Automated Cyber Defence: A Review

SANYAM VYAS, Cardiff University, United Kingdom
JOHN HANNAY, Cardiff University, United Kingdom
ANDREW BOLTON, Cardiff University, United Kingdom
PETE BURNAP, Cardiff University, United Kingdom

Within recent times, cybercriminals have curated a variety of organised and resolute cyber attacks within a range of cyber systems, leading to consequential ramifications to private and governmental institutions. Current security-based automation and orchestrations focus on automating fixed purpose and hard-coded solutions, which are easily surpassed by modern-day cyber attacks. Research within Automated Cyber Defence will allow the development and enabling intelligence response by autonomously defending networked systems through sequential decision-making agents. This article comprehensively elaborates the developments within Automated Cyber Defence through a requirement analysis divided into two sub-areas, namely, automated defence and attack agents and Autonomous Cyber Operation (ACO) Gyms. The requirement analysis allows the comparison of automated agents and highlights the importance of ACO Gyms for their continual development. The requirement analysis is also used to critique ACO Gyms with an overall aim to develop them for deploying automated agents within real-world networked systems. Relevant future challenges were addressed from the overall analysis to accelerate development within the area of Automated Cyber Defence.

CCS Concepts: • **Computer systems organization** → Artificial Intelligence; Reinforcement Learning; • **Security and privacy** → Network Security; Malware Mitigation.

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

ACM Reference Format:
Sanyam Vyas, John Hannay, Andrew Bolton, and Pete Burnap. 2023. Automated Cyber Defence: A Review. *Proc. ACM Meas. Anal. Comput. Syst.* 37, 4, Article 111 (February 2023), 32 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Individuals, organisations and governments across the world are facing an exponentially increasing digital involvement within their daily activities and operations. While this has efficiently connected the world together in terms of information awareness and communication, all entities mentioned above are continuously facing cyber attacks from a range of attackers such as opportunists, criminals and hostile states. The exponential growth of such Information Technology (IT) and Operational Technology (OT) devices within homes and industry, along with an increasing skills shortage has led to cybersecurity infrastructures and organisations being overwhelmed in lieu of the increasing amounts of cyber attacks that have occurred. While there is a desperate requirement of skilled cybersecurity practitioners, the level of recent novel and automated cyber attacks surpass the ability of

Authors' addresses: Sanyam Vyas, vyass3@cardiff.ac.uk, Cardiff University, Cardiff, Wales, United Kingdom, CF24 4AG; John Hannay, hannayj1@cardiff.ac.uk, Cardiff University, Cardiff, Wales, United Kingdom, CF24 4AG; Andrew Bolton, boltona2@cardiff.ac.uk, Cardiff University, Cardiff, Wales, United Kingdom, CF24 4AG; Pete Burnap, burnapp@cardiff.ac.uk, Cardiff University, Cardiff, Wales, United Kingdom, CF24 4AG.

2476-1249/2023/2-ART111 $15.00
https://doi.org/XXXXXXX.XXXXXXX

humans manually defending against them [19]. Therefore, there is now a desperate requirement for automated defence solutions [63, 77] to be implemented within IT and OT infrastructures in order to manage such threats against such systems. Consequently, while there have been automated defence solutions within literature, their inability to defend against recent cyber attacks require more research to be conducted to limit the monetary and intellectual property based damages.

As a solution to this problem, we introduce Automated Cyber Defence, an area that focuses on automated decision-making agents for networked systems to mitigate highly complex cyber attacks. This paper defines ACD and analyses literature within different divisions of Automated Cyber Defence. The analysis is conducted through an overall requirement analysis with the vision of focusing on the real-world deployment of such automated decision-making agents within networked systems. Overall, very few publications have highlighted the requirement of automated decision-making agents for defending against cyber-attacks within networked systems. Recent publications include include [61] which provides a detailed review on Reinforcement Learning (RL) solutions for moving target defence, cyber defence and honeypots. The publication also provides a detailed development of RL solutions within cybersecurity through optimal control-theoretic principles and latest AI developments. However, the review does not focus on terrains for the rapid development of such automated decision-making agents for network defence and attack. Wang et al [123] also focus on the development of RL solutions for network defence and attack, along with addressing future challenges similar to [61]. However, the paper did not exhaustively analyse the terrains on which such agents could be developed, which is an imperative part of Automated Cyber Defence. Authors from [22] provide a review on Machine Learning (ML) solutions within cybersecurity, specifically focusing on data sets that accelerate research within intrusion detection systems. While this implementation forms a part of automation approaches within cybersecurity, intrusion detection approaches do not apply to the Automated Cyber Defence term defined in this paper as they do not involve an automated cyber response. Burke et al [23] provide an in-depth review on the type of potential projects within Active Cyber Defence (AcCD), some of which also apply to the term Automated Cyber Defence mentioned in this paper. However, the report does not focus on the ACO gyms for development, and nor does it provide a detailed comparison of automated decision-making algorithms used for automated network attack and defence.

The rest of this article includes section 2 that firstly defines the different terms used frequently in this paper. Section 1 then addresses the methodology utilised to find relevant ACD publications. Subsequently, section 4 then elaborates the curated terminology of Automated Cyber Defence and its differentiation from similar terminologies used within recent literature. The section then provides the importance of the area in National Strategies. Lastly, the section provides a comprehensive Requirement Analysis that will be used to evaluate the selected publications recognised to be as part of Automated Cyber Defence. Section 5 elaborates and critiques on the automated defence and attack (blue and red) agents in custom ACO Gyms through the Requirement Analysis in Section 4. Section 6 elaborates an exhaustive list open-source and closed-source ACO Gyms and assesses them using the Requirement Analysis in Section 4. Section 5 elaborates a list of published automated agents within ACO Gyms and evaluates them using the Requirement Analysis in section 4. Section 8 provides a discussion identifying the challenges and gaps within Automated Cyber Defence literature using the assessments conducted in the previous sections. Lastly, section 9 concludes the article by summarising the area of Automated Cyber Defence. The contributions of this paper include:

- Complete definition of the term Automated Cyber Defence and distinguishes it's research as compared to other related terms.
- Development of a Requirement Analysis for the defined field of Automated Cyber Defence, highlighting the requirements of two important areas within Automated Cyber Defence, namely, the development criteria

of Automated Blue and Red Agents, and the development criteria of ACO Gyms to facilitate Automated Cyber Defence capabilities.
- Assessment of the publications within Automated Cyber Defence literature through the requirement analysis.
- Identification of novel and realistic challenges within the literature to highlight future novel research directions.

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.

Automated 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. [27]. In the context of this paper, automated red teaming refers to an autonomous agent possessing a set of operations (to uncover vulnerabilities and exploitative within the networked system) as their action space. The overall aim of automated red teaming is to ensure the robustness of the automated blue team agent (definition elaborated below) in terms of defending the system against known vulnerabilities and exploits.
Automated Blue Teaming: Blue teaming is a technique responsible for defending a networked systems by maintaining its security posture against a group of mock attackers that aim to exploit gaps and vulnerabilities of the networked system. Typically the Blue Team and its supporters 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, Automated 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 [112]. ACO Gyms aim to address and reduce the 'reality gap' of potential networked systems, used in [114] by combining learning on simulations with testing in a real environment.
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.[73]. In the context of this paper, sequential decision-making algorithms are considered in this paper as Automated Blue and Red Teaming agents due to the complexity of the network that requires navigation before a target action is taken by the automated 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)*.

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

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. [83]
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, which are used to create a computer networked system [83].

3 REVIEW METHODOLOGY

A methodology inspired by [67] was implemented to find all relevant articles for this review. In order to curate the overall ACD definition and the research questions for this article, papers from national and international government institutions and private organisations (mentioned in section 4.1) were utilised. These papers addressed the need for autonomous response solutions in networked systems within a variety of different areas. This allowed us to categorise areas where autonomous response could be utilised within the existing areas of Automated Cyber Defence terminology, specifically, Automated Red and Blue Teaming.

3.1 Research Questions

In order to utilise the ideas suggested for Automated Cyber Defence, research questions were created in order to identify a search strategy. These also allowed us to find relevant literature for this article. The third research question was identified by the search strategy (elaborated in the next subsection), leading to the updating of the overall search strategy.

The research questions (RQs) included:

- **RQ1**: What is Automated Cyber Defence?
- **RQ2**: What are the most suitable algorithmic approaches that have been used within the Automated Cyber Defence terminology defined through **RQ1**?
- **RQ3**: What are the best possible environments in which the most suitable algorithmic approaches could be developed?

3.2 Search Terminology Strategy

After identifying all research questions overall, the next step involved searching for relevant primary studies. Popular digital libraries including IEEE, ACM Digital Library, Springer and Science Direct are utilised along with Google Scholar in order to not overlook significant relevant work. A list of strings grouped within 3 overall themes of Automated Cyber Defence were collectively identified (shown in Table 1). 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:

- **i**: allow us to explore and rank the performance of identified algorithmic families in Automated Blue Teaming (**RQ1, RQ2**).
- **ii**: allow us to explore and rank the performance of identified algorithmic families in Automated Red Teaming (**RG1, 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 Automated Cyber Defence gaining popularity only recently, backward snowballing [65] for several searches was conducted in order to find publications identified as Automated Cyber Defence that were not listed in the search strategy. For example, with areas such as "autonomous cyber operations gym" being a recently

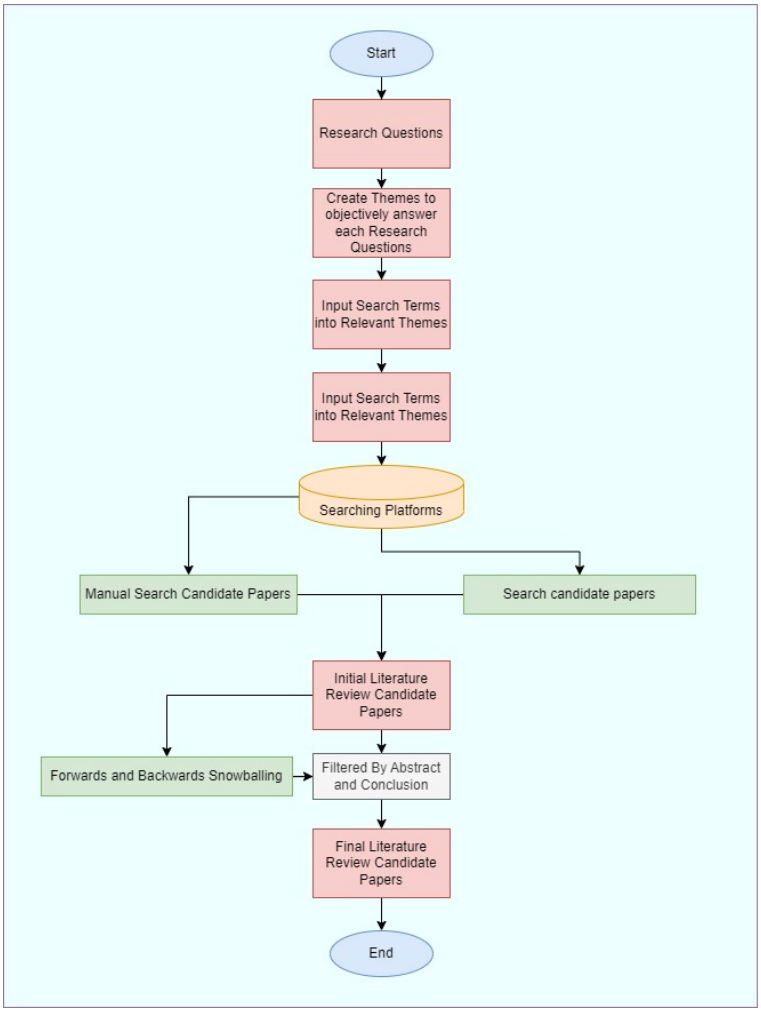

Fig. 1. Research Methodology

created terminology within this area, backward snowballing aided us to identify other popular publications (and 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 Automated Cyber Defence related papers (along with papers highlighting further potential areas within the domain) that cited the publications identified through the search strategy. All papers selected were passed through another screening process based on the papers abstract in order to align the scope of papers based on the definition overall definition of Automated Cyber Defence. Lastly, the remaining papers were then fully read and analyzed for further screening. Figure 1 suggests the overall steps included within this search methodology.

Table 1. Suggests the overarching themes for search terminology. **a.** includes algorithms and terminologies that incorporate an automated response (a requirement for Automated Cyber Defence agents). **b.** includes all terminologies that are a part of the defined Automated Blue Teaming terminology. **c.** includes all terminologies that are a part of the defined Automated Red Teaming terminology.

a. Algorithmic Approaches	b. Automated Blue-Teaming	c. Automated Red-Teaming
- "artificial intelligence"	- "autonomous cyber operations gym"	- "malware"
- "machine learning" OR "deep learning"	- "process killing"	- "process"
- "open ai" 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"

4 AUTOMATED CYBER DEFENCE

Automated Cyber Defence (ACD) 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. In order to define this term, a brief review was completed.

Rege et al [98] provided a high-level description of ACD algorithms as a decision-making system with expert-level ability inspired by how humans reason and learn, citing a publication [16] producing an automated blue agent within a custom networked system. Ko et al [68] provided a terminology for ACD when elaborating the purpose of the Defense Advanced Research Projects Agency (DARPA) grand challenge [2], where it described ACD as systems that are able to self-discover, prove, and correct software vulnerabilities in real-time – without human intervention. In 2016, Baah et al [15] provided a generalised overview of an ACD system. The paper described ACD 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 [16] define the ACD

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

term through their project called Cognitive Support for Intelligent Survivability Management (CSISM), where the authors implement an automated cyber defence decision-making mechanism with expert level ability. The ACD system interprets alerts and observation, 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 automated reasoning and learning through the use of cognitive architectures for cyber defence operations.

Burke et al [23] 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 ACD is the inclusion of Automated Security Planner within AcCD, while ACD strictly focuses automated red and blue-teaming, primarily for the overall development of automated blue teaming agents. Overall, the paper explains that intelligent automation is essential to enable system defenders to manage the risk posed by highly automated future threats and attack, and defend the systems at cyber-relevant national scale. The white paper also elaborated the need for automated red and blue teaming. However, it only provided high-level information on the research directions within the area. 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 [13] highlight the terminology, Autonomous Cyber Defense, within their paper that utilises *autonomous cyber defence* (ACD) agents based on tabular Q-learning. The terminology suggested that autonomous cyber defence is the leveraging of ML techniques to train an agent that is able to autonomously defend a system, minimising self-damage from responses that use noisy sensor data. While this definition aligns to our definition of ACD, the definition is very brief and must be expanded to understand the whole area of ACD that includes a parallel development of Autonomous Cyber Operations Gyms along with automated red and blue team agents. In addition, the defined terminology does not incorporate the role of automated red team agents, which is addressed in this paper as an imperative part of ACD.

The definition of Autonomous Cyber Operations (ACO) will also need to be addressed relative to ACD in order to clarify specific research directions within ACD as compared to ACO. [112] defines ACO as the parallel development of automated red (attacker) and automated blue (defender) agents within a networked system that combat one another in a game-playing scenario. ACD differs from ACO through ACD's focus being on the overall development of automated blue agent through the automated red agents being particularly designed as an automated penetration testing agent that also facilitates adversarial training. The development of ACO Gyms in the lens of ACD also differs to the development of AI ACO Gyms in that they must be designed to specifically for the development of automated blue teaming agents.

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

4.1 Automated Cyber Defence Importance within National Strategy Documents

Private and government-based organisations have made it clear that AI will soon be forefront within cybersecurity in terms of detecting and responding to attacks. Table 2 elaborates the importance of ACD within different countries.

Table 2. National Strategy Paper's on ACD

Country/Alliance	Department/Strategy	Reference to ACD
Australia	Department of Defence [111]	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 [9]	Aims to develop 30 more AI-based challenges for researchers, including the TTCP CAGE Autonomous Cyber Defence Challenge to produce AI-based automated 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 [120]	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 [90]	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 automated response to threats"
NATO	Cooperative Cyber Defence Centre of Excellence [85]	Suggest the need for Nation States to adopt and explore AI-enabled Cyber Defence.
	NATO AI Strategy [84]	The strategy includes "collaboration on AI technologies for Cyber Defence.

4.2 Automated Cyber Defence Requirements

The North Atlantic Treaty Organisation (NATO) outlined requirements for Autonomous Cyber Agents by producing a reference architecture and technical roadmap, AICA [69]. A specific part of the document focuses on the strategic deployment and the ethical concerns on the battlefield of autonomous agents. The key points in AICA relevant for this paper have been included along with additional ACD requirements in a summarised requirement analysis for ACD in Table 3 below. The table further elaborates essential requirements of automated red and blue agents **(A)** along with ACO Gyms requirements **(G)** which could allow the usage of automated red and blue agents. The requirements within this table should act as a checklist for researchers within ACD, allowing for the development of eventual deployment of ACD operations within real-world networked systems.

Table 3. Requirement for ACD

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 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 without configuration issues - **(A.1.1)** Automated agent will need to generalise their decisions relevant to the agent type it represents - **(A.1.2)** Automated agent will have to generalise and adapt to structural changes within the ACO Gyms (addition and removal of subnets and endpoints) - **(A.1.3)** Automated red and blue agents must be designed to sustain their high performance from simulation to real-world.
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 automated 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 ACD. - **(A.2.3)** The designed automated agents will require a "deep" architecture to sustain the complexity of the ACO Gyms - **(A.2.4)** Additionally, agents will need to be able to be explainable [21, 66, 97], 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 [69] 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 automated agents and for action/strategy negotiation. [3]. AICA combined with a MARL survey produced by [125], 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 [23] that coincides with ACD. 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 automated 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 automated red agent to adversarially train the automated blue agent to reduce the number of incorrect actions - **(A.6.1)** To improve performance of automated blue team agent (the sole purpose of ACD), adversarial training through an automated red agent must be encouraged. - **(A.6.2)** Automated red agents must be provided with a wide variety of cyber attacks (specified within the MITRE ATT&CK framework) - **(A.6.3)** along with a variety of algorithmic attacks [59] to address systems vulnerabilities. - **(A.6.4)** Automated blue and red agent must be able to launch deception defence and attacks respectively.

5 ACD ALGORITHMS USED WITHIN CUSTOM ACO GYMS

As mentioned in the section 4.2, a typical ACD system comprises of a mode of terrain i.e. networked system, which possesses the provision to allow automated red and blue team game-playing scenarios. Recent publications within ACD have utilised automated decision-making algorithms such as Game Theory (GT), Machine Learning (ML) and Reinforcement Learning (RL) for automated blue and red teaming within custom ACO Gyms.

ML-based solutions (along with RL-based solutions [87, 99, 107]) have also been utilised solely for quick incident and intrusion response over the years [50, 89, 118]. Specifically, Zago et al [126] 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 do not cater to the rapid response of zero-day attacks due to their single-step response characteristic. 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 [24] 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 ACD-based solutions within networked systems. Therefore, the publications selected for this section focus on sequential response that is required for automated agent to stop cyber attacks within an overall networked system.

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

5.1 Automated Blue Team Solutions

The automated 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 assymmetric scenario between cyber attackers and defenders, the defender with limited resources cannot afford to prepare for all possible attacks.

The problem area in focus for this subsection is the mitigation of Posture-related vulnerabilities (PrV) i.e. the defender 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 within the networked system. Due to this disadvantage in security posture, the defender with limited resources cannot afford to prepare for all possible attacks.

Table 4 below evaluates the automated blue teaming solutions along with their custom ACO Gyms that were published within literature.

Table 4. Automated Blue Team Solutions within custom networked systems

	Automated Blue Team Custom Networked System Publications														
Requirements	[131]	[18]	[60]	[88]	[78]	[46]	[37]	[25]	[29]	[24]	[122]	[47]	[121]	[110]	[101]
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			+	+			+		+						

Table 4 shows relevant ACD automated blue teaming publications within networked systems designed solely for their respective automated 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 [24], 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 a MDP/POMDP (G.2.2), which allows automated agents to take sequential response through the transitioning of

states, that signify 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 Deep RL (DRL) implementations were able to fill this requirement, making them more suitable. Dhir et al [35] also suggested the use of Causal Inference algorithms [48, 64, 94, 100, 127] 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 primal for utilisation of any automated agents within 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 automated 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 automated blue agent through adversarial training against a variety of cyber attacks. Moreover, the lack of implementations that fill the A.4.1 requirement also hinders the development of automated blue agents against algorithmic attacks mentioned in A.4.3, an area in which no publications highlighted in Table 4 have concentrated on.

Requirement A.6.4 in the context of automated 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 slowdown 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 [122] 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 banding together of 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 ACD, 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 [71] 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 [121] draws attention to the prospect of augmenting ACD environments with defensive cyber deception components by adapting the source code of an existing open-source ACO Gym called CyberBattleSim [115]. 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 automated 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 automated 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.

5.2 Automated Red Team Solutions

The existing literature on automated 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 [1] or Nessus [2] to design a POMDP (G.2.2) representing a corporate network. The Common Vulnerability Scoring System (CVSS) scores [3] 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 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 [28] 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 [4] offer attack path planning focusing on Active Directory weaknesses, without utilising ML.

To automate penetration testing, one can extend the RL game defined in the paragraph above to incorporate actions of penetration testing or red teaming tools (A.6.2). In fact, [129] did so to automate penetration testing with the Metasploit framework [5], whereas [76] utilised the PowerShell Empire framework [6] to automate post exploitation activities. Furthermore, researchers have analysed specific tasks of red teaming and attempted to automate them. For example, [70] automated privilege escalation through RL. One could envision multiple cells of the MITRE ATT&CK matrix [7] being automated in this fashion, such as defence evasion as seen in [39].

Given that research into RL for automated 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. For example, [113] build Deep RL agents in the Network Attack Simulator Gym [106]. 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 [117] did so in the CyBORG Gym environment [112] where they proposed a Hierarchical DQN algorithm. Research in the open-source gyms are summarised in 8.

Finally, it should be noted that GT Models (A.2.2) have also been explored (an example is provided by [30]), but in this case they are utilised to aid decision makers, such as in cyber war gaming.

Table 5. Automated Red Team solutions within custom networked systems

	Automated Red Team Custom Networked System Publications		
Requirement	[76]	[70]	[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			

6 AUTONOMOUS CYBER OPERATIONS GYM

As shown in the previous section, the lack of common open-source ACO Gyms prevent the possibility for a separate accelerated development of automated blue and red agents (and ACO Gyms). This section provides a detailed overview of literature that have recently developed ACO Gyms along with the automated agents developed and published within literature and websites. Such ACO Gyms are simulated and/or emulated networked systems designed specifically for the development of automated 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 possible.

6.1 Training strategies

The most common approach to train and test an agent involves validating its policies on the same environment in which it was trained. This applies to both simulated and emulated environments. Unfortunately, this strategy prevents making statements on the automated agent's ability to generalise (i.e. meet requirement A.1.1, A.1.2 and in Table 3). Additionally, the automated agent will not be able to fully utilise the benefits of using different types of environments (i.e. simulation for scalability and emulation to delve closer to realism) and meet requirement G.1.3.

Several research papers have strode to make progress in the domain of generalisation. For example, [113] built Deep RL agents in the Network Attack Simulator Gym [106]; a simulated environment to conduct research in automated penetration testing. Automated 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 automated agents on scenarios of lower complexity, the impact on performance in larger complexity scenarios was experimented with, where the PPO provided superior generalisation. The cutting-edge platforms built to conduct research in ACD designed by [83], [112] or [72] 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 automated 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, [70] enumerated all possible privilege escalation techniques from the MITRE ATT&CK matrix [7] 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 automated agent outperformed human experts.

6.2 Existing Autonomous Cyber Operations Gyms

For the acceleration of research within the domain of automated 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 automated 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 environments that are simulations, and then delves into other closed-source emulated (and other simulated) environments that have been published. Each part compares ACO Gyms amongst the other open-source/closed-source ACO Gyms using the requirement analysis for ACO Gyms.

6.2.1 Open-source Gyms. Firstly, The Cyber Battle Sim [115] (CBS) environment is created for training automated 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 access 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 "blue agent" training. In fact, [121] have done so to incorporate blue team deception into the environment. The developers ensured sufficient complexity exists in the environment to abstract the cells of the MITRE ATT&CK matrix [7] 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, thereby 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.

- A reconnaissance action
- Or an attack of type 1...m

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

- A monitoring action
- Or a 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 [79] 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.

The Network Attack Simulator environment [106], is purely built for training 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. While this implementation provides the ability to bolster research progress within AI-based red agent training, it to only provides simulations like CBS, Gym IDS Game and Gym Threat Defence, reducing the realism of the implementation.

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 CyBORG environment [112] 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 a 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 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 automated agent is trained, it is ready tested in the emulator, which comprises of AWS virtual machines to create a high fidelity cybersecurity environment in which the automated agent interacts with. The purpose of the environment is to act as a platform for research in ACD, whereby challenges are open to the public. Namely, the TTCP Cage Challenge 1, 2 and 3. The challenges are enterprise network environments with ascending complexity (in terms of the observation and action space for the red and blue agent).

In the TTCP CAGE challenge 2 (the most recent challenge), 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. This environment appears to fit all our needs for experimentation and details the high-level desired actions of an autonomous blue agent. On top of this simulated environment CybORG extends to an emulation (which is closed source), which can be spun up on AWS to validate the trained agents.

YAWNING TITAN [31] 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.

Table 6. ACO Gyms (Open-source)

	Automated Cyber Operations Gym (Open-source)						
Requirement	CBS [115]	GIG	GTD	OIR	CybORG [112]	NaSim	YT [11]
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							

6.2.2 Closed-source Gyms. The rest of the ACO Gyms have been analysed in Table 7 through the requirement analysis shown in Table 3. While the ACO Gyms highlighted are not open-source, they can provide important insights within the ACD community, particularly for researchers who can take inspiration when designing or making modifications to the existing ACO gyms.

Table 7. ACO Gyms (Closed-source)

	Automated Cyber Operations Gym (Closed-source)								
Requirement	[44]	[81]	[43]	[20]	[102]	[103]	[72]	[83]	[38]
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					+	+		+	

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 automated 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 automated red and blue agents (A.2.1 and A.2.2 respectively), and have structured the ACO Gym as a MDP in order to facilitate such agents. However, while such ACO Gyms are highly scalable (G.1.4) and allow the development of relevant automated agents, the environments utilised in all implementations are simulations to 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 automated agents, which will essential for utilisation within current networked systems.

While the rest of the analysis apply to those of automated agents, the design of the current state of the ACO Gyms could be used to assess the quality of automated agents that could be designed within the ACO Gyms. Overall, only one ACO Gym (CybORG [112] Cage Challenge 3 [53]) has recognised the need for automated multi-agent algorithms (A.4.1) as automated blue team solutions. [78] and [37] publications (specifically focusing on using RL for defending against DDoS attacks) environments could be a potential inspiration for structuring the ACO Gyms to facilitate multi-agent automated red and blue teaming collaboration (requirement G.4.1). Very few ACO Gyms facilitate adversarial training (G.6.1 and A.6.1), which could potentially utilised to strengthen the automated blue agent against a variety of cyber attacks (A.6.2). No open-source ACO Gyms currently available have recognised the need of incorporating algorithmic cyber attacks (A.6.3) within the action space of automated red agents against automated blue agents. Inspiration can be taken from a closed-source ACO Gym [83] to incorporate algorithmic attacks such as evasion and poising of automated agents such as DRL algorithms.

7 ACD ALGORITHMS WITHIN OPEN-SOURCE ACO GYMS

Out of the open-source ACO Gyms mentioned in the previous section, several automated decision-making algorithm's have been utilised for training and testing as automated agents. The ACO Gym creators and automated 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 automated agents through currently published implementations. Such gaps will require being met before the algorithms can be deployed into real-world networked systems for cybersecurity. Out of the current ACO Gyms, only two open-source ACO Gyms have been utilised in the publications of automated 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 [112] released three 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 automated blue agents, while the development of automated red agents (comprising of two different types of cyber attacks) is also possible. NaSim [106] authors made their code open-source for the development of automated red agents and a few publications and implementations have utilised the simulated networks for the development of such agents.

7.1 Automated Blue Team Solutions

Out of the two ACO gyms mentioned above, CybORG has published its results for the challenges [8] released, and has listed and ranked the RL-based algorithms that were used in Cage Challenge 1 [52] and Cage Challenge 2 [51] (Cage Challenge 3 results will be soon released [53]) through performance metrics set by the authors. Several approaches taken by different teams, and multiple unique strategies that were implemented by the automated agents. The best performing approaches across the challenges have been selected in this article and have been compared against the requirement analysis in Table 3.

From Cage Challenge 1, Team Mindrake [40] won the challenge and produced a Hierarchical RL algorithm that included proximal policy optimisation [104] with curiosity. The hierarchical [56] component of the algorithm is utilised through a controller to take relevant action according to the type of adversary that is deployed against the automated 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 [92], improving the reward achieved by nearly double. While the automated 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 [52] 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 [4]) 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 requirement analysis, the automated 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.

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

Overall, as shown in both challenges, variations of hierarchical PPO agents have shown most optimal performance (also suggested and algorithmically proven in [124]) as compared to other approaches. While the automated 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 and algorithmic attacks (A.6.2, A.6.3) for the automated 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 automated agents against a greater pool of attacks. In contrast, no automated agent implementations in both challenges provided any form of explainability (A.2.4) regarding their incoming actions that they will take.

7.2 Automated Red Team Solutions

Unfortunately, unlike for the Automated 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.

Automated Red Teaming Solutions have so far largely been performed through Reinforcement Learning in ACD gym environments such as CyBORG [112], Network Attack Simulator [106] and CyberBattleSim [115], 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 Automated 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 [112] 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 automated agents in the CyBORG emulator (G.1.3). Most of the automated 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 [113]. 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 Automated Red Teaming Solutions, even though these have recently been shown to be promising for the Blue Teaming side [12].

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

	Automated Red Team			
Papers	[117]	[112]	[86]	[113]
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

8 DISCUSSION

This main purpose of this paper was identify an imminent research area, ACD, within cybersecurity in order to mitigate cyber attacks in the future. Automated response to cyber attacks will need to be addressed through the research and development of automated 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 through the requirement analysis (Table 3) in this paper for the future development within the mentioned areas. Over 40 publications were analysed and compared through the requirement analysis in Table 3. While development of ACO Gyms and automated 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 within the specific requirement addressed, it is encouraged for researchers to build on this document to further address and develop areas within ACD that could further catalyse it's development into industrial use.

8.1 Challenges and their Importance

The direct mapping of the requirement analysis in Table 3 to the publications identified as ACD has addressed that there are evident challenges that need to be filled for ACD systems to be further developed before they are implemented into real-world systems. This section outlines the areas of further research and development that were identified, and links the areas back to the specific requirements within the requirement analysis. Requirements have been added for each challenge in a descending order of importance.

8.1.1 AI-based Attack Robustification of Automated Blue Agents (A.6.3, G.6.1, A.6.1). This focuses on the area includes the robustification of Deep RL algorithms against poisoning and evasion attacks that aim to attack the algorithmic functionalities of the automated agent. While very low number of publications have focused on such attacks for Deep RL algorithms, it is evident that the future cyber attackers will implement such attacks in the future through Deep RL and neural network based research within other domains [14, 26, 108, 130]. If this challenge is not addressed, future networked systems could be vulnerable to algorithmic attacks that could potentially take control of the automated blue agent, and eventually the entire network.

8.1.2 Continual evolution of action space for the Automated Red Agents (A.3.1, G.6.1, A.6.1, A.6.2, A.6.3). 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 automated penetration testing agents will need to be able to handle a large growing database of exploits." [105] Therefore the red agents and the gyms they are trained in would need to consider this challenge (G.1.1). While this challenge is reliant on challenges *7.1.1 and 7.1.5*, the development and addition of cyber attacks automated red agents based within a continual learning setting are yet to be explored. Failure to implement on these challenges will keep the automated blue agent agent outdated from latest cyber and algorithmic attacks.

8.1.3 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). Further inspiration could be taken from relevant survey papers and implementations [10, 49, 74, 75, 80, 82, 91, 95, 96, 109, 116]. Failure to address this challenge will lead to the automated blue agent not being certified by industrial employees within networked systems since the trust towards the agent will be low.

8.1.4 Multi-agent RL (G.4.1). Another research area within automated blue teaming for ACD is the utilisation of Multi-agent RL algorithms as opposed to using single RL algorithms for implementation. This will be particularly more beneficial within enterprise networks environments which are highly complex. While [112] authors have proposed the implementation of multi-agent RL within their third Cage Challenge [5], more research areas could emerge with more research within this domain. Using single automated 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.

8.1.5 Cybersecurity Attack Robustification of Automated Blue Agents (A.6.2, A.6.1). An area for improvement for future automated blue agents (and ACO Gyms) is the implementation of more types of cyber-attacks that could occur within an enterprise network. A useful framework for this could be the use of different cyber-attacks that have been listed within the MITRE ATT&CK matrix. Similar to software updates and patches, systems could be designed in such a way so that more attacks could be added to a knowledge base once they are listed within

[5] https://github.com/cage-challenge/cage-challenge-3

frameworks like MITRE ATT&CK matrix [6]. Failure to address this issue within training will lead to the automated blue teaming agent ignoring specific cyber attacks, eventually leading to a breach within the network.

8.1.6 Robustification of Deception Techniques in Automated Blue Agents (A.6.4). It's also important to highlight the necessity for research areas which utilise deception technology for ACD 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. Existing literature rarely considers the complexity of this challenge, underlining the infancy of deception as a tool for ACD. Research that falls into this category [47, 121, 122] typically prioritise the use of honey-x methods [93] or 'lures' to analyse adversary behaviours through intelligent deployment strategies. A useful framework to encourage diversity within deceptive assets is the MITRE ENGAGE matrix, which identifies numerous deception techniques that can be leveraged at different areas of ACD to optimise adversary engagement [7]. 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's ever-present between blue and red agents 5.

8.1.7 Realism of ACO Gyms (G.1.3, A.3.1, A.3.2, G.1.4, G.1.1, G.1.2). Another challenge 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) 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 [128]. Environments such as CyBORG [112] 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 unlike most RL tasks, 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 [17, 62, 119]. Such issues must be addressed, else the agent will not recognise the mode of terrain it is operated within, leading to incorrect actions being taken by the agent.

8.1.8 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 Operational Technology (OT) environments, where researchers must find methods to emulate devices in convincing ways without replicating the network in its entirety. There needs to be new methods for creating decoy profiles for assets which embody the attributes of the network component. Researchers can also look to deceptive techniques which already consider or enhance the fidelity of integrated-lures. 'Honeyshills' [57] 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 problems 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.

[6] https://attack.mitre.org/matrices/enterprise/
[7] https://engage.mitre.org/

8.1.9 Impact of Incorrect Action (G.6.1, G.1.3) [41]. The above issue also leads on a gap within the ACD literature for automated decision-making agents. Appropriate evaluation and metrics will need to be explored. Additionally, approaches such as Hierarchical RL, Neurosymbolic RL and other explainable implementations [58, 82, 109] must be explored for superior forensic evaluation of the automated agents. Not addressing this area will result in the automated blue team RL agent potentially eliminating important processes within the network, which could even lead to high monetary losses.

8.1.10 Action and Observation Spaces (G.2.1 ,G.2.2, A.2.3). The first difficulty relates to the huge action and observation spaces: Existing research in ACD 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 Cyber Security setting where agents may be deployed on thousands of hosts (in a single corporate network), each with huge action sets (kill any process, re-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 automated 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 PT can have very different effects such as attacking hosts in different subnets or different method of exploits" [117].

9 CONCLUSION

This article provided awareness to the area of Automated Cyber Defence by defining its terminology through research publications, government strategy reports and cybersecurity training organisations. Subsequently, the terminology allowed the segmentation of different sub-topics of research and development that exist within the area, namely, Automated Agents and Autonomous Cyber Operations (ACO) Gyms. The recognition of the sub-topics allowed the creation of a Requirement Analysis that is used as a metric to assess the publications that were recognised to be a part of the Automated Cyber Defence (ACD) literature. Through an extensive review of existing literature within automated blue and red teaming algorithms within custom ACO Gyms, it was discovered that Deep Reinforcement Learning (DRL) solutions were the most optimal algorithms for automated blue and red teaming as compared to Game Theoretic and Machine Learning solutions. This was primarily due to their ability to take sequential response for long-term and short-term goals. In addition, the review suggested the need for a parallel research and development of automated (blue and red) agents and ACO Gyms in order to accelerate research in both domains respectively. An extensive review was also conducted on existing open and closed-source gyms along with automated red and blue teaming implementations within them. The publications and implementations were assessed through the requirement analysis to find areas of further development within the literature. Through the requirement analysis of all publications of automated agents and ACO Gyms, specific challenges and gaps were discovered and elaborated. The challenges for the area of ACD included research within cybersecurity attack robustification of automated blue agents, realism of ACO Gyms, impact of incorrect actions and action/observation spaces within ACO Gyms. In contrast, the gaps included research within algorithmic attacks on the automated blue agent, explainable RL as automated blue agents and multi-agent automated blue agents. The aim of the challenges and gaps address the areas of future research and development within ACD in order for the transition of automated blue agents from ACO Gyms to networked systems deployed in the real world.

REFERENCES

[1] 2022. *Nmap* (2022). https://nmap.org/
[2] 2022. *Nessus* (2022). https://www.tenable.com/products/nessus
[3] 2022. *CVSS* (2022). https://nvd.nist.gov/vuln-metrics/cvss
[4] 2022. *Bloodhound* (2022). https://github.com/BloodHoundAD/BloodHound

[5] 2022. *Metasploit* (2022). https://www.metasploit.com/

[6] 2022. *Empire* (2022). https://github.com/EmpireProject/Empire

[7] 2022. *Mitre* (2022). https://attack.mitre.org/matrices/enterprise/

[8] 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.

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

[10] Prithviraj Ammanabrolu and Mark Riedl. 2019. Playing Text-Adventure Games with Graph-Based Deep Reinforcement Learning. 3557–3565. https://doi.org/10.18653/v1/N19-1358

[11] Alex Andrew, Sam Spillard, Joshua Collyer, and Neil Dhir. 2022. Developing Optimal Causal Cyber-Defence Agents via Cyber Security Simulation. In *Workshop on Machine Learning for Cybersecurity (ML4Cyber)*.

[12] 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).

[13] 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.

[14] 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. https://doi.org/10.1109/TNSM.2020.3031843

[15] George K. Baah, Thomas Hobson, Hamad Okhravi, Shannon C. Roberts, William W. Streilein, and Sophia Yuditskaya. 2015. A Study of Gaps in Cyber Defense Automation.

[16] 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. https://doi.org/10.1109/BLISS.2008.17

[17] 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).

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

[19] Robert A Bridges, Ashley E Rice, Sean Oesch, Jeff A Nichols, Cory Watson, Kevin Spakes, Savannah Norem, Mike Huettel, Brian Jewell, Brian Weber, et al. 2022. Testing SOAR Tools in Use. *arXiv preprint arXiv:2208.06075* (2022).

[20] 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. https://doi.org/10.1109/MILCOM.2016.7795375

[21] 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).

[22] 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.

[23] 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

[24] 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. https://doi.org/10.1117/12.2559319

[25] 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. https://doi.org/10.1109/DSC50466.2020.00065

[26] 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. https://doi.org/10.1109/ACCESS.2021.3068768

[27] 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. https://doi.org/10.1145/1276958.1277345

[28] 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.

[29] Ankur Chowdhary, Dijiang Huang, Abdulhakim Sabur, Neha Vadnere, Myong Kang, and Bruce Montrose. 2021. SDN-based Moving Target Defense using Multi-agent Reinforcement Learning. Autonomous Intelligent Cyber-defence Agents Conference.

[30] 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.
[31] 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.
[32] Richard Dazeley, Peter Vamplew, and Francisco Cruz. 2021. Explainable reinforcement learning for broad-xai: A conceptual framework and survey. *arXiv preprint arXiv:2108.09003* (2021).
[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] Susannah Kate Devitt and Damian Copeland. 2021. Australia's Approach to AI Governance in Security and Defence. https://doi.org/10.48550/ARXIV.2112.01252
[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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1145/3488932.3527286
[41] Making AI Work for Cyber Defense: The Accuracy-Robustness Tradeoff. 2021. https://doi.org/10.51593/2021CA007
[42] Keith Frankish and William Ramsey. 2014. *The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press. https://doi.org/10.1017/CBO9781139046855
[43] Angelo Furfaro, Antonio Piccolo, Andrea Parise, Luciano Argento, and Domenico Saccà. 2018. A Cloud-based platform for the emulation of complex cybersecurity scenarios. *Future Generation Computer Systems* 89 (2018), 791–803. https://doi.org/10.1016/j.future.2018.07.025
[44] Ariel Futoransky, Fernando Miranda, José Orlicki, and Carlos Sarraute. 2009. Simulating cyber-attacks for fun and profit. *Computing Research Repository - CORR*, 4. https://doi.org/10.1145/1537614.1537620
[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. https://doi.org/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 (10 2021). https://doi.org/10.1155/2021/6378218
[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. 2021. A Survey on Interpretable Reinforcement Learning. *arXiv preprint arXiv:2112.13112* (2021).
[50] Ross Gore, Saikou Diallo, Jose Padilla, and Barry Ezell. 2018. Assessing cyber-incidents using machine learning. *International Journal of Information and Computer Security* 10 (01 2018), 341. https://doi.org/10.1504/IJICS.2018.095298
[51] TTCP Cage Working Grouo. 2022. TTCP CAGE Challenge 2. https://github.com/cage-challenge/cage-challenge-2.
[52] TTCP CAGE Working Group. 2021. CAGE Challenge 1. arXiv.
[53] TTCP CAGE Working Group. 2022. TTCP CAGE Challenge 3. https://github.com/cage-challenge/cage-challenge-3.
[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] Bernhard Hengst. 2010. *Hierarchical Reinforcement Learning*. Springer US, Boston, MA, 495–502. https://doi.org/10.1007/978-0-387-30164-8_363
[57] 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.

[58] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. https://doi.org/10.48550/ARXIV.1812.04608

[59] Wyatt Hoffman. 2021. Making AI Work for Cyber Defense. (2021).

[60] 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. https://doi.org/10.1145/3305218.3305239

[61] Yunhan Huang, Linan Huang, and Quanyan Zhu. 2021. Reinforcement Learning for Feedback-Enabled Cyber Resilience.

[62] Yunhan Huang and Quanyan Zhu. 2019. Deceptive Reinforcement Learning Under Adversarial Manipulations on Cost Signals. In *GameSec*.

[63] 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. https://doi.org/10.32604/iasc.2021.016240

[64] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J. Kusner, and Ricardo Silva. 2022. Causal Machine Learning: A Survey and Open Problems. https://doi.org/10.48550/ARXIV.2206.15475

[65] 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.

[66] 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

[67] Barbara Kitchenham. 2004. Procedures for Performing Systematic Reviews. *Keele, UK, Keele Univ.* 33 (08 2004).

[68] 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.

[69] Alexander Kott, Paul Théron, Martin Drašar, Edlira Dushku, Benoît LeBlanc, Paul Losiewicz, Alessandro Guarino, Luigi Mancini, Agostino Panico, Mauno Pihelgas, et al. 2018. Autonomous intelligent cyber-defense agent (aica) reference architecture. release 2.0. *arXiv preprint arXiv:1803.10664* (2018).

[70] 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.

[71] 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. https://doi.org/10.1049/ise2.12050 arXiv:https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/ise2.12050

[72] 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).

[73] Michael Littman. 2009. Algorithms for Sequential Decision Making. (08 2009).

[74] Daoming Lyu, Fangkai Yang, Bo Liu, and Steven Gustafson. 2019. SDRL: Interpretable and Data-Efficient Deep Reinforcement Learning Leveraging Symbolic Planning. *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (07 2019), 2970–2977. https://doi.org/10.1609/aaai.v33i01.33012970

[75] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. Explainable Reinforcement Learning Through a Causal Lens. https://doi.org/10.48550/ARXIV.1905.10958

[76] Ryusei Maeda and Mamoru Mimura. 2021. Automating post-exploitation with deep reinforcement learning. *Computers & Security* 100 (2021), 102108.

[77] 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.

[78] Kleanthis Malialis and Daniel Kudenko. 2013. Large-Scale DDoS Response Using Cooperative Reinforcement Learning.

[79] 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.

[80] Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. 2022. A Survey of Explainable Reinforcement Learning. *arXiv preprint arXiv:2202.08434* (2022).

[81] 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)*. 1–7. https://doi.org/10.1109/THS.2010.5655108

[82] 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.

[83] Andres Molina-Markham, Cory Miniter, Becky Powell, and Ahmad Ridley. 2021. Network Environment Design for Autonomous Cyberdefense. *ArXiv* abs/2103.07583 (2021).

[84] NATO. 2021. Artificial Intelligence and Autonomy in the Military. https://ccdcoe.org/uploads/2021/12/Strategies_and_Deployment_A4.pdf

[85] NATO. 2022. Cooperative Cyber Defence Centre of Excellence. https://ccdcoe.org/library/publications/
[86] 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.
[87] Thanh Thi Nguyen and Vijay Janapa Reddi. 2021. Deep reinforcement learning for cyber security. *IEEE Transactions on Neural Networks and Learning Systems* (2021).
[88] 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. https://doi.org/10.1109/TNNLS.2018.2885530
[89] 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. https://doi.org/10.1109/COMM48946.2020.9141989
[90] Cabinet Office. 2022. Government Cyber Security Strategyy. https://www.gov.uk/government/publications/government-cyber-security-strategy-2022-to-2030
[91] 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.
[92] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. 2017. Curiosity-driven Exploration by Self-supervised Prediction. https://doi.org/10.48550/ARXIV.1705.05363
[93] Jeffrey Pawlick, Edward Colbert, and Quanyan Zhu. 2017. A Game-Theoretic Taxonomy and Survey of Defensive Deception for Cybersecurity and Privacy. https://doi.org/10.48550/ARXIV.1712.05441
[94] Shaohui Peng, Xing Hu, Rui Zhang, Ke Tang, Jiaming Guo, Qi Yi, Ruizhi Chen, Xishan Zhang, Zidong Du, Ling Li, Qi Guo, and Yunji Chen. 2022. Causality-driven Hierarchical Structure Discovery for Reinforcement Learning. https://doi.org/10.48550/ARXIV.2210.06964
[95] 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
[96] 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.
[97] Hamon R, Junklewitz H, and Sanchez Martin JI. 2020. Robustness and Explainability of Artificial Intelligence. KJ-NA-30040-EN-N (online) (2020). https://doi.org/10.2760/57493(online)
[98] Manjeet Rege and Raymond Blanch K Mbah. 2018. Machine learning for cyber defense and attack. *Data Analytics* 2018 (2018), 83.
[99] 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 (09 2022). https://doi.org/10.1038/s41598-022-19366-3
[100] Danilo J. Rezende, Ivo Danihelka, George Papamakarios, Nan Rosemary Ke, Ray Jiang, Theophane Weber, Karol Gregor, Hamza Merzic, Fabio Viola, Jane Wang, Jovana Mitrovic, Frederic Besse, Ioannis Antonoglou, and Lars Buesing. 2020. Causally Correct Partial Models for Reinforcement Learning. https://arxiv.org/abs/2002.02836
[101] 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.
[102] 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. https://doi.org/10.1145/2739482.2768429
[103] 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.
[104] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. https://doi.org/10.48550/ARXIV.1707.06347
[105] Jonathon Schwartz and Hanna Kurniawati. 2019. Autonomous Penetration Testing using Reinforcement Learning. *ArXiv* abs/1905.05965 (2019).
[106] Jonathon Schwartz and Hanna Kurniawatti. 2019. NASim: Network Attack Simulator. (2019).
[107] 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. https://doi.org/10.1007/978-3-030-97532-6_4
[108] 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
[109] Tianmin Shu, Caiming Xiong, and Richard Socher. 2017. Hierarchical and interpretable skill acquisition in multi-task reinforcement learning. *arXiv preprint arXiv:1712.07294* (2017).

[110] 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.
[111] Ben Spencer and Steve Cooper. [n. d.].
[112] 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:2108.09118 [cs.CR]
[113] 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.
[114] 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. https://doi.org/10.1109/IROS.2016.7759424
[115] Microsoft Defender Research Team. 2021. (2021).
[116] Ilaria Tiddi and Stefan Schlobach. 2022. Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence* 302 (2022), 103627. https://doi.org/10.1016/j.artint.2021.103627
[117] 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:2109.06449 [cs.AI]
[118] 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). https://doi.org/10.3389/fpsyg.2020.01049
[119] 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.
[120] Ben Wallace. 2022. Defence Artificial Intelligence Strategy. https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy/defence-artificial-intelligence-strategy
[121] Erich Walter, Kimberly Ferguson-Walter, and Ahmad Ridley. 2021. Incorporating deception into cyberbattlesim for autonomous defense. *arXiv preprint arXiv:2108.13980* (2021).
[122] 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. https://doi.org/10.1109/ACCESS.2020.2974786
[123] 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.
[124] Melody Wolk, Andy Applebaum, Camron Denver, 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).
[125] Annie Wong, Thomas Bäck, Anna V Kononova, and Aske Plaat. 2021. Multiagent deep reinforcement learning: Challenges and directions towards human-like approaches. *arXiv preprint arXiv:2106.15691* (2021).
[126] Mattia Zago, Víctor Sánchez, Manuel Pérez, and Gregorio Martinez Perez. 2017. Tackling Cyber Threats with Automatic Decisions and Reactions Based on Machine-Learning Techniques.
[127] Matej Zečević, Devendra Singh Dhami, Petar Veličković, and Kristian Kersting. 2021. Relating Graph Neural Networks to Structural Causal Models. https://doi.org/10.48550/ARXIV.2109.04173
[128] 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.
[129] 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.
[130] 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
[131] 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. https://doi.org/10.1109/DSN.2009.5270307