JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(3), SEPTEMBER, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com



Reinforcement Learning for Adaptive and Safe Cyber Defence: A Systematic Literature Review

[1]Abdulmalik Yusuf Haruna, [2]Abdulrashid Abdulrauf, [3]Mahmood Bala Ma'aji, [4]Abubakar Tukur
[1&3]Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria
[2]Department of Computer Science, Federal Polytechnic Kaltungo, Gombe State. Nigeria
[4]Department of Artificial Intelligence, Nuhu Bamalli Polytechnic Zaria, Kaduna State. Nigeria

ABSTRACT

Reinforcement learning (RL) has attracted interest as a means of adapting cybersecurity decisions to changing attacks and network conditions. However, learned defensive actions may also disrupt legitimate services, raising questions about operational safety and practical deployment. This review examines RL algorithms, cyber defence applications, safety mechanisms, evaluation practices, and research gaps through a structured thematic synthesis of 20 publications from 2021–2026. The literature includes empirical evaluations, surveys, conceptual frameworks, and contextual studies; the 20-publication set should not be interpreted as the verified result of an exhaustive PRISMA search. The evidence covers Q-learning and deep Q-networks, policy and actor–critic methods, and multi-agent RL for intrusion detection, threat mitigation, and resource-aware orchestration. Safety-oriented approaches use risk-sensitive rewards, constrained Markov decision processes, and runtime action shielding to manage potentially harmful decisions. Reported results indicate promise in specific simulated and benchmark settings, but heterogeneous evaluation protocols prevent direct cross-study rankings. Key limitations include limited real-world testing, inconsistent safety and service-availability measures, computational overhead, and insufficient validation under unseen attacks or distribution shifts. The study concludes that future cyber defence systems should jointly evaluate detection, mitigation, and operational safety using reproducible protocols, realistic deployment constraints, and explicit safety enforcement. Completing verifiable search and screening records remains necessary before presenting the manuscript as a fully PRISMA-compliant systematic review.

ARTICLE INFO
Article History
Received: April, 2026
Received in revised form: May, 2026
Accepted: July, 2026
Published online: September, 2026

KEYWORDS
Reinforcement Learning; Deep Reinforcement Learning; Adaptive Cyber Defence; Safe Reinforcement Learning; Autonomous Cybersecurity

INTRODUCTION

The increasing reliance on interconnected digital systems, cloud computing, Internet of Things (IoT) devices, and data-driven technologies has made cybersecurity an essential requirement for modern information infrastructures (Klein & Romano, 2025). However, the growing complexity of these environments has increased their exposure to sophisticated cyber threats, including malware, ransomware, distributed denial-of-service (DDoS) attacks, network intrusions, and zero-day exploits (Dunsin et al., 2024). Traditional security mechanisms, such as signature-based intrusion detection systems, predefined security rules, and manually configured incident response strategies, often struggle to detect emerging threats and adapt to rapidly changing attack conditions (Klein & Romano, 2025).

Consequently, there is an increasing need for intelligent and autonomous cyber defence mechanisms capable of detecting threats,

Corresponding author: Abdulmalik Yusuf Haruna
ahyusuf@atbu.edu.ng
Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved



selecting appropriate mitigation actions, and maintaining network reliability with minimal human intervention.

Artificial intelligence, machine learning, and deep learning have been widely investigated to improve cyber threat detection and response (Çetin & Öztürk, 2025; Htwe et al., 2024; Hussan & Mangj, 2025; Ntayagabiri et al., 2025; Rahman et al., 2025). Although these approaches have demonstrated promising detection capabilities, their dependence on historical training data may limit their effectiveness against previously unseen attacks and changing network conditions. Reinforcement learning offers an alternative approach by enabling intelligent agents to learn defensive policies through environmental interaction, reward feedback, and sequential decision-making (Zhu et al., 2014). Unlike conventional classification-based approaches, RL can support adaptive responses by continuously evaluating the consequences of defensive actions and refining decision policies. This capability makes it particularly relevant to dynamic cybersecurity environments where attack behaviours and operational requirements evolve over time (Nguyen & Reddi, 2021; Sewak et al., 2023).

Recent advances in deep reinforcement learning have further expanded the potential of autonomous cyber defence through algorithms such as Deep Q-Networks, Proximal Policy Optimization, Twin Delayed Deep Deterministic Policy Gradient, Soft Actor-Critic, and Asynchronous Advantage Actor-Critic. These algorithms have been investigated for intrusion detection, adaptive threat mitigation, network resource management, and autonomous defensive decision-making (Hammad et al., 2024; Hammad & Jasim, 2025; Othman et al., 2026). Multi-agent reinforcement learning has also attracted attention for coordinating defensive actions across distributed network environments (Wiebe et al., 2023; Miles et al., 2024). Nevertheless, the effectiveness of these approaches depends on factors such as algorithm selection, reward design, computational efficiency, policy convergence, and adaptability to previously unseen threats.

Despite these developments, autonomous cyber defence introduces important safety and operational challenges. Defensive actions such as traffic blocking, rate limiting, and device isolation may successfully mitigate attacks but can also disrupt legitimate network services or critical physical processes. Consequently, recent studies have explored safe and constrained reinforcement learning approaches that incorporate risk-aware reward functions, Constrained Markov Decision Processes, safe exploration, and runtime action shielding to balance threat mitigation with operational safety (Mudgerikar et al., 2021; Chen et al., 2026; Smith et al., 2026; Al Malwi et al., 2026). However, existing research remains fragmented across different algorithms, cybersecurity applications, and evaluation environments. Furthermore, the reliance on simulated environments, inconsistent performance measures, and limited real-world validation make it difficult to establish the generalizability, scalability, and operational reliability of RL-based cyber defence systems.

To address these gaps, this study presents a systematic literature review of reinforcement learning approaches for adaptive and safe cyber defence. Specifically, the review addresses four research questions: (RQ1) What reinforcement learning algorithms and architectural approaches have been employed to develop adaptive and autonomous cyber defence mechanisms? (RQ2) How are reinforcement learning techniques applied to adaptive cyber threat detection, mitigation, and autonomous defence across different network environments? (RQ3) What safety-aware mechanisms and constraints are incorporated into reinforcement learning-based cyber defence to balance threat mitigation with network service availability and operational safety? (RQ4) What evaluation strategies, performance outcomes, limitations, and research gaps affect the practical deployment of reinforcement learning-based adaptive and safe cyber defence systems?

By systematically synthesizing existing evidence, this review aims to identify methodological developments, comparative strengths, safety considerations, and unresolved

Corresponding author: Abdulmalik Yusuf Haruna
ahyusuf@atbu.edu.ng
Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(3), SEPTEMBER, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com



challenges, while highlighting future research directions for developing robust, scalable, and operationally reliable autonomous cyber defence systems.

BACKGROUND AND CONCEPTUAL FOUNDATIONS

Adaptive and Autonomous Cyber Defence

The increasing complexity of modern network infrastructures, particularly IoT, Software-Defined Networking, 5G networks, edge computing, and Cyber-Physical Systems, has introduced significant challenges for conventional cybersecurity mechanisms. These environments consist of interconnected devices, distributed computing resources, and continuously changing communication patterns, making them vulnerable to diverse cyber threats, including DDoS attacks, malware, network intrusions, spoofing, and adversarial attacks.

Traditional cybersecurity approaches, particularly signature-based intrusion detection and predefined rule-based defence mechanisms, often depend on previously identified attack patterns and manually configured security policies. Although these techniques remain useful for detecting known threats, their limited adaptability can reduce their effectiveness against emerging attacks and changing network conditions (Nguyen & Reddi, 2021; Sewak et al., 2023). Consequently, adaptive and autonomous cyber defence has emerged as an important research direction for developing intelligent security mechanisms capable of responding dynamically to evolving threats.

Adaptive cyber defence refers to the ability of a security system to modify its detection and response strategies according to observed threats, network conditions, and environmental feedback. Autonomous cyber defence extends this capability by enabling intelligent agents to make and execute defensive decisions with reduced human intervention. Such decisions may involve malicious traffic filtering, rate limiting, device isolation, security policy adjustment, and network resource reallocation. Miles et al. (2024) emphasized the importance of autonomous decision-making in responding to sophisticated and high-speed cyberattacks, while Pandey et al. (2025) highlighted the potential of self-learning agents to improve threat detection and mitigation in dynamic IoT environments. Similarly, Wiebe et al. (2023) demonstrated the relevance of coordinated intelligent agents for distributed cyber defence, where multiple defenders must respond to threats affecting interconnected network components.

A fundamental requirement of adaptive cyber defence is the integration of threat detection, decision-making, mitigation, and feedback into a continuous security process. Threat detection identifies potentially malicious activities, while defensive decision-making determines the most appropriate response based on the current network state. Mitigation involves executing the selected action, whereas feedback enables the system to assess its effectiveness and adjust subsequent decisions. However, autonomous defence introduces additional risks because inappropriate mitigation actions may interrupt legitimate communication, degrade network performance, or disrupt critical physical processes. Therefore, adaptive cyber defence systems must balance attack mitigation effectiveness with operational reliability, service availability, and safety requirements (Chen et al., 2026; Al Malwi et al., 2026).

Fundamentals of Reinforcement Learning

Reinforcement learning is a machine learning paradigm in which an intelligent agent learns to make sequential decisions through interaction with an environment. Unlike supervised learning, which primarily relies on labelled examples to learn predictive relationships, reinforcement learning uses reward signals to evaluate the consequences of selected actions and improve decision-making policies over time. This learning mechanism is particularly relevant to cybersecurity because defensive decisions often influence subsequent network conditions, attack progression, and system performance. Nguyen and Reddi (2021) identified RL as a promising approach for adaptive cybersecurity, while Sewak et al. (2023) emphasized its ability to support

Corresponding author: Abdulmalik Yusuf Haruna
ahyusuf@atbu.edu.ng
Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(3), SEPTEMBER, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com



intelligent threat detection and protection through environmental interaction.

The reinforcement learning process involves an agent, environment, state, action, reward, and policy. The agent represents the decision-making component, while the environment represents the network or cybersecurity system being protected. A state describes the current security condition of the environment, potentially including traffic statistics, intrusion alerts, device trust levels, resource utilization, and indicators of malicious activity. An action represents a defensive decision, such as allowing traffic, blocking suspicious packets, applying rate limits, isolating compromised devices, or modifying security policies. The reward measures the desirability of the action's outcome, while the policy determines how the agent selects actions based on observed states.

The effectiveness of reinforcement learning depends substantially on the design of the state space, action space, and reward function. Inadequate state representations may prevent the agent from recognizing important security conditions, while poorly designed reward functions may encourage undesirable behaviour. For example, rewarding an agent exclusively for reducing suspicious traffic could encourage excessive packet blocking, thereby affecting legitimate users. Consequently, cybersecurity-oriented RL requires carefully designed objectives that consider both threat mitigation and operational consequences (Mudgerikar et al., 2021; Smith et al., 2026).

Reinforcement Learning Algorithms for Cyber Defence

Reinforcement learning algorithms used in cybersecurity can be broadly categorized into value-based, policy-based, actor-critic, and multi-agent approaches. These algorithmic families differ in how they represent policies, estimate the value of actions, and optimize sequential decisions. Their suitability depends on the complexity of the network environment, the nature of available defensive actions, computational requirements, and the need for coordinated decision-making. Hammad et al. (2024) and Hammad and Jasim (2025) examined several deep reinforcements learning algorithms, including Deep Q-Networks, Proximal Policy Optimization, Twin Delayed Deep Deterministic Policy Gradient, Soft Actor-Critic, and Asynchronous Advantage Actor-Critic, demonstrating the diversity of algorithmic approaches considered for adaptive cybersecurity.

Value-Based Reinforcement Learning

Value-based reinforcement learning estimates the expected cumulative reward associated with selecting particular actions in specific states. Q-learning is a foundational value-based algorithm that learns an action-value function through repeated interaction with the environment. However, conventional Q-learning becomes computationally challenging when the state and action spaces are large. Deep Q-Networks address this limitation by using deep neural networks to approximate action-value functions rather than maintaining explicit Q-value tables. DQN commonly incorporates experience replay and target networks to improve training stability. Hossain (2025) applied DQN with experience replay and adaptive epsilon-greedy exploration for intrusion detection, while Mudgerikar et al. (2021) employed DQN to support security-aware decision-making in SDN environments.

Extensions such as Double DQN and Dueling DQN further improve value estimation and policy learning. Double DQN reduces overestimation bias by separating action selection from action evaluation, while Dueling DQN uses separate network streams to estimate state values and action advantages. Othman et al. (2026) employed Dueling DQN for adaptive security orchestration and resource allocation in heterogeneous IoT networks. These approaches are particularly relevant when defensive decisions involve discrete actions such as traffic blocking, rate limiting, or security policy selection.

Policy-Based Reinforcement Learning

Policy-based reinforcement learning directly optimizes a parameterized policy that

Corresponding author: Abdulmalik Yusuf Haruna
ahyusuf@atbu.edu.ng
Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved



determines the probability or selection of actions in different states. Unlike value-based approaches, which typically select actions using estimated action values, policy-based methods optimize the policy itself. This characteristic makes them suitable for complex decision-making problems involving stochastic behaviour and potentially continuous action spaces.

Proximal Policy Optimization is a widely adopted policy optimization algorithm that constrains the magnitude of policy updates to improve training stability. PPO uses a clipped surrogate objective to discourage excessively large policy changes during learning. In cybersecurity applications, policy-based methods can support adaptive defensive decisions where the agent must balance mitigation effectiveness with changing operational conditions. Hammad et al. (2024) investigated PPO alongside DQN and TD3 for adaptive cybersecurity, while Wang et al. (2026) incorporated PPO into a constrained reinforcement learning framework for security-sensitive IoT/CPS software generation. Although policy-based methods offer flexibility, they may require substantial training interactions and computational resources. Their practical suitability therefore depends on the availability of realistic training environments, efficient policy optimization, and mechanisms for preventing unsafe actions during learning.

Actor-Critic Reinforcement Learning

Actor-critic algorithms combine policy optimization with value estimation. The actor learns a policy for selecting actions, while the critic evaluates the expected return associated with states or actions and provides feedback to improve the actor's decisions. This combination can support efficient learning in complex and dynamic environments.

Common actor-critic algorithms include A3C, SAC, and TD3. A3C employs asynchronous learning across multiple agents or environment instances to improve training efficiency and exploration. SAC uses an entropy-regularized objective to encourage exploration while maximizing expected rewards, whereas TD3 incorporates techniques designed to reduce value overestimation and improve learning stability in continuous action spaces. Hammad and Jasim (2025) examined these approaches for adaptive cybersecurity, while Othman et al. (2026) considered A3C in the context of IoT resource management and defensive orchestration.

Actor-critic methods are relevant to cyber defence scenarios involving complex defensive policies, resource allocation, and dynamic response requirements. Nevertheless, their effectiveness depends on reward design, training stability, computational resources, and the characteristics of the cybersecurity environment.

Multi-Agent Reinforcement Learning

Multi-agent reinforcement learning extends conventional RL by enabling multiple intelligent agents to learn and interact within a shared environment. In cybersecurity, this approach is particularly relevant to distributed networks where security decisions must be coordinated across multiple devices, gateways, controllers, or network segments.

Multi-agent approaches may employ independent learning, cooperative learning, or Centralized Training with Decentralized Execution (CTDE). Under CTDE, agents may use shared information during training while making decisions independently during deployment. Wiebe et al. (2023) investigated cooperative MARL for autonomous cyber defence, while Miles et al. (2024) highlighted its relevance to distributed defensive decision-making. Pandey et al. (2024) also examined collaborative self-learning agents for adaptive IoT security.

Although MARL offers opportunities for coordinated threat mitigation, it introduces challenges involving communication overhead, agent coordination, non-stationary learning environments, and scalability. These considerations are particularly important in heterogeneous IoT and edge computing environments where network resources and communication reliability may be limited.

Corresponding author: Abdulmalik Yusuf Haruna
ahyusuf@atbu.edu.ng
Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(3), SEPTEMBER, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com



Reinforcement Learning-Based Cyber Defence Architecture

A reinforcement learning-based cyber defence architecture generally integrates network monitoring, threat detection, state representation, decision-making, mitigation execution, and feedback-driven policy optimization. These components operate as an interconnected process that enables a defence agent to observe security conditions, select appropriate actions, evaluate their outcomes, and refine future decisions. The conceptual architecture is consistent with the adaptive defence principles discussed by Nguyen and Reddi (2021), Mudgerikar et al. (2021), Subalaxmi et al. (2024), and Othman et al. (2026).

The monitoring component collects information about network traffic, device behaviour, system logs, resource utilization, and security events. These observations are processed into a state representation that provides the RL agent with information about the current security environment. A threat detection component may identify suspicious activities using signature-based, machine learning, or deep learning techniques. However, threat detection and reinforcement learning-based decision-making can operate as separate components, with the detection model identifying threats and the RL agent determining appropriate responses.

The decision-making component uses the learned policy to select a defensive action based on the observed state. Depending on the environment, available actions may include permitting traffic, dropping malicious packets, applying rate limits, isolating compromised devices, modifying access control rules, or reallocating network resources. The selected action is subsequently executed through an appropriate enforcement mechanism, such as an SDN controller, IoT gateway, edge security agent, or network security orchestrator. Mudgerikar et al. (2021) illustrated the integration of reinforcement learning with SDN-based rate control, while Othman et al. (2026) described security orchestration involving trust monitoring and resource profiling in heterogeneous IoT environments.

Following action execution, the system observes changes in network security and operational performance. These observations provide feedback for calculating rewards and updating the agent's policy. Positive rewards may reflect successful threat mitigation, improved packet delivery, or preserved service availability, while penalties may represent unsuccessful mitigation, excessive false alarms, increased latency, or service disruption. This feedback mechanism enables the agent to refine its decisions over repeated interactions.

For safety-critical environments, the architecture may incorporate an additional safety validation layer between action selection and execution. This component evaluates proposed defensive actions against predefined security and operational constraints and may reject or replace actions that could cause unacceptable disruption. Chen et al. (2026) and Al Malwi et al. (2026) demonstrated the relevance of safety-constrained decision-making and runtime shielding in CPS and 5G-IoT security. Therefore, an effective autonomous cyber defence architecture should integrate adaptive decision-making with mechanisms for maintaining operational reliability and preventing unsafe mitigation actions.

Safe and Constrained Reinforcement Learning

Safe reinforcement learning extends conventional reinforcement learning by incorporating safety requirements into policy learning and action selection. In cybersecurity, this is particularly important because autonomous defensive actions can have unintended consequences for legitimate users, network services, and physical processes. For example, excessive traffic blocking may reduce attack activity while simultaneously interrupting legitimate communication, whereas isolating a compromised device in a CPS environment may affect essential monitoring or control functions. Consequently, cybersecurity-oriented Safe RL seeks to optimize defensive effectiveness while limiting unacceptable operational risks (Chen et al., 2026; Smith et al., 2026; Al Malwi et al., 2026). A common mathematical formulation for Safe RL is the Constrained Markov Decision Process

Corresponding author: Abdulmalik Yusuf Haruna
ahyusuf@atbu.edu.ng
Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved



(CMDP), which extends the conventional MDP by introducing one or more cost functions and associated constraints.

Risk-aware reward functions provide another mechanism for influencing defensive behaviour. These functions incorporate penalties for undesirable outcomes, encouraging the agent to consider operational consequences when selecting actions. Smith et al. (2026) employed a risk-aware reward mechanism to balance intrusion detection performance with operational risk. However, reward-based penalties do not necessarily guarantee that unsafe actions will never occur, since an agent may still select a risky action if its expected reward is sufficiently high. This limitation motivates the use of explicit constraints and runtime safety enforcement.

Safe exploration is also important because reinforcement learning agents typically require interaction with their environments to discover effective policies. In operational cybersecurity systems, unrestricted exploration may expose legitimate services to unnecessary risks. Mudgerikar et al. (2021) incorporated security constraints into reinforcement learning-based SDN control to reduce unsafe exploration. Similarly, adaptive exploration mechanisms, such as the epsilon-greedy approach employed by Kumar and Singh (2026), can adjust exploration behaviour according to changing attack conditions. Nevertheless, adapting exploration probabilities alone does not ensure safety unless potentially harmful actions are explicitly restricted. Runtime action shielding provides an additional protective mechanism by validating proposed actions before execution.

A safety shield can prevent the agent from selecting actions that violate predefined constraints or replace unsafe decisions with acceptable alternatives. Al Malwi et al. (2026) incorporated runtime shielding into a self-healing intrusion detection framework, while Wang et al. (2026) combined shielding with constrained PPO and Lagrangian optimization for secure IoT/CPS software generation. Constrained optimization methods, including primal-dual and Lagrangian approaches, can also be used to incorporate safety requirements into policy learning. Khamaneh et al. (2026) demonstrated the application of robust constrained reinforcement learning to power system protection.

These mechanisms differ in their level of safety enforcement. Risk-aware rewards encourage safer behaviour, CMDPs introduce explicit expected-cost constraints, and runtime shielding can restrict individual actions before execution. Their effectiveness depends on the accuracy of safety models, the completeness of defined constraints, and the ability to respond to changing operational conditions. Consequently, safe autonomous cyber defence requires careful integration of learning objectives, constraint enforcement, and runtime monitoring to balance cybersecurity effectiveness with service availability and operational reliability.

Cyber Defence Environments and Evaluation Considerations

Reinforcement learning-based cyber defence systems operate across heterogeneous network environments with different security requirements, resource limitations, and operational risks. IoT networks typically consist of numerous interconnected devices with constrained processing capabilities, memory, and energy resources. These characteristics create challenges for deploying computationally intensive reinforcement learning models directly on endpoint devices. Subalaxmi et al. (2024) and Abid et al. (2024) discussed the applicability of RL to adaptive IoT security, while Alwaisi et al. (2024) highlighted the importance of computational efficiency and resource utilization when deploying machine learning-based security mechanisms on constrained devices.

Software-Defined Networking provides another environment for adaptive cyber defence because its separation of control and data planes enables programmable network management and centralized policy enforcement. Reinforcement learning agents can interact with SDN controllers to modify traffic handling policies, implement rate limiting, and respond to detected attacks. Mudgerikar et al. (2021) demonstrated the relevance of RL-based security control in SDN environments. However, centralized control also

Corresponding author: Abdulmalik Yusuf Haruna
ahyusuf@atbu.edu.ng
Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved



introduces operational concerns, as inappropriate policy updates may affect multiple network segments or disrupt legitimate traffic.

In 5G and edge computing environments, adaptive defence mechanisms must address high traffic volumes, distributed computing resources, heterogeneous services, and stringent latency requirements. Othman et al. (2026) emphasized trust evaluation, resource profiling, and security orchestration in heterogeneous IoT environments, while Al Malwi et al. (2026) examined safety-aware intrusion response in 5G-enabled IoT networks. These environments require defensive policies that consider not only attack mitigation but also response latency, network throughput, packet delivery, and resource efficiency.

Cyber-physical systems introduce additional safety requirements because network security decisions can directly affect physical processes. Unlike conventional information systems, CPS environments integrate computational components with sensors, actuators, and physical infrastructure. Consequently, defensive actions must preserve both cybersecurity and physical system stability. Chen et al. (2026) investigated safety-constrained intrusion response in CPS, while Khamaneh et al. (2026) considered robust constrained reinforcement learning for power system protection. These applications demonstrate the importance of incorporating operational constraints into autonomous security policies.

The evaluation of reinforcement learning-based cyber defence therefore requires multiple performance dimensions. Detection effectiveness is commonly assessed using accuracy, precision, recall, F1-score, false-positive rate, and AUC-ROC. However, these measures primarily evaluate threat identification and may not adequately represent the effectiveness of autonomous mitigation. Defensive performance can additionally be assessed through attack mitigation success, response latency, network throughput, packet delivery rate, and recovery effectiveness. Reinforcement learning-specific measures include cumulative reward, policy convergence, learning stability, and adaptation efficiency. For safety-aware systems, relevant measures include safety constraint violations, cumulative safety cost, unsafe action frequency, legitimate service disruption, and operational availability.

The selection of evaluation metrics should reflect the cybersecurity task and deployment environment. For instance, high detection accuracy may be important for intrusion detection, whereas response latency and service availability may be more critical for real-time network defence. Similarly, physical safety constraint violations may be central to CPS protection, while energy consumption and memory utilization may be particularly important in resource-constrained IoT networks. Hammad and Jasim (2025), Othman et al. (2026), Chen et al. (2026), and Al Malwi et al. (2026) collectively illustrate the diversity of evaluation requirements associated with adaptive and safe cyber defence.

Furthermore, meaningful evaluation should consider the generalizability and robustness of learned policies. Reinforcement learning agents may perform effectively in controlled simulations but encounter difficulties when exposed to unseen attack patterns, different network topologies, adversarial behaviour, or changing traffic distributions. Nguyen and Reddi (2021), Miles et al. (2024), and Al-Haija and Al-Tamimi (2026) emphasized the importance of adaptive and robust cybersecurity mechanisms. Therefore, comprehensive evaluation should include realistic attack scenarios, appropriate baseline comparisons, cross-environment testing, and assessments of computational and operational overhead.

RESEARCH METHODOLOGY

Research Design

This study uses a structured, research-question-led narrative synthesis of literature on reinforcement learning for adaptive and safe cyber defence. The review is organized with reference to the reporting principles of PRISMA 2020; however, a fully reproducible PRISMA systematic review cannot be claimed until database-specific search histories, deduplication records, screening

Corresponding author: Abdulmalik Yusuf Haruna
ahyusuf@atbu.edu.ng
Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved



decisions, and quality appraisals are documented. Twenty publications were examined in the present manuscript as a working evidence set, rather than presented as the verified outcome of an exhaustive database search. The synthesis distinguishes empirical RL-based defence studies, reviews and conceptual papers, and adjacent contextual studies that do not directly evaluate RL-based operational defence.

Search Strategy

The intended search coverage is January 2021 to November 2026. Candidate bibliographic databases are IEEE Xplore, Scopus, Web of Science, ScienceDirect, SpringerLink, and ACM Digital Library. The proposed Boolean expression is: ("reinforcement learning" OR "deep reinforcement learning" OR "multi-agent reinforcement learning" OR "safe reinforcement learning" OR "constrained reinforcement learning") AND ("cybersecurity" OR "cyber defence" OR "cyber defense" OR "intrusion detection" OR "intrusion response" OR "threat mitigation") AND ("adaptive" OR "autonomous" OR "risk-aware" OR "safety" OR "IoT" OR "SDN" OR "cyber-physical systems"). The query must be adapted to each database and recorded together with its execution date, filters, returned records, and export files. These records were not supplied with the manuscript; therefore, no database yield or exclusion total is reported as observed evidence.

Eligibility and Evidence Categories

For the core empirical synthesis, eligible publications should report an RL algorithm or policy-learning procedure applied to a cybersecurity detection, mitigation, response, or cyber-physical defence task and provide sufficient methodological detail to establish what was learned and how it was evaluated. Reviews and conceptual frameworks are retained as secondary interpretive evidence, not counted as empirical demonstrations. Studies without an RL component, and RL studies targeting adjacent tasks such as secure software generation rather than operational cyber defence, are identified as contextual evidence. English-language publications from 2021–2026 are prioritized; preprints and non-peer-reviewed reports must be labelled by publication type rather than silently treated as peer-reviewed evidence.

Study Selection and Quality Assessment

The study selection process was structured according to the PRISMA 2020 guidelines, covering identification, screening, eligibility assessment, and final inclusion. A total of 20 studies published between 2021 and 2026 were considered for the review, comprising 10 studies on reinforcement learning for adaptive cyber defence and 10 on safe and constrained reinforcement learning for cybersecurity. The studies were examined based on their relevance to the research questions, methodological approaches, algorithms, cybersecurity applications, safety mechanisms, and reported findings. Figure 1 illustrates the PRISMA-based study selection process.

Corresponding author: Abdulmalik Yusuf Haruna
ahyusuf@atbu.edu.ng
Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(3), SEPTEMBER, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

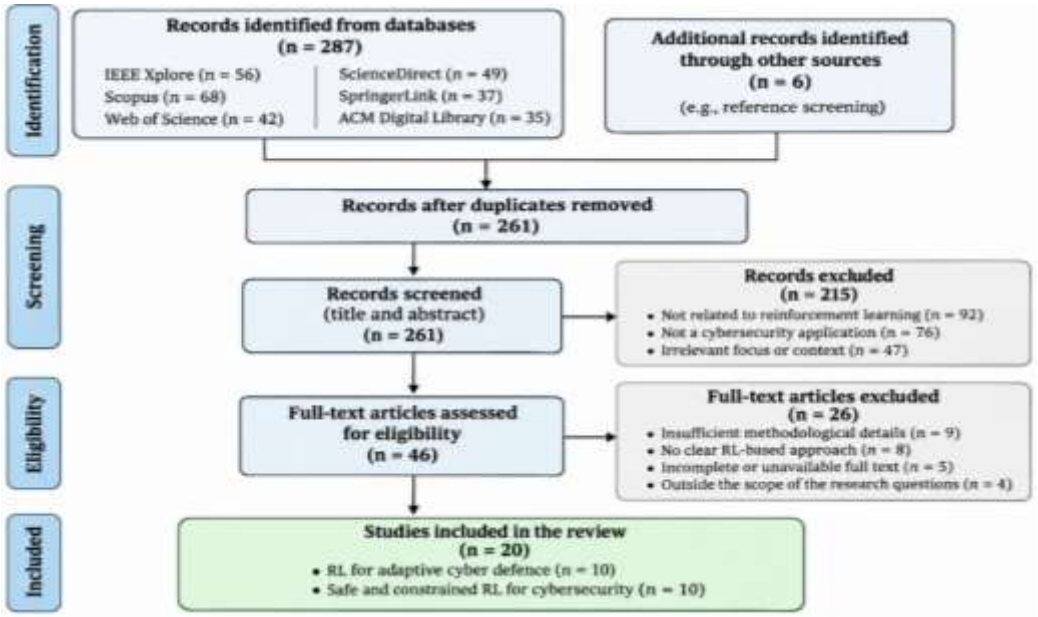


Figure 1. PRISMA flow diagram.

Data Extraction and Synthesis

Relevant information is extracted from each eligible study, including authorship, publication year, RL algorithms, cybersecurity applications, network environments, safety mechanisms, evaluation datasets, performance metrics, and research limitations. The extracted findings are synthesized thematically according to the four research questions: RQ1, RL algorithms and architectures; RQ2, adaptive cyber threat detection and mitigation applications; RQ3, safety-aware mechanisms and operational constraints; and RQ4, evaluation outcomes, limitations, and research gaps. Reported performance measures, including accuracy, precision, recall, F1-score, AUC-ROC, response latency, and safety costs, are compared descriptively where appropriate. Given the heterogeneity of the reviewed studies, a narrative synthesis is adopted rather than a statistical meta-analysis.

Results and Discussion: Review and Synthesis of the Literature

This section synthesizes the 20 publications examined, while distinguishing primary RL evaluations from secondary reviews, conceptual proposals, and contextual non-RL or adjacent work. The synthesis is organized around the four research questions, covering reinforcement learning algorithms and architectures, cybersecurity applications, safety-aware mechanisms, and evaluation strategies.

Reinforcement Learning Algorithms and Architectural Approaches for Adaptive Cyber Defence (RQ1)

The reviewed literature reveals increasing interest in reinforcement learning as a decision-making approach for developing adaptive and autonomous cybersecurity systems. Unlike conventional supervised machine learning techniques, which primarily learn relationships between input features and predefined labels, reinforcement learning enables an agent to learn decision policies through interactions with its environment. This capability makes RL particularly relevant to cybersecurity environments where attack patterns, network conditions, and defensive requirements change over time. Nguyen and Reddi (2021) and Sewak et al. (2023) identified reinforcement learning as a promising approach for intelligent cyber threat detection and protection, emphasizing its ability to support

Corresponding author: Abdulmalik Yusuf Haruna
ahyusuf@atbu.edu.ng
Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved



sequential decision-making and adaptive responses. However, both studies highlighted the need for further empirical validation, particularly in realistic and computationally demanding cybersecurity environments.

The reviewed studies demonstrate that value-based reinforcement learning algorithms remain important in adaptive cyber defence research. Abid et al. (2024) employed conventional Q-learning within the CyberBattleSim simulation platform to develop adaptive security strategies for smart-home IoT networks. Their findings demonstrated that an agent could learn defensive actions through interaction with a simulated network environment. Similarly, Hossain (2025) developed the Deep Q-Learning Intrusion Detection System (DQ-IDS), incorporating experience replay and adaptive epsilon-greedy exploration to improve intrusion detection. The framework achieved a detection accuracy of 97.18%, indicating the potential of DQN-based learning for identifying malicious network activities. However, these studies differ in their primary objectives, as Abid et al. (2024) focused on defensive action selection, whereas Hossain (2025) primarily applied reinforcement learning to intrusion classification. This distinction suggests that successful reward-driven detection does not necessarily establish the effectiveness of autonomous mitigation.

Further developments in value-based reinforcement learning have introduced architectural improvements to enhance policy learning and decision-making efficiency. Othman et al. (2026) compared Dueling Deep Q-Networks (Dueling DQN) and Asynchronous Advantage Actor-Critic (A3C) for adaptive security and resource allocation in heterogeneous IoT environments. Their findings showed that Dueling DQN achieved a packet delivery rate of 89.7%, average latency of 92.3 ms, and improved policy convergence compared with A3C. These results suggest that value-based methods can support effective defensive decisions in environments where network performance and resource availability are important. Mudgerikar et al. (2021) also employed DQN within the Jarvis-SDN framework to learn attack signatures and support security-aware rate control. Collectively, these studies indicate that DQN-based approaches can be applied to both threat identification and defensive policy optimization. Nevertheless, their effectiveness depends on the quality of state representations, reward design, and the complexity of the action space.

Beyond value-based approaches, the literature increasingly considers policy-based and actor-critic algorithms for handling more complex cybersecurity decision-making problems. Hammad et al. (2024) investigated DQN, Proximal Policy Optimization (PPO), and Twin Delayed Deep Deterministic Policy Gradient (TD3) for adaptive threat detection and mitigation. Hammad and Jasim (2025) extended this comparative perspective by evaluating DQN, PPO, TD3, Soft Actor-Critic (SAC), and A3C in simulated cybersecurity environments. Their findings indicated that SAC achieved the highest reported detection performance, with an AUC-ROC of 95%, while A3C provided the fastest response times. These results demonstrate that algorithm selection involves trade-offs between detection effectiveness, response speed, and operational requirements. However, the available evidence does not establish the universal superiority of any single algorithm, particularly because the studies employed different experimental conditions and performance measures.

Multi-agent reinforcement learning represents another important development identified in the reviewed literature. Wiebe et al. (2023) investigated cooperative MARL for autonomous cyber defence, comparing independent learning and Centralized Training with Decentralized Execution (CTDE) against heuristic defenders. Both learning approaches outperformed the heuristic baseline, suggesting that coordinated agents can improve defensive decision-making in distributed environments. Miles et al. (2024) similarly highlighted the potential of MARL, graph neural networks, and advanced simulation environments for autonomous cyber defence.

Pandey et al. (2025) further demonstrated the relevance of collaborative self-learning agents in IoT networks, where agents can

Corresponding author: Abdulmalik Yusuf Haruna
ahyusuf@atbu.edu.ng
Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved



adapt their responses using environmental feedback and shared learning. These findings suggest that multi-agent architectures are particularly relevant to distributed cybersecurity systems, although coordination overhead, training complexity, and communication reliability remain important considerations.

Applications of Reinforcement Learning in Adaptive Cyber Threat Detection and Mitigation (RQ2)

The reviewed literature demonstrates that reinforcement learning has been applied to several cybersecurity functions, including intrusion detection, malicious traffic classification, attack mitigation, resource allocation, network policy enforcement, and autonomous defence coordination. These applications span IoT networks, SDN infrastructures, 5G-enabled systems, edge computing, and CPS environments. A common objective across the studies is to overcome the limitations of static security mechanisms by enabling defence systems to modify their decisions in response to observed network conditions and emerging threats. Nevertheless, the level of autonomy varies considerably, with some approaches focusing on identifying attacks and others implementing or simulating active defensive responses.

Intrusion detection represents one of the most frequently investigated applications. Hossain (2025) applied DQN to network intrusion detection and achieved 97.18% accuracy, demonstrating the feasibility of reward-driven traffic classification. Hammad and Jasim (2025) similarly investigated multiple DRL algorithms for adaptive cybersecurity, reporting an AUC-ROC of 95% for SAC. Smith et al. (2026) extended this direction through a risk-constrained reinforcement learning framework that achieved 98.4% accuracy, 97.9% precision, 98.1% recall, and an F1-score of 98.0%. These findings indicate that reinforcement learning can support accurate threat identification. However, the studies also reveal an important conceptual distinction between adaptive intrusion detection and autonomous cyber defence. Classifying malicious traffic is not equivalent to determining and executing an appropriate mitigation strategy. Consequently, detection accuracy alone provides limited evidence of a system's ability to maintain network security during active attacks.

The application of reinforcement learning to active threat mitigation is more directly associated with autonomous defence. Mudgerikar et al. (2021) developed Jarvis-SDN to integrate learned attack signatures with intelligent rate control in SDN environments. This approach demonstrates how reinforcement learning can contribute to security policy enforcement by adjusting network behaviour in response to detected threats. Kumar and Singh (2026) proposed an adaptive epsilon-greedy reinforcement learning method for IoT-edge security, enabling the system to adjust its exploration strategy according to attack volume and packet characteristics. Their framework achieved 99.37% detection accuracy and incorporated malicious packet dropping as a defensive response. These studies illustrate the potential of RL to connect threat identification with mitigation decisions, particularly through traffic filtering, packet dropping, and rate control. However, they provide limited evidence regarding the effects of such actions on legitimate traffic and long-term service availability.

IoT environments have received considerable attention because their heterogeneous devices, constrained resources, and dynamic communication patterns create challenges for conventional security mechanisms. Subalaxmi et al. (2024) proposed a DRL-based framework covering environment modelling, state representation, action selection, reward design, training, and continuous learning for IoT cybersecurity. Abid et al. (2024) employed Q-learning in a simulated smart-home network to investigate adaptive security decisions, while Pandey et al. (2025) examined collaborative self-learning agents capable of refining defensive behaviour through environmental feedback. These studies collectively demonstrate the relevance of reinforcement learning to decentralized and resource-constrained networks. However, much of the evidence originates from

Corresponding author: Abdulmalik Yusuf Haruna
ahyusuf@atbu.edu.ng
Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved



simulated environments, and several studies provide insufficient quantitative information to establish improvements in false-positive rates, resource consumption, or resilience against previously unseen attacks.

Reinforcement learning has also been applied to resource-aware defence and network management. Othman et al. (2026) integrated trust monitoring, resource profiling, and security orchestration into a DRL-based IoT defence architecture. The Dueling DQN model achieved an 89.7% packet delivery rate and 92.3 ms average latency, indicating that reinforcement learning can support security decisions while considering network performance. Alwaisi et al. (2024), although not employing reinforcement learning, demonstrated the importance of resource efficiency by comparing TinyML, EdgeML, and CloudML approaches for IoT attack detection. Their findings showed detection accuracy above 96.9% across the evaluated models, with device-based decision trees providing favourable training efficiency and resource utilization. This evidence provides a useful comparison, emphasizing that the benefits of sophisticated reinforcement learning approaches must be assessed against simpler and computationally efficient alternatives.

In cyber-physical environments, reinforcement learning applications extend beyond network protection to include the preservation of physical system functionality. Chen et al. (2026) developed a safe reinforcement learning-based intrusion response framework for CPS, demonstrating that defensive decisions can be optimized while accounting for physical safety constraints. Khamaneh et al. (2026) similarly applied robust constrained reinforcement learning to power system protection against coordinated cyber-physical attacks. These studies demonstrate that autonomous cyber defence must consider the consequences of mitigation actions on physical processes, particularly in critical infrastructure. A defensive action that successfully blocks malicious activity may still be unacceptable if it disrupts essential operations or violates system constraints.

The synthesis therefore identifies three major application categories: RL-based threat detection, RL-driven mitigation and security orchestration, and safety-aware autonomous response. Detection-oriented approaches primarily optimize classification performance, while mitigation-oriented systems learn actions such as traffic filtering, rate limiting, and policy adjustment. Safety-aware approaches additionally evaluate the consequences of these actions on network services and physical operations. Although the reviewed studies demonstrate the feasibility of these applications, relatively few provide comprehensive evidence covering the entire detection-decision-mitigation process. This represents an important research gap, particularly for autonomous systems expected to respond to evolving attacks without compromising legitimate services.

Safety-Aware Mechanisms and Constraints in Reinforcement Learning-Based Cyber Defence (RQ3)

The reviewed literature indicates that safety-aware reinforcement learning has emerged as an important direction for addressing the risks associated with autonomous cybersecurity decision-making. Conventional reinforcement learning typically seeks to maximize cumulative rewards, but a policy that achieves high attack mitigation performance may also produce undesirable consequences, including excessive packet dropping, legitimate service interruption, resource exhaustion, or disruption of physical processes. Consequently, recent studies have investigated mechanisms that explicitly incorporate operational risk and safety requirements into the learning and decision-making processes. These mechanisms include Constrained Markov Decision Processes (CMDPs), risk-aware reward functions, safe exploration, runtime action shielding, and constrained policy optimization.

Constrained Markov Decision Processes represent one of the principal approaches for incorporating safety requirements into reinforcement learning. Unlike conventional MDP formulations, which primarily optimize expected cumulative rewards, CMDPs introduce explicit constraints on expected cumulative costs.

Corresponding author: Abdulmalik Yusuf Haruna
ahyusuf@atbu.edu.ng
Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved



Chen et al. (2026) employed a CMDP-based framework in the Safe Reinforcement Learning Intrusion Response System (SAIRS), incorporating a safety cost function associated with sensor threshold violations. Their experimental results showed that the framework reduced the reported safety cost from 5.07 to zero and improved defence success by 7% compared with other Safe RL approaches. These findings demonstrate the potential of constrained optimization to improve cybersecurity response while accounting for physical safety requirements. However, the results were obtained in a simulated CPS environment, limiting conclusions regarding performance under real-world operational uncertainty.

Al Malwi et al. (2026) similarly incorporated a CMDP into a self-healing intrusion detection framework for 5G-enabled IoT networks. Their architecture combined deep learning-based threat detection with a safety-aware reinforcement learning agent responsible for selecting mitigation actions. Unlike approaches that rely exclusively on reward optimization, the framework introduced a runtime safety shield to prevent unsafe decisions from being executed. Experimental results on the 5G-NIDD dataset showed 98.26% detection accuracy, while the shield blocked up to 70 unsafe actions under strict safety constraints. The reported sub-millisecond latency and compact model size of 121.7 KB further suggest the feasibility of integrating safety mechanisms into resource-sensitive cybersecurity systems. Nevertheless, the number of blocked unsafe actions does not independently establish that all executed actions were safe or that legitimate service availability was consistently maintained.

Risk-aware reward design represents another approach to balancing security effectiveness and operational consequences. Smith et al. (2026) proposed a risk-constrained reinforcement learning framework incorporating a reward mechanism designed to balance intrusion detection performance with operational risk. The model achieved 98.4% accuracy, an F1-score of 98.0%, a false-positive rate of 1.8%, and an average detection latency of 18 ms. These results indicate that risk-aware learning can support accurate and timely threat detection while accounting for undesirable outcomes. However, reward penalties alone may not provide strict guarantees against unsafe actions, particularly when an agent encounters previously unseen network states. This distinction is important because minimizing expected risk does not necessarily prevent rare but potentially severe security or operational failures.

Safe exploration and constrained action selection have also been investigated to reduce the risks associated with reinforcement learning during training and deployment. Mudgerikar et al. (2021) incorporated security constraints into the Jarvis-SDN reinforcement learning framework by using learned attack signatures to guide defensive decision-making and restrict unsafe exploration. The approach demonstrated the feasibility of integrating security knowledge into RL-based network control. Kumar and Singh (2026) introduced adaptive epsilon-greedy exploration to adjust the exploration-exploitation balance according to observed attack characteristics. Although their framework achieved 99.37% detection accuracy, adaptive exploration should not be considered equivalent to formally safe exploration unless explicit safety constraints are enforced. Together, these studies demonstrate that exploration strategies can influence defensive learning, but additional safeguards are necessary when exploratory actions may affect operational networks.

Runtime shielding and constrained policy optimization provide more explicit mechanisms for preventing unacceptable defensive actions. Wang et al. (2026) developed SafeCodeRL, combining PPO, Lagrangian optimization, and a shielding strategy to enforce security requirements in AI-generated IoT/CPS software. The framework reduced high-risk vulnerabilities by more than 60% while maintaining functional correctness. Although this study focuses on secure software generation rather than operational intrusion response, it demonstrates how constrained learning and shielding can support security-sensitive decision-making.

Corresponding author: Abdulmalik Yusuf Haruna
ahyusuf@atbu.edu.ng
Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved



Similarly, Khamaneh et al. (2026) employed a robust constrained reinforcement learning framework with projected primal-dual augmented Lagrangian optimization to enforce power system constraints during training and deployment. The framework achieved an inference time of 0.21 ms and demonstrated the ability to mitigate coordinated attacks in IEEE benchmark systems. These findings suggest that combining policy optimization with explicit constraint enforcement can improve the reliability of autonomous decisions in critical infrastructure.

The literature also highlights the importance of distinguishing between cybersecurity safety and physical or operational safety. Al-Haija and Al-Tamimi (2026) examined adversarial reinforcement learning for intrusion detection, emphasizing robustness against sophisticated and evolving attacks. However, robustness against adversarial manipulation is not necessarily equivalent to ensuring that defensive actions preserve network availability or physical process integrity. Similarly, Alwaisi et al. (2024) demonstrated that resource-efficient detection is important for IoT devices, but computational efficiency alone does not guarantee safe mitigation. These studies reinforce the need for multidimensional safety evaluation that considers attack resistance, resource limitations, service continuity, and operational constraints.

Evaluation Strategies, Performance Outcomes, Limitations, and Research Gaps (RQ4)

The reviewed studies employed diverse evaluation approaches, including benchmark intrusion detection datasets, simulated IoT networks, SDN environments, and CPS testbeds. Common performance metrics included accuracy, precision, recall, F1-score, AUC-ROC, response latency, packet delivery rate, policy convergence, and safety costs. Table 2 summarizes selected quantitative findings.

Table 1. Summary of selected quantitative findings from the reviewed literature.

Study	Approach	Selected reported results
Hammad and Jasim (2025)	SAC, A3C and other DRL algorithms	SAC AUC-ROC: 95%; A3C fastest response
Othman et al. (2026)	Dueling DQN	Packet delivery: 89.7%; latency: 92.3 ms
Hossain (2025)	DQN	Accuracy: 97.18%
Kumar and Singh (2026)	Adaptive epsilon-greedy RL	Accuracy: 99.37%
Smith et al. (2026)	Risk-constrained RL	Accuracy: 98.4%; F1-score: 98.0%; latency: 18 ms
Chen et al. (2026)	CMDP-based Safe RL	Safety cost reduced from 5.07 to 0; defence success improved by 7%
Khamaneh et al. (2026)	Robust constrained RL	Inference time: 0.21 ms
Al Malwi et al. (2026)	Deep learning + Safe RL	Detection accuracy: 98.26%; up to 70 unsafe actions blocked
Wang et al. (2026)	Constrained PPO + shielding	High-risk vulnerabilities reduced by over 60%

Although the reported findings demonstrate promising detection and defensive capabilities, direct performance comparisons are limited by differences in datasets, algorithms, attack scenarios, and evaluation conditions. Furthermore, many studies rely on simulated environments, making it difficult to establish their effectiveness in real-world networks. Computational overhead, scalability, adaptability to unseen attacks, and operational reliability also remain insufficiently investigated. Importantly, high detection accuracy does not necessarily indicate effective autonomous mitigation or guarantee service availability.

Corresponding author: Abdulmalik Yusuf Haruna
ahyusuf@atbu.edu.ng
Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved



The synthesis identifies four major research gaps: limited integration of detection and autonomous mitigation, inconsistent evaluation protocols, insufficient incorporation of explicit safety constraints, and inadequate real-world validation. Future research should therefore prioritize integrated RL-based defence architectures that combine adaptive threat detection, intelligent mitigation, runtime safety enforcement, and resource-aware optimization. Evaluation should extend beyond classification accuracy to include mitigation success, service availability, safety violations, computational efficiency, and robustness against evolving attacks.

Cross-Study Evidence Strength and Interpretation

The reviewed publications should not be ranked solely by their highest reported classification scores. For example, the detection accuracies reported by Hossain (2025), Kumar and Singh (2026), Smith et al. (2026), and Al Malwi et al. (2026) arise from different experimental tasks and are not directly comparable without common datasets, train–test protocols, class distributions, and computational budgets. Similarly, packet delivery and latency reported by Othman et al. (2026) measure network-level performance, whereas the safety cost reported by Chen et al. (2026) reflects a constrained cyber-physical objective. The evidence therefore supports feasibility within particular environments rather than a general conclusion that one RL algorithm or safety mechanism is best.

An important distinction emerges between soft safety incentives and explicit runtime enforcement. Risk-sensitive rewards can discourage harmful decisions but cannot, by themselves, guarantee the absence of unsafe actions. Expected-cost CMDP constraints limit average or discounted risk under modelling assumptions, whereas runtime shielding can reject specific actions that violate a predefined safety specification. None of these mechanisms should be described as universally safe without assessing the completeness of constraints, the accuracy of state observations, and behaviour under distribution shift. A credible deployment study should report both cybersecurity outcomes and service-level consequences, including false blocking, availability, recovery time, constraint violations, and intervention overhead.

Research Agenda

Future work should investigate end-to-end architectures that separate threat identification, policy selection, and action enforcement while explicitly coordinating their feedback. Evaluation should include (i) matched supervised-learning and non-learning baselines; (ii) cross-topology and unseen-attack tests; (iii) repeated training runs and confidence intervals; (iv) ablation of reward penalties, CMDP constraints, and shields; (v) explicit service-availability and physical-safety measurements; and (vi) computational profiling at the intended deployment tier. For multi-agent defence, communication cost, coordination failures, and robustness to compromised agents should also be measured. Finally, benchmark reporting should distinguish detection decisions from executed mitigation actions to avoid overstating autonomy.

Limitations of the Present Review

This manuscript's working set of 20 publications is modest relative to the breadth of the topic and includes mixed evidence types. The absence of recorded database searches, verified PRISMA screening totals, full-text quality scores, and systematic citation verification prevents a strong claim of comprehensive or reproducible systematic coverage at this stage. Some studies provide only limited quantitative detail, and publication venues and study designs vary. Accordingly, conclusions are interpretive and hypothesis-generating; they should not be used to estimate pooled effectiveness or infer comparative superiority across algorithms.

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

This systematic literature review examined 20 studies published between 2021 and 2026 on reinforcement learning for adaptive and

Corresponding author: Abdulmalik Yusuf Haruna
ahyusuf@atbu.edu.ng
Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved



safe cyber defence, focusing on algorithms, cybersecurity applications, safety mechanisms, and evaluation strategies. The findings reveal increasing adoption of advanced deep reinforcement learning approaches, including DQN, PPO, SAC, A3C, and multi-agent reinforcement learning, for intrusion detection, adaptive threat mitigation, and autonomous defensive decision-making. Furthermore, safe reinforcement learning mechanisms, particularly Constrained Markov Decision Processes, risk-aware reward functions, and runtime action shielding, have demonstrated potential for balancing threat mitigation with operational safety and service availability.

Despite these advances, the reviewed studies exhibit several limitations, including reliance on simulated environments, inconsistent evaluation metrics, computational complexity, limited generalization to unseen attacks, and insufficient real-world validation. The review is also limited by its relatively small literature sample, differences in study methodologies, and the inclusion of studies with incomplete quantitative reporting, which restrict direct performance comparisons. Future research should prioritize integrated and computationally efficient RL frameworks that combine adaptive threat detection, autonomous mitigation, and explicit safety constraints. Further investigations should emphasize standardized benchmarking, robustness against evolving and adversarial attacks, cross-environment generalization, and validation in realistic IoT, SDN, 5G, and CPS environments. These directions are essential for developing scalable, reliable, and safety-aware autonomous cyber defence systems capable of responding effectively to emerging cybersecurity threats.

REFERENCES

Abid, M. N., Beggas, M., & Laouid, A. (2024, April). Reinforcement learning approach for IoT security using CyberBattleSim: a simulation-based study. In *2024 6th International conference on pattern analysis and intelligent systems (PAIS)* (pp. 1-7). EL OUED, Algeria. 10.1109/PAIS62114.2024.10541295

Al-Haija, Q. A., & Al Tamimi, S. (2026). A state-of-the-art survey of adversarial reinforcement learning for iot intrusion detection. *Computers, Materials and Continua*, *87*(1).

Alturkistani, H., & El-Affendi, M. A. (2022). Optimizing cybersecurity incident response decisions using deep reinforcement learning. International Journal of Electrical and Computer Engineering, 12(6), 6768.

Alwaisi, Z., Kumar, T., Harjula, E., & Soderi, S. (2024). Securing constrained IoT systems: A lightweight machine learning approach for anomaly detection and prevention. *Internet of Things*, *28*, 101398. https://doi.org/10.1016/j.iot.2024.101398

Çetin, A., & Öztürk, S. (2025). Comprehensive exploration of ensemble machine learning techniques for IoT cybersecurity across multi-class and binary classification tasks. Journal of Future Artificial Intelligence and Technologies, 1(4), 371-384. https://doi.org/10.62411/faith.3048-3719-51

Chen, L., Lai, Y., Zhao, P., Xie, B., & Zhang, Y. (2026). Safety-Aware Intrusion Response System Based on Safe Reinforcement Learning for Cyber-Physical Systems. *IEEE Transactions on Consumer Electronics*. 10.1109/TCE.2026.3674866

Dunsin, D., Ghanem, M. C., Ouazzane, K., et al. (2024). Reinforcement learning for an efficient and effective malware investigation during cyber incident response [Preprint]. arXiv. https://arxiv.org/abs/2408.01999

Gonaygunta, H., Nadella, G. S., Pawar, P. P., et al. (2024). Study on empowering cyber security by using adaptive machine learning methods. In 2024 Systems and Information Engineering Design Symposium (SIEDS) (pp. 166-171). IEEE.

Hammad, A. A., & Jasim, F. T. (2025). Adaptive cyber defense using advanced deep reinforcement learning algorithms: A real-time comparative analysis. *Journal of Computing Theories and Applications*,

Corresponding author: Abdulmalik Yusuf Haruna
ahyusuf@atbu.edu.ng
Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved



2(4), 523-535. https://doi.org/10.62411/jcta.12560

Hammad, A. A., Ahmed, S. R., Abdul-Hussein, M. K., Ahmed, M. R., Majeed, D. A., & Algburi, S. (2024). Deep reinforcement learning for adaptive cyber defense in network security. In Proceedings of the Cognitive Models and Artificial Intelligence Conference (pp. 292-297). https://doi.org/10.1145/3660853.3660930

Hammad, A. A., Ahmed, S. R., Abdul-Hussein, M. K., Ahmed, M. R., Majeed, D. A., & Algburi, S. (2024, May). Deep reinforcement learning for adaptive cyber defense in network security. In *Proceedings of the Cognitive Models and Artificial Intelligence Conference* (pp. 292-297).

Hammad, A. A., Falih, M. A., Abd, S. A., & Ahmed, A. R. (2025). Detecting cyber threats in IoT networks: A machine learning approach. International Journal of Computing and Digital Systems, 17(1), 1-25. https://doi.org/10.12785/ijcds/1571020041

Hossain, M. A. (2025). Deep Q-learning intrusion detection system (DQ-IDS): A novel reinforcement learning approach for adaptive and self-learning cybersecurity. *ICT Express*. 10.1016/j.icte.2025.05.007

Htwe, C. S., Myint, Z. T. T., & Thant, Y. M. (2024). IoT security using machine learning methods with features correlation. Journal of Computing Theories and Applications, 2(2), 151-163. https://doi.org/10.62411/jcta.11179

Hussan, P. H., & Mangj, S. M. (2025). BERTPHIURL: A teacher-student learning approach using DistilRoBERTa and RoBERTa for detecting phishing cyber URLs. Journal of Future Artificial Intelligence and Technologies, 1(4). https://doi.org/10.62411/faith.3048-3719-71

Khamaneh, S. M., Wu, T., Sun, W., & Chen, C. (2026). Real-Time Defense Against Coordinated Cyber-Physical Attacks: A Robust Constrained Reinforcement Learning Approach. *IEEE Transactions on Power Systems*. 10.1109/TPWRS.2026.3724022

Khlif, N., Khraief, N., & Belghith, S. (2024). Comparative analysis of modified Q-learning and DQN for autonomous robot navigation. Journal of Future Artificial Intelligence and Technologies, 1(3), 296-308. https://doi.org/10.62411/faith.3048-3719-49

Klein, T., & Romano, G. (2025). Optimizing cybersecurity incident response via adaptive reinforcement learning. *Journal of Advances in Engineering and Technology*, *2*(1).

Kumar, A., & Singh, D. (2026). Reinforcement learning based security method for IoT-Edge computing. *Wireless Networks*, 1-21. https://doi.org/10.1007/s11276-026-04123-5

Kurt, M. N., Ogundijo, O., Li, C., et al. (2019). Online cyber-attack detection in smart grid: A reinforcement learning approach. IEEE Transactions on Smart Grid, 10(5), 5174-5185.

Miles, I., Farmer, S., Foster, D., Harrold, D., Palmer, G., Parry, C., ... & Heartfield, R. (2024). Reinforcement Learning for Autonomous Resilient Cyber Defence. *Black Hat USA*.

Mudgerikar, A., Bertino, E., Lobo, J., & Verma, D. (2021, June). A security-constrained reinforcement learning framework for software defined networks. In *ICC 2021-IEEE International Conference on Communications* (pp. 1-6). Montreal, QC, Canada. 10.1109/ICC42927.2021.9500763

Nguyen, T. T., & Reddi, V. J. (2021). Deep reinforcement learning for cyber security. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(8), 3779-3795. 10.1109/TNNLS.2021.3121870

Ntayagabiri, J. P., Bentaleb, Y., Ndikumagenge, J., & El Makhtoum, H. (2025). OMIC: A bagging-based ensemble learning framework for large-scale IoT intrusion detection. Journal of Future Artificial Intelligence and Technologies, 1(4), 401-416. https://doi.org/10.62411/faith.3048-3719-63

Nugroho, S., Setiadi, D. R. I. M., & Islam, H. M. M. (2024). Exploring DQN-based

Corresponding author: Abdulmalik Yusuf Haruna
ahyusuf@atbu.edu.ng
Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved



reinforcement learning in autonomous highway navigation performance under high-traffic conditions. Journal of Computing Theories and Applications, 1(3), 274-286. https://doi.org/10.62411/jcta.9929

Othman, A. A., Al-khazraji, S., Qadir, M. S., & Almomani, O. (2026). Secure Resource Allocation in IoT Environments Using Deep Reinforcement Learning-Based Adaptive Defenses. *Journal of Soft Computing and Data Mining*, *7*(2), 144-162.

Pandey, A., Bepari, K., Tiwari, R., Prakash, V., Singh, J., Singh, R., & Singh, H. (2025). Self-learning AI agents for adaptive cyber defense in Internet of Things ecosystems. *Networks*, *53*(54), 56.

Pathirana, A., et al. (2024). A reinforcement learning-based approach for promoting mental health using multimodal emotion recognition. Journal of Future Artificial Intelligence and Technologies, 1(2), 124-142. https://doi.org/10.62411/faith.2024-22

Rahman, M. A., Francia, G. A., & Shahriar, H. (2025). Leveraging GANs for synthetic data generation to improve intrusion detection systems. Journal of Future Artificial Intelligence and Technologies, 1(4), 429-439. https://doi.org/10.62411/faith.3048-3719-52

Setiawan, M. A., Setiadi, D. R. I. M., Astuti, E. Z., Sutojo, T., & Setiyanto, N. A. (2024). Exploring deep Q-network for autonomous driving simulation across different driving modes. Journal of Future Artificial Intelligence and Technologies, 1(3), 217-227. https://doi.org/10.62411/faith.3048-3719-31

Sewak, M., Sahay, S. K., & Rathore, H. (2023). Deep reinforcement learning in the advanced cybersecurity threat detection and protection. *Information Systems Frontiers*, *25*(2), 589-611. https://doi.org/10.1007/s10796-022-10333-x

Smith, O. J. M., Kantor, K. N., Zaky, A. A., & Freire, G. F. (2026). Risk-Constrained Reinforcement Learning for Online Detection of Previously Unseen Network Attacks. *Electronics, Communications, and Computing Summit*, 34-44. https://eccsubmit.com/index.php/congress/article/view/252

Subalaxmi, T., Akalya, R., Kaviya, R., Santhosh, V., & Praveen, R. (2024, April). Enhancing cyber-attack detection in iot networks through deep reinforcement learning. In *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)* (Vol. 1, pp. 1-6). Chikkaballapur, India. 10.1109/ICKECS61492.2024.10616565

Wang, Z., Chen, J., Wei, Z., Lin, L., & Tong, G. (2026). SafeCodeRL: Security-Constrained Multi-Agent Reinforcement Learning for Trustworthy LLM-Generated IoT/CPS Software. *Sensors*, *26*(11), 3502. https://doi.org/10.3390/s26113502

Wiebe, J., Mallah, R. A., & Li, L. (2023). Learning cyber defence tactics from scratch with multi-agent reinforcement learning. *arXiv preprint arXiv:2310.05939*. https://doi.org/10.48550/arXiv.2310.05939

Zhu, M., Hu, Z., & Liu, P. (2014). Reinforcement learning algorithms for adaptive cyber defense against Heartbleed. In Proceedings of the First ACM Workshop on Moving Target Defense (pp. 51-58). Association for Computing Machinery.

Corresponding author: Abdulmalik Yusuf Haruna
ahyusuf@atbu.edu.ng
Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi State. Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved