Reinforcement Learning for Autonomous Resilient Cyber Defence[1]

Ian Miles[2], Sara Farmer[3], David Foster[4], Daniel Harrold[5], Gregory Palmer[5], Chris Parry[5], Chris Willis[5], Marco Casassa Mont[6], Lisa Gralewski[6], Ryan Menzies[6], Neela Morarji[6], Esin Turkbeyler[6], Alec Wilson[6], Alfie Beard[7], Pedro Marques[7], Jonathan Francis Roscoe[7], Samuel Bailey[8], Madeline Cheah[8], Mark Dorn[8], Peter Haubrick[8], Matthew Lacey[8], David Rimmer[8], Jack Stone[8], Demian Till[8] , Ryan Heartfield[9], Andy Harrison[10], James Short[10], Tom Wilson[11], John H[12]

Abstract

Future cyber threats will include high volumes of sophisticated machine speed cyber-attacks that are able to evade and overwhelm traditional cyber defenders. In support of social good and global security we take the exceptional approach of summarising a large body of Defence research applying Reinforcement Learning (RL) to automated cyber defence decision making i.e., what action(s) do we take when a cyber-attack is detected? Promising concepts include two contrasting Multi Agent RL (MARL) approaches, deep RL combined with heterogenous Graph Neural Networks (GNNs), and a Cyber First Aid demonstrator. To achieve this we have matured simulators and tools including development of advanced adversaries to improve defender robustness. We have demonstrated that autonomous cyber defence is feasible on 'real' representative networks and plan to quadruple the number of high fidelity projects in the next year.

1. Introduction

Cyber-attackers are increasingly using Machine Learning (ML) approaches to launch high volumes of sophisticated machine speed cyber-attacks that can evade and overwhelm traditional cyber defenders (Kaloudi et al., 2020), (Guembe, et al., 2022). Furthermore, human cyber defenders are in high demand and cannot be located with all cyber systems. ML is a mature technology for anomaly detection, and commercial Security Orchestration and Automated Response (SOAR) platforms have begun implementing ML driven cyber defence decision making capability (i.e., what action do we take when an attack is detected). However, they are not mission or context aware, which is of particular concern in a Defence application, where it is often impractical to deploy large numbers of skilled cyber defenders to the front line.

Autonomous Resilient Cyber Defence (ARCD) is a Defence research programme, funded by the Defence Science and Technology Laboratory (Dstl). We aim to use ML technologies to develop self-defending, self-recovering cybersecurity concepts for generation after next military operational platforms and technologies. An example of the benefits of this project are increased cyber resilience and reducing the time taken to respond to cyber incidents in systems without co-located cyber responders. Our intended outcomes are:

- Demonstrators for autonomous response to cyber-attacks in military-relevant contexts.
- Greater understanding of the strengths and limitations of ML and their application to cyber defence.
- Improved national skills and knowledge to support combined ML and cyber innovation.

Our research focus is the respond and recover functions within the National Institute of Standards and Technology cybersecurity framework (Fig 1).

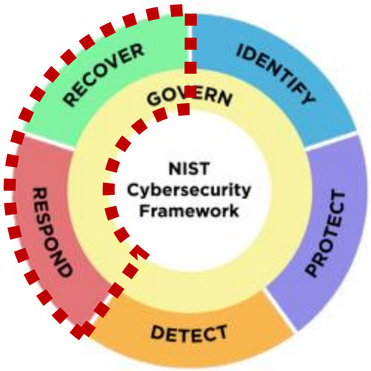

Figure 1 - NIST Cybersecurity functions. This paper's focus is post-detection response and recovery.

RL has emerged as a highly relevant approach for cyber defence (discussed in more detail at Section 3.1). This paper summarises a body of our high maturity RL cyber defence

[1] Presented at Black Hat USA, August 2024
[2] Frazer-Nash Consultancy, Leatherhead, UK
[3] Dstl, Porton Down, UK
[4] Applied Data Science Partners, London, UK
[5] BAE Systems Digital Intelligence, Chelmsford, UK
[6] BMT, Bath, UK
[7] BT, Ipswich, UK
[8] Cambridge Consultants, Cambridge, UK
[9] Exalens, London, UK
[10] QinetiQ, Farnborough, UK
[11] Smith Institute, Oxford, UK
[12] Trustworthy AI, London, UK

research, focussed on accelerating real-world application[13]. Our main contributions can be summarised as follows:

- We provide a largely unpublished body of knowledge[14], exploring the research question: “To what extent can RL driven cyber defence decision making capabilities be demonstrated in a representative environment[15], and how might this apply or be transferred to the real-world?”. We take the exceptional approach of publishing this large body of Defence research with the intent of contributing to social good and security in the face of a rapidly evolving cyber threat landscape.
- We summarise six maturing research projects and provide discussion exploring overall findings, recommendations, and future challenges. Our headline outcomes include a successful proof of concept for RL driven autonomous defence against cyber-attacks in a range of representative environments; Multi agent approaches outperforming single agents; Outperforming rules-based cyber defenders developed by human experts; End-to-end machine speed cyber defence against a ‘real’ cyber-attack on a ‘real’ network; RL red agents learning realistic cyber-attack strategies to support training of robust cyber defenders and reflect future threats; and, generalising to be able to defend network topologies that were not seen in training.

2. Related Work

MLsec, the intersection of ML and cybersecurity has two main areas: the use of ML in security applications, and the security of machine learning systems and algorithms. MLsec applications include malware classification, Domain Name System (DNS) analysis, vulnerability detection, cyber defence, and penetration testing (Ford et al., 2014), (Bilge et al., 2011) (Antonakais, et al., 2012) . These applications generally detect and highlight suspicious patterns, artefacts, and potential incursion, except for cyber defence and penetration testing applications that suggest or implement actions that change the state of a network or system. (Standen, et al., 2022) defines Autonomous Cyber Operations (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” (Vyas et al., 2023). Autonomous Cyber Defence (ACD) focusses on training blue agents, e.g., to autonomously defend a system against cyber attacks (Applebaum, et al., 2022). Early ACD work includes (Beaudoin et al., 2009).

Programmes supporting ACO work include: DARPA Grand Challenge[16] (2014-2016); NATO IST-152 (Pechoucek et al., 2017) (Kott, et al., 2019) (2016-2019); Autonomous Intelligent Cyber Agents (AICA) (Blakely et al., 2023) (2020-2022); The Technical Cooperation Program (TTCP) working group on Cyber Autonomy Gym for Experimentation (CAGE)[17]; Autonomous Resilient Cyber Defence (ARCD)[18,19] (2021-2025); Cyber Agents for Security Testing and Learning Environments (CASTLE)[20] (2022-2025). ICML-22 included the ML4Cyber workshop “to build a mutual comprehensive awareness of the problem and solution spaces across the greater ML community and the Cybersecurity/ML for Cybersecurity communities”.

2.1. Cyber Defence Frameworks

The MITRE ATT&CK® Tactics, Techniques and Procedures (TTP) and D3FEND® frameworks describe and label the behaviours of cyber adversaries (Strom, et al., 2018), (Alexander et al., 2020) and defenders (Kaloroumakis et al., 2021) respectively. We use these frameworks to provide structure and consistency between projects. The MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems[21] (ATLAS™) (NIST, 2024) and Microsoft STRIDE frameworks are useful in assessing potential cyber hardening needs for RL systems (Shevchenko, 2018).

3. Application

We first characterise the ML considerations of the automated cyber defence problem, and detail the suitability of RL (Section 3.1), we then outline our evaluation criteria (Section 3.2) and summarise two studies exploring the challenges and opportunities relating to high dimensionality and combinatorial action spaces in cyber defence (Section 0).

3.1. Characterisation

There are limitations in applying supervised or unsupervised learning to cyber defence as the overwhelming majority of large, high quality cyber security datasets (CICIDS, 2018) (KDD, 1999) (NDsec-1, 2016) relate to detection rather than response and recovery (Figure 1). Cyber defence suits RL as it involves interaction with a dynamic environment, potentially containing adversaries, along with an ability to solve challenges which require sequential decision making

[13] Low maturity state of the art ML techniques, autonomous cyber threat intelligence, and anomaly detection are subject to their own research areas and will inform our future work.

[14] Three projects have already published. We include their key findings in this paper to support our overall conclusions and provide an updated position.

[15] Representative environments may be high-fidelity simulations, emulations, or deployment to realistic hardware & software.

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

[17] https://github.com/cage-challenge

[18] https://www.fnc.co.uk/arcd

[19] https://www.qinetiq.com/en/what-we-do/services-and-products/autonomous-resilient-cyber-defence

[20] https://www.darpa.mil/news-events/2022-10-24

[21] https://atlas.mitre.org/

(Figure 2). Previous work had highlighted Proximal Policy Optimisation (PPO) as an approach showing early promise for autonomous cyber defence (Foley et al., 2022).

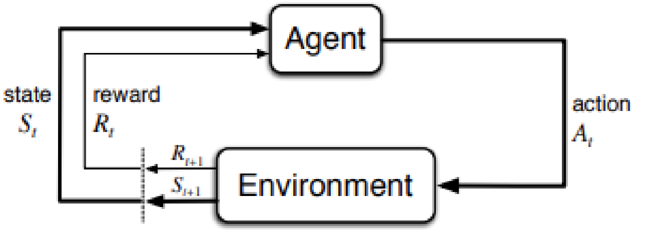

Figure 2 – The agent–environment interaction in a Markov decision process (Sutton & Barto, 2020)

Due to the nature of network traffic, the defence context and agent actions, cyber defence is a high dimensional action and observation space. Therefore, an ability to scale is a key challenge for successful cyber defence concepts.

Cyber Defence agents typically do not have full visibility of the state space so the problem can be formulated as a Partially Observable Markov Decision Process (POMDP) (Oliehoek et a., 2016). Further, centralised decision-making during runtime will be near impossible on a large cyber network due to permissions, boundaries, and risk of high impact vulnerabilities. Multi-agent reinforcement learning (MARL) approaches show great promise within a Decentralised POMDP (Dec-POMDP) or Partially Observable Markov Game if the red agent is also learning. Multi-agent action spaces will need to be heterogeneous due to variable factors such as agent(s) location and visibility (Sengupta et al., 2020).

Cyber-attacks typically present as a combination of bursts of high impact activity, and longer campaigns of stealthy activity, within extended periods of otherwise benign system behaviour. Cyber defence therefore features sparse rewards and difficulties with credit assignment (Arulkumaran et al., 2017). Attack detection inputs are also imperfect, with high levels of uncertainty e.g., false positive anomaly detection alerts. This means that large amounts of data generation and curation is required to ensure optimal coverage of observation-action spaces to learn from and respond to.

3.2. Evaluation

Evaluation of ML approaches is a complex, multi-faceted challenge. Evaluation criteria definition is covered elsewhere in our research programme[10] but is informed most notably by (NIST, 2023). High priority considerations include:

- Tractability (ability to manage real-world large combinatorial action & vast high dimensional observation spaces)
- Mission-level rewards (focus on real-world outcomes, rather than low level rewards)
- Generalisability (ability to deploy to unknown networks, or against new, and potentially learning, adversaries)
- Avoidance of new vulnerabilities (limit exploitability by attacks on AI systems.
- Explainability (logging and understanding reasoning for actions is essential to build trust).
- Scalability in terms of compute power and data handling.

While the above are still considered open problems, the RL literature provides many solutions towards solving them.

3.3. High Dimensionality Inputs and Combinatorial Action Spaces

Two studies were completed that explored the high dimensionality cyber defence problem. The first explored emerging novel techniques for managing high dimensionality and scalability (Morarji, 2023)[6]. Recommended RL dimensionality reduction approaches were offline RL, function approximation (e.g., neural networks), action masking (Tang et al., 2020) and sample-based planning (sampling). Recommended RL scaling approaches were parameter sharing (Christianos et al., 2021) (Gronauer et al., 2022), transfer learning (Zhuang, et al., 2020), hierarchical learning (Botvinick, 2012) (Yang et al., 2020), curriculum learning (Narvekar et al., 2020) (Gronauer et al., 2022) , mean-field methods (Yang, et al., 2018)., and meta-RL (Hafner et al., 2023).

We also explored the challenges for cyber defence at scale (Palmer et al., 2023)[5]. The study explored cyber defence benchmarking environments, and broader environments with similar challenges. The most scarce of these challenges, but one which is core to cyber defence, is the presence of graph-based dynamics. Approaches for combinatorial action spaces were also explored. Recommended approaches were proto actions (Dulac-Arnold et al., 2015), action decomposition (Tavakoli et al., 2018), action elimination (Zahavy et al., 2018, hierarchical RL (Wei et al. 2018), and curriculum learning (Farquhar, et al., 2020). Finally, the non-stationary adversarial learning challenge was considered, where red and blue agents adjust their approach over time. Further work was recommended relating to limiting blue agent exploitability, including Approximate Best Response (ABR) techniques (Lanctot, et al., 2017).

4. Results

Twenty research projects have been undertaken in our mature concepts area, covering different ML technologies and military use cases. Here we summarise the approaches and results for six notable projects that use RL.

4.1. RL Agent Proof of Concept[4, 6]

A PPO agent was deployed in a simple denial of service attack scenario within the PrimAITE cyber training environment[10]. Results were benchmarked against a random action defender and a rules-based defender, developed with a human cyber expert with >10 years' experience. Results provided a strong early proof of concept for RL cyber defenders:

- The PPO agent average episode reward outperformed the rules-based agent in ~50% of cases, and the random agent in 100% of cases;
- The PPO agent was quickly able to achieve near-perfect performance following an upgrade to the environment's defensive action space. The effort required to adapt the rules-based agent was prohibitive, even for a simple problem.
- The PPO agent adapted to an unintentional network misconfiguration, which is a plausible risk that was missed by the rules-based agent.

4.2. Co-operative Decision Making for Cybersecurity[8]

The 'Co-Decyber' project applies cooperative MARL (co-MARL) to a military platooning use case, where the lead logistics vehicle is manned and the remaining logistics vehicles autonomously follow the leader (Cheah, et al., 2023). The attack-defence tree methodology has been employed to scope the action space (Figure 2) (Jhawar et al., 2016) (Kordy et al., 2010). This approach translates conceptually to the decision-making breakdown and underpins the multi-agent architecture (Figure 3). The Co-Decyber architecture maps to a vehicle architecture, to aid deployment into a real system. Recent scenarios exploring denial of service on inter-vehicle communications include approximately thirty defence actions, some requiring sequential ordering.

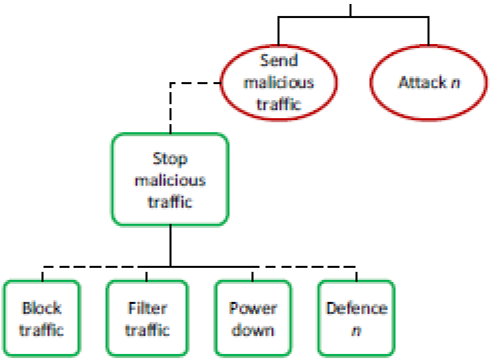

Figure 3 – Attack-Defence tree model for a generic cyber-attack. Attack action in red, defence actions in green (see Figure 4).

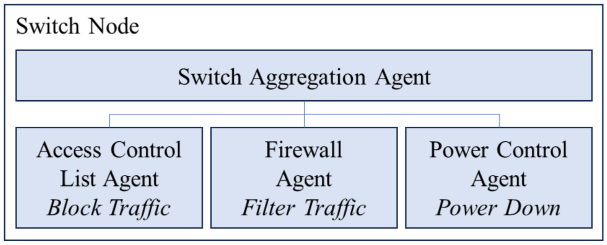

Figure 4 – A Co-Decyber Node with defence actions assigned to individual, but interconnected RL agents (defence actions in italics).

Rewards are assigned at mission-level (Table 1), which have the advantage of avoiding human bias in determining agent policies but increases the challenge of sparse rewards and credit assignment. Rewards are currently relatively abstract but will be matured with stakeholders in the coming months.

Table 1 – Co-Decyber rewards scheme.

Reward	Timeliness	Cargo Status
5	For every vehicle on time	With cargo
2	For every vehicle on time	Without cargo
4	For every vehicle arriving late	With cargo
2	For every vehicle arriving late	Without cargo
0	No arrival	-

Agents are trained against stochastic scripted attacks (e.g., varying entry points). To overcome the challenge of sparse rewards and training at scale the following approaches are used:

- Offline RL (Fujimoto et al., 2019).
- Curriculum Learning (Soviany, 2022).
- Deep Q-learning (Mnih, et al., 2015).

Offline RL separates training and exploration, which offers scaling benefits vs. online RL, which would require full re-running of the exploration process if a scenario was to change. A challenge with offline RL is the trade-off of exhaustive coverage of experience vs. high quality training data that provides the most benefit during training.

Results show that Co-Decyber agents are outperforming random reference agents in the cyber-attack scenarios that were tested (Figure 5). Further results are presented in the full paper for this Project (Cheah, et al., 2023).

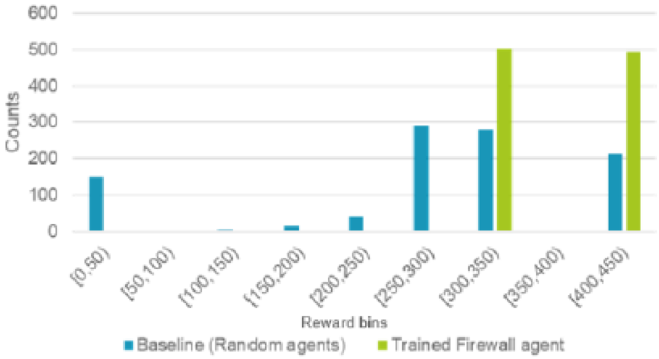

Figure 5 – Histogram of total rewards achieved in trained and random firewall agents, vs. a false fire alarm attack.

Future work on this project will explore:

- More complex attack scenarios to improve agent sophistication, including sequential problem solving.
- Training more agents to defend more vehicle components (i.e., a broader attack surface).
- Inclusion of more realistic, noisy, sensing data to increase the realism of simulation and agent robustness.
- Reward shaping to aid agent training, without losing mission focus.
- Engagement with military stakeholders inform future system requirements, enabling research exploitation and integration of autonomous cyber defence capability.

4.3. Cyber Defence for Maritime Systems[6, 4]

MARL was applied to cyber defence of an Integrated Maritime Platform Management System (IPMS) (Wilson, et al., 2024). These systems include Operational Technology (OT) components, which are relatively vulnerable to cyber-attacks as security controls are typically inferior to Information Technology (IT) equivalents. An IPMS simulator was developed, which included network connectivity, and local and remote controls of two critical subsystems (propulsion and chilled water). Training was conducted against combinations of two cyber-attack strategies; viral, representing an uninformed attack (random movement), or targeted, which represented a sophisticated attack requiring system knowledge. The cyber defenders:

- Have a restricted view of the network.
- Collect awareness of infection status, primarily from alerts received from a simulated Security Information and Event Management (SIEM) system based on ATT&CK Industrial Control System (ICS) tactic levels.
- Can apply the following defence actions to a node: contain (prevent lateral movement), eradicate (remove infection), recover (restore the node to an operational state).

A literature review compared eleven MARL algorithms with the cyber defence application (see *Characterisation*). Experimentation selected Multi-Agent PPO[22] (MAPPO), with a centralised critic, and Independent PPO (IPPO[22]), with independent critics, as the most performant options. During tuning the hyperparameters found to be the most sensitive to change were the learning rate, batch size, lambda, and clip parameter values. Reward shaping (e.g., assigning costs to actions) was also found to accelerate learning, mitigating the otherwise sparse global rewards (win/lose). Experimentation with MAPPO and IPPO showed:

- Multi agent defenders out-performing single agents, which were rarely able to win an episode (Wilson, et al., 2024).
- The defender was able to develop winning strategies over 95% of the time when only 75% of alerts on the network were observed by the defender (Figure 6).
- The defenders could still develop successful defensive policies even when defence action success probability was as low as 50% (Figure 7).
- Emergent behaviour where different agents would adopt different cyber defence roles (Figure 8).

In the next year the project aims to train in a more realistic simulator, and deploy to a ‘real’ test rig, defending against real cyber-attacks on representative maritime IPMS hardware and software. This would represent a significant step towards demonstrating the ability to deploy RL cyber defenders to a real system. Action masking and curriculum learning will be applied to support scaling. A parallel project is exploring sophisticated attack paths on an IPMS environment, which would improve agent robustness.

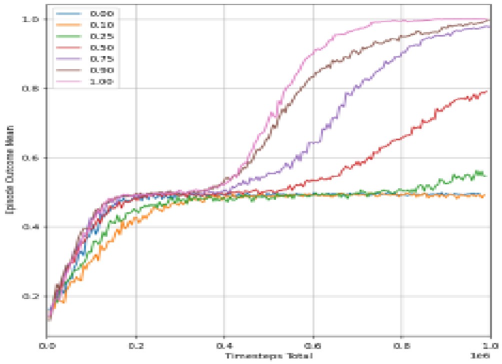

Figure 6 – Mean episode outcome for varied cyber alert success probabilities. An outcome of 1 means the defender wins, 0 is an attacker win, 0.5 is a draw (Wilson, et al., 2024).

[22] https://marllib.readthedocs.io

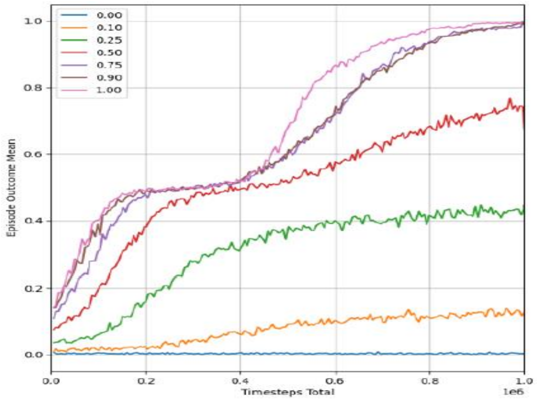

Figure 7 – Mean episode outcome for varied cyber defence action success probabilities (Wilson, et al., 2024).

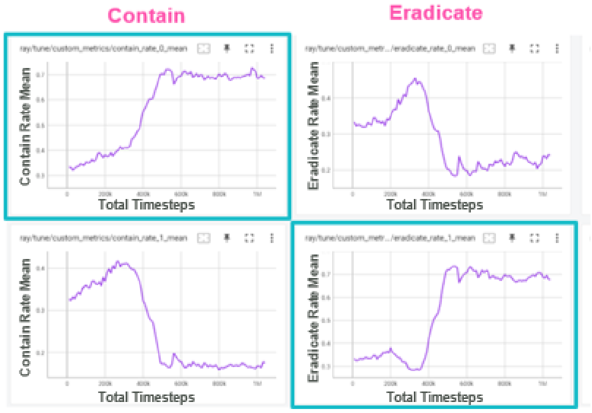

Figure 8 – Agents assumed defence roles. Agent 1 (top) became the 'container' and Agent 2 (bottom) became the 'eradicator'.

4.4. A Virtual Incident Response Analyst (VIRA)[9]

VIRA is a RL-based system aiming to offer first-aid to cyber-physical targets in the field, analysing threat alerts, and triggering or recommending understandable containment actions to a non-expert operator, to provide crucial time for detailed remediation and recovery, thus reducing operational risk and impact.

VIRA employs a novel method of system environment abstraction for efficient and accurate model training. Environment simulation parameters, such as network, asset, attacker, and detection source profiles are instantiated into a set of in-memory objects that are continuously updated during simulation to represent realistic system interactions and dependencies. The abstraction is underpinned by creating a system-agnostic state space in which its RL models train. The state space uses a configurable taxonomic schema for translating attack and detection telemetry into a fixed set of criteria for tractable RL training and response action convergence. The abstraction method also supports transferability to different networks/target systems.

Defender agents were trialled on a physical testbed including IoT sensors, cyber-physical OT (a ROSbot), and IT systems (servers, laptops, etc.). During trials the agent was deployed against a 'real' cyber-attack, and machine speed defensive actions were observed to be comparable to those a human cyber defender (Table 2).

Table 2 – Example defensive actions during a cyber-attack. The agent took a precise, low-risk approach by blocking the connectivity from the offending hosts to the target hosts.

Attack Stage	Defence Action	Suitable Response?
Command and Control (C2) Channel Implant	"dismiss" alert followed by *block_destination_proto_port_default"*	Yes –precise block on attack destination from source for specific attack vector
Reconnaissance (Network Scanning)	*"block_proto_port_extended"*	Yes – given the host already exhibited malicious C2 activity prior to scanning activity
SSH Bruteforce Login	"dismiss" alert followed by "*block_destination_proto_port_default"*	Yes – precise block on attack destination from source for specific attack vector

The VIRA prototype was successfully demonstrated on a real cyber-physical system including initial training of an agent through to live deployment & testing. VIRA achieved machine-speed, automated, detection, response and recovery against a real cyber-attack. Future work will explore automated environment learning and deployment to a representative military network and edge device.

4.5. Automating Generalised Cyber Defence[12]

Our research projects had typically been limited to ≤5 network topologies or adversaries, which led to overfitting. Generalisability is a desirable cyber defence property to provide robustness against factors including unseen adversaries and varying TTPs, network topologies, initial network states and starting information (e.g., information gathered through Open-Source Intelligence (OSINT) or from previous breaches, or even zero-days).

A population of red and blue cyber agents were built to generalise to unseen network topologies using coevolution within a new cyber defence simulator. The network to defend was treated as a heterogeneous graph where training occurs by combining Deep Reinforcement Learning with Heterogeneous GNNs. There were two distinct graphs: one for red and one for blue. The red agent needs to explore and discover the graph, whereas the blue agent had full visibility excluding any undetected red activity. A large language model (GPT-4) was used to produce 80 synthetic networks, 60 for training and 20 for evaluation, with outputs reviewed by cyber security specialists to ensure suitable realism for our intended application of tactical networks.

A curriculum learning approach was adopted for red agents as it has shown promise in Meta RL. A simple curriculum was created; red agent training always started at step 3 to allow blue agents to place decoys, with only positive rewards for reading or exfiltrating files to explore non-zero-sum games in population training. Red agents learned reconnaissance, scanning, and how to read files, exploit services and create sessions while competing against the AI trained blue agent. Increasing the number of network topologies used for training was shown to have a positive impact on performance, demonstrating promise of our approach for improving generalisability (Figure *9*). Future work will investigate the benefits of Meta RL (Beck, et al., 2023), Domain Adaptation (Yadav et al., 2022) and Domain Randomisation (Chen et al., 2022) to support scaling to a virtualised network.

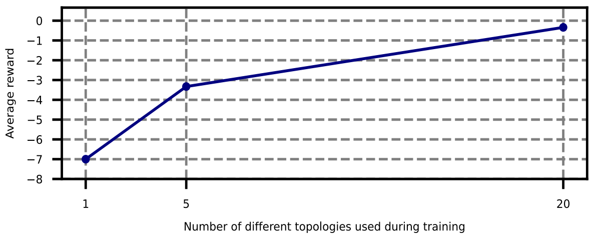

Figure 9 – Rudimentary sensitivity analysis exploring average blue rewards for agents trained on 1, 5 and 20 network topologies, then evaluated in 20 unseen topologies against a common population of red agents.

4.6. Project Odin[7]

Early experimentation identified generalisability limitations, where blue agents overfit to hard-coded adversaries. These adversaries also required re-work when changing variables such as network topology. An exploratory project was initiated to understand the benefits and challenges in developing an abstract RL-based red agent as an adversary that would be used to train blue agents.

Red agent strategies were developed in partnership with defensive cyber teams from a national telecom provider (Figure 10). These were used to design reward functions that would consider aspects such as stealth, effort and persistence. At this stage red agent actions were limited to network discovery, system discovery and exploit.

Originally attacks targeted the same node. To improve generalisability the "fastest time to target" agent trained against random target nodes. Initially, the agent became stuck in local optima and repeatedly selected the same action. The introduction of learning rate schedules, which decayed linearly over time overcame this issue. On a 10-node network, training was shown to improve performance of a PPO red agent by reducing both the number of invalid actions and steps taken to reach the target node (Figure 11).

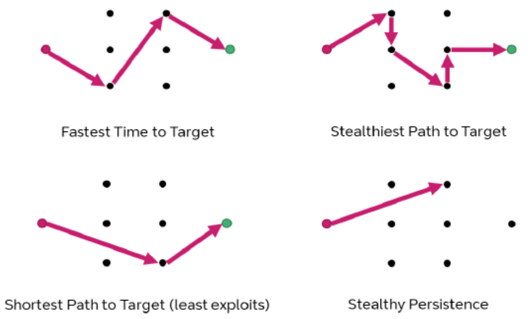
Fastest Time to Target
Stealthiest Path to Target
Shortest Path to Target (least exploits)
Stealthy Persistence

Figure 10 – Example strategies showing a red agent moving from the entry point (left) to a target (right). The 'steathy persistence' goal is to remain undetected.

Action No.		1				5					10					15					20					25					30
Training Episodes	0	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟩
	100k	🟩	🟩	🟩	🟨	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟥	🟩	🟨	🟥	🟨	🟩	🟨	🟨	🟩	🟨	🟨	🟨	🟩	🟨	🟩 *				
	200k	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟨	🟩 *																

Figure 11 – Evaluation of red agent actions at different levels of training during a 30-step episode. Actions are classed as invalid (red), duplicated (amber), valid (green), target reached (*). The untrained agent failed to reach its target.

The project is now looking to scale the red agent (in terms of network size and variety, sophistication of action and strategies), curriculum learning (Narvekar, et al., 2020) and train blue agents using co-evolution (Klijn et al., 2021).

5. Discussion

5.1. Findings

The headline outcome against our research question is a successful proof of concept for RL driven autonomous defence against cyber-attacks across a range of representative environments. Our research has demonstrated instances of:

- Multi agent approaches outperforming single agents, and independently adopting cyber defence roles e.g., container and eradicator (Figure 8).
- Out-performing rules-based cyber defenders, which reflect human expert input, including mitigation of network misconfiguration, which is a realistic risk within IT and OT networks.
- End-to-end machine speed cyber defence against a 'real' cyber-attack on a 'real' network.
- RL red agents learning realistic cyber-attack strategies to support training of robust cyber defenders and reflect future threats.
- Generalising to be able to defend network topologies that were not seen in training.

We consider several of these findings to be novel with limited evidence of such outcomes in current ML and cyber security literature (Section 2).

5.2. Limitations, Challenges & Opportunities

We acknowledge considerable limitations to our findings. To date most research has been conducted in simulators of varying sophistication for an initial proof of concept. Moving to an emulated or real network requires a step change in scaling and dimensionality of the observation and action spaces. We have summarised our review of emerging approaches to overcome this challenge, but within our research curriculum learning, reward shaping (and controls) (Devlin et al., 2014), learning rate schedules and action masking have been used as practical mitigations to high dimensionality, with all projects utilising at least one of these approaches. More state-of-the-art research elsewhere in our programme has shown Meta RL (Beck et al. 2023) and World Models (Hafner et al., 2023) have promise and will continue to be explored in future work.

In the next year we are aiming to deploy at least three of the concepts described in this paper into representative ‘real’ environments, with little abstraction, including virtualised networks and test rigs comprising hardware and software equivalent to that used in operational platforms. Transfer learning, where agents are trained in simulators, then deployed or refined in real systems, has emerged as the most promising method for this challenge. The move to higher fidelity environments will also support exploration of the real-world deployment challenges (i.e., how will an agent actuate actions and best capture observation data within a real system). Another scaling issue is network size, the largest network our trained blue agents have successfully defended is in the order of twenty nodes, whereas real networks could be hundreds or thousands of nodes. Projects will also need to increase realism by accounting for uncertainty in real world observations such as false positive rates from intrusion detection systems (Alahmadi et al., 2022).

Whilst generalisability has been explored in depth in one project (Section 4.5), most projects are still limited in terms of generalisability e.g., the number of network topologies or adversary TTPs. Further work is progressing in this space from consideration of ‘super’ agents, which are highly generalisable, to ensembles of expert agents (Foley et al., 2022), that fingerprint a cyber-attack, or consider multiple recommendations and deploy the most suitable defence action. Most projects will also explore more variable network topologies i.e., deployment to a wide variety of applications and topologies, or realistic dynamic runtime changes, such as the adding or removal of nodes. Our state-of-the-art focus area has active research implementing and adapting DeepMind’s Gato (Reed, et al., 2022), which is aiming to develop an environment agnostic generalist cyber defence agent. Near-term future work is also commencing to explore the application of foundation models such as transformers and large language models as alternative approaches to RL.

Decentralised approaches, such as MARL with communication (comm-MARL) (Zhu et al., 2022) offer great promise. Our early research into Adaptive Social Learning (ASL) (Bordignon et al., 2021) has shown exciting early results, particularly in terms of generalisability and scaling[11]. Agents with ASL beliefs used as input to a PPO trained policy model (ASL/PPO agents) were trained in a 9-node network, and deployed into a new 9-node network with a different topology, achieving comparable win rates to ASL/PPO agents that were trained in the new network (93.5% vs. 94.4%). Indicative results in January 2024 showed ASL/PPO agents trained in a 9-node network achieving an 88.4% win rate when transferred to a 50-node network. Training time and performance level benefits over PPO were also observed and increased as the network sizes grew. Further work is being considered to ratify these initial results.

Current work has focussed on autonomous approaches. Within a safety-critical Defence application, human involvement will be required to build trust in an RL-based system. Future work will explore human machine teaming, indeed Doctrine states “The winner of the robotics revolution will not be who develops this technology first or even who has the best technology, but who figures out how to best use it.” (Ministry of Defence, 2018). Explainability is required to help build trust and is a key focus elsewhere in our programme, with approaches such as Theory of Mind (Akula, et al., 2019), Structured Causal Modelling (Pawlowski, et al., 2020) and Shapley Q-Values (Wang, et al., 2022) under active research (Revell et al., 2023). Findings from this work will continue to influence our direction.

Another difficulty is in comparing performance of concepts training and evaluating in different environments. Our evaluation framework is under development (Section 3.2). Training logs have been provided by the Co-Decyber project for beta testing of the evaluation approach (Section 4.2).

A parallel programme is also researching ML techniques to autonomously generate Cyber Threat Intelligence (CTI) (Little, et al., 2023). This aims to generate enriched CTI observations to blue agents above network traffic including dynamic identification of critical assets, attribution of attacks, and prediction of targets.

6. Conclusion

Future cyber threats will include high volumes of sophisticated machine speed cyber-attacks that can evade and overwhelm traditional cyber defenders. Within this paper we summarise our body of research exploring RL as an approach to automate cyber defence decision making (i.e., what action do we take when an attack is detected). We do so with the intent of contributing to global social good and security in the face of a rapidly evolving cyber threat landscape.

Our research provides an increasingly high-fidelity proof of concept that RL can be applied to cyber defence. We present RL applications and outcomes that we consider to be novel in the field of cyber defence: notably two sophisticated, but contrasting, MARL approaches and combining deep RL with heterogenous GNNs. We have also conducted end-to-end machine speed cyber defence against a 'real' cyber-attack on a 'real' network using RL. As we approach our final demonstration for this phase of research in February 2025, our future direction includes more detailed evaluation in highly representative cyber-physical environments. We recommend further research, but also considerations for adoption, particularly relating to human aspects.

6.1. Recommendations

Our simplified recommendations are as follows:

- Continue maturing autonomous cyber defence concepts and identify rapid exploitation routes to mitigate emerging machine speed cyber threats.
- Incorporate emerging ML approaches to address known limitations and improve performance.
- Continue open knowledge sharing of autonomous cyber defence research to encourage social good outside of defence. This includes building our body of publications, providing the deep technical content required to replicate results and enhance approaches.
- Build understanding of how such systems can be deployed e.g., Human-Machine Teaming.

Acknowledgements

Research funded by Frazer-Nash Consultancy Ltd. on behalf of the Defence Science and Technology Laboratory (Dstl) which is an executive agency of the UK Ministry of Defence providing world class expertise and delivering cutting-edge science and technology for the benefit of the nation and allies. The research supports the Autonomous Resilient Cyber Defence project within the Dstl Cyber Defence Enhancement programme.

References

Akula, Liu, Saba-Sadiya, Lu, Todorovic, Chai, & Zhu. (2019). X-ToM: Explaining with Theory-of-Mind for Gaining Justified Human Trus.

Alahmadi, Axon, & Martinovic. (2022). 99% False Positives: A Qualitative Study of SOC Analysts' Perspectives on Security Alarms. *Proceedings of the 31st USENIX Security Symposium.*

Alexander, Belisle, & Steele. (2020). *MITRE ATT&CK® for Industrial Control Systems: Design and Philosophy.* The MITRE Corporation.

Antonakais, Perdisci, Nadji, Vasiloglou, Abu-Nimeh, Lee, & Dagon. (2012). From Throw-Away Traffic to Bots: Detecting the Rise of DGA-Based Malware. *21st USENIX Security Symposium (USENIX Security 12)*, (pp. 491-506).

Applebaum, Dennler, Dwyer, Moskowitz, Nguyen, Nichols, . . . Wolk. (2022). Bridging Automated to Autonomous Cyber Defense: Foundational Analysis of Tabular Q-Learning. *AISec'22: Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security.*

Arulkumaran, Deisenroth, Brundage, & Bharath. (2017). A Brief Survey of Deep Reinforcement Learning. *IEEE Signal Processing Magazine, Special Issue on Deep Learning for Image Understanding (arXiv extended version)*.

Beaudoin, Japkowicz, & Matwin. (2009). Autonomic Computer Network Defence using Risk State and Reinforcement Learning. *Cryptology and Information Security Series*.

Beck, Vuorio, Liu, Xiong, Zintgraf, Finn, & Whiteson. (2023). A Survey of Meta-Reinforcement Learning.

Bilge, Kirda, Kruefel, & Balduzzi. (2011). Exposure: Finding malicious domains using passive DNS analysis. *Network and Distributed System Security Symposium*, (pp. 1-17).

Blakely, Horsthemke, Evans, & Harkness. (2023). Case Study A: A Prototype Autonomous Intelligent Cyber-Defense Agent. In Kott, *Autonomous Intelligent Cyber Defence Agent (AICA) - A Comprehensive Guide* (pp. 395–408). Springer.

Bordignon, Matta, & Sayed. (2021). Adaptive Social Learning. *IEEE TRANSACTIONS ON INFORMATION THEORY VOL 67*.

Botvinick. (2012). Hierarchical reinforcement learning and decision making.

Cheah, Stone, Haubrick, Bailey, Rimmer, Till, . . . Dorn. (2023). CO-DECYBER: Co-operative Decision Making for Cybersecurity using Deep Multi-agent Reinforcement Learning. *Proceedings of the 28th European Symposium on Research in Computer Security (ESORICS) Workshop on Security and Artificial Intelligence (SECAI).*

Chen, Hu, Jin, Li, & Wang. (2022). Understanding Domain Randomization for Sim-to-Real Transfer.

Christianos, Papoudakis, Rahman, & Albrecht. (2021). Scaling Multi-Agent Reinforcement Learning with Selective Parameter Sharing. *38th International Conference on Machine Learning (ICML).*

CICIDS. (2018). Retrieved from https://www.unb.ca/cic/datasets/ids-2018.html

Devlin, Yliniemi, Kudenko, & Tumer. (2014). Potential-based difference rewards for multiagent reinforcement learning. *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, (pp. 165-172).

Dulac-Arnold, Evans, Sunehag, & Coppin. (2015). *Reinforcement Learning in Large Discrete Action Spaces.*

Farquhar, Gustafson, Lin, Whiteson, Usunier, & Synnaeve. (2020). Growing Action Spaces. *Proceedings of the 37th International Conference on Machine Learning*, (pp. 3040-3051).

Foley, Hicks, Highnam, & Mavroudis. (2022). Autonomous Network Defence using Reinforcement Learning. *ASIA CCS '22: Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security.*

Ford, & Ambareen. (2014). Applications of machine learning in cyber security. *Proceedings of the 27th international conference on computer applications in industry and engineering. Vol. 118.*

Fujimoto, Meger, & Precup. (2019). Off-policy deep reinforcement learning without exploration. *International conference on machine learning*, (pp. 2052-2062).

Gronauer, & Diepold. (20222). Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review, Volume 55*, pp. 895-943.

Guembe, Azeta, Misra, Osamur, Fernandez-Sanz, & Pospelova. (2022). The Emerging Threat of Ai-driven Cyber Attacks: A Review. *Applied Artificial Intelligence, Vol 36, Issue 1*.

Hafner, Pasukonis, Ba, & Lillicrap. (2023). Mastering Diverse Domains through World Models.

Jhawar; Mauw; Zakiuddin. (2016). Automating cyber defence responses using attack-defence trees and game theory. *European Conference on Cyber Warfare and Security* , (p. 163).

Kaloroumakis, & Smith. (2021). *Toward a Knowledge Graph of Cybersecurity Countermeasures.* The MITRE Corporation.

Kaloudi, & Li. (2020). The AI-Based Cyber Threat Landscape: A Survey. *ACM Computing Surveys, Volume 53, Issue 1*, 1-34.

KDD. (1999). Retrieved from https://archive.ics.uci.edu/ml/datasets/kdd+cup+1999+data

Klijn, & Eiben. (2021). A Coevolutionary Approach to Deep Multi-Agent Reinforcement Learning.

Kordy, Mauw, Melissen, & Schweitzer. (2010). Attack–Defense Trees and Two-Player Binary Zero-Sum Extensive Form Games Are Equivalent. *Decision and Game Theory for Security. GameSec.*

Kott, Theron, Drasar, Dushku, LeBlanc, Losiewicz, . . . Gaspari, D. (2019). *Autonomous Intelligent Cyber-defense Agent (AICA) Reference Architecture Release 2.0.*

Lanctot, Zambaldi, Gruslys, Lazaridou, Tuyls, Pérolat, . . . Graepel. (2017). A unified game-theoretic approach to multiagent reinforcement learning. *Advances in Neural Information Processing Systems*, (pp. 4190-4203).

Little. (2023). Applying Machine Learning to Attribute Cyber Attacks (Poster). *Conference on Applied Machine Learning for Information Security (CAMLIS).* Retrieved from https://www.fnc.co.uk/arcd

Ministry of Defence. (2018). Joint Concept Note 1/18: Human-Machine Teaming.

Mnih, Kavukcuoglu, Silver, Rusu, Veness, Bellemare, . . . others. (2015). p. 529.

Morarji, & Casassa-Mont. (2023). Managing High Dimensionality in AI Driven Cyber Defence Decision Making (Poster). *Conference on Applied Machine Learning for Information Security (CAMLIS).*

Narvekar, Peng, Leonetti, Sinapov, Taylor, & Stone. (2020). Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research 21(181):*, (pp. 1-50).

Narvekar, Peng, Leonetti, Sinapov, Taylor, & Stone. (2020). Curriculum Learning for Reinforcement Learning Domains: A Framework and Survey. *Journal of Machine Learning Research 21(181)*, 1-50.

NDsec-1. (2016). Retrieved from https://www2.hs-fulda.de/NDSec/NDSec-1/

NIST. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0).*

NIST. (2024). *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations.*

NIST. (2024). *The NIST Cybersecurity Framework V2.0.*

Oliehoek, & Amato. (2016). *A Concise Introduction to Decentralized POMDPs.*

Palmer, Parry, Harrold, & Willis. (2023). Deep Reinforcement Learning for Autonomous Cyber Operations: A Survey.

Pawlowski, Castro, & Glocker. (2020). Deep Structural Causal Models for Tractable Counterfactual Inference. *Proceedings of 34th Conference on Neural Information Processing Systems (NeurIPS).*

Pechoucek, & Kott. (2017). Intelligent Autonomous Agents for Cyber Defence and Resilience. *Proceedings of the NATO IST-152 Workshop on Intelligent Autonomous Agents for Cyber Defence and Resilience.*

Qinetiq Training & Simulation Ltd. (2023). *PrimAITE*. Retrieved from https://github.com/Autonomous-Resilient-Cyber-Defence/PrimAITE

Reed, Zolna, Parisotto, Colmenarejo, Novikov, & al., e. (2022). A Generalist Agent.

Revell. (2023). Can We Trust Autonomous Cyber Defence for Military Systems? (Poster). *Conference on Applied Machine Learning for Information Security (CAMLIS).*

Sengupta, S., & Kambhampati, S. (2020). Multi-agent Reinforcement Learning in Bayesian Stackelberg Markov Games for Adaptive Moving Target Defense.

Shevchenko, Chick, O'Riordan, Scanlon, & Woody. (2018). *Threat Modelling: A Summary of Available Methods.*

Soviany, Ionescu, Rota, & Sebe. (2022). Curriculum learning: A survey. *International Journal of Computer Vision*, (pp. 1526-1565).

Standen, Bowman, Hoang, Richer, Lucas, Tassel, V., . . . Collyer. (2022). *Cyborg: a gym for the development of autonomous cyber agents*. Retrieved from https://github.com/cage-challenge/CybORG

Strom, Applebaum, Miller, Nickels, Pennington, & Thomas. (2018). *MITRE ATT&CK: Design and Philosophy.* The MITRE Corporation.

Sutton, & Barto. (2020). *Reinforcement Learning, An Introduction (Second Edition).* The MIT Press.

Tang, Liu, Chen, & You. (2020). Implementing action mask in proximal policy optimization (PPO) algorithm.

Tavakoli, Pardo, & Kormushev. (2018). Action branching architectures for deep reinforcement learning. *Proceedings of the AAAI Conference on Artifcial Intelligence*, (pp. 4131–4138).

Vyas, Hannay, Bolton, & Burnap. (2023). Automated Cyber Defence: A Review. *Proceedings of the ACM on Measurement and Analysis of Computing Systems, Vol. 37, No. 4, Article 111*.

Wang, Zhang, Kim, & Gu. (2022). *Shapley Q-value: A Local Reward Approach to Solve Global Reward Games.* Retrieved from https://arxiv.org/abs/1907.05707

Wei, Wicke, & Luke. (2018). Hierarchical approaches for reinforcement learning in parameterized action space.

Wilson, Menzies, Foster, Mont, C., Morarji, Turbeyler, & Gralewski. (2024). Multi-Agent Reinforcement Learning for Maritime Operational Technology Cyber Security. *Proceedings of the Conference on Applied Machine Learning in Information Security (CAMLIS) 2023.*

Yadav, Mishra, Lee, & Kim. (n.d.). A Survey on Deep Reinforcement Learning-based Approaches for Adaptation and Generalization.

Yang, Borovikov, & Zha. (2020). Hierarchical Cooperative Multi-Agent Reinforcement Learning with Skill Discovery. *Proc. of the 19th International Conference on Autonomous Agents and Multiagent System.*

Yang, Luo, Li, Zhou, Zhang, & Wang. (2018). Mean Field Multi-Agent Reinforcement Learning. *ICML.*

Zahavy, Haroush, Merlis, Mankowitz, & Mannor. (2018). Learn What Not to Learn: Action Elimination with Deep Reinforcement Learning.

Zhu, Dastani, & Wang. (2022). A Survey of Multi-Agent Reinforcement Learning with Communication. *Proceedings of ACM Conference.*

Zhuang, Qi, Duan, Xi, Zhu, Zhu, . . . He. (2020). A Comprehensive Survey on Transfer Learning.