Towards the Deployment of Realistic Autonomous Cyber Network Defence: A Systematic Review

SANYAM VYAS, Cardiff University, United Kingdom
VASILIOS MAVROUDIS, The Alan Turing Institute, United Kingdom
PETE BURNAP, Cardiff University, United Kingdom

In the ongoing network cybersecurity arms race, the defenders face a significant disadvantage as they must detect and counteract every attack. Conversely, the attacker only needs to succeed once to achieve their goal. To balance the odds, Autonomous Cyber Network Defence (ACND) employs autonomous agents for proactive and intelligent cyber-attack response. This article surveys the state of the art of Autonomous Blue and Red Teaming agents, as well as cyber operations environments. We begin by presenting a detailed set of criteria for ACND algorithms and systems that evaluate the preparedness of integrating autonomous agents into real-world networked environments. Our analysis identifies critical research gaps and challenges within the ACND landscape, including issues of autonomous agent explainability, continuous learning capability under evolving threats, and the development of realistic agent training environments. Based on these insights, we discuss promising research directions and open challenges that need to be addressed for the deployment of ACND agents in real-world networks.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; • **Security and privacy** → **Network security**.

Additional Key Words and Phrases: autonomous cyber network defence, reinforcement learning, autonomous response, network security

ACM Reference Format:
Sanyam Vyas, Vasilios Mavroudis, and Pete Burnap. 2025. Towards the Deployment of Realistic Autonomous Cyber Network Defence: A Systematic Review. *ACM Comput. Surv.* 1, 1, Article 1 (January 2025), 36 pages. https://doi.org/10.1145/3729213

1 Introduction

As government and private sectors worldwide shift towards fully digitised systems for essential operations, they become increasingly vulnerable to a range of cyber adversaries, from individual cyber-attackers to organised hostile states. With this worldwide digital transition, coupled with an increasing global deficit in cybersecurity expertise [30], cyber defence mechanisms are becoming increasingly outdated in terms of defending against novel cyber-attacks. Therefore, there exists a need for the incorporation of advanced autonomous defence architectures and techniques [66, 84] within all digital infrastructures such as enterprise networks and operational technology (OT) networks. Despite existing technical publications and white papers addressing autonomous defence solutions, their limitations in countering novel and realistic cyber-attacks call for more structured and streamlined research and development. Such efforts are essential to hasten their deployment

Authors' Contact Information: Sanyam Vyas, vyass3@cardiff.ac.uk, Cardiff University, Cardiff, Wales, United Kingdom; Vasilios Mavroudis, vmavroudis@turing.ac.uk, The Alan Turing Institute, London, , United Kingdom; Pete Burnap, burnapp@cardiff.ac.uk, Cardiff University, Cardiff, Wales, United Kingdom.

ACM 1557-7341/2025/1-ART1
https://doi.org/10.1145/3729213

within real-world digital infrastructures. As a solution, this paper provides a comprehensive systematic literature review on Autonomous Cyber Network Defence (ACND), an area that focuses on autonomous decision-making agents for networked systems to mitigate a plethora of existing and potential cyber-attacks. To provide a holistic understanding of the area, the paper defines ACND and analyses literature within different divisions of the area. The analysis is conducted through an overall ACND Requirements Table 3 (in Section 4.2), which pinpoints potential objectives to accelerate the real-world deployment of such autonomous decision-making agents within networked systems through efficient environment and algorithmic design.

Overall, very few surveys have highlighted the requirement of autonomous decision-making agents for defending against cyber-attacks within networked systems. Recent publications include [63] which provides a detailed review on Reinforcement Learning (RL) solutions for moving target defence, cyber defence and honeypots. The publication also provides the development of RL solutions within cybersecurity through control-theoretic principles. However, the review does not address the development process of networked systems that enable the integration of autonomous decision-making agents for cyber defence and offence. In contrast, this paper presents a detailed development pipeline—from simulation to emulation—through the proposed Autonomous Cyber Operations Gym.. Wang et al. [133] also focus on the development of RL solutions for network defence and attack, along with addressing future challenges similar to [63]. However, their work does not thoroughly analyse the networked systems on which such agents could be developed, which is an imperative part of ACND. The authors of [18] provide a review on Machine Learning (ML) solutions in cybersecurity, specifically focusing on datasets that accelerate research in intrusion detection systems. While the implementations mentioned by this survey form a part of automation approaches in cybersecurity, intrusion detection approaches do not apply to ACND as they do not involve an autonomous decision-making component for network attack or defence. Burke et al.'s white paper [19] provide an in-depth review on the type of potential projects within Active Cyber Defence (AcCD). This was conducted by creating multiple ML-based abstract candidate projects revolving around automated defence and automated attack, and overall automated security planning. While some of them apply to Autonomous Cyber Network Defence mentioned in this paper, the white paper does not focus on the potential of utilising autonomous agent training environments for the transition towards their real-world deployment, nor does it provide a comparison of autonomous decision-making algorithms utilised for autonomous network attack and defence. Yu et al. [137] developed a white paper on Multidisciplinary University Research Initiative (MURI) for Adaptive Cyber Defence, which addressed the importance of adaptation and adversarial techniques which are synonyms to autonomous blue and red teaming respectively. In addition, they developed several high-level frameworks and projects covering topics also discussed by Burke et al.'s white paper, AcCD [19]. However, the paper did not address the need for the development of network simulation and emulation systems that will be pivotal for the transition towards real-world deployment of autonomous decision-making agents. DARPA introduced the Cyber Agents for Security Testing and Learning Environments (CASTLE) program [2], which uses RL agents for autonomous defence (blue team) and assessment (red team) in a realistically simulated and emulated networked system gym. The program is segmented into three specialisations: red team RL, blue team RL, and purple teaming, each undergoing parallel development. The overall project is structured into three phases aimed at the real-world deployment of the RL agents and environments. This program aligns with the high-level motivations outlined in this paper, and we provide a detailed technical requirements analysis within these phases, focusing on the technical development of the three specialisations, allowing for the overall real-world deployment of ACND.

This work identifies the necessity for a structured and streamlined technical pathway for the real-world implementation of robust autonomous defence algorithms within ACND. To facilitate this,

the domain is segmented into two concurrent research and development sub-areas: the development of Autonomous Blue and Red Teaming, and the simultaneous advancement of the Autonomous Cyber Operations Gym. Detailed definitions of these and other relevant terms can be found in Section 2. Overall, the contributions of this paper include:

- Organisation of concepts within the area of Autonomous Cyber Network Defence through a Requirements Table. Highlighting the need of two important areas within Autonomous Cyber Network Defence. Namely, Autonomous Blue and Red Agent development, and the development criteria of Autonomous Cyber Operation Gyms to facilitate Autonomous Cyber Network Defence capabilities.
- Assessment of publications classified as Autonomous Blue Teaming, Autonomous Red Teaming and Autonomous Cyber Operations Gyms within Autonomous Cyber Network Defence literature through the Requirements Table.
- Identification of novel and realistic open problems and their corresponding challenges within the Autonomous Cyber Network Defence literature. Facilitating future streamlined research directions for the eventual real-world deployment of Autonomous Cyber Network Defence agents within networked systems.

This article is organised as follows: We first introduce important terms used frequently in this paper (Section 2). Then, Section 3 addresses the research methodology utilised to find relevant ACND publications. Subsequently, Section 4 elaborates the curated terminology of ACND and its differentiation from similar terminologies used within recent literature. This section then provides the importance of the area in National Strategies. Lastly, the section provides a comprehensive Requirements Table that will be used to evaluate the selected publications recognised to be as part of ACND. Section 5 elaborates and critiques on the autonomous defence and attack (defined as autonomous blue and red team) agents in custom ACO Gyms through the ACND Requirements in Section 4. Section 6 elaborates an exhaustive list open-source and closed-source ACO Gyms and assesses them using the ACND Requirements in Section 4. Section 5 elaborates a list of published autonomous agents within ACO Gyms and evaluates them using the ACND Requirement in Section 4. Section 8 provides a discussion identifying the open research areas and their corresponding challenges within ACND literature using the analysis conducted in the previous sections. Lastly, Section 9 concludes the article by summarising the area of ACND.

2 Key Definitions

This article comprises of several technical terminologies that are commonly used within the fields of cybersecurity and artificial intelligence. This section will define the key terminologies used within this document.

Autonomous Red Teaming: Red Teaming is a technique used within military and industry operations to uncover networked system vulnerabilities or to find exploitable gaps in operational concepts, with the overall goal of reducing surprises, improving and ensuring the robustness of the networked system [25]. In the context of this paper, autonomous red teaming refers to an autonomous agent possessing a set of operations (to uncover vulnerabilities and exploits within the networked system) as their action space. In the context of this paper, the overall aim of autonomous red teaming is to ensure the robustness of the autonomous blue team agent (definition elaborated below) in terms of defending the system against known vulnerabilities and exploits.

Autonomous Blue Teaming: Blue Teaming is a technique responsible for defending a networked system by maintaining its security posture against a set of mock attackers that aim to exploit gaps and vulnerabilities of the networked system. Typically the Blue Team must defend against real or simulated attacks 1) over a significant period of time and 2) in a representative operational

context (e.g., as part of an operational exercise)[1]. In the context of this paper, Autonomous Blue Teaming refers to an autonomous agent possessing a set of operations as their action space to destroy malicious processes from entering the networked system through its nodes/endpoints.
Autonomous Cyber Operations Gym: Autonomous Cyber Operations (ACO) is concerned with the defence of computer systems and networks through autonomous decision-making and action. It is particularly required where the deployment of security experts to cover every network and location is becoming increasingly untenable, and where systems cannot be reliably accessed by human defenders, either due to unreliable communication channels or adversary action. ACO Gyms are networked system environments that facilitate the use of autonomous red and blue teaming agents in order to further strengthen the networked systems of the future from ever-evolving cyber-attacks [121]. ACO Gyms aim to address and reduce the 'reality gap' of potential networked systems, used in [124] by combining learning on simulations with testing in a real environments.
Simulated Network: A Simulated Network is an ACO Gym (or a part of the ACO Gym's training-testing strategy) that is designed as a finite state machine. The creation is usually completed in the form of code that includes objects that correspond to the components, agents and actions within the simulated network. [90]
Emulated Network: An Emulated Network is an ACO Gym (or a part of the ACO Gym's training-testing strategy) that is designed through a group of virtual machines (or a docker container with several network drivers), which are used to create a computer networked system [90].
Sequential Response: Sequential response, or sequential decision-making refers to algorithms that take the dynamics of the world into consideration, thus delaying segments of the problem until it is solved [42]. It is a fundamental task faced by any intelligent agent in an extended interaction with its environment which demands a set of decisions that are concerned with short and long-term decisions in order to reach a state that acts as an overall target within the environment [80]. In the context of this paper, sequential decision-making algorithms are considered in this paper as Autonomous Blue and Red Teaming agents due to the complexity of the network that requires navigation before a target action is taken by the autonomous agent (e.g. launching an exploit in a host within a different subnet).
Single-step Response: Single-step response algorithm refers to decision-making actions that only focus on the short-term outcomes. For example, in temporal context, the algorithm at time *t(n)* will perform calculations solely for a solution at time *t(n + 1)*.

3 Review Methodology

A methodology inspired by [71] was implemented to find all relevant articles for this review. In order to interpret the overall definition of ACND and the research questions for this article, an initial set of white papers [19, 73, 91] from national and international government institutions and organisations (mentioned in section 4.1) were utilised. Backward snowballing [68] was utilised to further find relevant similar and relevant papers. These papers addressed the need for autonomous response solutions in networked systems within a variety of different areas, allowing us to categorise areas where autonomous response could be utilised within the existing areas of ACND terminology, specifically, Autonomous Red and Blue Teaming.

3.1 Research Questions

To harness the concepts proposed for ACND, research questions were developed to establish a search strategy and utilise the scrutinised literature to delineate forthcoming research trajectories and challenges. Addressing the research questions articulated herein will equip future AI and

[1]https://csrc.nist.gov/glossary/term/blue_team

Cyber Security researchers to undertake studies within ACND, pinpointing essential research gaps necessary for publishing significant work. This requires employing the most effective algorithms in optimal ACO Gyms, specifically targeting a research gap identified in this study. Such a strategy promises to significantly streamline research and development efforts within the field of ACND and enhance the broader landscape of cyber security. The research questions (RQs) are as follows:

- **RQ1**: What is the role of Autonomous Cyber Network Defence in the current and projected cyber landscape?
- **RQ2**: What are the most promising algorithmic approaches used in Autonomous Cyber Network Defence?
- **RQ3**: What are the most suitable environments in which better algorithmic approaches could be developed?
- **RQ4**: What future research directions and challenges must be undertaken to enable the real-world deployment of Autonomous Cyber Network Defence solutions?

The first research question (answered in Section 4.1) investigates the importance of ACND and its projected role in safeguarding networks. We will address this question by primarily reviewing government and funding agency strategy documents (highlighted in Table 2), as ACND is currently predominantly supported by state-sourced funding [2].

The next two research questions aim to investigate distinct but complementary aspects of ACND. RQ2 seeks to identify the most promising algorithmic approaches by surveying past works identified as ACND literature (in Section 5), while RQ3 explores the most suitable environments for the development of these advanced algorithmic approaches (in Section 6). It aims to understand where (in terms of infrastructure, technology, and support systems) these algorithms can be best developed and refined to enhance their effectiveness in real-world applications. By addressing these two aspects, researchers can contribute to building more resilient and adaptive systems that are capable of defending against the increasingly sophisticated landscape of cyber threats.

Finally, the fourth research question aims to map out the necessary research paths and challenges (shown in Section 8) that need addressing to enable the effective real-world application of ACND solutions. Deploying such advanced systems in actual operational environments goes beyond theoretical research and development. The outcome of this inquiry will provide a comprehensive blueprint for transitioning ACND from a research and development phase to full-scale operational deployment, thus closing the gap between theoretical possibilities and practical usability in defending against cyber threats.

3.2 Search Terminology Strategy

After identifying the initial set of research questions, the next step involves searching for relevant primary studies. As elaborated in this section, RQ4 was developed only after the initial research questions were answered. In order to optimise our search for relevant papers, popular digital libraries including IEEE, ACM Digital Library, Springer Science Direct and Google Scholar were utilised. A list of strings grouped within 3 overall themes of ACND were collectively identified (shown in Table 1 as themes **a**, **b** and **c**). The strings from all different overall themes are then grouped together in 3 different groups of permutation combinations as an aim to identify publications in digital libraries that:

[2]https://www.thinkdigitalpartners.com/uncategorised/2022/09/29/the-22-billion-future-of-the-uks-cybersecurity-insights-for-suppliers/

- **i**: allow us to explore and rank the performance of identified algorithmic families in Autonomous Blue Teaming (**RQ1, RQ2**).
- **ii**: allow us to explore and rank the performance of identified algorithmic families in Autonomous Red Teaming (**RQ1, RQ2**).
- **iii**: allow us to discover the best possible environments in which the most suitable algorithms could be developed, trained and tested (**RQ1, RQ3**).

3.3 Overall Relevant Content Extraction

Due to the area of ACND gaining popularity only recently, backward snowballing [68] and forward snowballing [134] were conducted for several searches in order to find publications and code repositories relevant to ACND that were not listed in the search strategy. For example, with areas such as "autonomous cyber operations gym" being a recently created terminology within this area, backward snowballing aided us to identify other popular publications (with respective implementations found in code repositories) that were created before this term was officially introduced. In addition, a manual search was conducted to identify the latest ACND related papers (along with papers highlighting further potential areas within the domain) that cited the publications identified through the search strategy. Through this search strategy, a total of 132 papers were shortlisted. The papers selected were passed through another screening process based on their abstract and conclusion in order to select the papers that align to the scope of ACND, reducing the selected relevant papers to 70. Lastly, the remaining papers were then fully read and analysed as further screening step, leading to 55 papers selected for this review overall. Figure 1 suggests the overall steps included within this search methodology.

4 Autonomous Cyber Network Defence

Autonomous Cyber Network Defence (ACND) is a topic that has recently been mentioned within a few publications and news articles over the last decade, in light of the increasing cyber-attacks that have occurred over the last few years. To define and interpret this term, a brief review was completed.

Rege et al. [106] provided a high-level description of ACND algorithms as a decision-making system with expert-level ability inspired by how humans reason and learn, citing a publication [12] producing an autonomous blue agent within a custom networked system. Ko et al. [72] provided a terminology for ACND when elaborating the purpose of the Defence Advanced Research Projects Agency (DARPA) grand challenge [3], where it described ACND as systems that are able to self-discover, prove, and correct software vulnerabilities in real-time without human intervention. In 2016, Baah et al. [99] provided a generalised overview of an ACND system. The paper described

[3]https://www.darpa.mil/program/cyber-grand-challenge

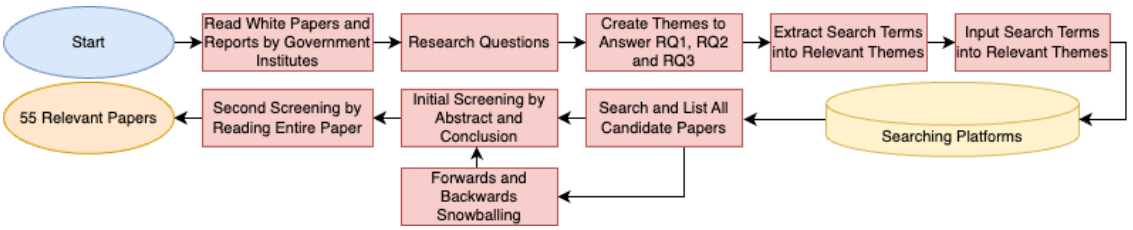

Fig. 1. A flowchart of the overall Research Methodology described in Section 3.

a. Algorithmic Approaches	b. Autonomous Blue-Teaming	c. Autonomous Red-Teaming
- "Artificial Intelligence" ‡	- "Autonomous Cyber Operations Gym"	- "Malware"
- "Machine Learning" OR "Deep Learning" ‡	- "Process Killing"	- "Process"
- "OpenAI" AND "Gym" ⋆	- "Cyber Defence" OR "Cyber Defense"	- "Penetration" AND "Testing"
- "Reinforcement Learning" ‡	- "Malware"	- "Offensive Cybersecurity"
- "Game Theory" ‡	- "Deception"	- "Autonomous Malware"
- "Generative Modelling" ⋆	- "Response"	- "Privilege Escalation"
- "Automated" OR "Automatic" OR "Autonomous" OR "Automation" ‡	- "Wargaming" OR "War-gaming"	- "Adversary Emulation"
- "Response" †	- "Cyber Resilience"	- "Wargaming" OR "War-gaming"
	- "Advanced Persistent Threats" OR "APT"	- "Red Team" OR "Red Teaming" OR "Red-teaming"
	- "Blue Team" OR "Blue-teaming" OR "Blue Teaming"	- "Reconnaissance"
	- "Cyber Threat Intelligence"	- "Autonomous Cyber Operations Gym" - "Cyber Defence" OR "Cyber Defense" - "Deception"

Table 1. This table outlines the overarching themes used for search terminology. **(a)** lists algorithmic approaches and terminologies that enable autonomous responses, which are essential for Autonomous Cyber Network Defence agents. **(b)** includes terms associated with Autonomous Blue Teaming, while **(c)** contains those related to Autonomous Red Teaming. Legend: ⋆ — The keyword has been individually combined with each term from column (a); † — The keyword has been individually combined with each term from column (b); ‡ — The keyword has been combined with all possible pairwise combinations of terms from both (a) and (b).

ACND as a response that begins with detection of an ongoing attack or an existing vulnerability in the network. The paper highlighted that speed and accuracy of detection is important in order to take action to mitigate threats before they can do damage to network assets or disrupt missions. It also illuminates a solution of machine learning analytics that can distinguish between suspicious and benign network activity, and automated fuzzing techniques that can discover previously unknown vulnerabilities in software. Benjamin et al. [12] define the ACND term through their project called Cognitive Support for Intelligent Survivability Management (CSISM), where the authors implement an Autonomous Cyber Network Defence decision-making mechanism with expert level ability. The ACND system observes and alerts the relevant users, and then takes defensive actions to ensure the survivability of the computing capability of the network. The authors realise that producing such an expert-level response in real-time with uncertain and incomplete information is a difficult target. However, they realise that there is a stepping-stone between the development of autonomous reasoning and learning through the use of cognitive architectures for cyber defence operations.

Burke et al. [19] from the Alan Turing Institute introduced a research initiative focusing on Active Cyber Defence (AcCD) through a white paper, which focuses on seeking increased automation within an enterprise to bolster network defenders and cybersecurity. Note, it is important to address the difference between the term AcCD and ACND lies in the inclusion of Automated Security Planners within AcCD, which are used to enhance *human decision-making*, while ACND strictly focuses on autonomous red and blue-teaming, primarily for the overall development of autonomous blue teaming agents. Overall, the paper explains that intelligent automation is essential to enable system defenders to manage the risk posed by highly autonomous future threats and attack, and defend the systems at cyber-relevant national scale. The white paper also elaborated the need for autonomous red and blue teaming. However, it only provided high-level information on the research directions within all areas without a further technical development pipeline. The use of

Artificial Intelligence has been suggested within such systems as a way to intelligently understand the terrain (i.e., networked system) for detecting and responding to complex cyber-attacks with minimal errors.

Applebaum et al. [7] introduce the term Autonomous Cyber Network Defence in the context of tabular Q-learning, defining it as the use of ML to train agents that autonomously defend systems while minimising self-damage from noisy sensor data. While conceptually aligned with our view of ACND, their definition is limited in scope. It overlooks the parallel development of simulation-to-emulation environments such as ACO Gyms and omits the critical role of autonomous red agents. This paper extends the definition by incorporating both red and blue agents within an integrated training and evaluation framework.

The definition of Autonomous Cyber Operations (ACO) will also need to be addressed relative to ACND in order to clarify specific research directions within ACND as compared to ACO. Standen et al. [121] define ACO as the parallel development of autonomous red (attacker) and autonomous blue (defender) agents within a networked system that combat one another in a game-playing scenario. ACND differs from ACO through its focus being on the overall development of autonomous blue agents, where autonomous red agents are particularly designed as an autonomous penetration testing agent facilitating holistic adversarial training. The development of ACO Gyms in the lens of ACND also differs to the development of ACO Gyms in that they must be designed to specifically for the development of autonomous blue teaming agents.

When compiling all the literature mentioned above, we interpret **Autonomous Cyber Network Defence** as a terminology focusing on the autonomous decision-making agents for cyber systems (such as enterprise network, industrial control systems) to mitigate highly complex cyber-attacks. The development of an ACND system could be conducted through a combination of different types of operations. This includes the development of autonomous blue-teaming agents within ACO Gyms as a mode of terrain (to replicate real-world cyber systems), where autonomous red teaming agents are used to adversarially validate, develop and strengthen the autonomous blue team agents for an overall goal of their deployment within networked systems.

4.1 ACND Importance within National Strategy Documents

From our initial set of white papers and search strategy, we discovered that several government-based organisations have made it clear that AI will soon be forefront within cybersecurity in terms of detecting, responding to attacks within networked systems, along with creating autonomous attacks to discover vulnerabilities. Table 2 elaborates the importance of ACND within different countries and organisations, allowing us to decipher our first research question.

4.2 ACND Requirements

The North Atlantic Treaty Organisation (NATO) and US Army Research Laboratory outlined requirements for Autonomous Cyber Agents by producing a reference architecture and technical roadmap, Autonomous Intelligent Cyber-defence Agent (AICA) Reference Architecture (AICA) [73]. A specific part of the document focuses on the high-level strategic deployment and the ethical concerns on the battlefield of autonomous agents. A few requirements in AICA relevant for this paper have been included in a domain-specific manner within the summarised Requirements Table for ACND (Table 3) due to their relevance within defending digital infrastructures against cyber-attacks through autonomous defence agents. Compiling the literature utilised within the initially collected white papers, the table includes a structured format of compiled essential requirements of autonomous red and blue agents **(A)** along with ACO Gym requirements **(G)**, which will incorporate the usage of autonomous red and blue agents.

Country/Alliance	Department/Strategy	Reference to ACND
Australia	Department of Defence [120]	Suggests the need to expand cybersecurity skills and integrating AI into it. DoD is coordinating research and investment in AI capabilities to strengthen capability across the information and cyber domains.
	AI for Decision-Making Initiative 2022 [4]	Aims to develop 30 more AI-based challenges for researchers, including the TTCP CAGE Autonomous Cyber Network Defence Challenge to produce AI-based autonomous decision blue teaming algorithms for instantaneous response against cyber-attacks.
	Royal Air Force of Australia [34]	Advises continuous evaluation in which decisions can be made by machines and which must be made by humans.
Canada	National Cybersecurity Strategy [33]	Specifically mentioned the importance of defence and security applications with autonomous decision support
	Defence Research and Development [36]	The publication suggests that a combination of deep learning and RL algorithms for accurate identification of evolving threats, and then recommend or execute an appropriate course of action.
United Kingdom	Defence Artificial Intelligence Strategy [130]	Discusses the new risks from AI-Enhanced Cyber Threats which operate at speeds and at scales preventing actions by human operators in a timely manner.
	Government Cybersecurity Strategy [98]	Described AI as an emerging technology to focus on. Proposes to explore AI in the context of detecting malicious activity and in some cases to "enable autonomous response to threats"
NATO	Cooperative Cyber Defence Centre of Excellence [92]	Suggest the need for Nation States to adopt and explore AI-enabled Cyber Defence.
	NATO AI Strategy [91]	The strategy includes "collaboration on AI technologies for Cyber Defence.
United States of America	DARPA CASTLE[2]	A long term strategy to develop autonomous Red, Blue and Purple Teaming for algorithmic development of autonomous defence, autonomous attack and ACO Gyms.
	Army Research Laboratory [73]	Designed a reference architecture providing an outline on development of autonomous agents within ACND

Table 2. Overview of the National Strategy Papers on ACND

The requirements in this table are grouped into six key categories, each representing a critical sub-area that demands focused research attention. Effective *generalisation* within ACO Gyms and ACND algorithms will enhance system flexibility and robustness, enabling them to adapt naturally to changes in networked environments. Strategic *high-level decision-making* empowers agents to operate with structure, transparency, trustworthiness, and adaptability in complex, dynamic scenarios. Investigating diverse algorithmic *learning* approaches will help researchers evaluate and uncover more effective methods for training and deploying autonomous agents. Enhancing *multi-agent collaboration* in ACND will facilitate strategic coordination among agents, strengthening defence capabilities and reducing the inherent asymmetry in cyber defence scenarios. To support progress, we advocate for open-source, well-documented *collaboration* across the ACND research community, helping to streamline and accelerate the broader deployment of ACND systems. Finally, ensuring agent *resilience* requires continuous exposure to a wide range of adversarial conditions throughout the training and deployment pipeline, fostering the development of more secure and reliable autonomous defenders. Overall, Table 3 contributes to ACND as a checklist for researchers

Requirement	Summary
Generalisation	- **(G.1.1)** ACO Gym will need to generalise to new settings and have the ability to seamlessly add components - **(G.1.2)** ACO Gym would need to be able to add different types of autonomous agents. - **(G.1.3)** Networked system training-testing must promote transfer from simulation to a real world design, including aspects like matching real networked system latency operations delays within networked systems. Examples include a hybrid of simulation and emulation within training-testing strategies. - **(G.1.4)** ACO Gym must have capability of scaling the network to larger sizes (additional subnets) without configuration issues - **(A.1.1)** Autonomous agent will need to generalise their decisions relevant to the autonomous agent type it represents. - **(A.1.2)** Autonomous agent will have to generalise and adapt to structural changes within the ACO Gyms (addition and removal of subnets and endpoints). - **(A.1.3)** Autonomous red and blue agents must be designed to sustain their high performance from simulation to real-world deployment.
High Level Decision-Making	- **(G.2.1)** ACO Gyms must be designed to explain their state after specific events occur within the networked system. - **(G.2.2)** ACO Gyms will need to be framed into MDP/POMDP format in order to allow for autonomous decisions to be made. - **(A.2.1)** For planning and collective response plans, sequential algorithms will need to be considered. - **(A.2.2)** AICA reference architecture argues that both Game Theory and Artificial Intelligence would be suitable for implementation within ACND. - **(A.2.3)** The designed autonomous agents will require a "deep" architecture to sustain their performance according to the complexity of the ACO Gyms. - **(A.2.4)** Additionally, agents will need to be able to be explainable [17, 56, 69], i.e., justify their real-time decisions made in order for them to be operational within real-world networked systems.
Learning	- **(A.3.1)** AICA [73] opens up on the possibility of enabling continual learning within ACO Gyms. - **(A.3.2)** But also argues the importance of training-testing approaches.
Multi-agent Collaboration	- **(G.4.1)** ACO Gyms must be designed in a way to allow for multi-agent reinforcement learning (MARL) to operate. - **(A.4.1)** Multi-Agent System representations would be required to train the autonomous agents and for action/strategy negotiation. [4]. AICA, combined with a MARL survey produced by [136], suggests utilising combinations of communication approaches and centralised training & Decentralising Execution solutions at a bear minimum.
Research Collaboration	A requirement is the need to explain and collaborate with other researchers within AcCD [19] that coincides with ACND. Thus: - **(G.5.1)** ACO Gym must be open-source for researchers to contribute further to implementations - **(G.5.2)** Documentation for ACO Gyms must be available for further development of gyms and ease of research and implementation of autonomous agents within them
Resilience	The AICA reference architecture highlights the need for resilience against differing malware samples and other algorithmic attacks. Therefore: - **(G.6.1)** ACO Gyms must be designed to allow for autonomous red agent to adversarially train the autonomous blue agent to reduce the number of incorrect actions. - **(G.6.2)** ACO Gyms must be able to incorporate cyber-attacks and algorithmic attacks (e.g. backdoor attacks on DRL agents [3, 138]) plausibly curated by an adversarial insider. - **(A.6.1)** To improve performance of autonomous blue team agent (the sole purpose of ACND), adversarial training through an autonomous red agent must be encouraged. - **(A.6.2)** Autonomous red agents must be provided with a wide variety of cyber-attacks (specified within the MITRE ATT&CK framework) - **(A.6.3)** Autonomous red agents must be provided with a variety of algorithmic attacks [41] (such as adversarial examples) on the trained autonomous blue agents to address autonomous blue agent's algorithmic vulnerabilities. - **(A.6.4)** Autonomous blue and red agent must be able to launch deception defence and attacks respectively.

Table 3. This table provides a list of Requirements for ACND to streamline its deployment within real networked systems.

to streamline their implementations and research contributions, which will expedite the eventual deployment of ACND operations within real-world networked systems.

5 ACND algorithms used within Custom ACO Gyms

As mentioned in the section 4.2, a typical ACND system comprises of a type of networked system, which possesses the provision to allow autonomous red and blue team game-playing scenarios.

Recent publications within ACND have utilised autonomous decision-making algorithms such as Game Theory (GT), Machine Learning (ML) and RL for autonomous blue and red teaming within custom ACO Gyms. A comprehensive overview on the fundamentals of GT and RL can be found in [119] and [123] respectively.

ML-based solutions (along with RL-based solutions [94, 107, 115]) have also been utilised solely for quick incident and intrusion response over the years [50, 96, 128]. Specifically, Zago et al [139] utilise ML techniques to analyse, detect and react against existing and upcoming cyber threats, including botnets. The proposed approach combines unsupervised and supervised approaches to create a scalable detection and reaction framework willing to decrease the error rate as well as increasing the efficiency in terms of computational resources. The approach uses dimensionality reduction algorithms and uses publicly available datasets for intrusion detection for its implementation. While sole ML-based implementations like this allow the mitigation of specific types of attacks, they lack the ability to defend against sophisticated attacks that require a multi-step response.

An example of a threat that requires a multi-step response is a ransomware attack that has already spread partially through the network. Once the attack is detected, containment actions are first taken to prevent further spread. This step might involve disconnecting infected machines, applying network segmentation, or temporarily shutting down network access. Following from this, the focus shifts to removing the ransomware from all infected machines and restoring data from backups. This step requires careful planning to avoid reinfection and to ensure data integrity. A rapidly acting autonomous defence system can potentially address the threat sooner.

Additionally, like zero-sum GT-based solutions, their performance does not scale to larger enterprise networks due to the algorithms not being complex enough to generalise state spaces further away from the scenario in operation. Cam et al. [21] also highlight how most ML-based solutions (which include supervised and unsupervised learning algorithms) provide solutions to a single-step learning problem, a feature of the algorithm that makes it infeasible for implementing it as ACND-based solutions within networked systems. Therefore, the publications selected for this section focus on sequential response that is required for autonomous agent to stop cyber-attacks within an overall networked system.

The rest of this section provides an overview of the recent publications within autonomous response for blue and red teaming respectively within custom networked systems, and analyses the publications based on their autonomous agents and custom ACO Gyms through the Requirements Table in section 4.2.

5.1 Autonomous Blue Team Solutions

The autonomous blue agent within a network system must be perpetually vigilant to defend the entire attacker surface in real-time, while the attacker only needs to succeed once within a single location. Due to this asymmetric scenario between cyber-attackers and defenders, the defenders with limited resources cannot afford to prepare for all possible attacks.

In this subsection, we focus on addressing Posture-related vulnerabilities (PrV), a concept introduced by Huang et al. [63] that highlights the inherent disadvantage faced by the blue team compared to the network attackers. Specifically, the blue team must continuously monitor and protect the entire attack surface from unauthorised access, while attackers only need to find and exploit a single vulnerability to succeed. Due to this disadvantage in security posture, a blue team with limited resources cannot afford to prepare for all possible attacks. Table 4 below evaluates the autonomous blue teaming publications along with their custom ACO Gyms.

Table 4 shows relevant ACND autonomous blue teaming publications within networked systems designed solely for their respective autonomous blue team agent implementations. The table highlights most publications meeting requirements A.1.1, A.1.2, A.2.1, A.2.2. This is specifically

because most publications highlight the need for a sequential blue agent response [21], as opposed to single-shot blue agent responses that are not feasible to defend the systems against modern day cyber-attacks. This is further shown by all publications framing the problem as an MDP/POMDP (G.2.2), which allows autonomous agents to take sequential response through the transitioning of states, that signifies a combination of actions taken within specific nodes of a networked system. However, while the requirements of A.1.2 are met within the specific publications, they are simulation based networked system implementations, which means that the system does not completely represent the complexity of configuration changes of the real-world networked systems. This is specifically highlighted in A.1.3 requirement which is not met by most publications in Table 4 that only test their algorithms within simulated networked systems. Most publications did not meet A.2.3 that is required within complex networked environments for appropriate generalisation of long-term actions for the agent. Only DRL implementations were able to fill this requirement, making them more suitable. Dhir et al. [35] also suggested the use of Causal Inference

	Autonomous Blue Team Custom Networked System Publications															
Requirements	[144]	[15]	[62]	[95]	[85]	[46]	[37]	[23]	[27]	[21]	[132]	[47]	[131]	[118]	[109]	
A.1.1	+			+	+	+	+	+	+	+	+	+		+	+	
A.1.2				+		+	+	+	+		+	+		+	+	
A.1.3								+		+	+					
A.2.1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
A.2.2	+			+	+	+	+	+	+	+	+	+	+	+	+	
A.2.3			+						+	+		+			+	
A.2.4	+	+														
A.3.1																
A.3.2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
A.4.1					+		+									
A.6.1				+		+			+							
A.6.2	+	+														
A.6.3																
A.6.4											+	+				
G.1.1	+	+	+	+		+	+		+		+	+		+	+	
G.1.2			+	+		+	+		+					+	+	
G.1.3		+			+					+	+					
G.1.4	+	+	+	+		+	+		+			+		+	+	
G.2.1		+											+	+	+	
G.2.2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
G.4.1					+		+		+							
G.5.1													+			
G.5.2													+			
G.6.1			+	+			+		+							
G.6.2																

Table 4. Autonomous Blue Team Solutions within custom networked systems

algorithms [48, 61, 67, 108, 140] that could maintain their performance within ACO Gyms. Most publications in Table 4 also do not meet explainability requirement of A.2.4, which is essential for utilisation of any autonomous agents within Security Operations Centre (SOC) environments, in which such agents will need to be certified before they are in operation. Only 2 of the selected publications met A.4.1, in which both publications implemented autonomous response against specific cyber-attacks (i.e., DDoS, as opposed to an agent that could detect and respond to a variety of cyber-attacks). Such requirement is highlighted in the form of A.6.1 and A.6.2, which suggests the need to continually develop the knowledge base of the autonomous blue agent through adversarial training against a variety of cyber-attacks. Moreover, the lack of implementations that fill the A.6.1 requirement also hinders the development of autonomous blue agents against algorithmic attacks mentioned in A.6.3, an area in which no publications highlighted in Table 4 have implemented solutions for.

Requirement A.6.4 in the context of autonomous blue teaming refers to defender agents which have the capacity to strategically launch deceptive elements that enhance the defence of a networked system through an increase in threat detection functions. Applications of Cyber Deception in literature seek to integrate high-fidelity deceptive assets into existing infrastructures with the purpose to mislead or slow down adversaries and ultimately thwart their cognitive processes. These assets are typically encapsulated inside virtual environments that resemble their physical counterparts; and have two overall aims: first, the defence of a system through the enhancement of threat detection functions such as lures and decoys, and second, the ability to misdirect and quarantine attackers to support the gathering of Cyber Threat Intelligence (CTI). Deception-based Cyber Defence (DCD) platforms counter classic attacker-defender asymmetries by executing and maintaining preventative cybersecurity tools that, unbeknown to an adversary, obfuscate the true security posture of a network. In fact, the use of DCD is becoming an increasingly prudent choice in the mitigation of PrV(s) on the account that adversaries must 'minesweep' through a sea of supposed vulnerabilities in order to execute a successful cyber-attack. Wang et al. [132] and Ghao et al. [47] both consider the notion of combining the use of intelligent algorithms with dynamic deployment strategies in order to analyse adversary behaviour. Both solutions succeed in training a blue agent to select optimal deployment strategies but fall short of many generalisation and resilience-based requirements due to the link between the attackers with the associated environment. As previously mentioned, solutions such as [47] which incorporate DRL typically meet the high-level decision-making requirement A.2.3. The use of DRL in this instance is sensible because the authors are aware of the impact that general attacker-defender scenarios have on the space complexity of typical RL algorithms. This is because Deep Neural Networks (DNNs) are introduced to make policy-based deployment decisions without the need to manually engineer the state space. In the context of ACND, determining a reward path through the trial and error of all possible states can often converge to computational intractability as the scale of the network environment grows; thus, by harnessing the predictive element of a DNN, knowledge becomes generalised by approximating each Q value rather than storing and looking up every distinct state. The authors in [47] utilise online learning to update defence models with newly collected attack information, although this is of a 'non-continual' variety, meaning continual learning techniques have not been implemented to address concerns regarding catastrophic interference, thereby failing to meet requirement A.3.1. Leveraging the approximations of DRL, Li et al. [78] proposes an optimal defensive deception framework by creating System Risk Graphs (SRG) which model adversary actions. The attack models are then used to train a DRL agent to generate optimal deployment strategies within micro-service architectures. Incorporating defensive deception into container-based cloud environments is sensible as, like the diversity and scale of typical OT networks, the virtualisation of technology and the dynamism of container services exposes a glut of additional

attack vectors to an already overwhelming issue. Through the intelligent deployment of deceptive assets, the expanding threat surface can be maintained and prevented. The authors highlight the issue of scalability when modelling network environments and threat models as high-dimensional input spaces, implementing a DRL framework that scaled up to 60 nodes. In a different light, Walter et al. [131] draws attention to the prospect of augmenting ACND environments with defensive cyber deception components by adapting the source code of an existing open-source ACO Gym called CyberBattleSim [125]. This paper falls short of many requirements as the solution does not necessarily create a dedicated blue agent. Instead, the aim of the paper was to gain insight by observing the impact of active cyber deception on attacker behaviours which can ultimately inform autonomous blue teaming agents.

In terms of the requirement of networked systems within the publications mentioned in Table 4, G.1.1 and G.1.2 were met within most simulated networked system publications. However, as mentioned previously, simulated systems do not represent the real-world systems accurately, hence the reason why very low number of mentioned implementations are able to meet the G.1.3 requirement. Similar to the requirement A.4.1, G.4.1 is an area in which networked systems will need to be developed in order to facilitate the inclusion of autonomous agents. Areas of research development also include G.4.1 and G.6.1, in which networked systems will need to be designed to allow such requirements. Overall, through our analysis, the most optimal autonomous blue teaming algorithm seen within literature is DRL due to it's ability to meet most algorithmic requirements in the ACND Requirements Table 3, answering our second research question. Using the analysis, it can be understood that many more requirements can be met by DRL algorithms, however, they (like all other algorithms) are highly dependent on the ACO Gym they are developed within.

5.2 Autonomous Red Team Solutions

The existing literature on autonomous red teaming solutions can be split into three categories: assistance to security analysts with attack planning, penetration testing or red teaming "automation", and red agent research conducted in gym environments. The later categories relate closely to ACO goals/objectives, whilst the former is an intermediary step towards it.

The attack path planning category utilises scanning information outputted from penetration testing tools such as Nmap or Nessus to design a POMDP (G.2.2) representing a corporate network. The Common Vulnerability Scoring System (CVSS) scores [5] from vulnerability scans are then utilised to define the transition probabilities. [45] also utilised the CVSS scores to inform the rewards (landing on the host as an administrator for instance). Researchers then utilise RL/DRL algorithms (A.2.2) on these environments to reach set objectives (while adding negative penalties at each step to avoid loops). For example, [45] and [26] utilised this approach to generate action plans to assist a human expert in reaching testing objectives with the DQN algorithm (A.2.3). Finally, it should be noted that tools such as Bloodhound [6] offer attack path planning focusing on Active Directory weaknesses, without utilising ML.

To automate penetration testing, one can extend the DRL game defined in the paragraph above to incorporate actions of penetration testing or red teaming tools (A.6.2). In fact, [142] did so to automate penetration testing with the Metasploit framework [7], whereas [83] utilised the PowerShell Empire framework [8] to automate post exploitation activities. Furthermore, researchers have analysed specific tasks of red teaming and attempted to automate them. For example, [74] automated

[5]https://nvd.nist.gov/vuln-metrics/cvss
[6]https://github.com/BloodHoundAD/BloodHound
[7]https://www.metasploit.com/
[8]https://github.com/EmpireProject/Empire

privilege escalation through RL. One could envision multiple cells of the MITRE ATT&CK matrix [9] being automated in this fashion, such as defence evasion as seen in [39]. Overall, there is a need for a system to continuously evolve the autonomous red teaming agent to append new types of attacks within its action space. A combination of open-source red-teaming such as Atomic Red Team [10], ATTPwn [11], Infection Monkey [12] and APTSimulator [13] follow the MITRE ATT&CK matrix and can be used by DRL agents to execute specialised scanning and attack techniques.

Given that research into RL for autonomous red team solutions can be abstracted into simulated environments (described in further detail in ACO Gyms, G.1.3), the literature also comprises of such research (making it the most relevant algorithm for autonomous red teaming aswell). For example, [122] build DRL agents in the Network Attack Simulator Gym [114]. The authors trained agents in five different scenarios of varied sizes and complexity, which were built with the PPO and DQN algorithms. They trained them on smaller scenarios to see how they performed in the larger ones at testing time, where PPO seemed to generalise slightly better. Given the exponential growth in action sets, researchers have begun analysing the use of Hierarchical RL in this setting, in fact [127] did so in the CyBORG Gym environment [121] where they proposed a Hierarchical DQN algorithm. Research in the open-source gyms are summarised through the list of requirements in Table 5.

Finally, it should be noted that GT Models (A.2.2) have also been explored (an example is provided by [28]), but in this case they are utilised to aid decision makers, such as in cyber war-gaming.

6 Autonomous Cyber Operations Gym

As shown in the previous section, the lack of common open-source ACO Gyms prevent the possibility for an independent, accelerated development of autonomous blue and red agents (and ACO Gyms). This section aims to answer the third research question and provides a detailed overview of literature that have recently developed ACO Gyms along with the autonomous agents developed and published within literature and websites. Such ACO Gyms are simulated and/or emulated networked systems designed specifically for the development of autonomous blue and red team solutions. Given the availability of several resources, different publications have produced different strategies for training and testing environments, algorithm development type, and the types of cyber-attacks.

6.1 Training strategies

The most common approach to training and testing involves validating agents within the same environment used for training, whether simulated or emulated. This limits the ability to assess generalisation (i.e., requirements A.1.1, A.1.2 in Table 3) and prevents agents from fully leveraging the complementary strengths of different environments—scalability in simulation and realism in emulation—thereby falling short of requirement G.1.3.

Several research papers have strived to make progress in the domain of generalisation. For example, [122] built DRL agents in the Network Attack Simulator Gym [114]; a simulated environment to conduct research in autonomous penetration testing. Autonomous agents were trained in five different scenarios (encompassing subnets, hosts, vulnerabilities) of varied sizes and complexity, where the authors adopted both the PPO and DQN algorithms. After training the autonomous agents on scenarios of lower complexity, the impact on performance in larger complexity scenarios

[9] https://attack.mitre.org/matrices/enterprise/
[10] https://github.com/redcanaryco/atomic-red-team
[11] https://github.com/Telefonica/ATTPwn
[12] https://github.com/guardicore/monkey
[13] https://github.com/NextronSystems/APTSimulator

Autonomous Red Team Custom Networked System Publications			
Requirement	[83]	[74]	[45]
A.1.1			
A.1.2			
A.1.3		+	
A.2.1	+	+	+
A.2.2	+	+	+
A.2.3	+	+	+
A.2.4			
A.3.1			
A.3.2			
A.4.1			
A.6.1			
A.6.2			
A.6.3			
A.6.4			
G.1.1	+		+
G.1.2			
G.1.3	+	+	+
G.1.4	+		+
G.2.1			
G.2.2	+	+	+
G.4.1			
G.5.1			
G.5.2			
G.6.1			
G.6.2			

Table 5. Autonomous Red Team solutions within custom networked systems

was experimented with, where the PPO provided superior generalisation. The cutting-edge platforms built to conduct research in ACND designed by [90], [121] or [79] all involve a simulated environment to train agents in a time efficient manner. In addition, emulations of the environment can be spun up on cloud providers with services running, actual malware performing malicious actions and autonomous blue agents with abilities to close ports or remove infections (mapping to the action spaces of the simulation). Another approach involves "real world" testing after training is performed in a simulated environment. One example worth mentioning are task specific agents, for example, [74] enumerated all possible privilege escalation techniques from the MITRE ATT&CK matrix and built an agent with DQN to perform this task. In order to speed up the learning process, they trained their agent in simulated environment built with Python and then conducted their testing in the "real world" (a Windows Virtual Machine). They measured its performance based on how many steps were needed to escalate privileges, for some cases/vulnerabilities, the autonomous agent outperformed human experts.

6.2 Existing Autonomous Cyber Operations Gyms

For the acceleration of research within the domain of autonomous red and blue teaming agents within networked systems, open-source networked systems, or Autonomous Cyber Operations Gym (ACO Gyms) will be required. The provision of ACO Gyms will allow researchers to streamline their focus on meeting the autonomous agent based requirements in Table 3. In addition, this allows researchers to also focus on developing more open-source ACO Gyms that meet the networked system requirements in Table 3. Below is a review of existing environments which are designed for cybersecurity research. The review begins with providing an overview of the existing open-source ACO Gym environments, and then delves into other closed-source emulated (and other simulated) ACO Gym environments that have been published within literature. Each part compares ACO Gyms amongst the other open-source/closed-source ACO Gyms using the ACND Requirements table (3) for ACO Gyms.

6.2.1 Open-source Gyms. Firstly, The Cyber Battle Sim [125] (CBS) environment is created for training autonomous red agents that focus on the lateral movement phase of a cyber-attack in an environment that simulates a fixed network with configured vulnerabilities. The red agent utilises exploits (specific code that remotely accesses a network and gain elevated privileges, or move deeper into the network) for lateral movement while a pre-defined blue agent aims to detect the red agent and obstruct access. The CBS environment can define the network layout and the list of vulnerabilities with their associated nodes. In CBS, the modelled cyber assets capture OS versions with a focus to illustrate how the latest operating systems and up-to-date patches can deliver improved protections. The implementation can also be extended due to its design for autonomous blue agent training. In fact, [131] have implemented this by incorporating blue teaming deception into the environment. The developers ensured sufficient complexity exists in the environment to abstract the cells of the MITRE ATT&CK matrix for vulnerabilities (to be exploited by red agents to get rewards). Overall, the documentation is sufficient to create new scenarios/networks, tweaking reward functions (values of compromised services and costs of exploitation) and adding vulnerabilities to services. While this allows users to extensively experiment with the environment, the code only exists for implementation within a simulated domain, thus, questioning the realism of the environment.

The Gym IDS Game [54] is a simplistic Markov game built upon the OpenAI gym environment. The attacker has two types of available actions:

- Reconnaissance action
- Attack of type 1...m

The defender also has two types of actions at his disposal:

- Monitoring action
- Defensive action of type 1...m

Different scenarios exist for either training a blue or red agent (or both). Unfortunately, the gym environment is overly simplistic and only provides a simulated environment, meaning that, like CBS, it also provides low realism. Similarly, to the Gym IDS Game described above, the Gym Threat Defence gym [86] is also a simulation-based system with a POMDP set-up. However, in this case, the authors have designed it as a purely defensive game where the defender has four different available actions.

- No action
- Blocking a service
- Disconnecting a machine
- Performing action 2 and 3 in parallel

One can define the probabilities of detection for each node, the attack probabilities, the spread probabilities, and the initial state.

Similar to the environments mentioned, the Optimal Intrusion Response Gym [55] is a Markov game built upon the OpenAI Gym libraries. The environment comprises of a simulated enterprise network with 6 subnets, with several hosts, each comprising of an IDS. Unfortunately, the game is overly simplistic for our use case as the defender can only select from two actions.

- ”Stop” will block the gateway. This will degrade the IT service and has a cost associated with it. However, it will also ensure the infection is contained.
- “Continue” is a non-action.

After doing some simulations/tests, [55] discovered that the blue agents they trained are more likely to “Stop" earlier when facing a stealthy attacker than against a noisier one.

The Network Attack Simulator environments [114], is purely built for training autonomous red agents (as there is no blue agent) to test AI systems in penetration testing tasks. This environment is built upon OpenAI gym and allows the ability to create scenarios by defining the number of hosts, services, the observability mode (fully observed for instance) and the asset criticality of the hosts in question. Finally, one can decide the vulnerabilities present on the network and define the cost of actions (cost of a subnet scan for instance). The red agent can select from seven different action types: Exploitation, Privilege escalation, Service scan, Operating system scan, Subnet scan, Process scan and No action. The goal of the project is to train red agents in performing penetration tests against simulated scenarios, while no blue agent interferes with the environment. Recently, the environment was extended to, NASimEmu, which included both simulation and emulation [65] [14]. The agent [15] that is developed within simulation can be seamlessly deployed within emulation. Novel inclusion within this include dynamic scenarios that represent prototypical situations, e.g., typical university or corporate networks. In these scenarios, some attributes are fixed (network topology, OSs, services and exploits), while some are left to change (network size and hosts’ configuration).

The CyBORG environment [121] is designed specifically for training blue agents. However similarly to CBS, it can simply be extended for red teaming use cases. The environment allows training and testing in simulated and emulated environments respectively. The simulated environment comprises of an agent interacting with a scenario modelled in a finite state machine (FSM), in which each state represents systems and networks. An action satisfying a respective pre-condition is required to move from one state to another. The state also provides specific details such as the

[14]https://github.com/jaromiru/NASimEmu
[15]https://github.com/jaromiru/NASimEmu-agents

creation and deletion of individual files, or the making or breaking of network connections. All combined, an ideal training environment is generated for both the defender and adversarial agent. Once the autonomous agent is trained, it can be tested in the emulator, which comprises of AWS virtual machines to create a high fidelity cybersecurity environment in which the autonomous agent interacts with. The purpose of the environment is to act as a platform for research in ACND, whereby challenges are open to the public. Namely, the TTCP Cage Challenge 1, 2, 3 and 4. The challenges are enterprise network environments with ascending complexity (in terms of the observation and action space for the red and blue agent). In Table 6, all CybORG challenges have been added encapsulated into one column to address the overall contributions provided by the contributors.

In the TTCP CAGE Challenge 2, which is an extension of CC1, the action sets for the blue agent are exhaustive.

- Remove - removes malware from a host.
- Restore - if malware has elevated privileges it cannot be removed, and the host must be restored from backup (with a cost associated with it).
- Analyse - monitoring does not always detect infection (5/100 times) but performing an analysis on the host will always detect it.
- Decoy service - sets up a decoy service on a specific host to delay and detect red agent activity (there are 7 different services available).
- No action - Monitoring occurs regardless of other actions.

Scenarios can be defined in YAML files (i.e network topology and asset criticality). In addition, the project comes with varying red agents utilising different strategies. Finally, the documentation is exhaustive and details the high-level desired actions of an autonomous blue agent. On top of this simulated environment, CAGE Challenge 2 extends to an emulation (which is closed source), which can be spun up on AWS to validate the trained agents.

TTCP CAGE Challenge 3 [52] requires participants to develop autonomous defences for a network of drones, pre-compromised by malware during manufacturing, to establish a necessary communication network. The challenge is set within the CybORG environment, focusing on a scenario with 18 drones at constant risk from dormant firmware malware, operating in a 100x100 area with a 30-unit communication radius and a maximum 100-unit bandwidth. Teams alternate in discrete steps to achieve their aims, with the environment automatically providing offensive (red) and neutral (green) teams, and researchers guiding the defensive (blue) team. The green team, representing one agent per drone, simulates ground operative bandwidth demand, while blue and red teams vie for drone control, totaling 18 active agents. Drone movements and network structure, dictated by a randomised swarming algorithm, remain constant, allowing researchers to focus on combating malware through software command and control tactics as a distinct challenge. The reward function, accessible at every timestep through the standard OpenAI gym interface, motivates the creation of Multi-Agent Reinforcement Learning (MARL) agents, evaluating their defensive performance by averaging scores over 1000 episodes, each up to 500 steps. The optimal score in the challenge is 0, indicating flawless message delivery, with -9000 as the minimum, reflecting complete message failure for an episode.

A recently released TTCP CAGE Challenge 4, the network architecture is divided into four subnetworks, including two deployed networks, the Headquarters (HQ) network, and the Contractor network, all interconnected via the internet. The deployed networks are further segmented into two security zones - a restricted and an operational zone, whereas the HQ network is organised into three security zones: a Public Access Zone, an Admin Zone, and an Office Network. The Contractor network, in contrast, comprises a singular UAV control zone. To foster the creation of sophisticated

agents, the composition of each security zone will be variable, with 1-6 servers and 3-10 user hosts, each equipped with 1 to 5 services, ensuring a dynamic and unpredictable environment. The network is defended by five network defenders (MARL): two per deployed network across security zones, one for the entire Headquarters, and none for the Contractor network, which remains undefended. The red team starts with access to the Contractor network, seeking to expand its reach. Red agents can multiply each turn either through opened phishing emails by the Green team or compromised service interactions, with a limit of one Red agent per zone capable of existing on multiple hosts. Although the blue team can eliminate Red's presence in a network, the red team retains a permanent foothold in the Contractor Network.

Yawning Titan [29] is a highly abstracted graph-based gym for training blue agents. The action spaces for both the blue and red agents do not map to realistic ones expected for cyber defence. Instead, it appears that the gym has been created to efficiently test and validate approaches/algorithms. The graph-based design also suggests it's true purpose is to explore computationally expensive approaches involving generalisation A.1.2 as networks can be defined as functions where the YAML file determines the behaviours and spaces. Table 6 has been used to summarise all open-source ACO gyms that can be experimented with.

Researchers from the KTH Royal Institute of Technology and DARPA have jointly developed an open-source platform named The Cyber Security Learning Environment (CSLE), as described in [97]. This framework features network simulation capabilities that facilitate the generation of Markov Decision Processes (MDPs) and enable the rapid learning of security strategies through the training of DRL algorithms for autonomous blue team operations. These strategies can be assessed within an emulated system that offers a realistic setting for evaluation without disrupting the workflow of the targeted system. CSLE includes comprehensive documentation for implementing autonomous blue team strategies within both simulated and emulated environments, enhancing its effectiveness for scalability and realism respectively.

Researchers from QinetiQ released PrimAITE [16] [17], which is an environment that provides the ability to model a customised networked system, while replicating real-world networked system intricacies (e.g., representation of connections, IP addresses, ports, OS's and services) in a way done by a static CybORG environment. The gym environment, made through OpenAI gym, is specifically incorporated to allow DRL functionalities as Autonomous Blue Teaming agents.

6.2.2 Closed-source Gyms. The rest of the ACO Gyms have been analysed in Table 7 through the ACND Requirements shown in Table 3. While the ACO Gyms highlighted are not open-source, they can provide important insights within the ACND community, particularly for researchers who can take inspiration when designing or making modifications to the existing ACO gyms. For example, no open-source ACO Gyms currently available have recognised the need of incorporating algorithmic cyber-attacks (G.6.2) within the action space of autonomous red agents. In addition, many closed-source gyms mention the need to scale the size of the network without configuration issues (G.1.4), an area which only one open-source gym implements and emphasises on. This feature within ACO Gyms incorporates enhanced realism within networked systems as networks and hosts in a corporate environment are non-stationary. Lastly, closed-source environments like [79] have provided more comprehensive cyber-attacks using the MITRE ATT&CK framework for autonomous red agents, allowing more open-source gyms to implement the features within their environments. Overall, similar to open-source gyms, closed-source gyms also provide us with key developments and research areas within ACND and can be utilised to further enhance ACO Gyms in the future.

[16] https://github.com/Autonomous-Resilient-Cyber-Defence/PrimAITE

[17] https://www.qinetiq.com/en/news/qinetiq-releases-primaite-software-to-support-evolution-of-cyber-defence-agents

Autonomous Cyber Operations Gym (Open-source)									
Requirement	CBS	GIG	GTD	OIR	CybORG	NaSim	YT	CSLE	PrimAITE
G.1.1							+	+	+
G.1.2				+	+			+	+
G.1.3					+			+	
G.1.4						+	+	+	+
G.2.1	+				+	+		+	+
G.2.2	+	+		+	+	+	+	+	+
G.4.1					+				
G.5.1	+	+	+	+	+	+	+	+	+
G.5.2	+	+		+	+	+	+	+	+
G.6.1					+			+	+
G.6.2									

Table 6. ACO Gyms (Open-source)

Specifically, their novel implementations could be treated as an open problem for future ACO Gym creators, leading to incremental progress towards the realism of ACO Gyms.

6.3 Combined Analysis of all ACO Gyms

As shown in Table 6, most authors have recognised the requirement of the seamless addition and removal of nodes and components (G.1.1). Authors also meet the requirement of the adding autonomous agents (G.1.2) that are able to generalise their decisions along with understanding the structural changes within the ACO Gyms (A.1.1 and A.1.2 respectively). Moreover, all publications have also understood the requirement of AI-based sequential decision-making autonomous red and blue agents (A.2.1 and A.2.2 respectively), and have structured the ACO Gym as an MDP in order to facilitate such agents. However, while such ACO Gyms are highly scalable (G.1.4) and allow the development of relevant autonomous agents, the environments utilised in all implementations are simulations of real networked systems, highlighting the lack of open-source emulated/real-world ACO Gyms (G.1.3). This results in the lack of "real-world" experience of autonomous agents, which will essential for utilisation within current networked systems.

While the rest of the analysis apply to those of autonomous agents, the design of the current state of the ACO Gyms could be used to assess the quality of autonomous agents that could be designed within the ACO Gyms. Overall, only one ACO Gym (CybORG Cage Challenge 3 [52, 121]) has recognised the need for autonomous multi-agent algorithms (A.4.1) as autonomous blue team solutions. Along with Cage Challenge 3, Malialis et al. [85] and Eghtesad et al.'s [37] publications (specifically focusing on using DRL for defending against DDoS attacks) environments could be a potential inspiration for structuring the ACO Gyms to facilitate multi-agent autonomous red and blue teaming collaboration (G.4.1). Very few ACO Gyms facilitate adversarial training (G.6.1 and A.6.1), which could potentially be utilised to strengthen the autonomous blue agent against a

	Autonomous Cyber Operations Gym (Closed-source)									
Requirement	[44]	[88]	[43]	[16]	[110]	[111]	[79]	[90]	[38]	[2]
G.1.1	+		+	+	+			+		+
G.1.2			+		+	+	+	+	+	+
G.1.3		+		+	+	+	+	+	+	
G.1.4	+			+	+	+			+	+
G.2.1	+	+		+		+	+	+		
G.2.2						+	+	+		
G.4.1										
G.5.1										
G.5.2										+
G.6.1					+	+		+		
G.6.2								+		

Table 7. ACO Gyms (Closed-source)

variety of cyber-attacks (A.6.2). No ACO Gyms currently open-source have recognised the need of incorporating algorithmic cyber-attacks (A.6.3) within the action space of autonomous red agents against autonomous blue agents. Inspiration can be taken from a closed-source ACO Gym [90] to incorporate algorithmic attacks such as evasion and poisoning of autonomous agents such as DRL algorithms.

6.4 Other Deployed Approaches

Several studies have focused on employing datasets and environments to enhance the detection and analysis of attacks. These environment-centric publications have been distinguished in this section from previous discussions, as they do not engage the use of wrappers for sequential decision-making algorithm frameworks, such as deep reinforcement learning (DRL). Such frameworks are crucial for autonomously addressing malicious alterations within the environment.

Researchers and engineers at Splunk create an open-source tool named Attack Range [18], designed for developing and testing the effectiveness of detection systems by simulating attacks in both cloud and local testbed environments. The detection development platform solves three challenges within the detection engineering domain, these include:

- The user being able to build a small lab infrastructure replicating a production environment
- Utilising attack simulation from different engines to generate highly realistic attack data
- Streamlined integration into Continuous Integration/Continuous Delivery pipeline to automate the detection rule testing process

The work therefore, allows the possibility to scale an implementation of a plethora of cyber attacks within the MITRE ATT&CK framework, and test the effectiveness of various detection methods. However, given the lack of autonomous response and sequential decision-making algorithmic

[18] https://github.com/splunk/attack_range

frameworks implemented within this environment, the work lies outside the scope of ACND research.

Landauer et al. [76, 77] developed simulations of user attack scenarios and shared multiple labelled datasets to assess and compare the efficacy of Intrusion Detection Systems (IDSs) based on their detection accuracy. Moreover, the simulation includes a transformation engine that can automatically generate testbeds with capabilities for parallel operations. The creation of these testbeds involves a level of abstraction, enhancing reproducibility, flexibility, and usability. The datasets within these testbeds are structured to analyse multi-step attacks on a host, with each step of the attack being logged and labelled. Landauer et al. [75] have further enhanced the realism of their work by creating the Kyoushi Environment, a testbed that emulates a small enterprise network. This environment utilises complex state machines to mimic typical user activities and to introduce additional multi-step attacks. The data is automatically generated and labelled according to the configuration of the testbed. The Kyoushi Environment is publicly accessible [19]. Although this testbed is one of the most effective setups for intrusion analysis, it currently lacks an autonomous response framework to counter multi-step attacks, thus relying heavily on numerous human operators to mitigate malicious threats within the network. Consequently, it remains outside the scope of ACND research. To date, no efforts have been made in this work to include wrappers or interfaces for integration with existing sequential decision-making frameworks such as DRL.

Chadha et al. [22] developed CyberVAN, a high-fidelity cyber environment specifically designed to counter evolving cyber threats. This tool is widely utilised by cybersecurity professionals for the effective evaluation and validation of cybersecurity technologies. CyberVAN offers a highly realistic representation of network environments, closely approximating the deployment of actual networks. It supports scalability, efficiently managing tens of thousands of varied cyber components such as hosts, routers, switches, firewalls, and communication infrastructures including Wi-Fi, LTE / 5G, and satellite networks. Additionally, CyberVAN is user-friendly, featuring advanced functionalities for the creation, implementation, and preservation of cyber scenarios used in experiments, operational planning, validation, and training. Despite its realism, CyberVAN relies on human analysts for threat mitigation and lacks the integration of sequential decision-making algorithms, thus not aligning with the scope of ACND.

As observed above, several projects explore the application of real-world approaches in cybersecurity research by developing flexible and abstract testbeds capable of adapting to various network environments. However, the current deployed state-of-the-art does not yet include the integration of sequential decision-making algorithms capable of quickly and universally detecting and mitigating multi-step attacks. Despite this, the methodologies used in developing real-world detection systems can inform the deployment of Autonomous Blue and Red Teaming algorithms within realistic networked systems, especially since their development is currently confined to simulation and emulation environments.

7 ACND Algorithms within open-source ACO Gyms

Out of the open-source ACO Gyms mentioned in the previous section, several autonomous decision-making algorithms have been utilised for training and testing as autonomous agents. The ACO Gym creators and autonomous blue and red team agent developers have recognised the need for DRL-based solutions within the domain due to their nature of sequential response. While many of the requirements are met through the use of DRL-based solutions, this section suggests several gaps that still exist within the design of the autonomous agents through currently published implementations. Such gaps will require being met before the algorithms can be deployed into real-world networked

[19] https://github.com/ait-aecid/kyoushi-environment

systems for cybersecurity. Out of the current ACO Gyms, only two open-source ACO Gyms have been utilised in the publications of autonomous red and blue agents. In addition, many algorithms have been developed and are released open-source to promote research and development within the domain. CybORG [121] released four challenges with simulated networked systems with varying ACO Gym complexity in terms of the actions and observation spaces. The challenges focus on the development of autonomous blue agents, while the development of autonomous red agents (comprising of two different types of cyber-attacks) is also possible. NaSim [114] authors made their code open-source for the development of autonomous red agents and a few publications and implementations have utilised the simulated networks for the development of such agents.

7.1 Autonomous Blue Team Solutions

Of the two ACO Gyms discussed, CybORG has published results for its challenges [1], ranking the RL-based algorithms used in Cage Challenge 1 [20] and Cage Challenge 2 [51], with results from Cage Challenge 3 forthcoming [52]. These rankings are based on performance metrics defined by the organisers. A variety of teams employed different approaches and implemented diverse strategies through their autonomous agents. This article selects the highest-performing methods from these challenges and assesses them against the ACND requirements presented in Table 3.

From Cage Challenge 1, Team Mindrake [40] won the challenge and produced a Hierarchical RL algorithm that included proximal policy optimisation [112] with curiosity. The hierarchical [57] component of the algorithm is utilised through a controller to take relevant action according to the type of adversary that is deployed against the autonomous agent (B_line and Meander APT agent). Models are pre-trained against both adversaries separately from the training phase and are then tested by the same adversaries at random episodes. The curiosity component allows exploration within the environment in the training phase via intrinsic reward [102], improving the reward achieved by nearly double. While the autonomous agent was victorious within the challenge, it does not meet the requirements A.1.3, A.2.4, A.3.1, A.4.1, A.6.3 and A.6.4. This is primarily due to the availability of the actions that could be taken amidst the two adversaries, along with the variety of attacks that could be conducted by the adversaries. Additionally, the environment [20] cannot facilitate A.4.1. Similarly, the other three submissions also met the same requirements as the winners of the challenge. From Cage Challenge 2, the team from Cardiff University (with GitHub code [20]) won the challenge and also produced a Hierarchical PPO similar to Team Mindrake in Cage Challenge 1. However, the team utilised the availability of deception within the 2nd challenge through the selection of decoys (when required within the scenario) in a greedy manner. Using the ACND Requirements, the autonomous agent was not able to meet the requirements A.1.3, A.2.4, A.3.1, A.4.1 and A.6.3, but met the requirement of using deception due to its availability within Cage Challenge 2. Bates et al. [10] utilise Cage Challenge 2 to study the effectiveness of reward shaping and intrinsic agent curiousity on the performance of their autonomous blue agent. While the autonomous agent met the same requirements as the implementation above, the authors managed to improve sample efficiency, which is an area critical within ACND when applied to emulated domains. From Cage Challenge 3, Hicks et al. [58] won the challenge by utilising a MARL PPO with curriculum learning [11] to efficiently manage large action spaces, meeting the key requirement of multi-agent collaboration (A.4.1).

As shown in the first two challenges, variations of hierarchical PPO agents have shown most optimal performance (also suggested and algorithmically proven in [135]) as compared to other approaches. While the autonomous agents are able to generalise the moves of the two adversaries, the environment in which they were trained on did not comprise of many different types of cyber

[20]https://github.com/john-cardiff/-cyborg-cage-2

and algorithmic attacks (A.6.2, A.6.3) for the autonomous agents to generalise a greater pool of algorithmic attacks. To meet these requirements within this ACO Gym, future implementations could modify the ACO Gym to increase their cyber and algorithmic attack capabilities to assess the quality of generalisation of the autonomous agents against a greater pool of attacks. In contrast, no autonomous agent implementations in both challenges provided any form of explainability (A.2.4) regarding their incoming actions that they will take. The third Cage Challenge aimed at resolving the requirement of MARL agents (A.4.1), leading to several implementations submitted to this challenge using the algorithm to deal with the challenge of large action spaces. However, the environment (and hence, the agents) lacks the adversarial training of a variety of cyber and algorithmic attacks (A.6.2, A.6.3).

7.2 Autonomous Red Team Solutions

Unfortunately, unlike for the Autonomous Blue Team Solutions, no public challenges have been proposed. As a result, research has been conducted in different gyms and under varying configurations. Therefore public comparable benchmarks are lacking.

Autonomous Red Teaming Solutions, as shown in Table 8 have so far largely been performed through Reinforcement Learning in ACO gym environments such as CyBORG [121], Network Attack Simulator [114] and CyberBattleSim [125], or in emulators or custom representations of IT networks. This intuitively makes sense as the problem is perfectly modelled for a Reinforcement Learning game (exploring a POMDP). Similarly to Autonomous Blue Teaming solutions, the Proximal Policy Optimisation algorithm has shown to be the most successful approach.

One example worth noting, involves research conducted in the CyBORG gym by [121] which presents the only known example of transferring a simulated red agents into an emulation. Researcher implemented DQN agents in the CyBORG simulator. They then validated the autonomous agents in the CyBORG emulator (G.1.3). Most of the autonomous agents successfully transferred to the emulator. Those which didn't likely failed due to over fitting to the observation in the simulator (moving from a discrete to continuous timed observations).
Another example from the Nasim gym, presents the first example of scaling generalisation (G.1.1) was conducted by [122]. They implemented Deep RL agents trained in small scenarios and validated on larger ones at testing time. Their research suggested that the Proximal Policy Optimisation algorithm seemed to generalise slightly better than other algorithms.

However, it remains an open-question if such algorithms are the most appropriate, indeed there appears to be a lack of research on casual approaches in Autonomous Red Teaming Solutions, even though these have recently been shown to be promising for the Blue Teaming side [6].

	Autonomous Red Team			
Papers	[127]	[121]	[93]	[122]
A.1.1				
A.1.2				
A.1.3		+		
A.2.1	+	+	+	+
A.2.2	+	+	+	+
A.2.3	+	+	+	+
A.2.4				
A.3.1				
A.3.2				
A.4.1				
A.6.1				
A.6.2				
A.6.3				
A.6.4				
Gym	CyBORG	CyBORG	Nasim	Nasim

Table 8. Autonomous Red Team solutions within open-source Gyms

8 Discussion

The main purpose of this paper is to identify an imminent research area, ACND, within cybersecurity in order to mitigate cyber-attacks in the future. Autonomous response to cyber-attacks will need to be addressed through the research and development of autonomous red and blue teaming agents that are sequential in the nature of their decision making. The development of such algorithms could be accelerated through a parallel research and development within the area of ACO Gyms. While recent advancements have developed the research area in particular directions, more challenges have been identified using the ACND Requirements (on existing literature) in this paper for the future development within the mentioned areas. Over 50 publications were analysed and compared through the ACND Requirements in Table 3. While the development of ACO Gyms and autonomous red and blue agent comprise of separate research and development strategies, the progress of one area is heavily dependent on the other, justifying the reasoning of having common research challenges. Since more challenges may exist in the specific requirement addressed, it is encouraged for researchers to build on this document to further address and develop areas within ACND that could further catalyse its development into industrial use.

The direct association of ACND Requirements in Table 3 with the publications identified as part of ACND has highlighted evident open problems and their corresponding challenges that must be addressed for the further development of ACND systems prior to their integration into real-world

applications. This section answers the fourth research question and delineates the identified areas for further research and development, connecting them to the specific requirements outlined in the ACND Requirements.

8.1 AI-based Attack Robustification of Autonomous Blue Agents (A.6.3, G.6.1, A.6.1, G.6.2)

This area focuses on enhancing the robustness of DRL algorithms against poisoning and evasion attacks, which target the algorithmic functions of autonomous agents. Such attacks could originate from an insider adversary or a supply chain cyber-attack that alters the training code. To date, there has been scant research on these types of attacks targeting DRL algorithms. Nonetheless, it is clear that future cyber-attackers will likely exploit these methods through DRL and neural network-based research in various domains [8, 9, 24, 31, 70, 116, 143]. Although a few defences have managed to cleanse such poisoned models successfully [14, 53], they are prone to being bypassed by more sophisticated poisoning techniques [9]. Thus, the overarching challenges for this open problem include defending against algorithmic poisoning and evasion attacks in baseline AI environments, as well as the implementation and defence against these attacks within an ACO Gym, which would involve more advanced and context-specific AI attacks. If this open problem remains unaddressed, future networked systems may be at risk of algorithmic attacks that could seize control of autonomous blue agents and eventually, the entire network.

8.2 Continual evolution of action space for the Autonomous Red Agents (A.3.1, G.6.1, A.6.1, A.6.2, A.6.3)

Autonomous red agents action spaces are constantly evolving. Indeed, new services are often added which may have vulnerabilities tied to them. In addition, "every year new exploits are found for software and so in order to be useful Autonomous penetration testing agents will need to be able to handle a large growing database of exploits." [113]. Overall, this open problem aims to develop autonomous red teaming agents and ACO Gyms that can continuously add new types of cyber-attacks autonomously. While this open problem is reliant on other open problems *8.1 and 8.6*, the development and addition of cyber-attacks autonomous red agents based within a continual learning setting are yet to be explored. The development of such system when utilising the current ACO Gyms requires the challenge of utilising different DRL algorithms that can continually add new sequential actions, while the challenge within ACO Gyms would be to convert the discrete configuration used in most gyms, into a continuous environment. Failure to implement on these challenges will keep the autonomous blue agent outdated from latest cyber and algorithmic attacks.

8.3 Explainable RL (A.2.4)

Explainable RL is more complicated than XAI, in fact "explainability for an RL agent, while clearly a subset of XAI and with similarities to IML (Interpretable ML), has distinct characteristics that requires its explicit separation from current XAI and IML research" [32]. Indeed, the first difficulty for XRL is due to the long-time horizons which determine the decisions/actions to take. The second one relates to the models not being built off labelled training data (which would simplify explainability). Therefore, this open problem currently relies on the development of AI research advancements, which can then permeate into the ACND domain. The challenge here involves the development of explainable and interpretable DRL algorithms within baseline AI environments, and then transferring their operations into ACO Gyms. Further inspiration could be taken from relevant survey papers and implementations [5, 49, 81, 82, 87, 89, 100, 104, 105, 117, 126]. Failure to address this challenge will lead to the autonomous blue agent not being certified by industrial employees within networked systems since the trust towards the agent will be low [82].

8.4 Multi-agent RL (G.4.1)

Another research area within autonomous blue teaming for ACND is the utilisation of multi-agent RL algorithms. This will be particularly more beneficial within enterprise networks environments which are highly complex. While [121] authors have proposed the implementation of multi-agent RL within their third and fourth Cage Challenge [21] [22], more research areas could emerge with increased research within this domain. Using single autonomous blue teaming agents will be useful, however, mistakes made by the agent within non-work hours will not be addressed unless there is another agent that evaluates the first agent and alerts it if a wrong decision is made.

8.5 Robustification of Deception Techniques in Autonomous Blue Agents (A.6.4)

It is highly important to highlight the necessity for research areas which utilise deception technology for ACND purposes. Their inclusion within ACO Gyms will allow the introduction more complex and proactive defensive deception techniques in order to study their effects in misdirecting and disrupting adversaries along the cyber kill chain. This is an open problem since existing literature rarely considers the complexity of this challenge, underlining the infancy of deception as a tool for ACND. Research that falls into this category [47, 131, 132] typically prioritise the use of honey-x methods [103] or 'lures' to analyse adversary behaviours through intelligent deployment strategies. This research challenge could make use of a useful framework for the challenge to encourage diversity within deceptive assets is the MITRE ENGAGE matrix, which identifies numerous deception techniques that can be leveraged at different areas of ACND to optimise adversary engagement [23]. Failure to address this challenge deflects from the key purpose of deception as adversaries can weaponise on the homogeneity of decoys and thus magnify the asymmetry that is ever-present between blue and red agents 5.

8.6 Realism of ACO Gyms (G.1.3, A.3.1, A.3.2, G.1.4, G.1.1, G.1.2)

Another open research area within the ACO gyms is the lack of realism of most of the environments that currently exist. A metric to classify the quality of the training-testing (or continual learning [58]) strategy as a research area is particularly important. Additionally, researchers generally would require building simulated environments and then transfer the learned policies to the real world (Sim-to-Real Transfer), this is often done in the case of robotics as pointed out by [141]. Environments such as CyBORG [121] attempt to address this challenge by supporting both simulation and emulation, however, both implementations comprise of areas which do not represent real networked systems (i.e., latency delays in simulation and network scalability in emulation). In addition, IT and OT networks, unlike traditional RL tasks, are continual and ever-changing environments which contrasts with most RL tasks. Moreover, networks and hosts in a corporate environment are non-stationary, whereas video games in which RL have been used would not expect an agent to perform well on an entirely new map [13, 64, 129]. Lastly, it should be noted that some networked systems like enterprise networks are multi-party network, which have a hierarchy of access levels depending on the user, future ACO Gyms should focus on designing such systems which incorporate this. The challenge here looks at developing new ACO Gyms to aid continual learning, an everchanging configuration and incorporating access restriction while training the DRL agents to maintain generalisability [101]. Such issues must be addressed, else the agent will not recognise the environment when implemented within real-world networked systems, leading to unwarranted actions being taken.

[21] https://github.com/cage-challenge/cage-challenge-3
[22] https://github.com/cage-challenge/cage-challenge-4
[23] https://engage.mitre.org/

8.7 Realism of Deception Techniques (A.6.4)

Deception fidelity is often overlooked and introduced as a part of a constraint or assumption in current literature. As virtualisation of physical assets becomes more commonplace in context of network emulation, the implementation of Deception-based Cyber Defence (DCD) platforms must have the capability to model and simulate physical processes to maintain system fidelity and not alert attackers of its use. However, it is difficult to strike a balance between system fidelity and a sizable attack surface, particularly when considering the complexity and scale of some networked systems such as corporate networked system and OT environments, where researchers must find methods to emulate devices in convincing ways without replicating the network in its entirety. This open problem requires new methods for creating decoy profiles for assets which embody the attributes of the network component. To solve this challenge, researchers can also look to deceptive techniques which already consider or enhance the fidelity of integrated-lures. 'Honeyshills' [59] are an example as they use real components or systems and configure them to communicate with decoys to further give the impression of realism. These encourage suggestions for scaling deception methods within simulation-based networks and ultimately the move towards the emulated domain. Failure to address these challenges may result in the exposure of deception to the attacker, nullifying the precedence of deception over an attacker's inadvertence to its use. Such a contradiction cancels-out the symmetric advantage that's provided by correctly implementing deception technology.

8.8 Impact of Incorrect Action (G.6.1, G.1.3) [41]

The issue of incorrect actions also leads to a wide open research gap within the ACND literature for autonomous decision-making agents. The impact of such actions could lead of a plethora of issues within a corporate organisation. Examples range from minor actions such as blocking benign user hosts from joining the network to major actions such as the deletion of mission-critical documents conducted due to a lack of data diversity within the data used for training the autonomous blue agents. Therefore, research challenges include appropriate evaluation and metrics for the maximal reduction of "damage control" done by the agent. Additionally, explainable approaches [60, 89, 117] must be prioritised for superior forensic evaluation of the autonomous agents. Not addressing this area will result in the autonomous blue team agent potentially eliminating important processes within the network, which could lead to high monetary losses.

8.9 Action and Observation Spaces (G.2.1 ,G.2.2, A.2.3)

Existing research in ACND significantly reduces the action and observation spaces by abstracting the action spaces to a point where they may no longer be usable in the "real world". Indeed, in a cybersecurity setting where agents may be deployed on thousands of hosts (in a single corporate network), each with huge action sets (kill any process, move/quarantine any file, change any firewall setting etc.) and essentially a continuous observation space, it would be challenging to sufficiently explore the space in training. This challenge applies to autonomous red agents also as "applying conventional DRL to automate penetration testing would be difficult and unstable as the action space can explode to thousands even for relatively small scenarios" where "each action in autonomous penetration testing can have very different effects such as attacking hosts in different subnets or different method of exploits" [127].

8.10 Development of new ACO Gyms (G)

In the current landscape of ACND, the availability of open-source ACO (Autonomous Cyber Operations) Gyms for researchers to test their autonomous blue and red team agents is markedly limited.

This scarcity presents an open problem, urging more collaboration among AI and Cybersecurity professionals to propel advancements in ACND by developing new ACO Gyms. The challenges associated with this open problem involve gaining proficiency with the OpenAI Gym framework and leveraging the foundational code present in existing ACO Gyms as a basis for further development. Additionally, researchers are encouraged to utilise the Table 3 and the open problems above for constructing the networked system and incorporating the suggested research enhancements for the ACO Gym outlined in both the table and the open problems. Inspiration could also be taken from the existing OpenAI benchmark games and custom Gyms used for different applications. Researchers can also make use of the Kyoushi Environment and incorporate host-based datasets for every machine as a way to improve realism of ACO Gym development.

9 Conclusion

This article advances the understanding of Autonomous Cyber Network Defence (ACND) by elucidating its terminology through research articles, government strategic reports, and cybersecurity training organizations. This clarification of terms facilitated the identification of specific ACND sub-areas, namely, Autonomous (Blue and Red) Agents and Autonomous Cyber Operations (ACO) Gyms, thus guiding the creation of ACND Requirements, a set of criteria used to evaluate the relevant literature. Through an extensive literature review on autonomous blue and red teaming algorithms within ACO Gyms it was revealed that Deep Reinforcement Learning (DRL) so far has outperformed Game Theoretic and conventional Machine Learning approaches. DRL's advantage lies in its ability to handle sequential decision-making for achieving short-term and long-term objectives. Moreover, an in-depth assessment of both open and closed-source gyms, along with their implementations of autonomous teaming, was conducted. These evaluations, guided by the ACND Requirements, pinpointed areas ripe for further research.

To leverage DRL's capabilities in practical cybersecurity applications, further advancements are necessary in autonomous agent technologies and ACO Gym environments. Our findings have pinpointed specific challenges and gaps in the current field, including improving the robustness of defences against autonomous blue agents, enhancing the realism of ACO Gyms, minimising the repercussions of erroneous actions, and refining ACO Gym designs. Additionally, critical issues with DRL defenders need addressing, such as safeguarding against adversarial policies targeting blue agents, enhancing the explainability of blue agents, and refining multi-agent systems. Tackling these unresolved problems is vital for the progression of autonomous agents from controlled simulations to real-world networked environments, ultimately steering future research and development efforts in ACND.

Acknowledgements

We would like to express our sincere gratitude to Andrew Bolton for his insightful contributions to the sections on Autonomous Blue Teaming. His expertise in the area of cyber deception enhanced the depth and quality of this publication.

References

[1] 2022. Cyber Operations Research Gym. https://github.com/cage-challenge/CybORG. Created by Maxwell Standen, David Bowman, Son Hoang, Toby Richer, Martin Lucas, Richard Van Tassel, Phillip Vu, Mitchell Kiely, KC C., Natalie Konschnik, Joshua Collyer.

[2] 2023. Cyber Agents for Security Testing and Learning Environments. https://sam.gov/opp/9c4593776a9b44e98b9bc734a3e16976/view#description. Created by Defense Advanced Research Projects Agency.

[3] Manoj Acharya, Weichao Zhou, Anirban Roy, Xiao Lin, Wenchao Li, and Susmit Jha. 2023. Universal Trojan Signatures in Reinforcement Learning. In *NeurIPS 2023 Workshop on Backdoors in Deep Learning-The Good, the Bad, and the Ugly.*

[4] Queensland Defence Science Alliance. 2022. Artificial Intelligence for Decision making initiative (2022). https://queenslanddefencesciencealliance.com.au/federal-and-state-defence-funding-opportunities-2/artificial-intelligence-for-decision-making-initiative-round-2022/

[5] Prithviraj Ammanabrolu and Mark O Riedl. 2018. Playing text-adventure games with graph-based deep reinforcement learning. *arXiv preprint arXiv:1812.01628* (2018).

[6] Alex Andrew, Sam Spillard, Joshua Collyer, and Neil Dhir. 2022. Developing optimal causal cyber-defence agents via cyber security simulation. *arXiv preprint arXiv:2207.12355* (2022).

[7] Andy Applebaum, Camron Dennler, Patrick Dwyer, Marina Moskowitz, Harold Nguyen, Nicole Nichols, Nicole Park, Paul Rachwalski, Frank Rau, Adrian Webster, et al. 2022. Bridging automated to autonomous cyber defense: Foundational analysis of tabular q-learning. In *Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security*. 149–159.

[8] Giovanni Apruzzese, Mauro Andreolini, Mirco Marchetti, Andrea Venturi, and Michele Colajanni. 2020. Deep Reinforcement Adversarial Learning Against Botnet Evasion Attacks. *IEEE Transactions on Network and Service Management* 17, 4 (2020), 1975–1987. doi:10.1109/TNSM.2020.3031843

[9] Chace Ashcraft and Kiran Karra. 2021. Poisoning deep reinforcement learning agents with in-distribution triggers. *arXiv preprint arXiv:2106.07798* (2021).

[10] Elizabeth Bates, Vasilios Mavroudis, and Chris Hicks. 2023. Reward Shaping for Happier Autonomous Cyber Security Agents. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*. 221–232.

[11] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*. 41–48.

[12] David Paul Benjamin, Partha Pal, Franklin Webber, Paul Rubel, and Mike Atigetchi. 2008. Using a Cognitive Architecture to Automate Cyberdefense Reasoning. In *2008 Bio-inspired, Learning and Intelligent Systems for Security*. 58–63. doi:10.1109/BLISS.2008.17

[13] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. 2019. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680* (2019).

[14] Shubham Bharti, Xuezhou Zhang, Adish Singla, and Jerry Zhu. 2022. Provable defense against backdoor policies in reinforcement learning. *Advances in Neural Information Processing Systems* 35 (2022), 14704–14714.

[15] Lashon Booker and Scott Musman. 2022. A Model-Based, Decision-Theoretic Perspective on Automated Cyber Response. *International Conference on Autonomous Intelligent Cyber-defence agents*.

[16] Scott Brown, Harold Brown, Michael Russell, Brian Henz, Michael Edwards, Frank Turner, and Giorgio Bertoli. 2016. Validation of network simulation model and scalability tests using example malware. In *MILCOM 2016 - 2016 IEEE Military Communications Conference*. 491–496. doi:10.1109/MILCOM.2016.7795375

[17] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, et al. 2020. Toward trustworthy AI development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213* (2020).

[18] Ricardo Buettner, Daniel Sauter, Jonas Klopfer, Johannes Breitenbach, and Hermann Baumgartl. 2021. A Review of Recent Advances in Machine Learning Approaches for Cyber Defense. In *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 3969–3974.

[19] A. Burke. 2017 [Online]. Robust Artificial Intelligence for Active Cyber Defence. Alan Turing Insitute. https://www.turing.ac.uk/sites/default/files/2020-08/publicaiacdtechreportfinal.pdf

[20] CAGE. 2021. CAGE Challenge 1. In *Proceedings of the IJCAI-21 1st International Workshop on Adaptive Cyber Defense*. arXiv. Available at https://arxiv.org/abs/placeholder.

[21] Hasan Cam. 2020. Cyber resilience using autonomous agents and reinforcement learning. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II*, Tien Pham, Latasha Solomon, and Katie Rainey (Eds.), Vol. 11413. International Society for Optics and Photonics, SPIE, 219 – 234. doi:10.1117/12.2559319

[22] Ritu Chadha, Thomas Bowen, Cho-Yu J Chiang, Yitzchak M Gottlieb, Alex Poylisher, Angello Sapello, Constantin Serban, Shridatt Sugrim, Gary Walther, Lisa M Marvel, et al. 2016. Cybervan: A cyber security virtual assured network testbed. In *MILCOM 2016-2016 IEEE Military Communications Conference*. IEEE, 1125–1130.

[23] Xinzhong Chai, Yasen Wang, Chuanxu Yan, Yuan Zhao, Wenlong Chen, and Xiaolei Wang. 2020. DQ-MOTAG: Deep Reinforcement Learning-based Moving Target Defense Against DDoS Attacks. 375–379. doi:10.1109/DSC50466.2020.00065

[24] Yu-Ying Chen, Chiao-Ting Chen, Chuan-Yun Sang, Yao-Chun Yang, and Szu-Hao Huang. 2021. Adversarial attacks against reinforcement learning-based portfolio management strategy. *IEEE Access* 9 (2021), 50667–50685.

[25] Chwee Seng Choo, Ching Lian Chua, and Su-Han Victor Tay. 2007. Automated Red Teaming: A Proposed Framework for Military Application. In *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation* (London, England) *(GECCO '07)*. Association for Computing Machinery, New York, NY, USA, 1936–1942. doi:10.1145/1276958.

1277345
[26] Ankur Chowdhary, Dijiang Huang, Jayasurya Sevalur Mahendran, Daniel Romo, Yuli Deng, and Abdulhakim Sabur. 2020. Autonomous security analysis and penetration testing. In *2020 16th International Conference on Mobility, Sensing and Networking (MSN)*. IEEE, 508–515.
[27] Ankur Chowdhary, Dijiang Huang, Abdulhakim Sabur, Neha Vadnere, Myong Kang, and Bruce Montrose. 2021. Sdn-based moving target defense using multi-agent reinforcement learning. In *Proceedings of the first International Conference on Autonomous Intelligent Cyber defense Agents (AICA 2021), Paris, France*. 15–16.
[28] Edward JM Colbert, Alexander Kott, and Lawrence P Knachel. 2020. The game-theoretic model and experimental investigation of cyber wargaming. *The Journal of Defense Modeling and Simulation* 17, 1 (2020), 21–38.
[29] Josh Collyer, Alex Andrew, and Duncan Hodges. 2022. ACD-G: Enhancing autonomous cyber defense agent generalization through graph embedded network representation. International Conference on Machine Learning.
[30] William Crumpler and James A Lewis. 2022. *Cybersecurity Workforce Gap*. JSTOR.
[31] Jing Cui, Yufei Han, Yuzhe Ma, Jianbin Jiao, and Junge Zhang. 2024. BadRL: Sparse Targeted Backdoor Attack Against Reinforcement Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 11687–11694.
[32] Richard Dazeley, Peter Vamplew, and Francisco Cruz. 2023. Explainable reinforcement learning for broad-xai: a conceptual framework and survey. *Neural Computing and Applications* 35, 23 (2023), 16893–16916.
[33] National Defence. 2021. Government of Canada. https://www.canada.ca/en/department-national-defence/programs/defence-ideas/element/innovation-networks/challenge/autonomous-systems-defence-security-trust-barriers-adoption.html
[34] S Kate Devitt and Damian Copeland. 2023. Australia's approach to AI governance in security and defence. In *The AI Wave in Defence Innovation*. Routledge, 217–250.
[35] Neil Dhir, Henrique Hoeltgebaum, Niall Adams, Mark Briers, Anthony Burke, and Paul Jones. 2021. Prospective artificial intelligence approaches for active cyber defence. *arXiv preprint arXiv:2104.09981* (2021).
[36] Maxwell Dondo and Natalia Nakhla. 2021. Towards a framework for autonomous defensive cyber operations in a Network Operations Centre. (2021). https://cradpdf.drdc-rddc.gc.ca/PDFS/unc382/p814083_A1b.pdf
[37] Taha Eghtesad, Yevgeniy Vorobeychik, and Aron Laszka. 2020. Adversarial Deep Reinforcement Learning Based Adaptive Moving Target Defense. *Decision and Game Theory for Security: 11th International Conference* (12 2020), 58–79. doi:10.1007/978-3-030-64793-3_4
[38] Thomas C. Eskridge, Marco M. Carvalho, Evan Stoner, Troy Toggweiler, and Adrian Granados. 2015. VINE: A Cyber Emulation Environment for MTD Experimentation *(MTD '15)*. Association for Computing Machinery, New York, NY, USA, 43–47. doi:10.1145/2808475.2808486
[39] Zhiyang Fang, Junfeng Wang, Boya Li, Siqi Wu, Yingjie Zhou, and Haiying Huang. 2019. Evading anti-malware engines with deep reinforcement learning. *IEEE Access* 7 (2019), 48867–48879.
[40] Myles Foley, Chris Hicks, Kate Highnam, and Vasilios Mavroudis. 2022. Autonomous Network Defence Using Reinforcement Learning. In *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security* (Nagasaki, Japan) *(ASIA CCS '22)*. Association for Computing Machinery, New York, NY, USA, 1252–1254. doi:10.1145/3488932.3527286
[41] Making AI Work for Cyber Defense: The Accuracy-Robustness Tradeoff. 2021. doi:10.51593/2021CA007
[42] Keith Frankish and William Ramsey. 2014. *The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press. doi:10.1017/CBO9781139046855
[43] Angelo Furfaro, Antonio Piccolo, and Domenico Sacca. 2016. Smallworld: A test and training system for the cybersecurity. *European Scientific Journal* (2016).
[44] Ariel Futoransky, Fernando Miranda, José Orlicki, and Carlos Sarraute. 2010. Simulating cyber-attacks for fun and profit. *arXiv preprint arXiv:1006.1919* (2010).
[45] Rohit Gangupantulu, Tyler Cody, Abdul Rahma, Christopher Redino, Ryan Clark, and Paul Park. 2021. Crown Jewels Analysis using Reinforcement Learning with Attack Graphs. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 1–6.
[46] Chungang Gao and Yongjie Wang. 2021. Reinforcement learning based self-adaptive moving target defense against DDoS attacks. *Journal of Physics: Conference Series* 1812 (02 2021), 012039. doi:10.1088/1742-6596/1812/1/012039
[47] Yazhuo Gao, Guomin Zhang, and Changyou Xing. 2021. A multiphase dynamic deployment mechanism of virtualized honeypots based on intelligent attack path prediction. *Security and Communication Networks* 2021, 1 (2021), 6378218.
[48] Maxime Gasse, Damien Grasset, Guillaume Gaudron, and Pierre-Yves Oudeyer. 2021. Causal reinforcement learning using observational and interventional data. *arXiv preprint arXiv:2106.14421* (2021).
[49] Claire Glanois, Paul Weng, Matthieu Zimmer, Dong Li, Tianpei Yang, Jianye Hao, and Wulong Liu. 2024. A survey on interpretable reinforcement learning. *Machine Learning* 113, 8 (2024), 5847–5890.
[50] Ross Gore, Saikou Y Diallo, Jose Padilla, and Barry Ezell. 2018. Assessing cyber-incidents using machine learning. *International Journal of Information and Computer Security* 10, 4 (2018), 341–360.

[51] TTCP Cage Working Grouo. 2022. TTCP CAGE Challenge 2. https://github.com/cage-challenge/cage-challenge-2.
[52] TTCP CAGE Working Group. 2022. TTCP CAGE Challenge 3. https://github.com/cage-challenge/cage-challenge-3.
[53] Junfeng Guo, Ang Li, Lixu Wang, and Cong Liu. 2023. Policycleanse: Backdoor detection and mitigation for competitive reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4699–4708.
[54] Kim Hammar and Rolf Stadler. 2020. Finding effective security strategies through reinforcement learning and Self-Play. In *2020 16th International Conference on Network and Service Management (CNSM)*. IEEE, 1–9.
[55] Kim Hammar and Rolf Stadler. 2021. Learning intrusion prevention policies through optimal stopping. In *2021 17th International Conference on Network and Service Management (CNSM)*. IEEE, 509–517.
[56] Ronan Hamon, Henrik Junklewitz, Ignacio Sanchez, et al. 2020. Robustness and explainability of artificial intelligence. *Publications Office of the European Union* 207 (2020), 2020.
[57] Bernhard Hengst. 2010. *Hierarchical Reinforcement Learning*. Springer US, Boston, MA, 495–502. doi:10.1007/978-0-387-30164-8_363
[58] Chris Hicks, Vasilios Mavroudis, Myles Foley, Thomas Davies, Kate Highnam, and Tim Watson. 2023. Canaries and Whistles: Resilient Drone Communication Networks with (or without) Deep Reinforcement Learning. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*. 91–101.
[59] William Hofer, Thomas Edgar, Draguna Vrabie, and Kathleen Nowak. 2019. Model-driven deception for control system environments. In *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*. IEEE, 1–7.
[60] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).
[61] Xing Hu, Rui Zhang, Ke Tang, Jiaming Guo, Qi Yi, Ruizhi Chen, Zidong Du, Ling Li, Qi Guo, Yunji Chen, et al. 2022. Causality-driven hierarchical structure discovery for reinforcement learning. *Advances in neural information processing systems* 35 (2022), 20064–20076.
[62] Linan Huang and Quanyan Zhu. 2019. Adaptive Strategic Cyber Defense for Advanced Persistent Threats in Critical Infrastructure Networks. *SIGMETRICS Perform. Eval. Rev.* 46, 2 (jan 2019), 52–56. doi:10.1145/3305218.3305239
[63] Yunhan Huang, Linan Huang, and Quanyan Zhu. 2022. Reinforcement learning for feedback-enabled cyber resilience. *Annual reviews in control* 53 (2022), 273–295.
[64] Yunhan Huang and Quanyan Zhu. 2019. Deceptive Reinforcement Learning Under Adversarial Manipulations on Cost Signals. In *GameSec*.
[65] Jaromír Janisch, Tomáš Pevnỳ, and Viliam Lisỳ. 2023. Nasimemu: Network attack simulator & emulator for training agents generalizing to novel scenarios. In *European Symposium on Research in Computer Security*. Springer, 589–608.
[66] Lawrence Awuah Johnson Kinyua. 2021. AI/ML in Security Orchestration, Automation and Response: Future Research Directions. *Intelligent Automation & Soft Computing* 28, 2 (2021), 527–545. doi:10.32604/iasc.2021.016240
[67] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. 2022. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475* (2022).
[68] Staffs Keele et al. 2007. *Guidelines for performing systematic literature reviews in software engineering*. Technical Report. Technical report, ver. 2.3 ebse technical report. ebse.
[69] Mi-Young Kim, Shahin Atakishiyev, Housam Khalifa Bashier Babiker, Nawshad Farruque, Randy Goebel, Osmar R. Zaïane, Mohammad-Hossein Motallebi, Juliano Rabelo, Talat Syed, Hengshuai Yao, and Peter Chun. 2021. A Multi-Component Framework for the Analysis and Design of Explainable Artificial Intelligence. *Machine Learning and Knowledge Extraction* 3, 4 (2021), 900–921. https://www.mdpi.com/2504-4990/3/4/45
[70] Panagiota Kiourti, Kacper Wardega, Susmit Jha, and Wenchao Li. 2020. Trojdrl: evaluation of backdoor attacks on deep reinforcement learning. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 1–6.
[71] Barbara Kitchenham. 2004. Procedures for performing systematic reviews. *Keele, UK, Keele University* 33, 2004 (2004), 1–26.
[72] Ryan KL Ko. 2020. Cyber autonomy: automating the hacker–self-healing, self-adaptive, automatic cyber defense systems and their impact on industry, society, and national security. In *Emerging technologies and international security*. Routledge, 173–191.
[73] Alexander Kott. 2023. *Autonomous intelligent cyber defense agent (aica)*. Springer.
[74] Kalle Kujanpää, Willie Victor, and Alexander Ilin. 2021. Automating Privilege Escalation with Deep Reinforcement Learning. In *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security*. 157–168.
[75] Max Landauer, Florian Skopik, Maximilian Frank, Wolfgang Hotwagner, Markus Wurzenberger, and Andreas Rauber. 2022. Maintainable log datasets for evaluation of intrusion detection systems. *IEEE Transactions on Dependable and Secure Computing* 20, 4 (2022), 3466–3482.
[76] Max Landauer, Florian Skopik, and Markus Wurzenberger. 2024. Introducing a new alert data set for multi-step attack analysis. In *Proceedings of the 17th Cyber Security Experimentation and Test Workshop*. 41–53.
[77] Max Landauer, Florian Skopik, Markus Wurzenberger, Wolfgang Hotwagner, and Andreas Rauber. 2020. Have it your way: Generating customized log datasets with a model-driven simulation testbed. *IEEE Transactions on Reliability* 70,

1 (2020), 402–415.
[78] Huanruo Li, Yunfei Guo, Penghao Sun, Yawen Wang, and Shumin Huo. 2022. An optimal defensive deception framework for the container-based cloud with deep reinforcement learning. *IET Information Security* 16, 3 (2022), 178–192. doi:10.1049/ise2.12050 arXiv:https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/ise2.12050
[79] Li Li, Raed Fayad, and Adrian Taylor. 2021. Cygil: A cyber gym for training autonomous agents over emulated network systems. *arXiv preprint arXiv:2109.03331* (2021).
[80] Michael Littman. 2009. Algorithms for Sequential Decision Making. (08 2009).
[81] Daoming Lyu, Fangkai Yang, Bo Liu, and Steven Gustafson. 2019. SDRL: interpretable and data-efficient deep reinforcement learning leveraging symbolic planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2970–2977.
[82] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2020. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 2493–2500.
[83] Ryusei Maeda and Mamoru Mimura. 2021. Automating post-exploitation with deep reinforcement learning. *Computers & Security* 100 (2021), 102108.
[84] Mohamad Imad Mahaini, Shujun Li, and Rahime Belen Sağlam. 2019. Building taxonomies based on human-machine teaming: Cyber security as an example. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*. 1–9.
[85] Kleanthis Malialis and Daniel Kudenko. 2013. Large-Scale DDoS Response Using Cooperative Reinforcement Learning.
[86] Erik Miehling, Mohammad Rasouli, and Demosthenis Teneketzis. 2015. Optimal defense policies for partially observable spreading processes on Bayesian attack graphs. In *Proceedings of the second ACM workshop on moving target defense*. 67–76.
[87] Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. 2022. A survey of explainable reinforcement learning. *arXiv preprint arXiv:2202.08434* (2022).
[88] Jelena Mirkovic, Terry V Benzel, Ted Faber, Robert Braden, John T Wroclawski, and Stephen Schwab. 2010. The DETER project: Advancing the science of cyber security experimentation and test. In *2010 IEEE International Conference on Technologies for Homeland Security (HST)*. IEEE, 1–7.
[89] Ludovico Mitchener, David Tuckey, Matthew Crosby, and Alessandra Russo. 2022. Detect, Understand, Act: A Neuro-symbolic Hierarchical Reinforcement Learning Framework. *Machine Learning* 111, 4 (2022), 1523–1549.
[90] Andres Molina-Markham, Cory Miniter, Becky Powell, and Ahmad Ridley. 2021. Network environment design for autonomous cyberdefense. *arXiv preprint arXiv:2103.07583* (2021).
[91] NATO. 2021. Artificial Intelligence and Autonomy in the Military. https://ccdcoe.org/uploads/2021/12/Strategies_and_Deployment_A4.pdf
[92] NATO. 2022. Cooperative Cyber Defence Centre of Excellence. https://ccdcoe.org/library/publications/
[93] Hoang Viet Nguyen, Hai Ngoc Nguyen, and Tetsutaro Uehara. 2020. Multiple Level Action Embedding for Penetration Testing. In *The 4th International Conference on Future Networks and Distributed Systems (ICFNDS)*. 1–9.
[94] Thanh Thi Nguyen and Vijay Janapa Reddi. 2021. Deep reinforcement learning for cyber security. *IEEE Transactions on Neural Networks and Learning Systems* 34, 8 (2021), 3779–3795.
[95] Zhen Ni and Shuva Paul. 2019. A Multistage Game in Smart Grid Security: A Reinforcement Learning Solution. *IEEE Transactions on Neural Networks and Learning Systems* 30, 9 (2019), 2684–2695. doi:10.1109/TNNLS.2018.2885530
[96] Constantin Nilă, Ioana Apostol, and Victor Patriciu. 2020. Machine learning approach to quick incident response. In *2020 13th International Conference on Communications (COMM)*. 291–296. doi:10.1109/COMM48946.2020.9141989
[97] KTH Royal Institute of Technology and DARPA. 2023. The Cyber Security Learning Environment. https://github.com/Limmen/csle.
[98] Cabinet Office. 2022. Government Cyber Security Strategyy. https://www.gov.uk/government/publications/government-cyber-security-strategy-2022-to-2030
[99] Hamed Okhravi, Thomas R. Hobson, William W. Streilein, George K. Baah, Shannon C. Roberts, and Sophia Yuditskaya. 2015. *Title of the Report.* Technical Report AD1034028. Defense Technical Information Center. https://apps.dtic.mil/sti/html/tr/AD1034028/index.html Accessed: 2025-06-03.
[100] Matthew L Olson, Roli Khanna, Lawrence Neal, Fuxin Li, and Weng-Keen Wong. 2021. Counterfactual state explanations for reinforcement learning agents via generative deep learning. *Artificial Intelligence* 295 (2021), 103455.
[101] Sindhu Padakandla. 2021. A survey of reinforcement learning algorithms for dynamically varying environments. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–25.
[102] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. 2017. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*. PMLR, 2778–2787.
[103] Jeffrey Pawlick, Edward Colbert, and Quanyan Zhu. 2019. A game-theoretic taxonomy and survey of defensive deception for cybersecurity and privacy. *ACM Computing Surveys (CSUR)* 52, 4 (2019), 1–28.

[104] XIANGYU PENG, Mark Riedl, and Prithviraj Ammanabrolu. 2022. Inherently Explainable Reinforcement Learning in Natural Language. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=DSEP9rCvZln

[105] Erika Puiutta and Eric M. S. P. Veith. 2020. Explainable Reinforcement Learning: A Survey. In *Machine Learning and Knowledge Extraction*, Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar Weippl (Eds.). Springer International Publishing, Cham, 77–95.

[106] Manjeet Rege and Raymond Blanch K Mbah. 2018. Machine learning for cyber defense and attack. *Data Analytics* 2018 (2018), 83.

[107] Kezhou Ren, Yifan Zeng, Zhiqin Cao, and Yingchao Zhang. 2022. ID-RDRL: a deep reinforcement learning-based feature selection intrusion detection model. *Scientific reports* 12, 1 (2022), 15370.

[108] Danilo J Rezende, Ivo Danihelka, George Papamakarios, Nan Rosemary Ke, Ray Jiang, Theophane Weber, Karol Gregor, Hamza Merzic, Fabio Viola, Jane Wang, et al. 2020. Causally correct partial models for reinforcement learning. *arXiv preprint arXiv:2002.02836* (2020).

[109] Ciaran Roberts, Sy-Toan Ngo, Alexandre Milesi, Sean Peisert, Daniel Arnold, Shammya Saha, Anna Scaglione, Nathan Johnson, Anton Kocheturov, and Dmitriy Fradkin. 2020. Deep reinforcement learning for der cyber-attack mitigation. In *2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, 1–7.

[110] George Rush, Daniel R. Tauritz, and Alexander D. Kent. 2015. Coevolutionary Agent-Based Network Defense Lightweight Event System (CANDLES) *(GECCO Companion '15)*. Association for Computing Machinery, New York, NY, USA, 859–866. doi:10.1145/2739482.2768429

[111] Kevin Schoonover, Eric Michalak, Sean Harris, Adam Gausmann, Hannah Reinbolt, Daniel Tauritz, Chris Rawlings, and Aaron Pope. 2018. Galaxy: A Network Emulation Framework for Cybersecurity.

[112] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[113] Jonathon Schwartz and Hanna Kurniawati. 2019. Autonomous Penetration Testing using Reinforcement Learning. *ArXiv* abs/1905.05965 (2019).

[114] Jonathon Schwartz and Hanna Kurniawatti. 2019. NASim: Network Attack Simulator.

[115] Mohit Sewak, Sanjay K. Sahay, and Hemant Rathore. 2022. Deep Reinforcement Learning for Cybersecurity Threat Detection and Protection: A Review. In *Secure Knowledge Management In The Artificial Intelligence Era*. Springer International Publishing, 51–72. doi:10.1007/978-3-030-97532-6_4

[116] Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. 2018. Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/22722a343513ed45f14905eb07621686-Paper.pdf

[117] Tianmin Shu, Caiming Xiong, and Richard Socher. 2017. Hierarchical and interpretable skill acquisition in multi-task reinforcement learning. *arXiv preprint arXiv:1712.07294* (2017).

[118] Ryan Silva, Cameron Hickert, Nicolas Sarfaraz, Jeff Brush, Josh Silbermann, and Tamim Sookoor. 2022. AlphaSOC: Reinforcement Learning-based Cybersecurity Automation for Cyber-Physical Systems. In *2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPS)*. IEEE, 290–291.

[119] William Spaniel. 2014. *Game theory 101: the complete textbook*. CreateSpace.

[120] Ben Spencer and Steve Cooper. 2021. $10 million to build defence's AI capability and Support Critical Tech for Australia. https://www.minister.defence.gov.au/media-releases/2021-11-18/10-million-build-defences-ai-capability-and-support-critical-tech-australia

[121] Maxwell Standen, Martin Lucas, David Bowman, Toby J Richer, Junae Kim, and Damian Marriott. 2021. Cyborg: A gym for the development of autonomous cyber agents. *arXiv preprint arXiv:2108.09118* (2021).

[122] Madeena Sultana, Adrian Taylor, and Li Li. 2021. Autonomous network cyber offence strategy through deep reinforcement learning. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III*, Vol. 11746. SPIE, 490–502.

[123] Richard S Sutton, Andrew G Barto, et al. 1998. *Reinforcement learning: An introduction*. Vol. 1. MIT press Cambridge.

[124] Jie Tan, Zhaoming Xie, Byron Boots, and C. Karen Liu. 2016. Simulation-based design of dynamic controllers for humanoid balancing. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2729–2736. doi:10.1109/IROS.2016.7759424

[125] Microsoft Defender Research Team. 2021. CyberBattleSim. https://github.com/microsoft/cyberbattlesim. Created by Christian Seifert, Michael Betser, William Blum, James Bono, Kate Farris, Emily Goren, Justin Grana, Kristian Holsheimer, Brandon Marken, Joshua Neil, Nicole Nichols, Jugal Parikh, Haoran Wei..

[126] Ilaria Tiddi and Stefan Schlobach. 2022. Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence* 302 (2022), 103627.

[127] Khuong Tran, Ashlesha Akella, Maxwell Standen, Junae Kim, David Bowman, Toby Richer, and Chin-Teng Lin. 2021. Deep hierarchical reinforcement agents for automated penetration testing. *arXiv preprint arXiv:2109.06449* (2021).
[128] Vladislav D Veksler, Norbou Buchler, Claire G LaFleur, Michael S Yu, Christian Lebiere, and Cleotilde Gonzalez. 2020. Cognitive models in cybersecurity: learning from expert analysts and predicting attacker behavior. *Frontiers in Psychology* 11 (2020), 1049.
[129] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.
[130] Ben Wallace. 2022. Defence Artificial Intelligence Strategy. https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy/defence-artificial-intelligence-strategy
[131] Erich Walter, Kimberly Ferguson-Walter, and Ahmad Ridley. 2021. Incorporating deception into cyberbattlesim for autonomous defense. *arXiv preprint arXiv:2108.13980* (2021).
[132] Shuo Wang, Qingqi Pei, Jianhua Wang, Guangming Tang, Yuchen Zhang, and Xiaohu Liu. 2020. An intelligent deployment policy for deception resources based on reinforcement learning. *IEEE Access* 8 (2020), 35792–35804.
[133] Wenhao Wang, Dingyuanhao Sun, Feng Jiang, Xingguo Chen, and Cheng Zhu. 2022. Research and Challenges of Reinforcement Learning in Cyber Defense Decision-Making for Intranet Security. *Algorithms* 15, 4 (2022), 134.
[134] Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*. 1–10.
[135] Melody Wolk, Andy Applebaum, Camron Dennler, Patrick Dwyer, Marina Moskowitz, Harold Nguyen, Nicole Nichols, Nicole Park, Paul Rachwalski, Frank Rau, et al. 2022. Beyond cage: Investigating generalization of learned autonomous network defense policies. *arXiv preprint arXiv:2211.15557* (2022).
[136] Annie Wong, Thomas Bäck, Anna V Kononova, and Aske Plaat. 2023. Deep multiagent reinforcement learning: Challenges and directions. *Artificial Intelligence Review* 56, 6 (2023), 5023–5056.
[137] Paul L Yu. 2023. Multidisciplinary University Research Initiative: Adversarial and Uncertain Reasoning for Adaptive Cyber Defense (Summary Technical Report, 2013-2021). (2023).
[138] Yinbo Yu, Jiajia Liu, Shouqing Li, Kepu Huang, and Xudong Feng. 2022. A Temporal-Pattern Backdoor Attack to Deep Reinforcement Learning. In *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 2710–2715.
[139] Mattia Zago, Víctor Sánchez, Manuel Pérez, and Gregorio Martinez Perez. 2017. Tackling Cyber Threats with Automatic Decisions and Reactions Based on Machine-Learning Techniques.
[140] Matej Zečević, Devendra Singh Dhami, Petar Veličković, and Kristian Kersting. 2021. Relating graph neural networks to structural causal models. *arXiv preprint arXiv:2109.04173* (2021).
[141] Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. 2020. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 737–744.
[142] Tian-yang Zhou, Yi-chao Zang, Jun-hu Zhu, and Qing-xian Wang. 2019. NIG-AP: A new method for automated penetration testing. *Frontiers of Information Technology & Electronic Engineering* 20, 9 (2019), 1277–1288.
[143] Chen Zhu, W. Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. 2019. Transferable Clean-Label Poisoning Attacks on Deep Neural Nets. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 7614–7623. https://proceedings.mlr.press/v97/zhu19a.html
[144] Saman A. Zonouz, Himanshu Khurana, William H. Sanders, and Timothy M. Yardley. 2009. RRE: A game-theoretic intrusion Response and Recovery Engine. In *2009 IEEE/IFIP International Conference on Dependable Systems Networks*. 439–448. doi:10.1109/DSN.2009.5270307