The Alan Turing Institute

Turing Technical Report

Autonomous cyber defence: Beyond games?

Chris Hicks and Vasilios Mavroudis

April 2024

Report number 2

https://doi.org/10.5281/zenodo.10974183

Abstract

Cyber defence is one side of a dynamic and ever-evolving confrontation with attackers who seek to exploit weaknesses and cause unintended behaviour in digital systems. Historically both attackers and defenders have relied on human creativity and intellect to outdo one another, learning from each other's tactics in a competitive, emergent, *ad hoc* curriculum. Modern cyber combatants increasingly rely on a variety of automated tools, machine learning; and perhaps more surprisingly, game playing in the pursuit of their goals. Whilst autonomous agents based on deep reinforcement learning have made tremendous breakthroughs playing complex strategy games such as Go, StarCraft, and Stratego, comparatively little progress has been witnessed in cyber defence games and the real-world problems they relate to. We highlight the current state of autonomous agents in cyber defence, presage their considerable future potential, and identify key research challenges both near and far.

1 Introduction

Autonomous cyber defence (ACD) is the use of intelligent autonomous agents for hardening, defending, and recovering computer networks and systems. ACD is urgently motivated by the significant gap in even basic cyber security skills [17, 26], the global shift towards digital and online economies, and substantial rises in the variety and quality of cyber attacks and their consequences [27]. Although it has been eight years since deep Q-networks (DQN) proved the potential for deep reinforcement learning (DRL) to match human-levels of performance in Atari 2600 games [47], and many cyber defence problems are also played as games, there are few to no examples of deep-learning–based autonomous agents actively defending real-world computer systems.

DRL has been famously successful in classic board games including Chess, Go and Shogi [60, 61]. These two-person, zero-sum, stochastic games are especially well suited to DRL because they have clear rules, relatively small action spaces, unambiguous reward functions, and provide both players with perfect information about the game. All of these conditions are problematic for solving real-world problems, however they have been significantly relaxed in more recent work [9, 48]. Vinyals et al. [70] present an autonomous StarCraft II agent that ranks above 99.8% of human players despite contending with an average of 10^{26} possible actions per time step, a planning horizon spanning thousands of decisions, and a vast space of cyclical non-transitive strategies to explore. In similarly groundbreaking work, Perolat et al. [51] accomplish human-expert-level performance at Stratego using self-play and model-free DRL. Compared with Go, the Stratego game tree is 10^{175} times larger, there are 10^{66} possible starting positions, players must make decisions using imperfect information, and deception is part of any good strategy.

The success of DRL and multi-agent RL (MARL) in mastering complex strategy games points to considerable potential in advancing autonomous decision making for cyber defence. Games are already used for a broad spectrum of cyber defence tasks ranging from cryptography proofs, where formal games are used to show the computational hardness of certain primitives [31], through traditional capture-the-flag (CTF) hacking competitions, to tabletop wargames that promote learning and facilitate the development of organisational strategies for dealing with cyber attacks [35]. Many of these games represent intuitive and bounded versions of tasks and challenges that currently demand the time of skilled cyber professionals. Moreover, many cyber games are comparable, or less than, in complexity to some of the strategy games already mastered by DRL. Juxtaposed with the astonishing successes of autonomous and multi-agent systems, the cyber defence community has yet to fully appreciate the potential of modern deep-learning-based decision making techniques applied to their problem sets.

2 Status Quo

ACD is not a new idea. The Defense Advanced Research Projects Agency (DARPA) Cyber Grand Challenge (CGC) launched in 2014 tasked participants with turning cyber defence into a task managed entirely by autonomous agents.

The CGC finale saw seven teams fight to autonomously detect, exploit and patch software vulnerabilities over 96 rounds of competition. Emulating the co-evolving conflict of real-world cyber defence, challenge moderators released the patches submitted during each round to other competitors [52]. At the conclusion of the CGC in 2016, the winning autonomous system was pitted against some of the best hackers in the world at the annual DefCon CTF. Although competitive, the autonomous system eventually ranked last and was beaten by all 14 human teams [10].

The CGC demonstrated the feasibility of autonomous software vulnerability patching, but mainly relied on scalability and brute force search; the role of 'knowing where to look' based on qualities such as 'intuition', 'abstraction' and 'creativity' was left to humans [10]. Indeed, the third-ranked autonomous CGC finalist team (Shellphish) independently qualified for the annual DefCon CTF, uniquely allowing them to enter a mixed human-machine team combining their autonomous CGC finalist model and their human CTF team. Shellphish ranked 10/15 overall, demonstrating that human-assisted automated software analysis outperformed the purely autonomous team [59].

Whilst the CGC gave an early indication of ACD capabilities, machine learning was notably absent from the competition. Between 2016 and 2019 only a small number of fragmented and low-impact academic publications flagged the potential for deep-learning-based autonomous agents to defend against cyber attacks [28, 79, 42, 58, 40, 12, 25]. Over the last three years, however, there has been an exponential rise in publications and, perhaps more importantly, the emergence of open source gyms for training cyber agents. High-quality gym environments are, in general, a key enabler of accessible and reproducible autonomous agent research [8].

The state-of-the-art in cyber gyms is surveyed by Vyas et al. [71]. There are at least seven open-source gyms, but only YAWNING-TITAN (YT) [1] and CybORG [62] focus specifically on training defensive agents. YT offers a highly simplified and abstracted environment which has mainly been used for validating a causal approach relying on an expert graph describing the causal relationships between endogenous variables. Because of the requirement for an expert causal graph, YT only features two defensive actions and lacks the fidelity necessary for real-world ACD impact. CybORG offers a comparatively realistic defensive environment that was initially validated using virtual machine emulation, and aims to allow realistic command-and-control (C2) strategies to be learned. CybORG has been used for three Kaggle-style competitions known as Cyber Autonomy Gym for Experimentation (CAGE) challenges. Similarly to the DARPA CGC, albeit on a reduced scale, introducing a competitive challenge has motivated considerable new research in the area [29, 30, 37, 4, 76, 2]. Furthermore, by providing standardised OpenAI Gym [8] and PettingZoo [66] interfaces, CybORG encourages a long-overdue focus on the ACD capabilities of deep-learning-based and multi-agent autonomous systems.

The first [18] and second [19] CAGE challenges differ only in that the second expands the discrete action space from 54 to 145. Beyond this, both challenges are turn-based, feature the same two adversaries, and use the same simulated enterprise network architecture with 13 hosts spanning three subnets. The reward function is quite heavily engineered and offers different negative penalties dependent on the location and activity of the adversary. The native CybORG observation space is a vector of 11 293 but all participants who open-sourced their models use an included CAGE wrapper reducing this to 52. The winning submissions [29, 34] share much in common: (1) They use proximal policy optimisation (PPO) [57] with modest actor-critic network sizes (e.g., 2–3 hidden layers and widths of 64–256), (2) generalising to the two adversary types is accomplished by training specialised policies and then employing a method of profiling which adversary is present, and (3) considerable expert analysis and engineering is used to redefine the action (and sometimes observation) spaces in a way that exploits specific features of the challenge.

The CAGE 3 challenge [20] is a vastly more complex ACD task which provides a multi-agent simulation with a PettingZoo interface. The network architecture comprises 18 drones tasked with providing an ad hoc communication network. Each drone is permanently infected with latent malware that stochastically activates and exhibits one of six different adversarial profiles. The drones include a secondary low-bandwidth communication channel intended to mitigate the observability limitations of each individual drone and allows for both expert and emergent communication strategies. When inter-drone communication is neglected the discrete action and observation spaces, 56 and 109 respectively, are comparable to the earlier challenges. However, including the 16 bit inter-drone message space grows these substantially to 2^{16} and 381. The reward function is somewhat simplified and issues a negative penalty for any message that is unable to reach the intended recipient. Benign user activity is introduced by autonomous green agents who simulate the demand for ad hoc communication through the network.

Compared with the previous challenges, CAGE 3 introduces and combines a plethora of individually challenging

problems for autonomous and multi-agent systems. Firstly the reward function is noisy [74] because the drones, which are flown automatically by the simulator, frequently form a fragmented ad hoc network that does not permit messages between certain sender-recipient combinations. Secondly, the need for communication between drones makes learning more difficult [11]. The move from two to six types of adversary further aggravates the severe difficulties in generalisability faced by DRL policies [76]. Finally, compared with single-agent RL, MARL is computationally more challenging [77] and has a less mature software ecosystem requiring more developer resources [66]. The winning submission to CAGE 3 [37] partially mitigated these difficulties by: (1) using an expert-defined inter-drone communication protocol; (2) engineering the observation space to include key temporal information; (3) removing noise from the reward function; and (4) using an expert agent to shape a curriculum for learning.

Apart from deep (MA)RL, large language models (LLMs) have also been investigated for their ACD decision making capabilities. LLMs, also known as generative or foundational AI models, have a transformer-based architecture [68] that allows capturing long-range dependencies and contextual relationships in sequential data. Although research on LLMs and decision making has been considerably disjointed, there is increasingly an interest at the intersection of these two communities [81, 14]. Rigaki et al. [55] evaluate pre-trained GPT-3.5-turbo and GPT-4 as offensive cyber decision making agents within the CyberBattleSim [65] environment. CyberBattleSim is similar to CybORG but focuses on the task of attacking the network rather than defending it. The authors find there is a large gap in performance between GPT-3.5-turbo and GPT-4. Only the latter combined with multiple-stage ReAct prompting [82] achieves near-optimal performance comparable to DQN.

LLMs have also been investigated for their professional security certification test-taking and real-world CTF-solving abilities. Tann et al. [64] assess GPT-3.5 and find it correctly answers 82% and 50% of factual and conceptual professional Cisco security certifications questions, respectively. The grading criteria is not public but this probably falls below the standard required to pass. Tann et al. furthermore evaluate GPT-3.5, PaLM 2 (Bard), and Prometheus (Bing) on 7 CTF challenge test cases; the models solve six, two, and one of the challenges respectively. Similarly Yang et al. [80] introduce a CTF benchmarking environment, InterCode-CTF, and find that GPT-4 is unable to perform multi-step cybersecurity tasks out of the box. Finally, considering the broader sociotechnical ACD landscape, Heiding et al. [36] investigate four popular LLMs including GPT-4 and find a human-level ability both to detect and, when combined with an advanced set of manual rules, generate successful phishing emails.

Overall, ACD has yet to see much real-world benefit from learning-based autonomous agents and multi-agent systems. There is a promising trend towards environments taking advantage of recent advances in DRL and LLMs, but there is little consistency between how the networks behave, the decision making roles they simulate, the granularity of decisions, the attacker and defender models assumed, or the goals defined. The competitive self-play which enabled super-human levels of performance in Go [61] and Stratego [51] is notably absent in the existing gyms. Self-play is likely an essential step towards achieving both human-level performance and the adversarial robustness required for real-world operation.

3 A Vision for ACD

Despite the significant challenges facing autonomous agents in the transition to solving real-world tasks, many cyber defence problems are sufficiently game-like that they may already be solvable using existing techniques. Below we focus on the gap between simulated and real-world ACD, and point to how it could be ameliorated in the near future.

Reality Gap?

For many real world tasks, collecting data for training autonomous agents is prohibitively expensive and potentially dangerous. Simulators help to solve this problem, but often limit performance owing to a lack of fidelity [13]. For ACD we can attenuate both the motivation and typical caveats of simulation. Firstly, data from computer systems is abundant and low-cost. We are less likely to need ACD simulators to reduce costs and e.g., competitive CTF hacking competitions are commonly run on real hardware rented ad hoc from cloud service providers [67]. There are circumstances where real data may be more challenging to access online, such as when seeking to defend critical systems, but techniques such as learning from offline data [15], or using digital twins [6] to offer a virtual representation coupled to real-world systems, offer much promise.

Regarding fidelity, all manner of computer systems and networks are frequently emulated (i.e., virtual machines and virtual networks), and provided as a service, such that they provide a complete functional substitute for their real world equivalents [16]. Unlike simulators that have been built for the sole purpose of training autonomous agents, virtual machines are a longstanding part of the digital ecosystem that have been refined for decades [53]. Beyond the

value of autonomous agents actively defending systems, the gamification of many cyber defence problems may present unique opportunities to tackle specific challenges such that the reality gap has already been addressed. For example, game-based proofs of security are typically used to provide formal assurances about cryptographic primitives such as encryption algorithms [32]. The standard framing is a competitive two-player zero-sum game between an attacker and a defender. These games are used for proofs showing that the probability of any adversary whatsoever winning the game is negligible, usually by reduction to the intractability of some well-known hard problem. These games, which have very precisely defined rules and goals, are readily playable by DRL agents that can antagonise or corroborate the theory of the proof. This may seem futile in the instance of algorithms proven reducible to intractably hard mathematical problems, but all too frequently devastating attacks are discovered in schemes that were previously proven secure [5].

Finally, attacks themselves can be thought of as a type of reality gap. Often, despite a principled approach to security being used during the design of a system, attacks emerge when reality offers greater fidelity than the theoretical security model. Concretely, side-channel attacks [54] emerge when real-world implementation phenomena (e.g., timing, power consumption and electromagnetic emissions) are utilised to undermine the security of a target. Microarchitectural attacks, for example, exploit the implementation of microprocessors to leak information from architecturally isolated processes running on shared physical resources [33]. Autonomous DRL agents have a high potential for impact in this area since they can use model-free approaches to discover vulnerabilities that may not have been considered possible during theoretical security modelling. Indeed, Luo et al. [44] formulate cache-timing attacks on microprocessors as a two player game and find, using standard DRL agents, a novel attack on real hardware, outperforming previous attacks with 71% higher information leakage. Their technique shows the potential for autonomous agents to conduct blackbox analysis of real systems, ultimately reducing the cost to manufacturers and improving security for end users.

Overall there is enormous potential for autonomous and multi-agent systems to advance the state of the art in cyber defence. This can be accomplished both through interaction with existing native-fidelity virtualised environments, and offline learning on real network data, as well as self-play in specialised multi-agent games such as those already used in cryptography.

4 Caveat Emptor

There are significant challenges that need to be addressed by both the cyber defence and autonomous agents communities, especially to ensure that the autonomous advantage does not disproportionately enrich attackers over defenders.

Complexity

The complexity of real-world ACD tasks could yield impossibly large state and action spaces that make it computationally infeasible, or at least prohibitively expensive, to learn useful policies. In a game like Chess or Stratego all of the information available to a player, however imperfect, can be put into the observation space of an autonomous agent. In contrast, there is vastly more information inside even a single computer than could possibly be used as a standard DRL observation space. Suppose an average computer has 8 GB of volatile memory that is shared by all running processes. Observing each bit discretely would require an observation space of $\approx 7 \times 10^{10}$ bits. This means that ACD agents will most likely omit potentially relevant information to make decision making tractable. In games where the spaces are very large by DRL standards, specifically StarCraft II [46], expert-human researchers specify the spaces used for decision making. This approach can be replicated for cyber defence tasks, and this will likely succeed to a point, but every decision that limits an agent's observation of the system will ultimately limit performance. It will be extremely challenging to determine in advance that information is irrelevant for a specific task, against a potentially unknown adversary, without having unintended consequences.

More optimistically, carefully defined state and action spaces in specialised environments, that make very weak assumptions about the adversary (i.e., assume a very powerful adversary [23]), could allow autonomous agents to discover meaningful, general-purpose techniques that can be applied to real-world systems. For example, Hicks et al. [37] conduct early work in this direction assuming a priori that the adversary has already compromised the target system. We do not necessarily have to be concerned with exactly how the adversary accomplishes a specific goal, and can instead focus on how to mitigate the impact and ensure overall resilience.

Time

Another key challenge for ACD agents is the variability and range of times taken to accomplish different tasks. Unlike the discrete turn-based simulations of existing cyber defence environments, most real-world computing environments fea-

ture variation in the time taken to complete specific tasks (e.g., caused by shared resources) and large differences between the duration of different tasks. Neither of these issues have yet been tackled in the ACD literature but there are encouraging fundamental results. Semi-Markov Decision Processes (SMDPs) [7] extend the standard MDP framework to handle actions with variable durations and allow for modeling continuous-time discrete-event systems, however DRL of SMDPs is not currently practical [3]. The options framework [63] allows reasoning at different levels of temporal abstraction by modelling courses of actions (e.g., restoring a damaged application) called options. A limiting precondition of the framework is that options must be expertly defined a priori, however Machado et al.[45] recently discovered that successor representation [21] can be used to automatically learn and continually refine appropriate options. Nevertheless, combining these fundamental advances with function approximation for DRL remains an outstanding problem with severe implications for real-world ACD agents.

Concerning ACD tasks that occur over long time frames, such as defending against an advanced persistent threat adversary [24], Lampinen at al. [41] show that standard memory architectures including LSTMs and transformers [69, 49] fail to recall important information even after a few minutes. The authors propose Hierarchical Chunk Attention Memory (HCAM) which provides state-of-the-art performance in simulated tasks occurring over long time horizons and including intervening distractions. Despite these encouraging results, HCAM requires storing every potentially relevant time step in memory and would be problematic to scale for many real-world ACD tasks. The ideal cyber defence agent would be capable of acting and perceiving on both very short and long time scales. It should also tolerate, and ideally would derive an advantage from [44], asynchronous and temporally variable actions.

Generalisation

Unlike humans, who readily generalise abstract concepts from games and apply them in entirely different tasks (i.e., transfer learning), DRL performance is extremely sensitive to even minor environmental perturbations [39, 56]. Furthermore, DRL agents are usually incapable of any significant degree of meta-learning [73] (i.e., generalising to a distribution of tasks with shared structure). The highly superficial nature of most DRL policies is unfortunately likely to favour attackers who can count themselves successful with the discovery of even a single vulnerability. Indeed, the main real-world impact of DRL on cyber defence to date has been the automated discovery of vulnerabilities in specific targets [72, 44]. Causal RL (CRL) is an approach that uses an explicit causal model [50] to inform decision-making [22]. CRL techniques could enhance the ability for autonomous agents to generalise, and might also improve interpretability and sample efficiency, but have yet to be applied in complex real-world environments. Some preliminary results are presented by Andrew et al. [1] who apply a CRL method to the YT cyber defence simulator environment (see Section 2). The authors find that CRL is able to efficiently take optimal decisions, but note that the causal model is likely to grow to unmanageable scales in any hostile real-world cyber environment.

In cyber defence, anything that we add to our system will also become a target for the adversary. The fragility of DRL policies is therefore particularly concerning. We can be certain that attackers will deliberately try to perturb the environment towards achieving their goals. The robustness of DRL policies is both an open problem [75] and a prerequisite for many ACD use cases. Robustness in ACD has not been significantly explored, likely hindered by the lack of realistic and problem-aligned two-player zero-sum environments.

Adaptability

A feature of cyber environments is that they are iteratively populated with variable numbers of discrete subunits representing e.g., services running on a host, hosts on a subnet, or subnets on a network. Since most standard RL algorithms assume fixed action and observation spaces, there does not appear to be a straightforward way to allow a pre-trained policy to adapt to variable real-world system and network configurations. Recent work by Hua et al. [38] demonstrate the feasibility of hybrid action spaces in a multi-agent particle environment [43], but it remains unclear how scalable to complex real-world problems such methods will prove. However, ACD may present an interesting middle ground where it could suffice to allow a variable range of common network configurations. In this setting graph neural networks [78] could enable more generalisable interfaces for autonomous agents.

5 Conclusion

There is a tremendous potential for autonomous and multi-agent systems to advance the state of the art in cyber defence tasks but this will require both communities to work together closely on key challenges. We have outlined how, in several respects, ACD is highly amenable to modern autonomous decision making techniques; in particular, the availability of

native-fidelity environments for training agents as well as pre-existing multi-agent games that are well-aligned with the RL paradigm (e.g., in cryptography). Furthermore we review the main limitations of deep-learning-based agents and how they might be anticipated to limit the advantage that autonomous agents might offer attackers over defenders.

6 Acknowledgments

Research funded by 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 (ARCD) project within the Dstl Cyber Defence Enhancement programme.

References

[1] Alex Andrew et al. "Developing Optimal Causal Cyber-Defence Agents via Cyber Security Simulation". In: *International Confernece on Machine Learning (ICML). Workshop on Machine Learning for Cybersecurity (ML4Cyber)*. July 2022.

[2] Andy Applebaum, Camron Dennler, Patrick Dwyer, et al. "Bridging Automated to Autonomous Cyber Defense: Foundational Analysis of Tabular Q-Learning". In: *Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security*. AISec'22. Association for Computing Machinery, 2022, pp. 149–159. ISBN: 9781450398800. DOI: 10.1145/3560830.3563732.

[3] Giacomo Ascione and Salvatore Cuomo. "A Sojourn-Based Approach to Semi-Markov Reinforcement Learning". In: *Journal of Scientific Computing* 92.2 (June 2022), p. 36. ISSN: 1573-7691. DOI: 10.1007/s10915-022-01876-x. URL: https://doi.org/10.1007/s10915-022-01876-x.

[4] Liz Bates, Vasilios Mavroudis, and Chris Hicks. "Reward Shaping for Happier Autonomous Cyber Security Agents". In: *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec '23), Copenhagen, Denmark* (2023).

[5] Dan Boneh. "Twenty Years of Attacks on the RSA Cryptosystem". In: *Notices of the American Mathematical Society* 46 (1999), pp. 203–212.

[6] Hugh Boyes and Tim Watson. "Digital Twins: An Analysis Framework and Open Issues". In: *Comput. Ind.* 143 (Dec. 2022). ISSN: 0166-3615. DOI: 10.1016/j.compind.2022.103763.

[7] Steven J. Bradtke and Michael O. Duff. "Reinforcement Learning Methods for Continuous-Time Markov Decision Problems". In: *Proceedings of the 7th International Conference on Neural Information Processing Systems*. NeurIPS'94. Cambridge, MA, USA: MIT Press, 1994, pp. 393–400.

[8] Greg Brockman, Vicki Cheung, Ludwig Pettersson, et al. *OpenAI Gym*. 2016. arXiv: 1606.01540 [cs.AI].

[9] Noam Brown and Tuomas Sandholm. "Superhuman AI for multiplayer poker". In: *Science* 365.6456 (2019), pp. 885–890. DOI: 10.1126/science.aay2400. URL: https://www.science.org/doi/abs/10.1126/science.aay2400.

[10] David Brumley. "The Cyber Grand Challenge and the Future of Cyber-Autonomy". In: *USENIX Login* 43.2 (2018), pp. 6–9.

[11] Rahma Chaabouni, Florian Strub, Florent Altché, et al. "Emergent Communication at Scale". In: *International Conference on Learning Representations*. ICLR '22. 2022.

[12] Moitrayee Chatterjee and Akbar-Siami Namin. "Detecting Phishing Websites through Deep Reinforcement Learning". In: *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*. Vol. 2. 2019, pp. 227–232. DOI: 10.1109/COMPSAC.2019.10211.

[13] Yevgen Chebotar et al. "Closing the Sim-to-Real Loop: Adapting Simulation Randomization with Real World Experience". In: *2019 International Conference on Robotics and Automation (ICRA)* (2018), pp. 8973–8979.

[14] Lili Chen et al. "Decision transformer: Reinforcement learning via sequence modeling". In: *Advances in neural information processing systems* 34 (2021), pp. 15084–15097.

[15] Ching-An Cheng et al. "Adversarially Trained Actor Critic for Offline Reinforcement Learning". In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 3852–3878.

[16] Marcello Cinque et al. "Virtualizing Mixed-Criticality Systems: A Survey on Industrial Trends and Issues". In: *Future Gener. Comput. Syst.* 129 (Apr. 2022), pp. 315–330. ISSN: 0167-739X. DOI: 10.1016/j.future.2021.12.002.

[17] Steve Coutinho et al. *Cyber security skills in the UK labour market 2023: findings report*. Tech. rep. Ipsos and Perspective Economics, July 2023.

[18] *Cyber Autonomy Gym for Experimentation Challenge 1*. https://github.com/cage-challenge/cage-challenge-1. Created by Maxwell Standen, David Bowman, Son Hoang, et al. 2021.

[19] *Cyber Autonomy Gym for Experimentation Challenge 2*. https://github.com/cage-challenge/cage-challenge-2. Created by Maxwell Standen, David Bowman, Son Hoang, et al. 2022.

[20] *Cyber Autonomy Gym for Experimentation Challenge 3*. https://github.com/cage-challenge/cage-challenge-3. Created by Maxwell Standen, David Bowman, Son Hoang, et al. 2022.

[21] Peter Dayan. "Improving Generalization for Temporal Difference Learning: The Successor Representation". In: *Neural Computation* 5.4 (1993), pp. 613–624. DOI: 10.1162/neco.1993.5.4.613.

[22] Zhihong Deng et al. *Causal Reinforcement Learning: A Survey*. 2023. arXiv: 2307.01452 [cs.LG].

[23] D. Dolev and A. Yao. "On the Security of Public Key Protocols". In: *IEEE Transactions on Information Theory* 29.2 (1983), pp. 198–208. DOI: 10.1109/TIT.1983.1056650.

[24] Feng Dong et al. "DISTDET: A Cost-Effective Distributed Cyber Threat Detection System". In: *USENIX Security Symposium*. 2023.

[25] Noah Dunstatter et al. "Solving Cyber Alert Allocation Markov Games with Deep Reinforcement Learning". In: *Decision and Game Theory for Security*. Springer International Publishing, 2019, pp. 164–183. ISBN: 978-3-030-32430-8.

[26] Organisation for Economic Co-operation and Development (OECD). *Building a Skilled Cyber Security Workforce in Five Countries: Insights from Australia, Canada, New Zealand, United Kingdom, and United States*. OECD, 2023. ISBN: 9789264381247.

[27] *ENISA Threat Landscape 2023*. Tech. rep. European Union Agency for Cybersecurity (ENISA), Oct. 2023. DOI: 10.2824/782573.

[28] Ming Feng and Hao Xu. "Deep reinforecement learning based optimal defense for cyber-physical system in presence of unknown cyber-attack". In: *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. 2017, pp. 1–8. DOI: 10.1109/SSCI.2017.8285298.

[29] Myles Foley et al. "Autonomous Network Defence using Reinforcement Learning". In: *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security*. 2022.

[30] Myles Foley et al. "Inroads into Autonomous Network Defence using Explained Reinforcement Learning". In: *Proceedings of the Conference on Applied Machine Learning in Information Security*. Vol. 3391. CEUR Workshop Proceedings. 2022, pp. 1–19.

[31] Oded Goldreich. *Foundations of Cryptography*. Vol. 1–2. Cambridge Uni. Press, 2004.

[32] Shafi Goldwasser and Silvio Micali. "Probabilistic encryption". In: *Journal of Computer and System Sciences* 28.2 (1984), pp. 270–299. ISSN: 0022-0000. DOI: https://doi.org/10.1016/0022-0000(84)90070-9.

[33] Ben Gras et al. "ABSynthe: Automatic Blackbox Sidechannel Synthesis on Commodity Microarchitectures". In: *NDSS*. Feb. 2020.

[34] John Hannay. *Winning Submission to the CAGE 2 Challenge (CardiffUni)*. https://github.com/john-cardiff/-cyborg-cage-2. 2022.

[35] Stephen Hart et al. "Riskio: A Serious Game for Cyber Security Awareness and Education". In: *Computers & Security* 95 (2020).

[36] Fredrik Heiding et al. *Devising and Detecting Phishing: Large Language Models vs. Smaller Human Models*. 2023. arXiv: 2308.12287 [cs.CR].

[37] Chris Hicks et al. "Canaries and Whistles: Resilient Drone Communication Networks with (or without) Deep Reinforcement Learning". In: *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*. 2023, pp. 91–101.

[38] Hongzhi Hua et al. "A Further Exploration of Deep Multi-Agent Reinforcement Learning with Hybrid Action Space". In: *Artificial Neural Networks and Machine Learning – ICANN 2023*. Springer Nature Switzerland, 2023, pp. 1–12. ISBN: 978-3-031-44223-0.

[39] Ken Kansky, Tom Silver, David A. Mély, et al. "Schema Networks: Zero-Shot Transfer with a Generative Causal Model of Intuitive Physics". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. 2017, pp. 1809–1818.

[40] Alexander Kott, Paul Théron, Martin Drašar, et al. *Autonomous Intelligent Cyber-defense Agent (AICA) Reference Architecture. Release 2.0*. 2018. arXiv: 1803.10664 [cs.CR].

[41] Andrew Lampinen et al. "Towards mental time travel: a hierarchical memory for reinforcement learning agents". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 28182–28195.

[42] Yandong Liu et al. "Deep Reinforcement Learning based Smart Mitigation of DDoS Flooding in Software-Defined Networks". In: *2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*. 2018, pp. 1–6. DOI: 10.1109/CAMAD.2018.8514971.

[43] Ryan Lowe et al. "Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NeurIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6382–6393. ISBN: 9781510860964.

[44] Mulong Luo et al. "AutoCAT: Reinforcement Learning for Automated Exploration of Cache-Timing Attacks". In: *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. 2023, pp. 317–332. DOI: 10.1109/HPCA56546.2023.10070947.

[45] Marlos C. Machado et al. "Temporal Abstraction in Reinforcement Learning with the Successor Representation". In: *Journal of Machine Learning Research* 24.80 (2023), pp. 1–69.

[46] Michael Mathieu et al. "StarCraft II Unplugged: Large Scale Offline Reinforcement Learning". In: *Deep RL Workshop NeurIPS 2021*. 2021.

[47] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. "Human-level control through deep reinforcement learning". In: *Nature* (2015).

[48] OpenAI et al. *Dota 2 with Large Scale Deep Reinforcement Learning*. 2019. arXiv: 1912.06680 [cs.LG].

[49] Emilio Parisotto et al. "Stabilizing Transformers for Reinforcement Learning". In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 7487–7498.

[50] J. Pearl. *Causality*. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2009. ISBN: 9780521895606.

[51] Julien Perolat, Bart De Vylder, et al. "Mastering the game of Stratego with model-free multiagent reinforcement learning". In: *Science* 378.6623 (2022), pp. 990–996. DOI: 10.1126/science.add4679.

[52] Benjamin Price et al. "House Rules: Designing the Scoring Algorithm for Cyber Grand Challenge". In: *IEEE Security & Privacy* 16.2 (2018), pp. 23–31. DOI: 10.1109/MSP.2018.1870877.

[53] Allison Randal. "The Ideal Versus the Real: Revisiting the History of Virtual Machines and Containers". In: *ACM Comput. Surv.* 53.1 (Feb. 2020). ISSN: 0360-0300. DOI: 10.1145/3365199.

[54] Mark Randolph and William Diehl. "Power Side-Channel Attack Analysis: A Review of 20 Years of Study for the Layman". In: *Cryptography* 4 (2020). DOI: 10.3390/cryptography4020015.

[55] Maria Rigaki et al. *Out of the Cage: How Stochastic Parrots Win in Cyber Security Environments*. 2023. arXiv: 2308.12086 [cs.CR].

[56] Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, et al. *Progressive Neural Networks*. 2022. arXiv: 1606.04671 [cs.LG].

[57] John Schulman et al. "Proximal Policy Optimization Algorithms". In: *CoRR* abs/1707.06347 (2017). arXiv: 1707.06347. URL: http://arxiv.org/abs/1707.06347.

[58] Jonathon Schwartz. *Autonomous Penetration Testing using Reinforcement Learning*. Bachelor's Thesis. Available at https://arxiv.org/abs/1905.05965. Nov. 2018.

[59] Yan Shoshitaishvili, Antonio Bianchi, Kevin Borgolte, et al. "Mechanical Phish: Resilient Autonomous Hacking". In: *IEEE Security & Privacy* 16.2 (2018), pp. 12–22. DOI: 10.1109/MSP.2018.1870858.

[60] David Silver, Aja Huang, and Chris J. others Maddison. "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529.7587 (Jan. 2016), pp. 484–489. ISSN: 1476-4687. DOI: 10.1038/nature16961.

[61] David Silver, Thomas Hubert, Julian Schrittwieser, et al. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". In: *Science* 362.6419 (2018), pp. 1140–1144. DOI: 10.1126/science.aar6404.

[62] Maxwell Standen, Martin Lucas, David Bowman, et al. "CybORG: A Gym for the Development of Autonomous Cyber Agents". In: *IJCAI-21 1st International Workshop on Adaptive Cyber Defense*. 2021.

[63] Richard S. Sutton, Doina Precup, and Satinder Singh. "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning". In: *Artificial Intelligence* 112.1 (1999), pp. 181–211. ISSN: 0004-3702. DOI: https://doi.org/10.1016/S0004-3702(99)00052-1.

[64] Wesley Tann et al. *Using Large Language Models for Cybersecurity Capture-The-Flag Challenges and Certification Questions*. 2023. arXiv: 2308.10443.

[65] Microsoft Defender Research Team. *CyberBattleSim*. https://github.com/microsoft/cyberbattlesim. Created by Christian Seifert and Michael Betser and William Blum and others. 2021.

[66] J Terry, Benjamin Black, Nathaniel Grammel, et al. "PettingZoo: A Standard API for Multi-Agent Reinforcement Learning". In: *Advances in Neural Information Processing Systems*. NeurIPS '21 34 (2021), pp. 15032–15043.

[67] Erik Trickel et al. "Shell We Play A Game? CTF-as-a-service for Security Education". In: *2017 USENIX Workshop on Advances in Security Education (ASE 17)*. USENIX Association, Aug. 2017.

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[69] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

[70] Oriol Vinyals, Igor Babuschkin, et al. "Grandmaster level in StarCraft II using multi-agent reinforcement learning". In: *Nature* 575.7782 (Nov. 2019), pp. 350–354. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1724-z.

[71] Sanyam Vyas et al. "Automated Cyber Defence: A Review". In: *Proceedings of the ACM on Measurement and Analysis of Computing Systems* (2023).

[72] Salim Al Wahaibi, Myles Foley, and Sergio Maffeis. "SQIRL: Grey-Box Detection of SQL Injection Vulnerabilities Using Reinforcement Learning". In: *32nd USENIX Security Symposium (USENIX Security 23)*. Anaheim, CA: USENIX Association, 2023, pp. 6097–6114. ISBN: 978-1-939133-37-3.

[73] Jane X Wang et al. "Alchemy: A benchmark and analysis toolkit for meta-reinforcement learning agents". In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2021.

[74] Jingkang Wang, Yang Liu, and Bo Li. "Reinforcement Learning with Perturbed Rewards". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.04 (2020), pp. 6202–6209. DOI: 10.1609/aaai.v34i04.6086.

[75] Tony Tong Wang et al. "Adversarial Policies Beat Professional-Level Go AIs". In: *Deep Reinforcement Learning Workshop NeurIPS 2022*. 2022.

[76] Melody Wolk, Andy Applebaum, Camron Dennler, et al. *Beyond CAGE: Investigating Generalization of Learned Autonomous Network Defense Policies*. 2022. arXiv: 2211.15557 [cs.LG].

[77] Annie Wong et al. "Deep multiagent reinforcement learning: challenges and directions". In: *Artificial Intelligence Review* 56.6 (June 2023), pp. 5023–5056. ISSN: 1573-7462. DOI: 10.1007/s10462-022-10299-x.

[78] Zonghan Wu et al. "A Comprehensive Survey on Graph Neural Networks". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.1 (2020), pp. 4–24.

[79] Liang Xiao et al. "A Secure Mobile Crowdsensing Game With Deep Reinforcement Learning". In: *IEEE Transactions on Information Forensics and Security* 13 (2018), pp. 35–47. DOI: 10.1109/TIFS.2017.2737968.

[80] John Yang et al. "Language Agents as Hackers: Evaluating Cybersecurity Skills with Capture the Flag". In: *Multi-Agent Security Workshop @ NeurIPS'23*. 2023.

[81] Sherry Yang, Ofir Nachum, Yilun Du, et al. *Foundation Models for Decision Making: Problems, Methods, and Opportunities*. 2023. arXiv: 2303.04129 [cs.AI].

[82] Shunyu Yao et al. "ReAct: Synergizing Reasoning and Acting in Language Models". In: *The Eleventh International Conference on Learning Representations*. 2023.