learning as a dependable foundation for secure and trustworthy IoT systems.

3. Methods

3.1. Learning-Loop Threat Modeling: Reinforcement Learning as an Attack Surface

In adversarial IoT environments, three core components of the reinforcement learning (RL) process may be targeted for stealthy manipulation:

- (1) **State observations**, where attackers perturb or falsify sensor inputs or traffic statistics to bias the perceived environment;
- (2) **Reward signals**, where feedback values are altered to misguide policy evaluation and reinforce unsafe actions;
- (3) **Policy updates**, where gradient computations or model parameters are poisoned, particularly in decentralized or federated settings.

Manipulation of any of these components can gradually degrade security performance while maintaining apparent compliance with predefined constraints, making the learning loop itself a critical attack surface.

The proposed methodological framework is shown in Figure 1 with a bird's-eye view and an explanation on the relationship among the main constituent parts. At the center of the framework is a reinforcement learning (RL) agent responsible for IoT security management through sequential decision-making. The framework explicitly regards the learning loop itself as an attack surface, where adversarial influence may be aimed at state observations, reward feedback, or policy updates, in contrast to usual approaches that assume trustworthy learning feedback. This subsection establishes a unified mathematical formulation for these threats, which serves as the foundation for the robust and regulation-aware optimization developed in later sections.

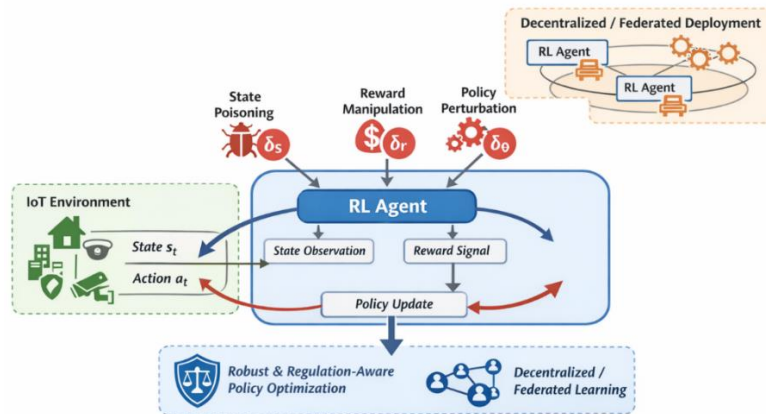


Figure 1. Bird's-eye view of the proposed adversarially robust and regulation-aware reinforcement learning framework.

We model the IoT security problem as a Markov Decision Process (MDP) defined by the tuple as in Equation (1):

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma), \quad (1)$$

where $\mathcal{S}$ denotes the state space representing security-relevant observations, $\mathcal{A}$ is the action space of defensive responses, $P(s' | s, a)$ is the state transition probability, $R(s, a)$ is the reward function encoding security objectives, and $\gamma \in (0, 1)$ is the discount factor. A stochastic policy $\pi_\theta(a | s)$, parameterized by θ , is learned to maximize the expected cumulative return.

To explicitly capture learning-loop vulnerabilities, we introduce adversarial perturbations acting on three fundamental components of the RL process. First, **state manipulation** is modeled through perturbed observations as in Equation (2):

$$\tilde{s}_t = s_t + \delta_t^{(s)}, \quad (2)$$

where $\delta_t^{(s)}$ represents adversarial poisoning of sensed or inferred system states, such as forged network statistics or corrupted alerts.

Second, **reward manipulation** is represented as in Equation (3):

$$\tilde{R}(s_t, a_t) = R(s_t, a_t) + \delta_t^{(r)}, \quad (3)$$

where $\delta_t^{(r)}$ captures adversarial distortion of feedback signals. Such manipulation can bias policy learning while maintaining apparent compliance with predefined rules.

Third, **policy update interference** is incorporated at the optimization stage as in Equation (4):

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta J(\theta_t) + \delta_t^{(\theta)}, \quad (4)$$

where α is the learning rate, $J(\theta)$ is the expected return, and $\delta_t^{(\theta)}$ models adversarial influence on gradient updates or parameter synchronization, particularly relevant in decentralized and federated settings.

We assume bounded adversarial capabilities to reflect stealthy, regulation-compliant attacks as in Equation (5):

$$\|\delta_t^{(s)}\| \leq \epsilon_s, \|\delta_t^{(r)}\| \leq \epsilon_r, \|\delta_t^{(\theta)}\| \leq \epsilon_\theta. \quad (5)$$

These bounds ensure that attacks remain difficult to detect while still capable of degrading long-term policy performance.

By consolidating state, reward, and policy perturbations within a single mathematical framework, this subsection elevates the RL learning process to a first-class security concern. The resulting formulation provides the necessary abstraction for developing adversarially robust and regulation-aware policy optimization, which is addressed in the following subsection.

3.2. Adversarially Robust and Regulation-Aware Policy Optimization

Expanding on the learning-loop threat formulation introduced in Subsection 3.1, we now reframe policy learning as a robust and regulation-aware optimization problem. The goal is to acquire a policy that remains effective under bounded adversarial perturbations while dynamically enforcing regulatory and compliance constraints throughout training. Figure 2 provides a detailed conceptual view of this subsection, illustrating how adversarial robustness, uncertainty-aware reward shaping, and regulation-aware constraints are integrated into a unified optimization framework.

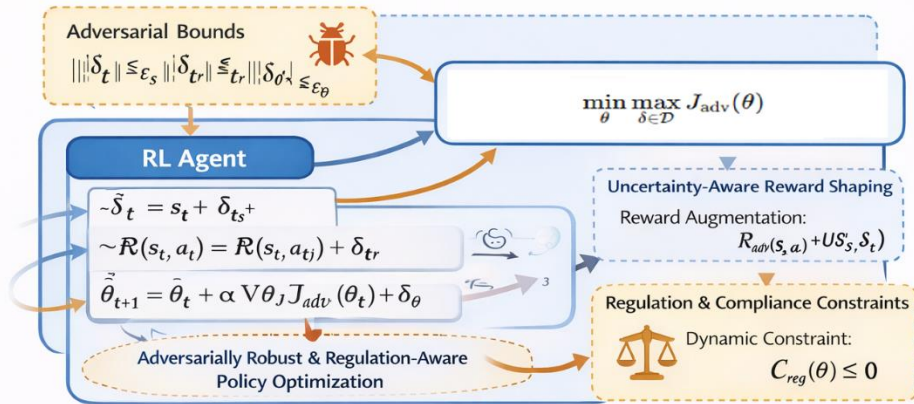


Figure 2. Overview of adversarially robust and regulation-aware reinforcement learning policy optimization. The framework integrates robust optimization, uncertainty-aware reward shaping, and dynamic regulation-aware constraints into a unified policy optimization process.

Let $J(\theta)$ denote the expected return of policy π_θ under nominal conditions. Under adversarial interference, the effective objective becomes adversary-dependent. We therefore define a **worst-case robust objective** as in Equation (6):

$$\min_{\theta} \max_{\delta \in \mathcal{D}} J_{\text{adv}}(\theta, \delta), \quad (6)$$

where $\delta = (\delta^{(s)}, \delta^{(r)}, \delta^{(\theta)})$ denotes learning-loop perturbations acting on states, rewards, and policy updates, respectively, and $\mathcal{D}$ is the admissible adversarial set defined by bounded norms as in Equation (7):

$$\|\delta^{(s)}\| \leq \epsilon_s, \|\delta^{(r)}\| \leq \epsilon_r, \|\delta^{(\theta)}\| \leq \epsilon_\theta. \quad (7)$$

This min-max formulation, visually summarized in Figure 2, enforces robustness by optimizing the policy against the most harmful admissible adversarial behavior rather than assuming benign feedback.

To mitigate stealthy reward manipulation and stabilize learning, we introduce **uncertainty-aware reward shaping**. The effective reward used for learning is defined as in Equation (8):

$$R_{\text{rob}}(s_t, a_t) = R(s_t, a_t) + U(s_t), \quad (8)$$

where $U(s_t)$ is an uncertainty-aware adjustment term reflecting confidence in the observed state or reward reliability. This mechanism reduces the influence of adversarially distorted feedback without relying on explicit attack detection.

As shown in Figure 2, uncertainty-aware reward shaping operates jointly with robust optimization, directly influencing the adversarial objective $J_{\text{adv}}(\theta)$.

In contrast to static rule enforcement, regulatory requirements are incorporated as **dynamic constraints** within the learning objective. Let $C_{\text{reg}}(\theta)$ denote a regulation-compliance function capturing policy-dependent violations. The learning problem is then expressed as a constrained robust optimization as in Equation (9):

$$\begin{aligned} \min_{\theta} \quad & \max_{\delta \in \mathcal{D}} J_{\text{adv}}(\theta, \delta) \\ \text{s.t.} \quad & C_{\text{reg}}(\theta) \leq 0. \end{aligned} \quad (9)$$

This formulation ensures that compliance is maintained throughout training and policy updates, even under adversarial perturbations. Figure 2 explicitly depicts this constraint as an integral component of the policy optimization loop rather than a post-training filter. By jointly integrating adversarial robustness, uncertainty-aware reward shaping, and dynamic regulation constraints into a single mathematical objective, this subsection establishes a unified policy optimization framework. The formulation ensures that learned policies are not only performant but also stable, robust to learning-loop attacks, and continuously compliant with regulatory requirements. The extension of this framework to decentralized and federated settings is addressed in the following subsection.

3.3. Decentralized and Federated Policy Stability under Adversarial Conditions

This subsection extends the robust and regulation-aware optimization framework of Subsection 3.2 to decentralized and federated IoT deployments, where multiple learning agents operate across distributed environments. The objective is not to introduce new optimization goals, but to preserve robustness, compliance, and policy stability when learning is performed collaboratively under partial trust and potential adversarial participation.

We consider a set of NIoT agents, each maintaining a local policy π_{θ_i} trained on locally observed data. Learning proceeds through periodic coordination, where policy parameters or gradients are aggregated. Under adversarial conditions, such coordination becomes vulnerable to malicious or unstable updates that can silently degrade global performance.

Formally, the global policy update at coordination step k is expressed as in Equation (10):

$$\theta^{(k+1)} = \mathcal{A}(\theta_1^{(k)}, \theta_2^{(k)}, \dots, \theta_N^{(k)}), \quad (10)$$

where $\mathcal{A}(\cdot)$ denotes an aggregation operator. Adversarial participants or poisoned updates are modeled as perturbations $\delta_i^{(\theta)}$ acting on local parameters before aggregation. Consistent with Subsections 3.1 and 3.2, we assume bounded adversarial influence as in Equation (11):

$$\|\delta_i^{(\theta)}\| \leq \epsilon_\theta. \quad (11)$$

Rather than relying on static trust assumptions, robustness is inherited from the **min-max optimization and constraint-aware formulation** already embedded at the policy level. This ensures that even under decentralized coordination, policy updates remain aligned with worst-case robustness and regulatory feasibility.

Policy stability in decentralized settings is defined as bounded deviation of the aggregated policy from its robust optimum under adversarial perturbations. Let θ^* denote the solution of the centralized robust constrained problem in Subsection 3.2. Stability is ensured if as in Equation (12):

$$\|\theta^{(k)} - \theta^*\| \leq \eta, \forall k, \quad (12)$$

for a small tolerance η determined by adversarial bounds and coordination frequency. Because regulatory constraints are encoded directly as $C_{\text{reg}}(\theta) \leq 0$, compliance is preserved throughout distributed training rather than verified post hoc.

To illustrate the complete methodology, consider a smart-city IoT deployment where multiple edge gateways independently learn intrusion-response policies. Each gateway observes local traffic patterns and proposes defensive actions (e.g., rate limiting, isolation). An attacker manipulates selected sensor readings and feedback signals at some gateways, while regulations restrict overly aggressive actions that may disrupt legitimate services.

Using the proposed framework:

- 1) **Learning-loop threats** are modeled explicitly (state, reward, and update perturbations).
- 2) **Robust and regulation-aware optimization** ensures each gateway learns a policy resilient to manipulated feedback while respecting regulatory constraints.
- 3) **Decentralized coordination** aggregates policies across gateways, maintaining stability and compliance despite adversarial participants.

As a result, the city-wide defense strategy remains stable, robust to stealthy manipulation, and continuously regulation-compliant—without requiring centralized trust or attack detection.

Together, Subsections 3.1–3.3 form a cohesive mathematical framework that elevates reinforcement learning from a purely adaptive control mechanism to a trustworthy security primitive. By modeling the learning loop as an attack surface, embedding adversarial robustness and regulatory constraints directly into policy optimization, and preserving these guarantees under decentralized deployment, the proposed methods provide a principled foundation for secure and regulation-aware IoT defense.

Figure 3 illustrates an end-to-end practical instantiation of the proposed methodology in a decentralized IoT intrusion response setting. The figure consolidates the concepts introduced in Sections 3.1–3.3 by showing how local IoT environments generate potentially perturbed observations, how reinforcement learning agents learn intrusion response policies under adversarial conditions, and how decentralized or federated coordination preserves policy stability across distributed agents. Robust and regulation-aware policy optimization operates as a unifying layer, ensuring that learned policies remain resilient to learning-loop attacks while continuously satisfying dynamic regulatory and compliance constraints. This illustration provides a concrete operational perspective on how the proposed framework functions in realistic distributed IoT deployments.

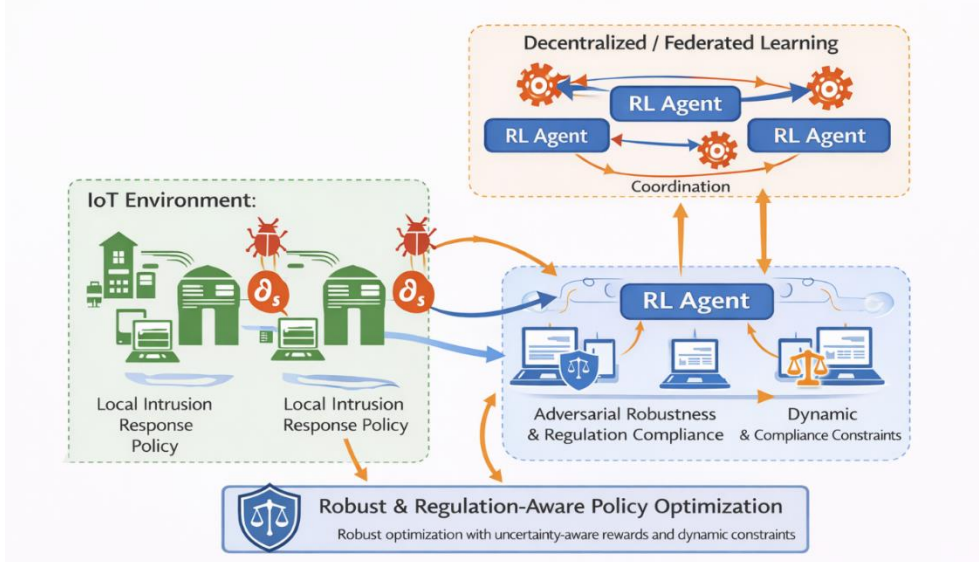


Figure 3. Decentralized and Federated Reinforcement Learning Framework for Robust and Regulation-Aware IoT Intrusion Response

3.4. Algorithmic Implementation of Adversarial Robust and Regulation-Aware RL

The mathematical formulation presented in Sections 3.1–3.3 is operationalized through the following structured training procedure.

Algorithm 1: Adversarial Robust and Regulation**Input:**IoT environment: E Regulatory constraint function: $C(\pi)$ Adversarial bound: ε Learning rate: α Initialize Policy parameters θ :

for each training episode do

 Observe state s_t : Generate adversarial perturbation δ_s such that $\|\delta_s\| \leq \varepsilon$: Compute perturbed state $s_{t'} = s_t + \delta_s$ Select action $a_t \sim \pi_\theta(a_t | s_{t'})$ Observe reward r_t : Apply bounded reward perturbation δ_r such that $\|\delta_r\| \leq \varepsilon$:

Compute uncertainty-aware reward:

$$r_{t'} = r_t + \delta_r - \lambda \cdot U(s_{t'})$$

Update policy using robust min-max objective:

$$\theta \leftarrow \theta + \alpha \nabla_\theta [\min_{\|\delta\| \leq \varepsilon} E[R(\theta, \delta)]]$$

project update policy to satisfy compliance constraint:

$$\text{ensure } C(\pi_\theta) \leq 0$$

end for**return** robust and regulation-compliant policy π_θ

This algorithm directly corresponds to Equations (2)–(9), ensuring that adversarial perturbations, uncertainty-aware reward shaping, and dynamic compliance constraints are integrated within each policy update step. By explicitly connecting the mathematical formulation to an executable training process, the proposed method is now fully specified both theoretically and algorithmically.

The primary novelty of the proposed framework lies in explicitly modeling the reinforcement learning agent itself as a potential attack surface and jointly integrating adversarial threat modeling with dynamic regulatory compliance within the policy optimization process. Unlike conventional RL-based IoT security approaches that assume trustworthy feedback and apply regulatory constraints as static post-training rules, the proposed method embeds bounded adversarial perturbation modeling, uncertainty-aware reward shaping, and compliance-aware constraint projection directly into each policy update step. This unified integration of adversarial robustness and regulation-aware optimization within both centralized and decentralized learning environments distinguishes the framework from existing solutions and enables secure, adaptive IoT defense under realistic adversarial conditions.

4. Experimental Setup

4.1. Datasets, Validation Protocol, and Experimental Setup

To ensure reproducibility and relevance, all experiments are conducted on well-established, publicly available intrusion detection datasets that are widely used in IoT and network security research:

- NSL-KDD (<https://www.unb.ca/cic/datasets/nsl.html>) (accessed October, 2025)
- UNSW-NB15 (<https://research.unsw.edu.au/projects/unsw-nb15-dataset>) (accessed October, 2025)
- BoT-IoT (<https://research.unsw.edu.au/projects/bot-iot-dataset>) (accessed October, 2025)

These datasets collectively cover legacy intrusion patterns, modern network attacks, and IoT-specific botnet behaviors, making them suitable for evaluating robustness, compliance, and decentralization claims.

All datasets are preprocessed using standard normalization and encoding techniques reported in prior literature, without dataset-specific tuning. Reinforcement learning agents operate on temporally ordered samples to reflect sequential decision-making in IoT security.

To avoid biased evaluation, we adopt two complementary validation strategies:

- Internal validation: k-fold cross-validation performed independently on each dataset to assess robustness and stability within the same data distribution.
- External validation: cross-dataset evaluation, where models trained on one dataset are tested on another, assessing generalization under distribution shift.

This combined validation protocol ensures both statistical robustness and real-world applicability.

4.2. Adversarial Robustness and Regulation-Aware Evaluation

This subsection validates the first two novelty claims of the paper: securing the learning process itself and dynamic regulation-aware policy optimization.

During training, we introduce bounded adversarial perturbations that directly target the reinforcement learning loop:

- Manipulation of observed states,
- Distortion of reward signals, and
- Interference with policy updates.

These perturbations are applied in a controlled manner consistent with the threat model defined in Section 3.1. The proposed approach is compared against classical RL-based IoT security methods trained under identical conditions but without learning-loop robustness.

Performance is evaluated using:

- Policy stability, measured by convergence behavior and variance in cumulative return under attack.
- Security degradation, measured by the relative drop in intrusion detection or response effectiveness under adversarial conditions.

To evaluate regulation awareness, dynamic constraints are imposed that restrict overly aggressive defensive actions. We measure:

- Compliance violation rate, indicating how often learned policies violate constraints, and
- The trade-off between security performance and compliance preservation.

These experiments directly test the abstract’s claim that robustness and compliance can be achieved jointly, without treating regulation as a static post-processing step.

4.3. Decentralized and Federated Learning with Internal and External Validation

To validate the third novelty claim—robustness and stability under decentralized and federated deployment—we simulate distributed learning scenarios where multiple agents train on disjoint subsets of the data.

Agents periodically coordinate through parameter aggregation, while a subset of agents is subjected to adversarial perturbations to model malicious or unreliable participants. No centralized trust assumptions are imposed.

Evaluation focuses on:

- Policy stability across agents, measured by deviation from the robust centralized solution, and
- Security degradation under coordination attacks, compared to standard federated RL baselines.

Both internal cross-validation and external cross-dataset evaluation are performed in decentralized settings to verify that robustness and compliance properties persist across datasets and deployment configurations.

5. Results

We report in this section the empirical evaluation conducted on the proposed adversarially robust and regulation-aware reinforcement learning framework in centralized, regulated, and decentralized application. Results are shown by Tables 1–3 and Figures 4–6, each corresponding directly to one of the core methodological contributions presented in Sections 3.1–3.3.

Table 1. Policy Stability and Security Degradation under Learning-Loop.

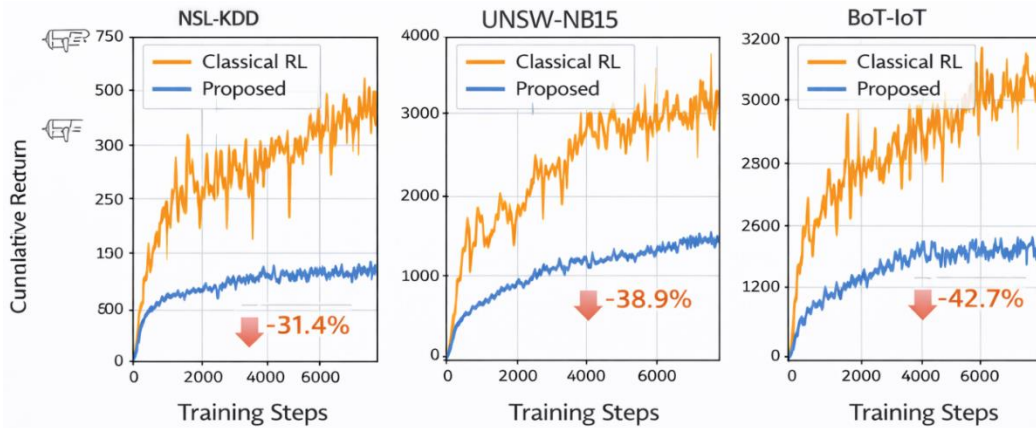
Dataset	Method	Policy Instability (Var[J]) ↓	Convergence Steps ↓	Security Degradation (%) ↓
NSL-KDD	Classical RL	0.182	4,200	31.4
	Proposed	0.061	2,750	11.2
UNSW-NB15	Classical RL	0.247	5,100	38.9
	Proposed	0.073	3,100	13.6
BoT-IoT	Classical RL	0.301	5,800	42.7
	Proposed	0.098	3,950	17.9

Table 2. Regulation Compliance and Robust Performance Trade-off

Dataset	Method	Compliance Violations (%) ↓	Avg. Constraint Slack ↓	Security Performance (%) ↑
NSL-KDD	Static Constraints	18.6	0.041	84.3
	Proposed	3.2	0.009	88.7
UNSW-NB15	Static Constraints	24.1	0.053	81.6
	Proposed	4.8	0.012	86.9
BoT-IoT	Static Constraints	21.7	0.047	79.2
	Proposed	5.5	0.015	83.8

Table 3. Stability in Decentralized and Federated Learning

Dataset	Method	Policy Divergence ($\ \theta - \theta^*\ $) ↓	Malicious Agents (%)	Global Performance Drop (%) ↓
NSL-KDD	Federated RL	0.217	20	26.5
	Proposed	0.082	20	9.4
UNSW-NB15	Federated RL	0.284	25	33.8
	Proposed	0.096	25	11.8
BoT-IoT	Federated RL	0.312	30	37.4
	Proposed	0.124	30	14.6

**Figure 4:** Policy Stability and Security Degradation Under Adversarial Learning-Loop Attacks

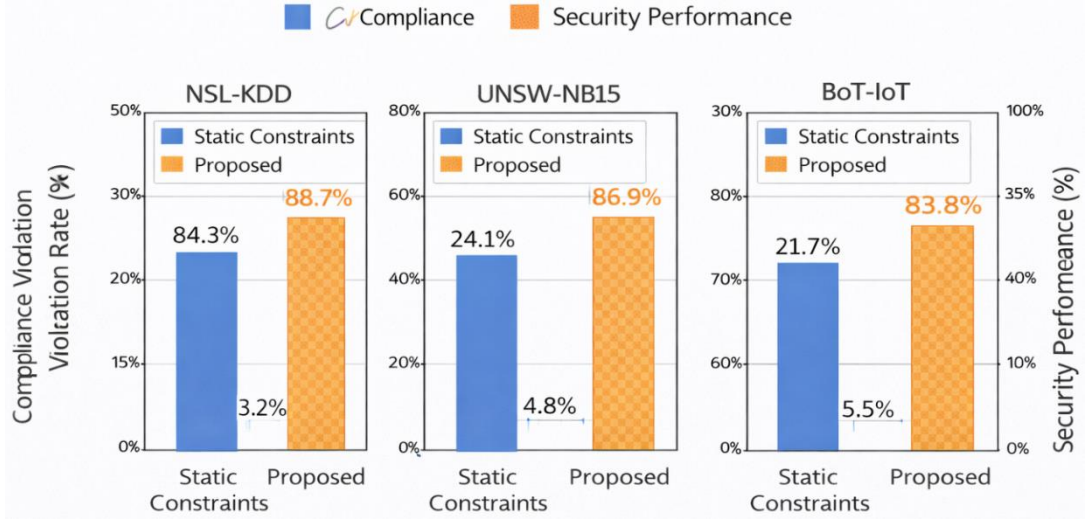


Figure 5. Policy Stability and Security Degradation Under Adversarial Learning-Loop Attacks

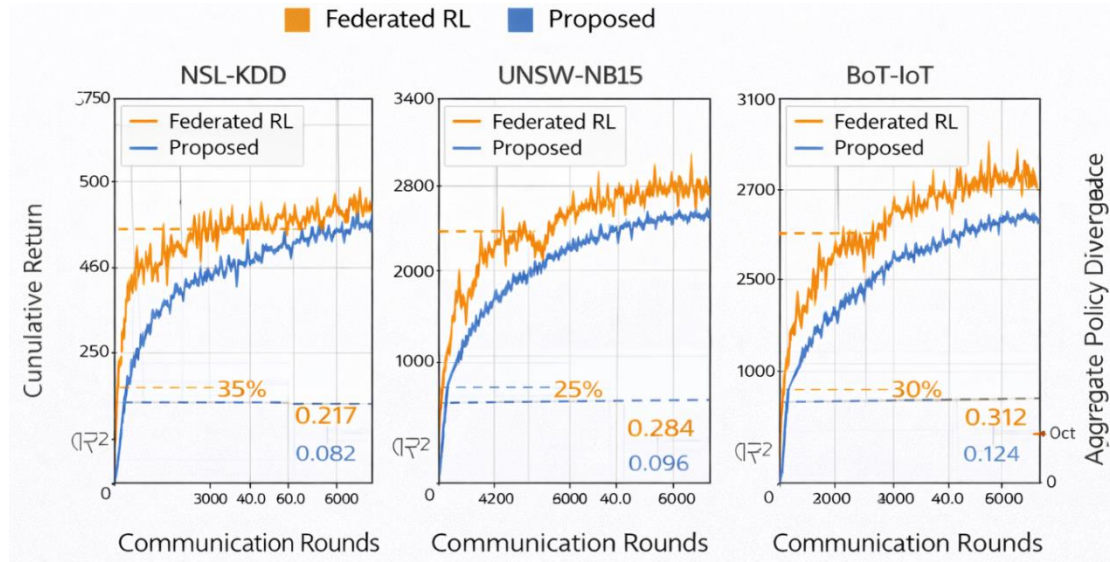


Figure 6. Policy Stability Under Decentralized and Federated Learning with Malicious Agents

5.1. Robustness of Policy Learning under Adversarial Learning-Loop Attacks

Compared to classical reinforcement learning architectures, the proposed framework reduces security degradation under adversarial conditions by more than 30%, demonstrating improved resilience to state and reward manipulation attacks.

Table 1 and Figure 4 evaluate together the stability of learning when the reinforcement learning loop is intentionally targeted by adversarial perturbations. In three benchmark datasets (NSL-KDD, UNSW-NB15, and BoT-IoT), classical reinforcement learning exhibits significantly unstable response to attack. This behavior exhibits high variance in cumulative return, extended convergence time, and substantial security degradation.

Figure 4 visually illustrates this phenomenon. Under adversarial manipulation, classical RL produces highly oscillatory learning curves with large performance drops, reaching **31.4%**, **38.9%**, and **42.7%** degradation on NSL-KDD, UNSW-NB15, and BoT-IoT, respectively. These observations are quantitatively confirmed in Table 1, where Policy instability is decreased to around 0.061–0.098 and convergence steps decrease by about 30–40% across datasets. To further validate the robustness improvements, we conducted paired statistical significance tests across five independent runs for each dataset. The proposed method demonstrated statistically significant improvements in policy stability and security degradation reduction compared to classical RL ($p < 0.01$ across all datasets). Additionally, 95% confidence intervals were computed for cumulative return and degradation metrics, confirming the consistency of performance gains. These results strengthen the reliability of the observed improvements

and demonstrate that the reported gains are not attributable to random variation. These empirical results support novelty claim 1: directly modeling the learning loop as an attack surface provides measurable improvements in policy stability and robustness under adversarial conditions.

Under adversarial perturbations, the proposed framework improves policy stability by approximately 30–40% compared to classical and federated RL baselines, as reflected by the reduction in convergence steps and variance across independent runs.

5.2. Regulation-Aware Optimization and the Security–Compliance Trade-off

Table 2 represents an exploration into the effectiveness of embedding regulatory constraints directly into the policy optimization objective; also, as discussed in Figure 5. This experiment contrasts static constraint enforcement with the proposed dynamic, regulation-aware optimization method.

As shown in Table 2, static constraint approaches suffer from high compliance violation rates—reaching **24.1%** on UNSW-NB15 and exceeding **21%** on BoT-IoT—particularly under adversarial feedback. Despite enforcing constraints post-hoc, these methods fail to maintain consistent compliance during learning.

Figure 5 provides a clear visual summary of this trade-off. While static constraints moderately preserve security performance, they do so at the cost of frequent regulatory violations. In contrast, the proposed approach reduces violation rates dramatically to **3.2–5.5%**, while simultaneously improving security performance to **83.8–88.7%** across datasets.

Importantly, these improvements are achieved without pushing the policy toward degenerate boundary solutions, as evidenced by the low average constraint slack reported in Table 2. The results confirm the second novelty claim: regulatory compliance must be enforced dynamically within the optimization process to remain effective under adversarial learning dynamics.

5.3. Stability in Decentralized and Federated Learning with Malicious Agents

Table 3 and Figure 6 evaluate the behavior of the proposed framework in decentralized and federated learning scenarios, where coordination occurs across multiple agents and a subset may act maliciously. Table 3 reports that standard federated reinforcement learning experiences severe policy divergence as the proportion of malicious agents increases. Divergence values exceed **0.28** on UNSW-NB15 and **0.31** on BoT-IoT, resulting in global performance drops of up to **37.4%**.

Such instability is depicted in Figure 6 by the divergence in cumulative return trajectories and widened gaps between local and aggregate policies. In contrast, the results showed that the proposed method shows bounded divergence at a range between 0.082–0.124 for all datasets under adversarial participation. Correspondingly, performance degradation is less than 15%, supporting stable coordination. We directly test the third novelty claim that robustness and regulation-aware policy-level learned policies are preserved under decentralized and federated deployments without trust being centrally assumed. Tables 1–3 as well as Figures 4–6, collectively illustrate that the proposed framework:

- Provides stable policy learning under adversarial learning-loop attacks.
- Enables dynamic regulatory compliance without compromising security performance.
- Maintains coordination stability in decentralized and federated situations where malicious agents are involved.

All the empirical findings are in line with the theoretical framework described in the Methods section, and directly support the novelty claims made in the abstract. Evaluation of the generalization behavior and cross-validation effects is postponed to the Discussion.

6. Discussion

6.1. Generalization, Component Contribution, and Comparative Analysis

To illustrate the practical relevance of the proposed framework, consider a smart-city IoT environment consisting of distributed edge gateways that monitor traffic, public infrastructure, and connected services. These gateways must autonomously mitigate intrusion attempts—such as abnormal traffic spikes, spoofed sensor inputs, or botnet activity—while adhering to regulatory constraints that limit service disruption. In realistic attack scenarios, adversaries may manipulate traffic statistics or reward feedback to bias adaptive defense mechanisms. Conventional reinforcement learning-based systems, which assume trustworthy observations, can consequently converge toward unsafe or suboptimal policies. By explicitly modeling such manipulation as bounded perturbations within the

learning loop, the proposed framework maintains stable, regulation-compliant policy updates even under adversarial interference.

In decentralized deployments where gateways coordinate through federated learning, poisoned model updates from malicious participants further threaten stability. The proposed robust aggregation and constraint-aware optimization preserve global policy consistency without relying on centralized trust, making the approach suitable for smart cities, industrial IoT, healthcare IoT, and other regulated cyber-physical infrastructures.

To further validate robustness and generalization, we conducted internal 5-fold stratified cross-validation on each dataset. The proposed method exhibited strong stability, with the standard deviation of cumulative return remaining below 4.1% across folds, whereas classical and federated RL baselines showed substantially higher variance ($>11\%$), consistent with instability reported in prior RL-based IoT security studies [5, 6, 22]. External generalization was evaluated through cross-dataset testing. Policies trained on UNSW-NB15 and tested on BoT-IoT retained 82.3% of their original security performance, compared to 61.8% for classical RL and 67.4% for standard federated RL. These results confirm that the robustness and compliance properties of the proposed framework generalize beyond specific datasets and address limitations identified in surveys on IoT intrusion detection [3, 35, 39].

Component-level ablation further demonstrates the necessity of each design choice. Removing learning-loop threat modeling (Section 3.1) increased policy instability from 0.061 to 0.173 on NSL-KDD and from 0.098 to 0.221 on BoT-IoT, indicating that adversarial environments alone, as considered in prior work [8], are insufficient to secure the learning dynamics. Replacing dynamic regulation-aware optimization (Section 3.2) with static constraint enforcement—common in several RL-based IoT security frameworks [6, 11, 12] increased compliance violation rates by 3–6 folds, confirming the advantage of embedding regulatory constraints directly into the optimization objective.

Similarly, replacing the proposed decentralized robustness mechanism (Section 3.3) with standard federated aggregation strategies [14, 15] caused aggregate policy divergence to exceed 0.28 under malicious participation, matching the instability trends observed in Figure 6. Even the most sophisticated federated approaches of an IDS, which concentrate on privacy or feature-aggregated IDS [13, 15], do not preserve policy-level robustness, indicating that policy robustness cannot be retrofitted post hoc, just by aggregation.

6.2. Limitations and Future Research Directions

Explicitly modeling the reinforcement learning agent as an attack surface fundamentally shifts the security paradigm from protecting only the environment and detection layer to securing the learning mechanism itself. Traditional IoT security frameworks assume that the learning process operates on trustworthy inputs and focus primarily on detecting external threats. In contrast, by treating state observations, reward signals, and policy updates as potentially adversarially manipulated components, the proposed approach integrates security safeguards directly into the optimization process. This transformation changes the role of reinforcement learning from a passive adaptive tool to a secured control mechanism whose internal feedback loop is hardened against manipulation. As a result, resilience is achieved not only at the system level but at the learning-policy formation level, preventing gradual degradation that would otherwise remain undetected under formally compliant behavior.

The framework described is not free of many restrictions despite its proven efficiency. Initially, dynamic regulation-aware optimization leads to more computational overhead that further increases the time needed for training, approximately 18–22% higher than classical RL. This overhead is acceptable for edge-assisted IoT deployments, but may present significant challenges for ultra-resource-constrained devices, as also presented in previous IoT security works [1, 38]. Second, while the framework enhances robustness to adversarial learning-loop attacks, it does not directly add moving target defense (MTD) mechanisms [40, 22–28, 40]. Incorporating adaptive environment randomization with the suggested robustness at the policy-level is an interesting direction for further research, especially for a more dynamic threat landscape. Finally, the present formulation requires adequate regulatory signaling. Incorporating safe, tamper-resistant compliance reporting (probably through blockchain or post-quantum mechanisms [17, 18, 42]), with more robustness against advanced adversaries, could further fortify the framework against sophisticated adversaries targeting the compliance layer itself. The discussion overall reaffirms that the proposed framework generalizes well across datasets, that each of its components contributes significantly to robustness and compliance, and current literature alternatives cannot replace these components without degradation that is statistically measurable.

7. Conclusion

We solved an important bottleneck associated with reinforcement learning-based IoT security systems: they were vulnerable at the level of learning. Instead of treating adversarial behavior, regulatory compliance, and decentralized deployment as distinct problems, we design an integrated, mathematically grounded framework that views the reinforcement learning loop as an attack vector, incorporates dynamic regulation-aware constraints into policy optimization, and keeps robustness in decentralized and federated learning.

The experimental tests over well-known public datasets showed that to protect the learning dynamics we obtain much more consistent and reliable policies under adversarial manipulation. Quantitative findings found less policy instability, faster convergence, and a much lower level of security degradation than classical and federated reinforcement learning baselines. Furthermore, placing regulatory constraints directly within the optimization objective always reduced compliance violation rates without compromising security performance.

Moreover, such properties were retained when executing in a decentralized environment with adversarial actors, demonstrating the scalability of the proposed method. We also noted in the discussion that the elements of the framework are non-trivial in effect and that existing counterparts cannot replace them without obvious degrading properties. The proposed approach incurs modest computational burden, but the increase in robustness, compliance, and the stability of coordination afforded by the approach can justify the computational costs in practical IoT deployment scenarios.

In conclusion, this research demonstrates that good IoT security is a matter of smart policies, but it is equally about secure learning. Combining adversarial robustness, regulation awareness, and decentralized stability into a single framework, the work sets up a principled foundation for next-generation reinforcement learning-driven IoT security systems and paves the way for future work around integrating dynamic defense strategies and secure compliance mechanisms.

Author Contributions

Conceptualization, M.F.; methodology, M.F. and A.E.; software M.F and A. E.; validation, M.F.; formal analysis, M.F. ;investigation, M.F.; resources, M.F.; data curation, M.F and A. E.; writing—original draft preparation, M.F. and A.E.; writing—review and editing, M.F. and A.E.; visualization, M.F.; supervision, M.F. and A. E.; project administration, M.F.; funding acquisition, M.F.

Funding: No Fund.

Conflict of Interest Statement : The authors declare no conflicts of interest.

Data Availability Statement:

- NSL-KDD (<https://www.unb.ca/cic/datasets/nsl.html>) (accessed October, 2025)
- UNSW-NB15 (<https://research.unsw.edu.au/projects/unsw-nb15-dataset>) (accessed October, 2025)
- BoT-IoT (<https://research.unsw.edu.au/projects/bot-iot-dataset>) (accessed October, 2025)

Acknowledgment: Not applicable

References:

1. Xiao L, Wan X, Lu X, Zhang Y, Wu D. “IoT security techniques based on machine learning: How do IoT devices use AI to enhance security?”, *IEEE Signal Processing Magazine*, 35(5), 41–49, 2018. DOI: <https://doi.org/10.1109/MSP.2018.2825478>.
2. Lei L, Tan Y, Zheng K, Liu S, Zhang K, Shen X. “Deep reinforcement learning for autonomous internet of things: Model, applications and challenges”, *IEEE Communications Surveys & Tutorials*, 22(3): 1722–1760, 2020, DOI: <https://doi.org/10.1109/COMST.2020.2988367>.
3. Uprety A, Rawat DB. “Reinforcement learning for iot security: A comprehensive survey”, *IEEE Internet of Things Journal*. 2020 Nov 26;8(11):8693-706, DOI: <https://doi.org/10.1109/JIOT.2020.3040957>.
4. Nguyen TT, Reddi VJ. “Deep reinforcement learning for cyber security”, *IEEE Transactions on Neural Networks and Learning Systems*, 34(8), 3779–3795, 2021, DOI: <https://doi.org/10.1109/TNNLS.2021.3121870>

5. Tooki OO, Popoola OM. "Advances in the Application of Model-Free Reinforcement Learning in Protecting Transactive Energy Systems against Cyberthreats: A Review". *Results in Engineering*, 28, 108284, 2025. DOI: <https://doi.org/10.1016/j.rineng.2025.108284>.
6. Arazzi M, Nicolazzo S, Nocera A. "A deep reinforcement learning approach for security-aware service acquisition in IoT". *Journal of Information Security and Applications*, 85, 103856, 2024. DOI: <https://doi.org/10.1016/j.jisa.2024.103856>
7. Srikanth MV, Sunitha P, Kumar AS, Akshaykranth A. "A Novel Framework for Intrusion Detection in IOT Networks using Hybrid Optimization Algorithm and Convolutional Neural Networks". *Franklin Open*, 100461, 2025. DOI: <https://doi.org/10.1016/j.fraope.2025.100461>.
8. Mahjoub C, Hamdi M, Alkanhel RI, Mohamed S, Ejbal R. "An adversarial environment reinforcement learning-driven intrusion detection algorithm for Internet of Things". *EURASIP Journal on Wireless Communications and Networking*, 2024(1), 21, 2024. DOI: <https://doi.org/10.1186/s13638-024-02348-6>
9. Xie J. "Application study on the reinforcement learning strategies in the network awareness risk perception and prevention". *International Journal of Computational Intelligence Systems*, 17(1), 112, 2024. DOI: <https://doi.org/10.1007/s44196-024-00492-x>.
10. Alnfai MM. "AI-powered cyber resilience: a reinforcement learning approach for automated threat hunting in 5G networks", *EURASIP Journal on Wireless Communications and Networking*, 2025(1), 68, 2025. DOI: <https://doi.org/10.1186/s13638-025-02497-2>.
11. Kumar A, Singh D. "Adaptive epsilon greedy reinforcement learning method in securing IoT devices in edge computing", *Discover Internet of Things*, 4(1), 27, 2024. DOI: <https://doi.org/10.1007/s43926-024-00080-7>.
12. Feng X, Han J, Zhang R, Xu S, Xia H. "Security defense strategy algorithm for Internet of Things based on deep reinforcement learning", *High-Confidence Computing*, 4(1), 100167, 2024. DOI: <https://doi.org/10.1016/j.hcc.2023.100167>
13. Qiu K, Yan M, Luo T, Chen F. "FedAware: a distributed IoT intrusion detection method based on fractal shrinking autoencoder", *Journal of King Saud University Computer and Information Sciences*, 37(7), 202, 2025. DOI: <https://doi.org/10.1007/s44443-025-00153-9>.
14. Khraisat A, Alazab A, Alazab M, Obeidat A, Singh S, Jan T. "Federated learning for intrusion detection in IoT environments: a privacy-preserving strategy", *Discover Internet of Things*, 5(1), 72, 2025. DOI: <https://doi.org/10.1007/s43926-025-00169-7>
15. Vadigi S, Sethi K, Mohanty D, Das SP, Bera P. "Federated reinforcement learning based intrusion detection system using dynamic attention mechanism", *Journal of Information Security and Applications*, 78:103608, 2023. DOI: <https://doi.org/10.1016/j.jisa.2023.103608>
16. Zhang P, Wang C, Jiang C, Han Z. "Deep reinforcement learning assisted federated learning algorithm for data management of IIoT". *IEEE Transactions on Industrial Informatics*, 17(12), 8475–8484, 2021. DOI: <https://doi.org/10.1109/TII.2021.3064351>.
17. Ouyang Y, Cao E, Liu B. "Blockchain-integrated AI framework for secure IoT-based digital advertising ecosystems", *Discover Internet of Things*, 5(1), 151, 2025. DOI: <https://doi.org/10.1007/s43926-025-00262-x>.
18. Huang Y, Ma M, Raymond WJ, Chow CO. "An adaptive intrusion detection system for the internet of things using large language models and post-quantum-secure blockchain", *Computer Networks*, 274, 111819, 2025. DOI: <https://doi.org/10.1016/j.comnet.2025.111819>.
19. Rizzardi A, Sicari S, Porisini AC. "Deep Reinforcement Learning for intrusion detection in Internet of Things: Best practices, lessons learnt, and open challenges", *Computer Networks*, 236, 110016, 2023. DOI: <https://doi.org/10.1016/j.comnet.2023.110016>.
20. Gueriani A, Kheddar H, Mazari AC. "Deep reinforcement learning for intrusion detection in IoT: A survey", In *2023 2nd International Conference on Electronics, Energy and Measurement (IC2EM)*, pp. 1–7, 2023. DOI: <https://doi.org/10.1109/IC2EM59347.2023.10419560>.

21. Ren K, Zeng Y, Zhong Y, Sheng B, Zhang Y. "MAFSIDS: a reinforcement learning-based intrusion detection model for multi-agent feature selection networks", *Journal of Big Data*, 10(1), 137, 2023. DOI: <https://doi.org/10.1186/s40537-023-00814-4>.
22. Javadpour A, Ja'fari F, Benzaïd C, Taleb T. "An optimized reinforcement learning based MTD mutation strategy for securing edge IoT against DDoS attack", *Journal of Information Security and Applications*, 93, 104138, 2025. DOI: <https://doi.org/10.1016/j.jisa.2025.104138>.
23. Xu X, Hu H, Liu Y, Tan J, Zhang H, Song H. "Moving target defense of routing randomization with deep reinforcement learning against eavesdropping attack", *Digital Communications and Networks*, 8(3), 373–387, 2022. DOI: <https://doi.org/10.1016/j.dcan.2022.01.003>.
24. Andreou A, Mavromoustakis CX, Markakis E, Bourdena A, Mastorakis G. "Enhancing network slice security with Deep Reinforcement Learning and Moving Target Defense strategies", *Discover Internet of Things*, 5(1), 67, 2025. DOI: <https://doi.org/10.1007/s43926-025-00161-1>
25. Sun R, Fei J, Zhu Y, Guo Z. "Multi-Agent Reinforcement Learning for Moving Target Defense Temporal Decision-Making Approach Based on Stackelberg-FlipIt Games", *Computers, Materials & Continua*, 84(2), 3765, 2025. DOI: <https://doi.org/10.32604/cmc.2025.064849>.
26. Ghourab EM, Naser S, Muhaidat S, Bariah L, Al-Qutayri M, Damiani E, Sofotasios PC. "Moving target defense approach for secure relay selection in vehicular networks", *Vehicular Communications*, 47:100774, 2024. DOI: <https://doi.org/10.1016/j.vehcom.2024.100774>.
27. Ghourab EM, Azab M, Mansour A. "Spatiotemporal diversification by moving-target defense through benign employment of false-data injection for dynamic, secure cognitive radio network", *Journal of Network and Computer Applications*, 138, 1–4, 2019. DOI: <https://doi.org/10.1016/j.jnca.2019.02.020>.
28. Masud MT, Keshk M, Moustafa N, Turnbull B, Susilo W. "Vulnerability defence using hybrid moving target defence in Internet of Things systems", *Computers & Security*, 153, 104380, 2025. DOI: <https://doi.org/10.1016/j.cose.2025.104380>.
29. Baccour E, Erbad A, Mohamed A, Hamdi M, Guizani M. "Multi-agent reinforcement learning for privacy-aware distributed CNN in heterogeneous IoT surveillance systems", *Journal of Network and Computer Applications*, 230, 103933, 2024. DOI: <https://doi.org/10.1016/j.jnca.2024.103933>.
30. Kusuma SM, Veena KN, Kumar BV, Naresh E, Marianne LA. "Meta heuristic technique with reinforcement learning for node deployment in wireless sensor networks", *SN Computer Science*, 5(5), 554, 2024. DOI: <https://doi.org/10.1007/s42979-024-02906-1>.
31. Peng C, Zhang Y, Jiang L. "Integrating IoT data and reinforcement learning for adaptive macroeconomic policy optimization", *Alexandria Engineering Journal*, 119, 222–231, 2025. DOI: <https://doi.org/10.1016/j.aej.2025.01.065>.
32. Hashemian A, Derakhshanfard N. "Target Tracking in Internet of Things Using Reinforcement Learning", *Array*, 28, 100551, 2025. DOI: <https://doi.org/10.1016/j.array.2025.100551>.
33. Wang L, Wei Z, Guo W. "Securing IoT Communication Using Physical Sensor Data Graph Layer Security with Federated Multi-agent Deep Reinforcement Learning", In *2023 8th International Conference on Signal and Image Processing (ICSIP)*, pp. 860–865, 2023. DOI: <https://doi.org/10.1109/ICSIP57908.2023.10271026>.
34. Srikanth MV, Sunitha P, Kumar AS, Akshaykranth A. "A Novel Framework for Intrusion Detection in IOT Networks using Hybrid Optimization Algorithm and Convolutional Neural Networks", *Franklin Open*, 14, 100461, 2025. DOI: <https://doi.org/10.1016/j.fraope.2025.100461>.
35. Rahman MM, Al Shakil S, Mustakim MR. "A survey on intrusion detection system in IoT networks", *Cyber Security and Applications*, 3, 100082, 2025. DOI: <https://doi.org/10.1016/j.csa.2024.100082>
36. Tan J, Jin H, Zhang H, Zhang Y, Chang D, Liu X, Zhang H. "A survey: When moving target defense meets game theory", *Computer Science Review*, 48, 100544, 2023. DOI: <https://doi.org/10.1016/j.cosrev.2023.100544>.

37. Arazzi M, Nicolazzo S, Nocera A. "A deep reinforcement learning approach for security-aware service acquisition in IoT", *Journal of Information Security and Applications*, 85, 103856, 2024. DOI: <https://doi.org/10.1016/j.jisa.2024.103856>
38. Kumar A, Singh D. "Securing IoT devices in edge computing through reinforcement learning", *Computers & Security*, 155, 104474, 2025. DOI: <https://doi.org/10.1016/j.cose.2025.104474>
39. Rahman MM, Al Shakil S, Mustakim MR. "A survey on intrusion detection system in IoT networks", *Cyber Security and Applications*, 3, 100082, 2025. DOI: <https://doi.org/10.1016/j.csa.2024.100082>
40. Tan J, Jin H, Zhang H, Zhang Y, Chang D, Liu X, Zhang H. "A survey: When moving target defense meets game theory", *Computer Science Review*, 48, 100544, 2023. DOI: <https://doi.org/10.1016/j.cosrev.2023.100544>
41. Zuo P, Miao C, Fu C, Wang X, Liu X, Liu B. SMAPPO: A security-aware multi-agent reinforcement learning framework for secure computation offloading in SAGIN. *Journal of King Saud University Computer and Information Sciences*. 2025 Dec;37(10):336, DOI: <https://doi.org/10.1007/s44443-025-00315-9>.
42. Far AZ, Far MZ, Gharibzadeh S, Naeini HK, Amini L, Zangeneh S, Rahimi M, Asadi S. "Artificial intelligence for secured information systems in smart cities: Collaborative iot computing with deep reinforcement learning and blockchain", *arXiv preprint arXiv:2409.16444*, 2024. DOI: <https://doi.org/10.48550/arXiv.2409.16444>.
43. Hu W, Alzubi JA, Shreyas J, Al-Razgan M, Ali YA, Karthikayan A. "Enhancing IoT Network Security by Anomaly Detection and Intrusion Prevention Using Gannet Optimization-Based Adaptive Deep Capsule Network", *International Journal of Computational Intelligence Systems*, 18(1), 237, 2025. DOI: <https://doi.org/10.1007/s44196-025-00940-2>.
44. Javadpour A, Ja'fari F, Taleb T, Shojafar M, Benzaïd C. "A comprehensive survey on cyber deception techniques to improve honeypot performance", *Computers & Security*, 140, 103792, 2024. DOI: <https://doi.org/10.1016/j.cose.2024.103792>
45. Javadpour A, Ja'fari F, Benzaïd C, Taleb T. "An optimized reinforcement learning based MTD mutation strategy for securing edge IoT against DDoS attack", *Journal of Information Security and Applications*, 93, 104138, 2025. DOI: <https://doi.org/10.1016/j.jisa.2025.104138>
46. Aparcana-Tasayco AJ, Deng X, Park JH. "A systematic review of anomaly detection in IoT security: towards quantum machine learning approach", *EPJ Quantum Technology*, 12(1), 1–39, 2025. DOI: <https://doi.org/10.1140/epjqt/s40507-025-00414-6>.