Context-Aware Multi-Agent Reinforcement Learning for Cyber Défense: Problem Formulation and Literature Review

Anuj Kumar[1], Deepak Kumar Verma[2],
[1] Lincoln University College, Malaysia; [2] Marwadi University Rajkot, Gujarat, India
pdf.anuj@lincoln.edu.my, pdfsv.deepak@lincoln.edu.my

Abstract: Computer networks face attacks that change quickly and come from many directions at once, and a single automated defender often cannot keep up. Multi-agent reinforcement learning (MARL) offers a way to place several learning defenders across a network so that they can act at the same time and still work as a team. However, most existing MARL cyber-defense agents make decisions from raw, low-level observations and do not explicitly use "context" information such as how valuable an asset is, how much traffic looks abnormal right now, or what stage an attack has reached. This paper looks closely at this gap. We give a clear, formal statement of the cyber-defense problem as a context-aware, decentralized, partially observable Markov decision process (Dec-POMDP), and we review the research that has been done on multi-agent reinforcement learning, cyber-defense simulators, and context-aware learning. We bring these three areas together, point out where they overlap and where they do not, and summarize the gap in a simple comparison table. We also list the main open problems, such as handling constantly changing attacker behavior, scaling to large networks, and making the learned policies easy to trust and explain. The paper does not present a solution or run experiments; instead, it lays the groundwork a precise problem definition and an organized picture of prior work for building context-aware MARL defenders in future studies.
Keywords: multi-agent reinforcement learning; cyber defense; context-aware reinforcement learning; Dec-POMDP; autonomous cyber operations; network security; game theory.

Introduction

Cyberattacks keep growing in number and in cleverness. Attackers often move step by step through a network, hiding their tracks and changing tactics as they go [5]. Security teams try to keep up using firewalls, rule-based intrusion detection, and human analysts, but these tools are slow to adapt and expensive to run at large scale [6]. Reinforcement learning (RL) offers an appealing alternative: instead of following fixed rules, an RL agent learns, through repeated trial and error, which defensive action works best in which situation. Because a real enterprise network usually needs more than one defender one per subnet, host group, or service researchers have turned to multi-agent reinforcement learning (MARL), where several learning agents act together, sometimes cooperating and sometimes only partly aware of what the others are doing [8], [9].

Public research testbeds such as the Cyber Operations Research Gym (CybORG) and the associated Cyber Autonomy Gym for Experimentation (CAGE) challenges have made it possible to train and compare defensive agents on realistic, simulated enterprise networks [1]-[4]. The most recent version, CAGE Challenge 4, places several defender agents on different subnets of a large network, each of which must cooperate with the others while only seeing part of the overall picture [1]. Building on top

of MARL foundations such as value-decomposition and actor-critic methods [17]-[20], these challenges have shown that MARL defenders can learn useful behavior, but they have also shown that current agents mostly react to whatever a fixed observation vector gives them, without explicitly reasoning about the wider situation they are in.

Problem Statement

The situation around a network change constantly: the value of the assets under attack, the current alert level, the time of day, the known behavior of a particular attacker group, and the topology of the network itself. Humans call this surrounding information "context," and human analysts use it naturally they treat an alert on a finance server very differently from the same alert on a test machine. Most MARL cyber-defense agents, however, do not receive this context in an explicit, structured form. They instead learn from raw or lightly processed observations, and any notion of context has to be re-learned implicitly and separately by every agent, which wastes data and makes coordination harder. At the same time, the general machine-learning field has produced formal tools for context-aware reinforcement learning that make context an explicit, first-class part of the decision process [26] but these tools have mostly been tested on robotics and control tasks, not on adversarial, multi-agent, cyber-defense problems. This leaves an open and, to our knowledge, under-explored question: how should context be formally added to a multi-agent cyber-defense problem, and what does the existing literature already tell us about how to do this well?

Research Objectives

This paper has three objectives. First, to give a precise, formal problem statement for context-aware multi-agent reinforcement learning in cyber defense, expressed as a context-augmented Dec-POMDP. Second, to review, organize, and compare the relevant literature across three areas that are usually studied separately: multi-agent reinforcement learning algorithms, cyber-defense simulation environments and applications, and context-aware reinforcement learning. Third, to identify, with the help of a simple comparison table and figures, exactly where these three areas overlap and where a genuine research gap remains, so that future work has a clear and well-supported starting point.

Contribution Statement

This paper contributes.

- A formal, context-augmented Dec-POMDP formulation for multi-agent cyber defense, written in plain terms and illustrated with two figures;
- An organized review of MARL algorithms, cyber-defense testbeds, and context-aware reinforcement learning, covering more than thirty verified, recent sources;
- Ataxonomy figure and a qualitative comparison figure that classify existing approaches by algorithm family and by practical properties such as scalability and context-adaptability;
- A gap-analysis table and Venn-style figure that make explicit why context-aware MARL for cyber defense remains an open problem, together with a discussion of the concrete challenges a future solution would need to address.

Paper Organization

The rest of the paper is organized as follows. Section II gives the formal problem formulation, including the context-augmented Dec-POMDP model and the main design challenges. Section III reviews the literature in four parts: MARL algorithms, cyber-defense testbeds and applications, context-aware reinforcement learning, and game-theoretic / moving-target-defense approaches, followed by a comparative synthesis and the research-gap analysis. Section IV discusses open challenges and implications. Section V concludes the paper and outlines future work.

PROBLEM FORMULATION

Consider an enterprise network made of several subnets, each containing several hosts and services. A team of N defender agents is deployed, typically one agent per subnet or per class of asset, and an adversary (the attacker) tries to compromise hosts, move across the network, and reach high-value targets. At every discrete time step, each defender agent receives a local, partial view of the network for example, alerts, traffic statistics, or the status of the hosts it is responsible for and must choose a defensive action, such as isolating a host, restoring a service, blocking traffic, or doing nothing. This setting is naturally modeled as a decentralized, partially observable Markov decision process (Dec-POMDP) [16], a well-established formal tool for cooperative multi-agent decision-making under uncertainty. In a standard Dec-POMDP, agents share a global state and a team reward but each only sees a local observation; solving the model well is known to be computationally very hard in general [16].

What is usually missing from this standard picture, when it is applied to cyber defense, is an explicit representation of context: information that is not part of the moment-to-moment state transition itself, but that changes what a "good" action means. For instance, the same suspicious login attempt calls for a stronger response on a financial database than on a disposable test server, and the right level of caution changes when threat intelligence reports an active campaign against the organization's sector. We therefore extend the plain Dec-POMDP with an explicit context variable, following the general idea of contextual reinforcement learning [12], in which context is any information that shapes the environment's dynamics or reward but is not itself controlled by the agents.\

Context-Aware Dec-POMDP Formulation

We formalize the problem as a tuple:

$$M = \{ N, S, C, (A_{\mathrm{i}}\}, \{O_{\mathrm{i}}), P, r, \gamma \}$$

where $N = \{1, \ldots, n\}$ is the set of defender agents; S is the set of network states (host status, service availability, compromise level of each asset); C is the context space, covering variables such as asset criticality, current threat level, time-of-day, and recent alert history; A_i is the action set available to agent i (e.g., monitor, patch, isolate, restore); O_i is the local observation set of agent i, drawn from a context-dependent observation function; $P(s' \mid s, a, c)$ is the state-transition function, which is allowed to depend on the context $c \in C$ as well as on the current state s and the joint action $a = (a_1, \ldots, a_{\mathrm{n}})$; $r(s, a, c)$ is a shared team reward, typically combining a security term (e.g., negative reward for compromised high-value assets, weighted by context) and a cost term (e.g., negative reward for unnecessary disruption); and $\gamma \in [0, 1)$ is the discount factor. Each agent follows a context-conditioned policy $\pi_{\mathrm{i}}(a_{\mathrm{i}} \mid o_{\mathrm{i}}, c)$, so that, unlike in a plain Dec-POMDP, the same local observation can lead to different actions when the surrounding context changes. Fig. 1 shows how such a system fits together in practice, and Fig. 2 shows the resulting interaction loop between context, state, agents, and reward.

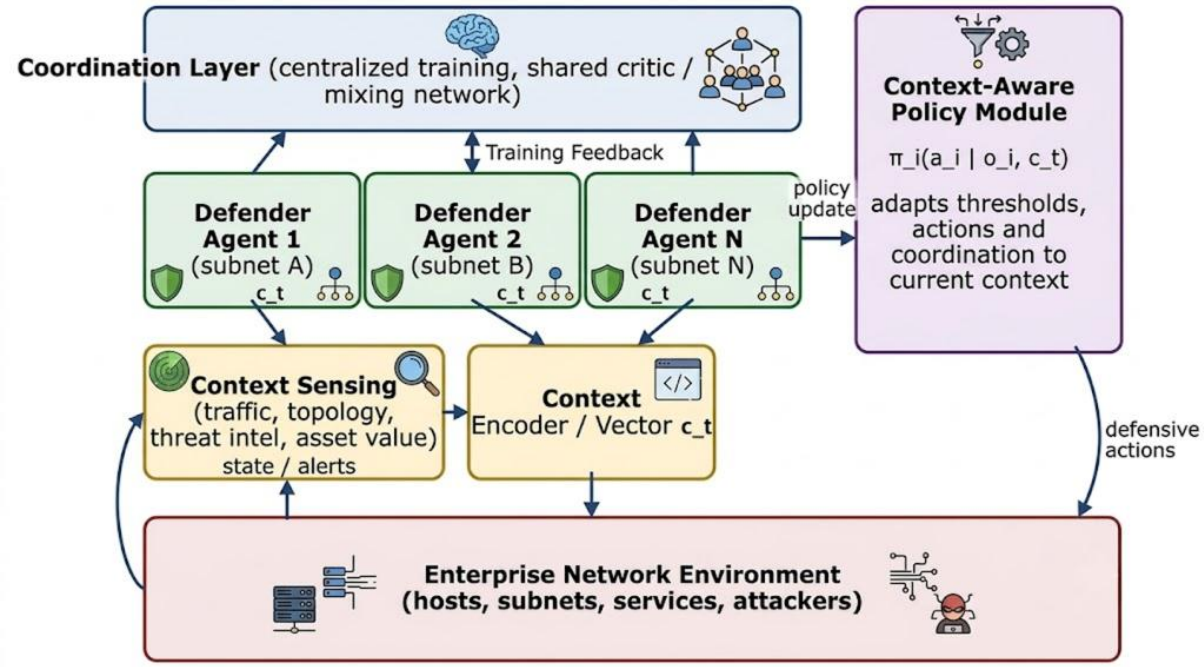


Fig. 1. *A context-aware MARL cyber-defense architecture. Context is sensed from the network, encoded into a vector c_t, and passed to every defender agent alongside its local observation, so that coordinated, context-conditioned actions are sent back to the network*

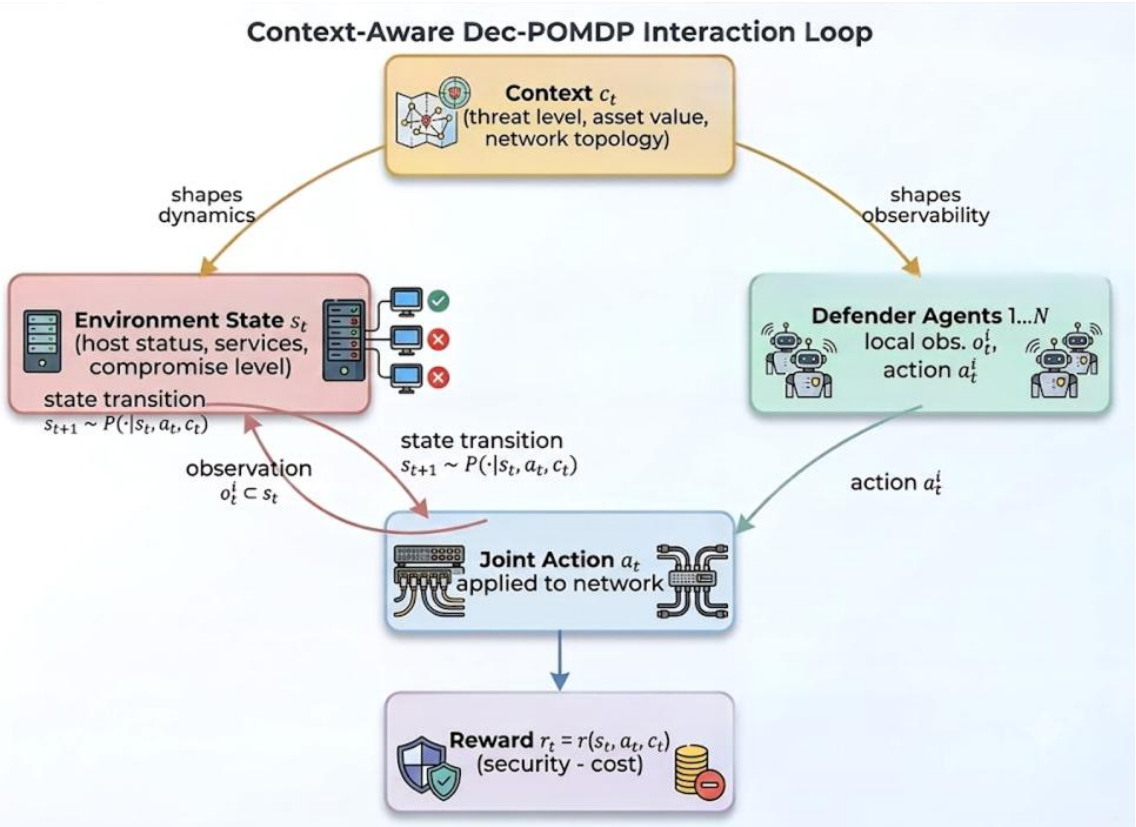


Fig. 2. *The context-aware Dec-POMDP interaction loop. Context c_t shapes both the state dynamics and what each agent can observe, while the joint action feeds back into the next state and the shared reward.*

Design Challenges

Formulating the problem this way immediately surfaces several concrete design challenges that any solution must confront. First, non-stationarity: because every agent is learning and changing its policy at the same time, the environment as seen by any one agent keeps shifting, which is a well-documented difficulty in MARL generally [24] and becomes worse in cyber defense because the attacker is also adapting. Second, partial and noisy observability: real monitoring data contains false positives and false negatives, and no single defender sees the whole network [4], [32]. Third, credit assignment: when a team reward changes, it is not obvious which agent's action was actually responsible, a problem that value-decomposition and counterfactual methods were designed to address [18]-[20]. Fourth, context representation: deciding what should count as context, how to measure or estimate it when it is not directly observable, and how to avoid simply hiding the entire problem inside an oversized context vector. Fifth, scalability: enterprise networks can have thousands of hosts, and the size of the joint action space grows quickly with the number of agents [1], [11]. Section III reviews how far existing work has gone toward solving each of these challenges, and Section III-G summarizes exactly which combinations remain unaddressed.

LITERATURE REVIEW

Foundational Multi-Agent Reinforcement Learning Algorithms

Modern MARL builds on a small number of foundational algorithm families, summarized in Fig. 3. Value-decomposition methods, such as Value-Decomposition Networks (VDN) [18] and QMIX [19], let a team of agents learn a single joint action-value function that is built out of each agent's individual value function, so that agents can be trained together but act independently afterward an approach known as centralized training with decentralized execution (CTDE). VDN simply adds the individual values together, while QMIX allows a richer, monotonic combination, which tends to work better on harder coordination tasks such as the StarCraft II multi-agent benchmark [25]. Policy-gradient and actor-critic methods take a different route: Multi-Agent Deep Deterministic Policy Gradient (MADDPG) [17] gives every agent a centralized critic that sees everything during training, while Counterfactual Multi-Agent Policy Gradients (COMA) [20] tries to solve the credit-assignment problem directly by asking what would have happened if only one agent had acted differently. General surveys of the field [21]-[23], along with foundational reinforcement-learning texts [15], describe these methods in more depth and catalog the many variants that have followed. A recurring and still unresolved theme across all of these surveys is non-stationarity: because every agent is learning at once, the guarantees that make single-agent RL algorithms converge do not automatically carry over to the multi-agent case [24].

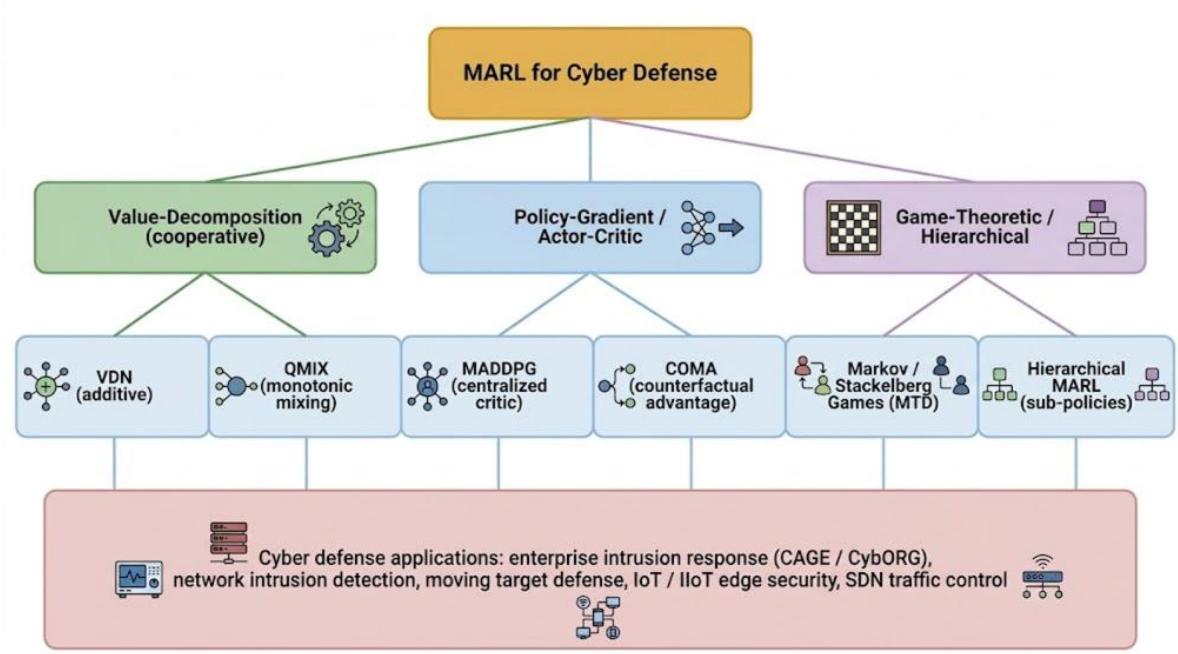


Fig. 3. *A simple taxonomy of MARL algorithm families used in cyber-defense research, grouped into value-decomposition, policy-gradient/actor-critic, and game-theoretic/hierarchical approaches, together with representative cyber-defense applications.*

MARL Testbeds and Applications in Cyber Defense

Progress in applying MARL to cyber defense has depended heavily on the availability of realistic, shared simulation environments. CybORG, the Cyber Operations Research Gym, provides a simulated enterprise network in which a defending ("blue") agent must respond to an attacking ("red") agent [3]. Building on CybORG, the CAGE challenge series has progressively increased in scale and realism: earlier CAGE challenges modeled a single defender against a single attacker [4], [5], while CAGE Challenge 4 introduces a genuinely multi-agent scenario in which several blue agents, each responsible for one subnet of a large enterprise network, must defend against multiple attackers while only partially observing the network and while their actions can help or hurt each other [1], [2]. Within this environment, hierarchical MARL approaches decompose the overall defense policy into reusable sub-policies that can be fine-tuned and transferred when the attacker's behavior or the network configuration changes [11]. Other CAGE-based studies have explored learning defensive tactics from scratch with multi-agent reinforcement learning [13], competitive reinforcement learning against multiple attacker strategies [14], and quantitative ways of measuring how resilient a learned policy is to unseen attacks [12]. Beyond CAGE, a large body of work applies deep reinforcement learning directly to network intrusion detection, treating detection thresholds and response actions as the agent's action space [32], [34], [35], including recent applications to software-defined networking [32] and IoT/IIoT edge devices where federated and privacy-preserving variants of reinforcement learning are increasingly used to avoid moving sensitive data off-device [33]. Several recent survey papers have started to map this whole space specifically from a multi-agent point of view [8]-[10], [22], and an emerging reference architecture for autonomous cyber-defense agents attempts to standardize how such agents should be structured and evaluated [7]. Fig. 4 places these milestones on a timeline to

show how the field has moved from single-agent, algorithm-focused work toward larger, more realistic multi-agent testbeds.

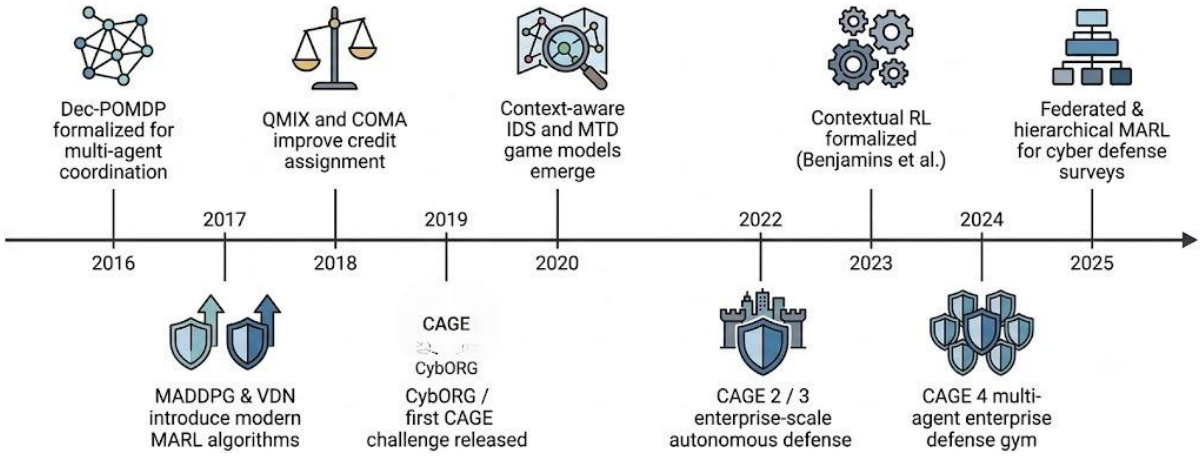


Fig. 4. *A timeline of developments feeding into context-aware MARL cyber defense, from early Dec-POMDP and MARL algorithm work through the CAGE challenge series to recent context-aware and federated MARL surveys.*

Context-Awareness in Reinforcement Learning

Outside of cyber security, the reinforcement-learning community has developed a formal notion of context: measurable information about the task or environment that is not part of the agent's usual state but that changes how the environment behaves or how the agent should act. Benjamins et al. formalize this as contextual reinforcement learning (cRL), in which the context is made an explicit, separate input to the policy rather than something the agent must infer purely from its raw observations [26]. This idea has been applied to robotics, autonomous racing, and multi-task control, where context might represent a robot's payload, surface friction, or a task identifier, and it is either directly available from sensors or must be estimated. Closer to security, Sethi et al. propose a context-aware intrusion detection system that uses reinforcement learning and explicitly incorporates contextual features about the network into its decision process, showing measurable gains over context-blind baselines [27]. However, this line of work is single-agent: it does not address how several defenders, spread across a network and only partly aware of each other, should share and act on a common notion of context. This is precisely the combination multiple agents, explicit context, and a realistic adversarial cyber environment that current research has not yet brought together, as illustrated in Fig. 6 and Table I.

Game-Theoretic and Moving-Target-Defense Approaches

A separate but related line of research models cyber defense as a game between an attacker and a defender rather than as a single-sided learning problem. Moving Target Defense (MTD) methods deliberately and continually change a system's configuration for example, its network addresses or open ports to confuse an attacker, and researchers have combined MTD with Markov games and Bayesian Stackelberg games, solved using multi-agent reinforcement learning, to decide when and how to reconfigure the system [28], [29]. These game-theoretic formulations naturally include a notion of the current security "situation" for instance, an attacker's inferred type or the belief a defender holds about ongoing reconnaissance which is philosophically close to context, but it is normally folded into the game's state rather than represented as a separate, explicit context variable

that could be reused across different networks or attack types [30], [31]. This body of work confirms that game-theoretic reasoning is valuable for cyber defense, but it also shows that existing formulations have not adopted the more general, explicit context-conditioning framework used in the contextual-RL literature [12].

Comparative Synthesis

Bringing the three areas together, Fig. 5 gives a simple, qualitative comparison of the three main MARL algorithm families discussed in Section III-A, rated on five properties that matter specifically for cyber defense: scalability to many agents and hosts, sample efficiency, robustness to the non-stationarity created by co-learning agents (and, in security, by an adapting attacker), context-adaptability, and interpretability for a human security analyst. This comparison is a qualitative synthesis drawn from the discussion in the reviewed papers, not a measured experimental result, and it is intended only to guide where future context-aware methods might be built. Value-decomposition methods tend to be efficient and reasonably scalable but were not designed with context-adaptability in mind. Policy-gradient and actor-critic methods handle continuous and mixed action spaces more naturally but can be harder to scale and just as context-blind. Game-theoretic and hierarchical methods tend to be more robust to an adapting adversary and easier for an analyst to interpret, because their structure keeps some notion of attacker/defender roles, but they are usually less sample-efficient and have not been combined with the general contextual-RL machinery described in Section III-C

The introduction should set the tone of the article and provide the reader with a good understanding of the problem statement. Please keep in mind that the problem statement should match the Abstract's motivation/problem statement[1].

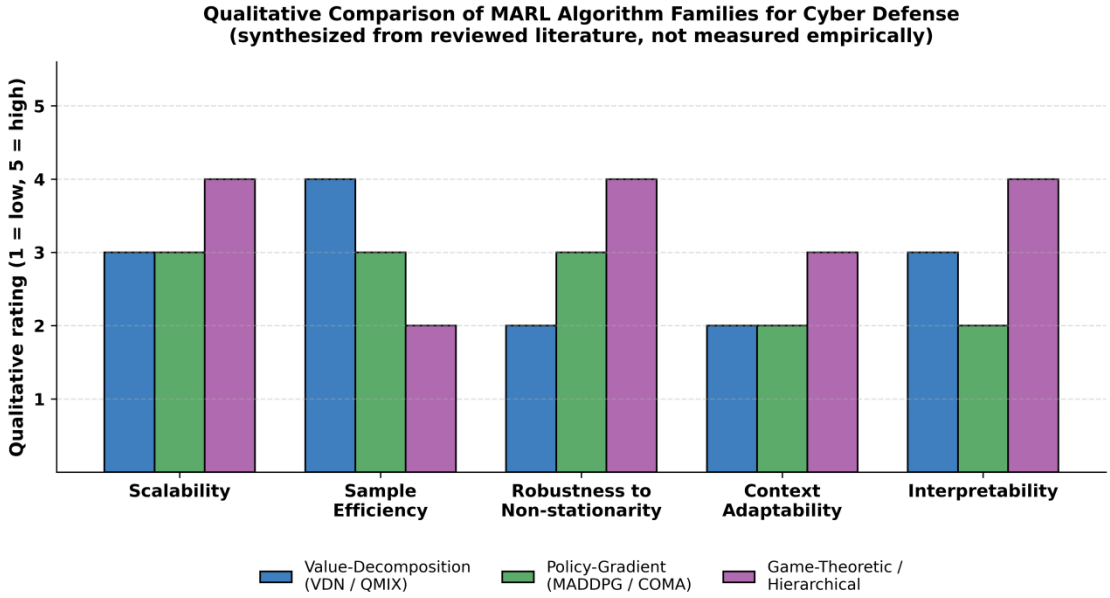


Fig. 5. *Qualitative comparison of the three MARL algorithm families discussed in this review, across five properties relevant to cyber defense. Ratings are a synthesis of the discussion in the reviewed literature and are not experimental measurements.*

Research Gap Identification

Table 1 summarizes the comparison directly: it lists representative lines of work from each of the three areas reviewed above and checks whether each one is multi-agent, whether it explicitly represents context, and whether it has been tested on a realistic cyber-defense testbed. No existing line of work

satisfies all three properties at once. CAGE/CybORG-based MARL studies are multi-agent and use realistic testbeds, but their agents do not receive an explicit context vector. Context-aware intrusion detection work explicitly represents context, but only for a single agent and mostly on offline datasets rather than an interactive, adversarial testbed. Contextual-RL theory formalizes context cleanly but has not been evaluated in adversarial, multi-agent settings at all. Game-theoretic MTD work is multi-agent and partly contextual through the game state, but it does not use the general, explicit, reusable context representation of contextual RL, and its testbeds are typically abstract rather than full enterprise-network simulations. Federated and IoT-security reinforcement learning is distributed and realistic, but its main focus is privacy and communication efficiency rather than context-conditioned coordination. Fig. 6 shows the same idea visually, as the overlap of three circles: multi-agent RL, context-awareness, and cyber defense. The triple overlap context-aware MARL evaluated on realistic, adversarial cyber-defense environments is the gap this review identifies as the central open opportunity for future work.

Table 1. Research Gap Summary Across Reviewed Lines of Work

Study / Line of Work	**Multi-Agent?**	**Explicit Context-Awareness?**	**Realistic Cyber Testbed?**	**Key Limitation**
CAGE / CybORG MARL studies [1]-[4], [11]-[14]	Yes	No (fixed observation, no explicit context vector)	Yes	Agents do not condition policy on threat level or asset value explicitly
Context-aware IDS [27]	No (single agent)	Yes	Partial (dataset-based)	Single-agent only; does not address multi-agent coordination
Contextual RL theory [26]	No (single agent)	Yes (formal cRL)	No (general benchmarks)	Not evaluated in adversarial, multi-agent cyber settings
MTD game-theoretic MARL [28]-[31]	Yes	Partial (game state, not explicit context vector)	Partial (abstract models)	Context limited to attacker/defender belief, not full situational context
Federated / IoT security RL [32], [33]	Partial (distributed agents)	No	Yes (IoT traffic)	Focuses on privacy and distribution, not context-conditioned coordination

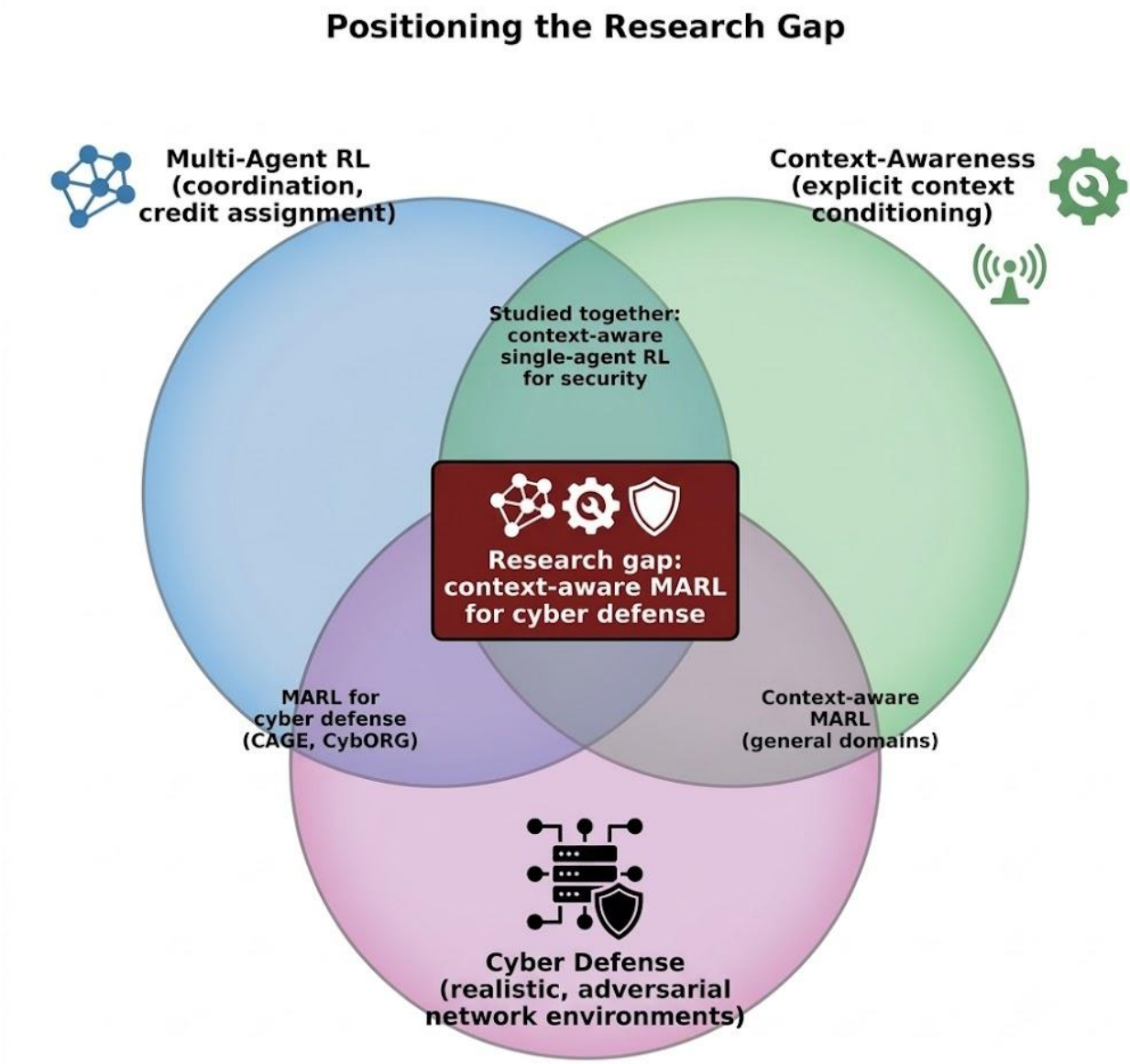


Fig. 6. *Positioning the research gap: context-aware MARL for cyber defense sits at the unstudied intersection of multi-agent reinforcement learning, context-awareness, and realistic adversarial cyber-defense environments.*

DISCUSSION

Open Challenges

Closing the gap identified in Section III-F raises several concrete open questions. The first is context definition and measurement: cyber-defense context includes things that are easy to log, such as time or asset labels, but also things that are hard to measure directly, such as an attacker's intent or an organization's true current risk appetite, and it is not yet clear which of these should be modeled explicitly versus inferred. The second is context sharing under partial observability: if each agent senses only local context, agents need a principled way to combine or communicate their local views into a shared picture without overwhelming limited bandwidth, building on existing MARL communication research [5]. The third is robustness to an adapting attacker: unlike the benchmark tasks used in most contextual-RL papers, in cyber defense the "environment" includes an adversary

who may deliberately try to manipulate the very context signals a defender relies on, for example by generating misleading traffic patterns; this connects to adversarial-robustness concerns already raised in the MARL cyber-defence literature [6], [11]. The fourth is scalable evaluation: testing context-aware MARL defenders convincingly will require testbeds at least as realistic as CAGE 4 [1], but extended so that context variables (asset value, threat intelligence, time-of-day patterns) are built into the simulator rather than added afterward. The fifth is interpretability and trust: because cyber-defense decisions can have serious consequences, security teams are unlikely to adopt a context-aware MARL system unless it can explain, in terms a human analyst understands, why a given context led to a given action.

Implications

If these challenges are addressed, context-aware MARL could let a defense system automatically adjust how cautious or aggressive it is, per asset and per moment, without a human having to hand-write separate rules for every situation. This could shorten the time between an intrusion and an effective response, and it could let smaller security teams manage larger networks by letting the learned agents absorb routine, context-dependent judgment calls that currently require a human in the loop. At the same time, because the systems reviewed here are trained through trial and error, deploying them in real networks will require careful, staged testing first in simulation, then in monitored pilot deployments to avoid situations where a poorly generalized context representation causes the system to under-react to a genuinely serious threat or over-react to a harmless one.

CONCLUSION AND FUTURE WORK

This paper set out to give a clear problem formulation and an organized literature review for context-aware multi-agent reinforcement learning in cyber defense. We modeled the problem as a context-augmented Dec-POMDP, in which an explicit context variable shapes both the network's dynamics and the reward that defender agents receive, and we illustrated this model with two figures showing the system architecture and the interaction loop. We then reviewed more than thirty recent, verified sources across four connected areas foundational MARL algorithms, cyber-defense testbeds and applications, context-aware reinforcement learning, and game-theoretic moving-target-defense methods and used a taxonomy figure, a timeline figure, a qualitative comparison figure, a gap-summary table, and a Venn-style figure to show exactly where these areas overlap and where they do not. The central finding is that, while multi-agent cyber defense and context-aware reinforcement learning have each matured substantially on their own, no existing line of work has combined explicit, reusable context-conditioning with multi-agent coordination inside a realistic, adversarial cyber-defense testbed. Future work should focus on building such context-aware MARL agents and testing them on extended versions of existing testbeds such as CAGE, on developing lightweight ways for agents to share partial context under bandwidth and observability limits, and on measuring how well these agents hold up against attackers who might try to manipulate the context signals themselves. We hope the formulation and gap analysis given here make that next step easier to plan and to evaluate.

REFERENCES

[1] M. Kiely, D. Bowman, M. Standen, and C. Moir, "CAGE challenge 4: A scalable multi-agent reinforcement learning gym for autonomous cyber defence," AI Magazine, Wiley, 2025.

[2] TTCP CAGE Working Group, "TTCP CAGE Challenge 4," 2024. [Online]. Available: https://github.com/cage-challenge/cage-challenge-4

[3] M. Standen, M. Lucas, D. Bowman, T. J. Richer, J. Kim, and D. Marriott, "CybORG: A gym for the development of autonomous cyber agents," in Proc. IJCAI-21 1st Int. Workshop on Adaptive Cyber Defense, 2021.

[4] M. Kiely, D. Bowman, M. Standen, and C. Moir, "On autonomous agents in a cyber defence environment," arXiv:2309.07388, 2023.

[5] S. Vyas, J. Hannay, A. Bolton, and P. Burnap, "Automated cyber defence: A review," arXiv:2303.04926, 2023.

[6] G. Palmer, C. Parry, D. J. B. Harrold, and C. Willis, "Deep reinforcement learning for autonomous cyber defence: A survey," arXiv:2310.07745, 2024.

[7] A. Kott et al., "Autonomous intelligent cyber-defense agent (AICA) reference architecture, release 2.0," arXiv:1803.10664, 2023.

[8] C. R. Landolt, C. Würsch, R. Meier, A. Mermoud, and J. Jang-Jaccard, "Multi-agent reinforcement learning in cybersecurity: From fundamentals to applications," arXiv:2505.19837, 2025.

[9] S. Finistrella, S. Mariani, and F. Zambonelli, “Multi-agent reinforcement learning for cybersecurity: Classification and survey,” Intelligent Systems with Applications, vol. 26, Art. no. 200495, 2025.

[10] S. Finistrella, S. Mariani, and F. Zambonelli, "Multi-agent reinforcement learning for cybersecurity: Approaches and challenges," in Proc. WOA 2024 – 25th Workshop "From Objects to Agents," Forte di Bard, Italy, Jul. 2024.

[11] A. V. Singh, E. Rathbun, E. Graham, L. Oakley, S. Boboila, P. Chin, and A. Oprea, “Hierarchical multi-agent reinforcement learning for cyber network defense,” Reinforcement Learning Journal, vol. 6, pp. 790–810, 2025.

[12] X. Cadet, S. Boboila, E. Koh, P. Chin, and A. Oprea, “Quantitative resilience modeling for autonomous cyber defense,” Reinforcement Learning Journal, vol. 6, pp. 894–908, 2025.

[13] J. Wiebe, R. Al Mallah, and L. Li, "Learning cyber defence tactics from scratch with multi-agent reinforcement learning," in Proc. 2nd Int. Workshop on Autonomous Cyber Defense, arXiv:2310.05939, 2023.

[14] G. McDonald, L. Li, and R. Al Mallah, “Finding the optimal security policies for autonomous cyber operations with competitive reinforcement learning,” IEEE Access, vol. 12, pp. 120292–120305, 2024.

[15] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[16] F. A. Oliehoek and C. Amato, A Concise Introduction to Decentralized POMDPs. Cham, Switzerland: Springer, 2016.

[17] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in Proc. NeurIPS, 2017.

[18] P. Sunehag et al., "Value-decomposition networks for cooperative multi-agent learning," in Proc. AAMAS, 2018.

[19] T. Rashid, M. Samvelyan, C. Schroeder de Witt, G. Farquhar, J. Foerster, and S. Whiteson, "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," in Proc. ICML, 2018.

[20] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in Proc. AAAI, 2018.

[21] S. Gronauer and K. Diepold, "Multi-agent deep reinforcement learning: A survey," Artificial Intelligence Review, vol. 55, no. 2, pp. 895–943, 2022.

[22] D. Huh and P. Mohapatra, "Multi-agent reinforcement learning: A comprehensive survey," arXiv:2312.10256, 2024.

[23] S. V. Albrecht, F. Christianos, and L. Schäfer, Multi-Agent Reinforcement Learning: Foundations and Modern Approaches. Cambridge, MA, USA: MIT Press, 2024.

[24] G. Papoudakis, F. Christianos, A. Rahman, and S. V. Albrecht, "Dealing with non-stationarity in multi-agent deep reinforcement learning," arXiv:1906.04737, 2019.

[25] O. Vinyals et al., "Grandmaster level in StarCraft II using multi-agent reinforcement learning," Nature, vol. 575, pp. 350–354, 2019.

[26] C. Benjamins, T. Eimer, F. Schubert, A. Mohan, S. Döhler, A. Biedenkapp, Bodo Rosenhahn, F. Hutter, and M. Lindauer, "Contextualize me: The case for context in reinforcement learning," Transactions on Machine Learning Research, Mar. 2023.

[27] K. Sethi, E. S. Rupesh, R. Kumar, P. Bera, and Y. V. Madhav, "A context-aware robust intrusion detection system: A reinforcement learning-based approach," International Journal of Information Security, vol. 19, no. 6, pp. 657–678, 2020.

[28] S. Sengupta and S. Kambhampati, "Multi-agent reinforcement learning in Bayesian Stackelberg Markov games for adaptive moving target defense," arXiv:2007.10457, 2020.

[29] T. Eghtesad, Y. Vorobeychik, and A. Laszka, "Adversarial deep reinforcement learning based adaptive moving target defense," in Proc. GameSec, Springer, 2020, pp. 58–79.

[30] S. Sengupta, A. Chowdhary, A. Sabur, A. Alshamrani, D. Huang, and S. Kambhampati, "A survey of moving target defenses for network security," IEEE Communications Surveys & Tutorials, 2020.

[31] M. H. Manshaei, Q. Zhu, T. Alpcan, T. Başar, and J.-P. Hubaux, "Game theory meets network security and privacy," ACM Computing Surveys, vol. 45, no. 3, pp. 25:1–25:39, 2013.

[32] R. Kanimozhi and P. S. Ramesh, "Deep reinforcement learning-based intrusion detection scheme for software-defined networking," Scientific Reports, vol. 15, art. 38827, 2025.

[33] M. Ragab, B. M. Alghamdi, L. A. Maghrabi, K. H. Allehaibi, et al., "Advanced artificial intelligence with federated learning framework for privacy-preserving cyberthreat detection in IoT-assisted sustainable smart cities," Scientific Reports, vol. 15, art. 4470, 2025.

[34] V. Mnih et al., "Human-level control through deep reinforcement learning," Nature, vol. 518, pp. 529–533, 2015.

[35] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv:1707.06347, 2017.