The Alan Turing Institute

Defence and Security Programme

Robust artificial intelligence for active cyber defence

Anthony Burke

Robust artificial intelligence for active cyber defence

March 2020

Anthony Burke

Contents

Executive Summary

The global environment for artificial intelligence (AI) research, technology and talent is highly competitive. The United States (US) and China are competing for global AI dominance, and it is within this context that other nations, including the United Kingdom (UK), seek AI advantage.

The UK National Cyber Security Centre (NCSC) and Defence Science and Technology Laboratory (Dstl) seek to leverage AI to develop enhanced Active Cyber Defence (ACD) capabilities. To support Dstl and NCSC research planning, the Alan Turing Institute (the Turing) undertook an analysis to create a research roadmap for applied AI in ACD.

Recent UK and US cyber security studies have defined ACD as a concept for scalable cyber defence systems to provide proactive, near real time cyber defence. Robust, intelligent, automated decision-making and action in cyber scenarios are central to this ACD concept. Fundamental AI research is required to create the technologies needed to protect against cyber threats at national scale and cyber-relevant speed.

The goal of this research initiative is to create radical advancements in AI powered autonomous cyber defence systems.

To achieve this, researchers must tackle technical challenges in following areas:

- **AI-enabled network defender**: can we create AI agents that automatically monitor, protect and heal our systems?
- **AI-enabled security planner**: can we create AI cyber planning systems that enhance human decision-making and action in complex, high-tempo operational scenarios?
- **AI-enabled penetration tester**: can we create AI agents that automatically identify system weaknesses and vulnerabilities before real-world adversaries?

Dstl, NCSC and the Turing should pursue the most valuable research, and constantly seek out new impactful ideas. The following initial set of candidate projects are proposed:

AI-enabled penetration tester	AI-enabled security planner	AI-enabled network defender
Adversarial machine learning agent attacks	Autonomous cyber agent & human teaming	Automated cyber defence course of action evaluation
Autonomous competitive cyber agents (attack)	Hypergame modelling cyber threat evolution	Autonomous competitive cyber agents (defence)
Defeating AI detection using generative modelling	Threat identification using causal modelling of weak indicators	Explainable agents & action justification
	Threat situational awareness at scale	Generative modelling for decoy defences

Four **success criteria** are intended to aid researchers and planners to judge progress:

- Demonstrate AI technology applied to realistic cyber defence challenges.
- Publish papers in top tier venues/journals to advance the state of the art.
- Publish high quality reusable software, data, infrastructure and documentation artefacts.
- Develop capability within Dstl and NCSC.

Facilities, data and domain knowledge are all important research enablers, without which progress risks being impeded. Dstl, NCSC and the Turing should review existing resources to determine if any may be offered to research teams to satisfy these needs.

It is recommended that Dstl, NCSC and the Turing continue to work in partnership to pursue the impactful AI research agenda presented in this report.

“The prospect of sophisticated AI enhancing cyber capabilities – both defence and attack – is not some far off theoretical concern.”

Introduction

Cyber security and artificial intelligence (AI) are pervasive issues of our current era. A deeply interconnected world, enabled by ever more powerful networks, computing infrastructure and software, fuels data generation and automation.

Citizens across the globe face an expanding digital existence, with all aspects of life being impacted by technology, communications and data. Citizens and organisations face relentless attacks from a range of cyber actors, including opportunists, criminals and hostile states causing real harm to individuals, families, businesses and society. Advances in machine learning and artificial intelligence have created entirely new economic eco-systems, reshaped sectors and created a rush by enterprises to realise benefits from automation and AI. Continuation along a trajectory of increasingly sophisticated AI may radically reshape society, politics, economics and warfare across the globe [1].

The prospect of sophisticated AI enhancing cyber capabilities – both defence and attack – is not some far off theoretical concern [2]. As AI research leads to new technologies, advances will be adopted by individuals and organisations to enhance protection and safety, while threat actors develop new attack capabilities[1].

The UK Defence Science and Technology Laboratory (Dstl) and National Cyber Security Centre (NCSC) commissioned the Alan Turing Institute (the Turing) to define projects to provide enhanced automation within the context of Active Cyber Defence (ACD) – to conduct AI research in order to demonstrate new technologies that enhance cyber security defence capabilities within a civil and Defence context[2].

As part of this project, the Turing undertook analysis to formulate a roadmap for applied AI research in ACD. This report presents the results and recommendations of this analysis.

[1] Cyber threat actors will look to use AI to develop intelligent malware, automated operations and enhanced vulnerability and exploit capabilities. They will also seek out new ways to use AI to execute old scams, cons and ‘phishing style’ attacks. A BBC article from July 2019 describes how cyber criminals use ‘deepfaked’ audio to trick senior finance executives to steal large sums of money: https://www.bbc.com/news/technology-48908736

[2] While the focus of this work is on the use of AI for cyber security, consideration of the security of AI itself is not ignored. Consideration of ‘AI Safety’ is discussed in later sections.

> **“Dstl and NCSC must ensure that military and civil cyber security AI is fully considered by the UK’s research and technology community, despite intense globally competitive environment and commercial pressures driving market applications.”**

Artificial intelligence global context

Global competition for AI dominance is underway, where the United States (US) and China are the two major players [3] [4] [5] [6]. The US views AI dominance as an urgent strategic concern, arguing that dominance in AI has the potential to challenge established values, international norms and global order [7]. Consequently, ambitious plans and initiatives are being undertaken across the US government with the intention of securing global dominance in AI [8] [9]. China views AI dominance as essential to national survival and a mechanism to create competitive advantage over the US and other nations [5] [6] [10]. Both the US and China are making strategic investments in AI for economic, military and societal gain in accordance with the seriousness and urgency with which they seek to succeed. China has made substantial, rapid progress in both fundamental research and commercialisation of AI in recent years, challenging US global leadership [10] [11].

Other nations also seek advantages from AI. Russia is pursuing AI capabilities for autonomous weapons, cyber weapon systems and disinformation capabilities [12] [13] [14]. The European Union (EU) seeks a coordinated approach to bolster European AI skills, expertise and capability [15] [16]. Recent analysis has identified systemic European weaknesses (including the United Kingdom) in technology commercialisation, AI adoption and early stage (‘start-up’) investment [11]. Despite these global efforts, currently the US still leads the world in technology commercialisation and monetisation, fuelled by a start-up eco-system that has the risk appetite and financial means to fuel such dynamic enterprise [17] [18].

As global competition progress races ahead, international organisations are seeking to understand how best to frame, enable and govern AI developments for societal benefit and prevent potential hazards associated with major advances in AI [19] [20].

Within this global context, the UK’s strategic reviews of AI [21] [22] have largely focused on sustainment of world-class AI ‘brain trusts’ and expert participation in the development of safe, ethical AI norms and standards. As a result, the UK has increased investment in in academic growth (increased AI doctorate & post-doctorate positions). UK academia, led by the Turing, is growing the UK talent base through more AI research positions, educational initiatives and new partnerships.

This strategy is not without risk – a recent analysis of AI in the UK highlighted a major risk to this approach resulting from the UK's ability to translate academic excellence into world-leading products and services and the loss of graduates and PhD holders to China and the US due to favourable conditions for investment, innovation and salaries [23]. This concern is particularly acute for defence and national security, where assured supply chains and national sovereign capability are vital.

AI expertise remains in short supply and in high global demand. Major global technology companies and world leading research institutes compete to hire or acquire AI experts at a rapid pace. Such investment is welcomed within UK strategy[3] despite pressures and market forces for AI experts and resources to relocate across the globe, or tackle issues that do not align with UK, Defence or cyber security needs.

Given UK AI talent is likely to be acquired by global competitors, the UK Ministry of Defence (MOD), Dstl and NCSC may be confronted with insufficient body of experts considering the unique cyber security and Defence needs for AI. This is an existing concern within MOD with a recent report noting concern over an inability to access AI expertise, skills and talent [24]. Dstl and NCSC must ensure that military and civil cyber security AI is fully considered by the UK's research and technology community, despite intense globally competitive environment and commercial pressures driving market applications.

Dstl and NCSC must remain proactive and committed to research and technology development if they are to successfully drive AI advances for operational benefit, influence the AI landscape and provide valued advice and assurance.

[3] [19] states: "*Attracting and retaining the investment and expertise of the global majors is a key part of making the UK the best environment for developing AI.*" and " *To date, when US majors have acquired UK AI companies, the companies and their expertise have largely stayed in the UK."*

"ACD can be viewed as a NCSC programme that builds and runs services to help users be safer online. ACD can also be viewed as a concept for highly automated, scaled, proactive network defence inside civilian and military systems"

What is Active Cyber Defence (ACD)?

UK and US cyber security strategies place heavy emphasis on the development of intelligent cyber systems, defensive artificial intelligence and understanding adversarial cyber security AI [25] [26].

This drive towards intelligent systems is reflected within the concept of Active Cyber Defence (ACD). ACD concept seeks increased automation within an enterprise to bolster network defenders and cyber security. The term first appears in the literature in articles published by the US National Security Agency (NSA) in 2014 [27]. Intelligent automation is essential to enable system defenders to manage the risk posed by highly automated future threats and attacks, and to defend systems at cyber-relevant speed national scale [26] [7] [27] [28].

The ACD concept is not only an NSA concept; the wider US Department of Defence (DOD) and NATO have both placed strong emphasis on autonomous systems operating within Defence enterprise systems in large-scale, regional military contexts [29] [30] [31]. Furthermore, the NCSC has undertaken a programme of work in Active Cyber Defence.

In 2016, The Technical Director of the NCSC published a high-level description of NCSC's ACD programme [32]:

"The ACD programme is intended to tackle, in a relatively automated way, a significant proportion of the cyber attacks that hit the UK. Automation means the measures scale much better."

The NCSC's ACD programme has established a suite of NCSC services and products to tackle high-volume, commodity cyber attacks, helping users to be safer online [33]. As well as aiding users, ACD services generate data that allows NCSC to analyse and report the impact of ACD interventions [34] [35]. A recent academic review of NCSC ACD highlighted programme success, and suggested that the programme can be viewed as a 'public good' provided by Government to provide benefit to society [36]. Such a scale-out of existing NCSC products and services beyond the public sector raises engineering challenges, but also presents opportunities for greater observational data at national scale. NCSC continues to grow applied data science efforts, seeking novel 'narrow AI' automation that will support and enhance NCSC ACD services and products.

ACD can be viewed as a NCSC programme that builds and runs services to help users be safer online. ACD can also be viewed as a concept for highly automated, scaled, proactive network defence inside civilian and military systems.

The ACD concept presents ambitious requirements for intelligent automation capabilities. In order to create these technologies, researchers and technologists must overcome difficult challenges in artificial intelligence, network security and human-machine teaming. Sustained research in fundamental AI topics applied to realistic cyber scenarios is needed to address the ambitious aims of the ACD concept [37] [28] [25].

> **“Applied data science projects alone will not drive progress towards highly automated intelligent systems for proactive defence of networks, at national scale and cyber-relevant speed. Breakthroughs in intelligence autonomous systems, knowledge representation and reasoning, AI safety, explainability and assurance and human machine teaming are required.”**

Artificial intelligence and cyber security

The AI research and technology landscape continues to grow at a rapid pace, with accelerating pace of progress across theory, technology and applications [18] [38].

AI has seen adoption within the cyber security market. Vendors and suppliers seek market advantage in protection and defence products and services and threat actors seek new attack capabilities. To date, AI within the cyber security domain has largely focused on important ‘narrow AI’ applications, the two primary areas being the automated detection and classification of malware and automated threat alerting on network traffic and endpoint sensor datasets [39] [40].

Continued applied data science efforts will likely yield near-term benefits for Dstl and NCSC in current services and products. Dstl and NCSC are encouraged to continue to explore useful large-scale analytics and automation, for example creating models to automatically perform predictions such as:

- Is this new TLS certificate for a malicious domain? *(Certificate Transparency data)*
- Is this new website/webpage/URL malicious? *(Common Crawl data)*
- Is there a new cyber attack underway within my network? *(NCSC Logging Made Easy data)*
- Is this user activity suspicious enough to intervene? *(NCSC Logging Made Easy data)*
- Is this malware novel (e.g. targets industrial/telecoms/military systems)? *(VirusTotal data)*
- Is this DNS traffic exfiltrating data? *(NCSC P-DNS data)*
- Is this new software vulnerability an urgent priority for action? *(CVE, CVSS data)*
- Is this email/attachment/URL part of a phishing scam? *(PhishTank data)*

While valuable and important, such applied data science projects alone will not drive progress towards highly automated intelligent systems for proactive defence of networks, at national scale and cyber-relevant speed. Breakthroughs in intelligence autonomous systems, knowledge representation and reasoning, AI safety, explainability and assurance and human machine teaming are required.

Research breakthroughs must be developed into prototypes, prototypes developed into reusable technologies, and sustained engineering effort spent to build real-world systems.

The UK hosts world-leading research institutions in both AI and cyber security. UK research in AI is consistently ranked as world leading, with the UK being ranked third behind the US and China for AI research [23]. UK academia also excellent cyber security research capability, with 19 UK universities as academic centres of excellence for cyber security research by NCSC in June 2019 [41]. Several universities host both leading AI and cyber security groups, with several collaborative efforts between AI researchers and cyber security researchers. Cardiff University's Centre for Cyber Research [42] is one example of a cross-disciplinary group focused on AI in cyber security.

As research and development creates better AI, human decision makers may be displaced from ever more complex, subtle and consequential decision-making settings. As the complexity and seriousness of AI applications increases, proving these AI systems are safe is essential. AI technologies are being targeted by focused adversarial attacks, presenting a new class of security risks. 'AI Safety' is a vital area of research, and exceptionally important in the Defence context.

In an environment where people and systems are relentlessly targeted and attacked by determined, capable, resourced hostile actors and where decisions and actions have the potential for grave human consequences, AI safety is a central concern. The recently published report on AI Ethics by the US DOD [43] sets out five principles to guide the development and adoption of AI within the US Defence enterprise, of which two principled directly relate to the safety and assurance of AI systems. Trust, safety, reliability and assurance are recognised as fundamental attributes for any real-world AI systems in these domains [7] [25] [37]. Dstl, NCSC and the Turing should seek to follow these guidelines across all efforts within this domain.

Efforts to discover new AI techniques for autonomous ACD capabilities is a core technical focus. However, the need to discover effective forms of integration between AI capabilities and human-centric organisation, processes and activities is also an essential element of research for autonomous ACD systems. UK MOD recently published a doctrine note placing great emphasis on understanding and exploiting optimal human-machine teaming for autonomous systems [24]. Within the context of ACD different organisations, military units and cyber defence teams will each need tailored configurations to enhance readiness, preparedness and reaction to cyber threats – not a centralised 'one size fits all' AI systems that risks hindering normal operations or impeding mission activities.

Research into human-machine teaming offers a clear collaboration opportunity for Dstl, NCSC and the Turing human sciences, AI and cyber security experts to explore radical, innovative concepts. Cognitive science research into prospection [44] and anticipatory thinking [45] has the potential to shape AI research into the formulation and evaluation of possible actions within a complex scenario and inform research into autonomous formation of distributed cyber situational awareness systems. Dstl and NCSC should seek to form collaborative engagement with their respective human sciences and socio-technical expert groups to ensure the very best cross-disciplinary research is delivered.

Several studies have highlighted the need for realistic, representative environments in which to develop and evaluate cyber security research [46]. The authors of the IEEE “Grand Challenge: Applying Artificial Intelligence and Machine Learning to Cyber Security” article propose a cyber arena/rodeo [46] initiative to fuel progress in cyber security AI. The paper states:

“In creating an arena/rodeo environment as a fully instrumented digital twin of both the resources to be protected as well as the SOC in which the cybersecurity analysts operate, we establish the environment needed to capture the in vivo behaviour of the analyst as well as add external expert analyst commentary. This platform and the standardization of logging form the basis for the training, simulation, and analysis of strategies analogous to those related to strategic competitive games, such as chess and poker.”

The 2016 DARPA Cyber Grand Challenge [47] is a famous example of a realistic, controlled, competitive environment that led to tangible advances, derived from long-standing research programmes within industry and academia in automated binary analysis, vulnerability discovery, exploit development and patch generation.

The general approach of research teams participating in competitive, realistic cyber security challenges offers a mechanism to close the gap between research and pull-through into real world applications. Dstl and NCSC should consider sponsoring sustained, multi-year team participation in key challenges to ensure constructive pressure on the creation of demonstrable prototypes that tackle hard problems in realistic environments. Sponsorship of sustained, multi-year team participation in key challenges provides a way to apply constructive pressure on research teams, placing emphasis on the creation of demonstrable prototypes that tackle hard problems in realistic environments. A step beyond these games/competitions is the opportunity to leverage research outcomes on human-machine teaming and autonomous defender action and threat identification within existing cyber exercises, possibly through the NCSC Exercise in a Box service [48] and the larger scale Cyber Warrior exercises [49].

AI research challenges for ACD are not new - AI researchers have been exploring fundamental approaches to intelligent autonomous systems for decades. More recently, researchers have been pursuing new ideas in explainable AI and adversarial machine learning. All of this research has relevance for ACD, and examples exist of research that is directly related to ACD challenges including the development of autonomous agents, capable of identifying, triaging and responding to threats without direct human intervention [40] [50] [51] [52].

The UK’s world class research environment, collaborative partnerships and facilities provide an excellent foundation upon which to create new civil and military ACD AI capabilities. This foundation provides an opportunity to Dstl and NCSC to lead AI research initiatives, and are encouraged to:

- Provide sustained investment in impactful, high-calibre research, technology projects and teams.
- Engage with academic research groups to influence research to consider ACD challenges.
- Work to exploit results to develop new technologies for real-world AI systems.

“Dstl, NCSC and the Turing should establish agile governance processes to enable rapid commission and review of research to maximise progress and impact in priority areas”

AI for ACD roadmap

This section defines a roadmap for AI research applied to ACD and is intended to support research planning within Dstl, NCSC and the Turing. The roadmap combines analysis from previous sections to set out a structure within which to define, deliver and iterate novel research ideas to progress towards ACD concept ambitions, in military and civil domains.

The goal of this roadmap is to create radical advancements in artificial intelligence (AI) powered autonomous cyber defence systems.

Success criteria

Conducting high-calibre research that demonstrates positive impact in real-world military and civilian ACD is essential. To aid reproducibility, adoption and future work, researchers should seek to publish results as both peer reviewed papers and as baseline implementations and prototypes, alongside exemplar datasets.

Researchers need to pressure test results, technologies and prototypes in realistic cyber defence scenarios, and strive to close the gap between research and application. Opportunities exist to prototype, trial and evaluate research within UK Government and Defence cyber security exercises, and broader cyber security challenges (such as `IEEE Arena initiative [46] and US Department for Homeland Security (DHS) Cybersecurity and Infrastructure Security Agency (CISA) Presidents Club event [53]).

High-calibre research should lead to positive impact on ACD capability development and positive academic impact. Dstl and NCSC should have access to a larger, more expert group of cyber security AI researchers and technologists, and also have grown internal AI knowledge and expertise.

Four high-level positive indicators are proposed to gauge progress across these key areas of success:

- **Demonstration of prototype AI technology** within cyber defence challenges, achieving cutting-edge results.
- **Growth in AI capability within Dstl and NCSC** by creating new technologies, knowledge and skills.
- **Publish papers** in leading venues/journals to advance the state of the art.
- **Publish high quality reusable software, data and documentation** artefacts to encourage future research.

These success criteria are aligned with the Turing Defence and Security (D&S) Programme priorities for impact [54]. The Turing D&S impact strategy paper states:

"It is assumed that all research under the programme will strive to be academically excellent. However, the priority of the programme is to translate research into operational impact which would fit under the technological, economic, political, societal impact types. Research that only achieves purely academic impact and is not translated into real-world impact in another category is a lower priority for the programme."

The Turing D&S Impact strategy defines four types of impact[4] (the type of change created as a result of research) and five categories[5] of impact (areas in which change will be realised). The goal and success criteria fully align with the impact priorates defined in [54], placing emphasis on excellent research that creates novel prototypes, and leads to positive advances in the state of the art in applied environments.

It is recommended that all projects, activities and tasks proposed within this roadmap define expected impact using the Turing D&S impact criteria. Dstl, NCSC and the Turing should seek to base prioritisation decisions in line with the Turing D&S programme's emphasis on operational impact.

Capability challenges

Challenges are presented to guide relevant and impactful research, describing possible futures in which AI is used to improve, augment, and advance current cyber security capabilities. If high calibre research ideas cannot be directly linked to at least one challenge, it should be noted as outside the scope of this roadmap. Capability challenges provide an additional focus within the roadmap on real-world impact and also support stakeholder desire to use participation in competitive cyber security events as a key proving ground for research outcomes.

AI-enabled network defender ('AI Blue Team').

The term 'network defender' is commonly used to refer to teams charged with proactive and responsive action to prevent, contain, heal cyber-attacks [55]. The term 'blue team' is commonly used to refer to human teams that adopt a defensive position within a scenario, exercise or operation and focus on the protection of a target system [56].

This challenge relates to the development of intelligent, autonomous and resilient AI systems to fend off attacks, and relates to systems that operate within an enterprise to detect, identify and predict threats, assess defensive options, decide on effective options then act to protect networks within a continuous feedback loop. Such systems will need to reason under uncertainty, reason over a range of complex courses of action, evolve their understanding of actions, options and consequence, evaluate potential consequences of action and provide an explainable justification for actions.

[4] Four types of impact are: Understanding; Attitudinal; Behavioural; and Capacity.

[5] Five categories of impact are: Academic; Economic; Cultural; Environmental; Health; Political; Technological and Societal.

To be viable within the cyber security domain, systems must: robust to sabotage, subversion; resilient to deliberate or accidental effects that repurpose the agent to damage, remove or destroy data and assets within a target system; and must operate at large scale under challenging compute, memory and networking constraints.

AI-enabled security planner (‘AI Cyber Situational Awareness’).

Some large organisations are able to sustain the necessary investment to run sophisticated Security Operations Centres (SOC) to enhance their cyber security posture. A SOC is a capability to create visibility across enterprise systems on cyber attack detection, triage and response [57]. SOCs provide decision-makers and planners with cyber situational awareness across their organisational systems, identifying threats, issues and the responses undertaken to defend and protect systems.

At present, SOC-like capabilities are dependent on deep human expertise, highly specialist knowledge and skills, and team capacity which represents a major hurdle to establishing a SOC, and serious constraint on speed and scale of operation. As such, the majority of organisations are unable to benefit from these kinds of capabilities. This challenge relates to the development of intelligent, autonomous and resilient AI systems to perform threat situational awareness, threat triage, security advice and defensive action tracking without human intervention at machine-speed and scale. For Defence, creating decentralised cyber situational awareness that can operate in harsh operating conditions and in resource constrained environments is a complex challenge. In the NCSC ACD context, creating situational awareness at national scale, across many thousands of organisations that can be used to generate advice and alerts is a complex challenge.

AI enabled automated systems for cyber operations situational awareness will need to reason under uncertainty, reason over a range of complex courses of action, explore the consequences of action and provide an explainable justification for actions. Such systems must be robust to sabotage, subversion, and deliberate or accidental effects that degrade system performance or cause damage or degradation within a target system.

AI-enabled penetration tester (‘AI Red Team’).

The term ‘penetration testing’ refers to a particular evaluation exercise undertaken by security experts to review the information assurance capabilities of an organisation and to test operational systems against known cyber-attacks [58]. The term ‘red team’ is commonly used to refer to a highly skilled human team, tasked with seeking to identify flaws, holes or vulnerabilities in a target network under controlled conditions to enable organisations to bolster defence [56]. This challenge relates to the development of intelligent, autonomous and resilient AI systems that can probe, profile, explore and target enterprise systems and identify points of weakness or possible compromise.

Such systems will need to be able to reason under uncertainty, reason about previously unknown circumstances, update its behaviour and reasoning as new knowledge is created about vulnerabilities and exploits and be able to communicate its results in a machine-readable manner for subsequent analysis and defender responses. Such systems will have to be robust against being intentionally or accidentally redirected for malicious ends or repurposed to cause harm or damage.

These challenges are not disjoint. Progress in one challenge area will be relevant to others, and a dynamic interplay of research and technology is anticipated. The emphasis and nature of each of the challenges is distinct enough to reflect distinct technical challenges and domains of expertise. Any future integrated active cyber defence system would draw upon the technological progress made across all challenges, which may define radically different approaches to defending data, systems and services in a civilian and military context.

It is recommended that Dstl, NCSC and the Turing actively engage with open competitions within the cyber security domain, in order to identify any relevant competition events, shape future events and create opportunities for participation.

Candidate projects

A set of possible AI research projects, designed to make demonstrable progress against capability challenges and satisfy success criteria and create impact. Specific project ideas will change and evolve as new research results and trends emerge. The list of candidate projects detailed below is not comprehensive and should not be viewed as static or final. New projects will start fresh, others will build on prior successes; stalled, closed or completed projects will be closed to ensure focus and tempo. Dstl, NCSC and the Turing should establish agile governance processes to enable rapid commission and review of research to maximise progress and impact in priority areas.

Threat situational awareness at scale

Cyber security situational awareness refers to efforts to generate a common, unified and shared representation of the state of an enterprise and it's security posture [59] [60]. This project seeks to develop large-scale threat models that enable decision-makers to understand observed threat events and anticipated/predictive threat events, using industry standard threat representation frameworks [61] [62] [63] to provide a near-real time, enterprise view on current and anticipated threats across the system. Specifically, this project should seek to explore models and technologies for distributed cyber threat situational awareness, situational awareness models and threat knowledge base integrations.

Automated cyber defence course of action evaluation

In order for cyber defences to be automated, some representation of the set of possible actions that can be taken within a network is required. Given a representation of complex sets of actions within a complex network environment, automated defender systems must be able to identify feasible, relevant, effective courses of action and be able to reason about the risks associated with each, if some notion of preferred course of action is to be identified and enacted.

This project seeks to explore data driven and knowledge driven approaches to network defender actions, providing representations aligned to industry standard models [64] and problem structuring tools [65], and build on existing works – specifically Dstl's Prototype for Automated Network Defence Actions (PANDA) project that seeks to build an implementation of OpenC2 [64] network defender actions [66].

This project would include reviews of the utility and feasibility of integrating cognitive modelling [67], probabilistic risk analysis [68] [69], reinforcement learning [50] and exploration of other agent development approaches to create AI-based cyber reasoning agents capable of effective near-real time automated cyber defence actions.

Autonomous competitive cyber agents

This project seeks to apply cutting edge AI adaptive control algorithms within a highly realistic, complex and dynamic simulation environment to train, develop and experiment with competitive, adaptive agents for cyber defence and cyber-attack [40] [52] [70] [71] [72]. In essence, this is a research project seeking to create AI-powered attacker agents capable of reasoning about complex attack sequences, minimising their observability, evading detection and defender agents able to seeking out threat actors within systems and responding to attacks and events with appropriate actions.

This project will explore simulated environments and cyber scenarios, developing techniques, models and agents that span multi-agent team cooperation and competitive multi-agent simulations [73] [74] [75]. It is likely that this project will build on prior work developing industry-standard representations of cyber threats (ATT&CK [61]) and defence (OpenC2 [64], IACD playbooks [65]), and also have linkages to broader automated cyber reasoning system relating to automated vulnerability, exploit and patch development [47].

Adversarial machine learning and agent attacks

Adversarial machine learning is an active area of research that seeks to develop techniques to create input examples that appear normal/benign to human observers, but cause machine learning classifiers to error. Adversarial machine learning attack and defence research has been particularly active in the field of deep neural networks. Recent research has created new adversarial techniques to attack agents trained using cutting edge reinforcement learning techniques [76]. The development of autonomous cyber defence agents using reinforcement learning techniques must be robust to such attacks. This project seeks to evaluate the state of the art, and develop techniques to defend for cyber agent training [52] [51] [77] [78].

Threat identification using causal modelling of weak indicators

Many current approaches to threat detection within cyber security products is based on anomaly detection concepts. This project seeks to look beyond these techniques and explore the application of novel models, algorithms and techniques to network and host based data to automatically characterise types and status of cyber attacks using industry standard models such as ATT&CK [61], allowing a cause and effect model of threat actor actions to be generated from data, integrating domain expertise. The goal is to explore the feasibility of developing causal models of threat operations from data, alongside casual models for threat identification in complex distributed systems [62] [69] [79] [80] [81] [82].

Explainable AI for agent action justification

AI safety, trust and accountability are critical factors for defence and security AI systems. Explainable AI as applied to intelligent agent decision-making explanation and justification is active area of research [83] [84] [85] [86]. This project seeks to explore the state of the art in explainable AI, applied to autonomous agents operating within a cyber security network defence scenario exploring models for capturing decision-making and planning and creating human-centric natural language, alongside machine readable structured action justifications.

Generative modelling for decoy defences

The use of decoy assets, honeypots and honeynets are widely discussed in the cyber security domain [87] including use of such techniques are often used in order to generate cyber threat intelligence (CTI) such as indicators and signatures for cyber attacks [88]. Decoy assets have also been explored as a mechanism to frustrate cyber attackers efforts to identify and steal valuable data within an enterprise, and provide a 'trip wire alert' mechanisms [89], and such approaches may become more relevant as ransomware attacks continue to grow in sophistication targeting high-value files, databases and backups. This project seeks to explore the use of generative modelling techniques to create decoy assets (e.g. datasets, files, accounts, network traffic) such that cyber threat actors are unlikely to differentiate decoys assets from true assets within an enterprise.

Defeating AI detection using generative modelling

This project seeks to explore generative artificial intelligence approaches to create models that output tailored, targeted synthetic data (network and host-based sensor data) that can defeat AI-based threat detection systems. Research will explore whether such generative modelling techniques may be implemented in a dynamic fashion, so as new AI-defences are developed, new tailored adversarial data is created.

Autonomous cyber agent and human teaming

UK MOD has highlighted human-machine teaming as a critical concern for the development of future capabilities [24]. Recent research has identified opportunities and challenges relating to human-machine teaming in the context of cyber security operations [90]. Wider cognitive and human science [45] [44] research will also be explored in a cross-disciplinary research effort to identify effective models for AI integration within realistic military and civil cyber security operations.

This project seeks to conduct workshops, trials and exercises that integrate human participants and automation technologies within a realistic cyber exercise environment (including realistic scenarios) in order to explore how AI systems performing blue team and red team functions can be harnessed for effective and efficient machine-human teaming strategies.

Hypergame modelling cyber threat evolution

The interplay between attack and defence within the cyber security domain is a complex dynamic system comprised of many multiple participants, each with individual intent and capabilities, expertise and resources and each with an imperfect, partial awareness of the global environment. Analysis has shown that as organisations successfully deny cyber threats the use of simple, routine or commodity attacks, threat actors seek new targets and seek new capabilities and approaches. This project to explore recent advances in game theory and hypergame theory (those where the rules of the games are themselves hyperparameters), to explore the utility of the technique to create large-scale complex dynamic cyber system simulations, to support long-term reasoning and planning [91] [92] [93].

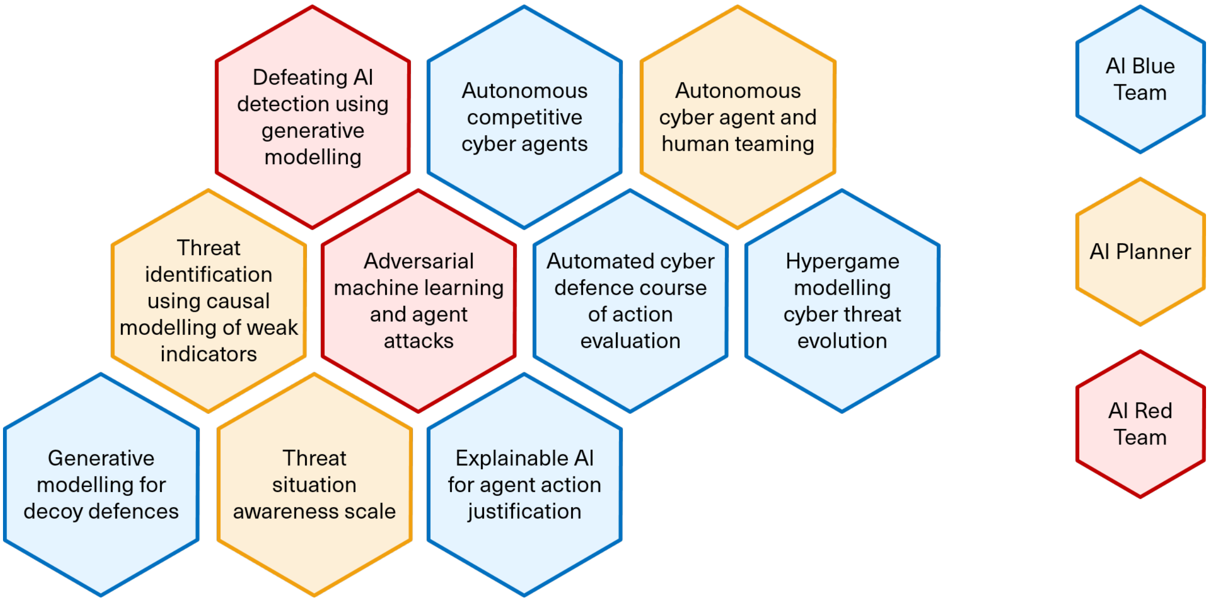

Defeating AI detection using generative modelling
Autonomous competitive cyber agents
Autonomous cyber agent and human teaming
AI Blue Team
Threat identification using causal modelling of weak indicators
Adversarial machine learning and agent attacks
Automated cyber defence course of action evaluation
Hypergame modelling cyber threat evolution
AI Planner
Generative modelling for decoy defences
Threat situation awareness scale
Explainable AI for agent action justification
AI Red Team

Figure 1: Visualisation of candidate project connections, colour-coded by challenge

Enablers

Resources, facilities and assistance to enable researchers and technologists to make progress against complex intellectual challenges, with the fewest practical possible obstacles and least possible amount of friction.

Cyber range experimentation infrastructure

Recent analysis from US Sandia National Laboratories [94] [95] highlights the needs for enhanced cyber range infrastructure and experimental processes to underpin next-gen cyber security research. Realistic environments representing target systems of interest - including Supervisory control and data acquisition (SCADA), Internet of Things (IOT), on-premises enterprise and Defence-unique systems are essential to prove real-world impact. Examples such as US Cyber Command's DreamPort facility [96] and NATO Locked Shields exercise infrastructure, scenarios and processes [97] may represent suitable infrastructure for scalable ACD experimentation [98]. Sustaining such facilities is a major undertaking but essential for examining research outputs in a life-like environment.

Cyber simulation environment

Credible simulation environment for exploration of AI adaptive control approaches to creating autonomous cyber agents. Simulation is required to enable scalable training, rapid iterative research that is unfeasible with replica cyber infrastructure. Alignment between simulation and 'cyber range' environments is an important enabler for rapid transition of research.

Data lab workbench

Secure, scalable and flexible data science environment to allow researchers to easily explore and experiment on existing datasets to create novel models, techniques and results. Such a capability can act as a bridge between ACD 'narrow AI' tasks, and broader AI experimentation and research projects.

Cyber scenarios and datasets

Define, capture and maintain a set of realistic cyber exercise scenarios and scenario datasets. These artefacts are essential to represent use cases, and at least one scenario should be created to represent the unique environment and constraints of military networked systems. Alignment with IACD playbooks and ATT&CK threat modelling ensures commonality with frameworks used in wider research and development.

Collaboration tools

Researchers should be able to collaborate remotely in a seamless way, to share ideas, to plan and to create as a collective. Commonly used tools and services should be adopted and shared amongst researchers and team to encourage collaborative working. Industry standard tools for software development (including team messaging, software development tools, version control, continuous integration, agile task tracking, collaborative wiki etc.) should be established to support project teams.

Dstl, NCSC and the Turing should review existing resources and facilities to determine if any can be provided to research teams to satisfy these needs.

These elements together – goal, challenges, projects, enablers and success criteria – make up the research roadmap. These elements are linked with dependencies and relationships between them. ANNEX A presents a visual summary of the roadmap, showing links between elements.

Recommendations

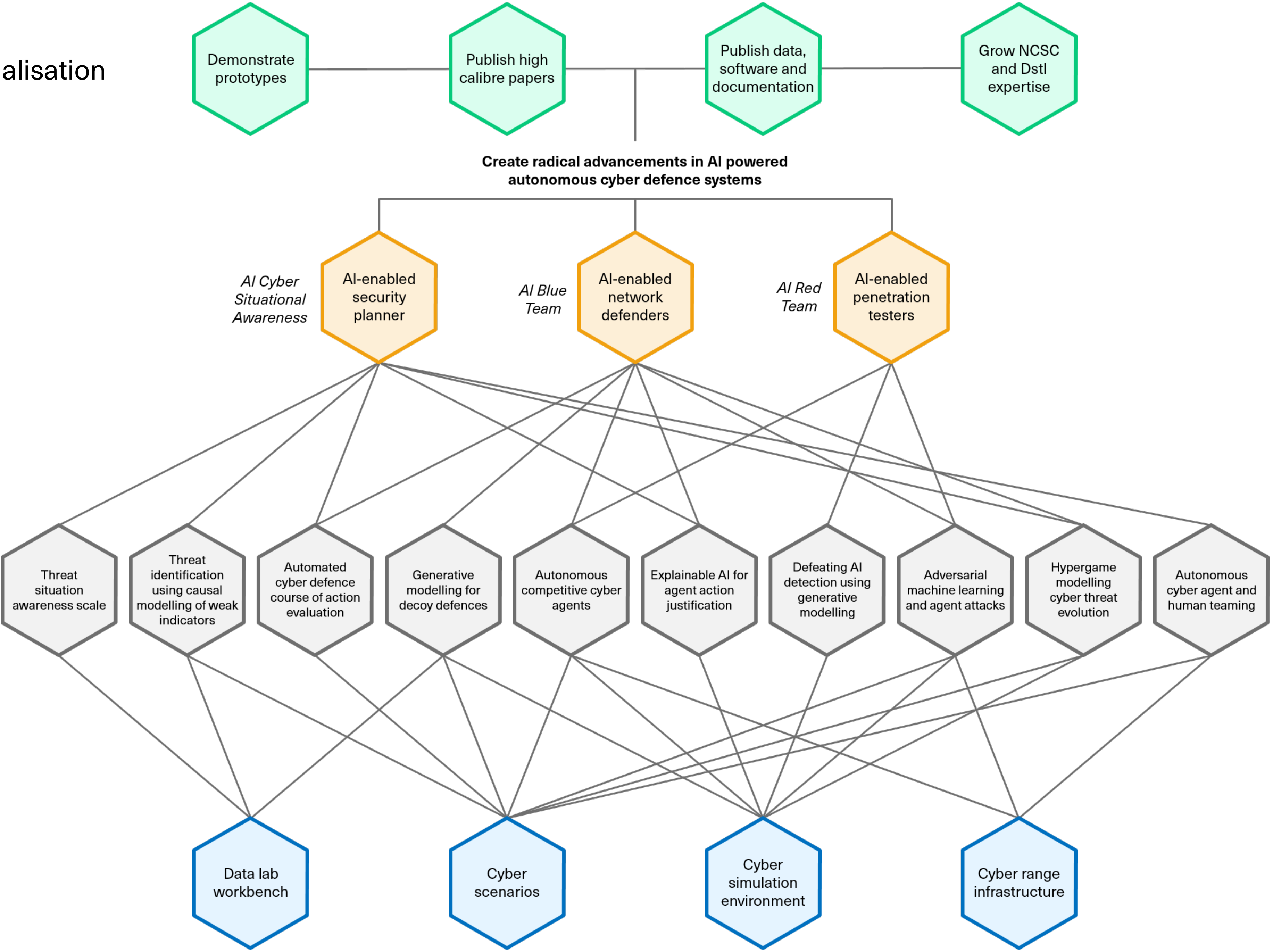
Dstl and NCSC operate in a globally competitive environment for AI. It is vital that Dstl and NCSC remain proactive and committed to AI research to ensure military and civil cyber security needs are fully considered, to access expertise to create new AI technology, and to provide assurance and advice on AI.

This roadmap provides a structure for research projects to advance the state of the art in AI for active cyber defence. The structure allows researchers and technologists to explore state of the art techniques, to be dynamic and agile with ideas and projects, while ensuring that work is targeted against ACD capability challenges in civil and military environments.

It is recommended that Dstl, NCSC and the Turing continue to work in partnership in order to pursue and exploit impactful AI research in order to seize future opportunities for active cyber defence.

Annex A

Roadmap visualisation



Demonstrate prototypes
Publish high calibre papers
Publish data, software and documentation
Grow NCSC and Dstl expertise
Create radical advancements in AI powered autonomous cyber defence systems
AI Cyber Situational Awareness
AI-enabled security planner
AI Blue Team
AI-enabled network defenders
AI Red Team
AI-enabled penetration testers
Threat situation awareness scale
Threat identification using causal modelling of weak indicators
Automated cyber defence course of action evaluation
Generative modelling for decoy defences
Autonomous competitive cyber agents
Explainable AI for agent action justification
Defeating AI detection using generative modelling
Adversarial machine learning and agent attacks
Hypergame modelling cyber threat evolution
Autonomous cyber agent and human teaming
Data lab workbench
Cyber scenarios
Cyber simulation environment
Cyber range infrastructure

Figure 2: Diagram showing relationship between enablers, projects, challenges and success criteria

References

[1] Cummings et al, "Artificial Intelligence and International Affairs Disruption Anticipated," 4 June 2018. [Online]. Available: https://reader.chathamhouse.org/artificial-intelligence-and-international-affairs#. [Accessed 15 March 2020].

[2] University of Oxford, "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," February 2018. [Online]. Available: https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads/MaliciousUseofAI.pdf?ver=1553030594217. [Accessed 11 February 2020].

[3] US National Security Commission on Artificial Intelligence, "NSC on AI Initial Report to Congress," 2019.

[4] US Select Committee on Artificial Intelligence of the National Science & Technology Council, "National AI R&D Strategic Plan: 2019 Update," 2019.

[5] Chinese State Council, "China's New Generation of Artificial Intelligence Development Plan," 30 July 2017. [Online]. Available: https://flia.org/notice-state-council-issuing-new-generation-artificial-intelligence-development-plan/.

[6] Allen, "Understanding China's AI Strategy," 06 February 2019. [Online]. Available: https://s3.amazonaws.com/files.cnas.org/documents/CNAS-Understanding-Chinas-AI-Strategy-Gregory-C.-Allen-FINAL-2.15.19.pdf?mtime=20190215104041.

[7] DOD, "Summary of the 2018 Department of Defence Artificial Intelligence Strategy," US DOD, 2019.

[8] Selman et al, "A 20-Year Community Roadmap for Artificial Intelligence Research in the US," Computing Community Consortium (CCC) and Association for the Advancement of Artificial Intelligence (AAAI), 2019.

[9] Russell, et al "Research Priorities for Robust and Beneficial Artificial Intelligence," Association for the Advancement of Artificial Intelligence, pp. 105-114, 2015.

[10] Ding, "Deciphering China's AI Dream," 14 March 2018. [Online]. Available: https://www.fhi.ox.ac.uk/wp-content/uploads/Deciphering_Chinas_AI-Dream.pdf.

[11] M. M. E. C. Daniel Castro, "Who Is Winning the AI Race: China, the EU or the United States?" 19 August 2019. [Online]. Available: https://www.datainnovation.org/2019/08/who-is-winning-the-ai-race-china-the-eu-or-the-united-states/.

[12] Future of Life Institute, "AI Policy - Russia," 11 January 2019. [Online]. Available: https://futureoflife.org/ai-policy-russia/.

[13] Konaev, "Thoughts on Russia's AI Strategy," 11 February 2019. [Online]. Available: https://cset.georgetown.edu/2019/10/30/russias-ai-strategy/.

[14] Polyakova, "Weapons of the weak: Russia and AI-driven asymmetric warfare," 11 November 2018. [Online]. Available: https://www.brookings.edu/research/weapons-of-the-weak-russia-and-ai-driven-asymmetric-warfare/.

[15] European Commission Joint Research Centre, "Artificial Intelligence - A European Perspective," 01 August 2018. [Online]. Available: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC113826/ai-flagship-report-online.pdf.

[16] Directorate-General for Communications Networks, Content and Technology, "The European Artificial Intelligence landscape," 01 April 2018. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/european-artificial-intelligence-landscape.

[17] Horowitz et al, "Strategic Competition in an Era of Artificial Intelligence," Centre for a New American Secuirty, 2018.

[18] Mateos-Garcia et al, "A Semantic Analysis of the Recent Evolution of AI Research," 15 November 2019. [Online]. Available: https://media.nesta.org.uk/documents/A_Semantic_Analysis_of_the_Recent_Evolution_of_AI_Research.pdf.

[19] Future of Life Institute, "AI Policy - United Nations," 2019. [Online]. Available: https://futureoflife.org/ai-policy-united-nations/. [Accessed 11 February 2020].

[20] OECD Council on Artificial Intelligence, "Recommendation of the Council on Artificial Intelligence," 01 May 2019. [Online]. Available: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[21] Hall and Pesenti, "Growing the artificial intelligence industry in the UK," UK Government, 2019.

[22] UK House of Commons Science and Technology Committee, "Robotics and artificial Intelligence," UK Parliment, 2016.

[23] Tortoise Media, "The Global AI Index," 2019. [Online]. Available: https://www.tortoisemedia.com/intelligence/ai/. [Accessed 11 February 2020].

[24] Development, Concepts and Doctrine Centre, "Human-Machine Teaming," UK Ministry of Defence, 2018.

[25] UK Cabinet Office, "Interim cyber security science and technology strategy," UK Cabinet Office.

[26] Cyber Security and Information Assurance Interagency Working Group, "Federal Cyber Secuity Research & Develpopment Strategic Plan," US National Science and Technology Council, 2019.

[27] Herring et al, "Active Cyber Defence: A Vision for Real-Time Cyber Defence," Journal of Information Warfare, pp. 46-55, 2014.

[28] US National Security Agency, "Active Cyber Defense," 01 August 2015. [Online]. Available: https://apps.nsa.gov/iaarchive/programs/iad-initiatives/active-cyber-defense.cfm.

[29] De Lucia et al, "Features and Operation of an Autonomous Agent for Cyber Defense," Journal of Cyber Security and Information Systems, pp. 6-13, 2019.

[30] Tonin, "2019 Artificial Intelligence: Implications for NATOs armed forces," 01 October 2019. [Online]. Available: https://www.nato-pa.int/sites/default/files/2019-10/REPORT%20149%20STCTTS%2019%20E%20rev.%201%20fin-%20ARTIFICIAL%20INTELLIGENCE.pdf.

[31] Kott, "Intelligent Autonomous Agents are Key to Cyber Defense of the Future Army Networks," The Cyber Defence Review, pp. 57-70, 2018.

[32] Levy, "Active Cyber Defence - tackling cyber attacks on the UK," 01 November 2016. [Online]. Available: https://www.ncsc.gov.uk/blog-post/active-cyber-defence-tackling-cyber-attacks-uk .

[33] NCSC, "Products & Services," 01 February 2019. [Online]. Available: https://www.ncsc.gov.uk/section/products-services/active-cyber-defence.

[34] Levy, "Active Cyber Defence - one year on," 04 February 2018. [Online]. Available: https://www.ncsc.gov.uk/blog-post/active-cyber-defence-one-year.

[35] Levy, "Active Cyber Defence (ACD) - The Second Year," 16 July 2019. [Online]. Available: https://www.ncsc.gov.uk/report/active-cyber-defence-report-2019.

[36] Stevens et al, "UK Active Cyber Defence: A public good for the private sector," 01 January 2019. [Online]. Available: https://www.kcl.ac.uk/policy-institute/assets/uk-active-cyber-defence.pdf.

[37] Office Director of National Intelligence, "The AIM Initiative: A Strategy for Augmenting Intelligence Using Machines," 16 January 2019. [Online]. Available: https://www.dni.gov/files/ODNI/documents/AIM-Strategy.pdf.

[38] Perrault et al, "The AI Index 2019 Annual Report," Human-Centered AI Institute, Stanford University, , Stanford, CA, USA, 2019.

[39] Amit et al, "Machine Learning in Cyber-Security - Problems, Challenges and Data Sets," 22 April 2019. [Online]. Available: https://arxiv.org/pdf/1812.07858.pdf.

[40] Nguyen et al, "Deep Reinforcement Learning for Cyber Security," 20 June 2019. [Online]. Available: https://arxiv.org/pdf/1906.05799.pdf.

[41] NCSC, "Academic Centres of Excellence in Cyber Security Research," 25 June 2019. [Online]. Available: https://www.ncsc.gov.uk/information/academic-centres-excellence-cyber-security-research. [Accessed 11 February 2020].

[42] Cardiff University, "Centre for Cyber Security Research," [Online]. Available: https://www.cardiff.ac.uk/centre-for-cyber-security-research/research. [Accessed 11 February 2020].

[43] US DOD Defence Innovation Board, "AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense," 31 October 2019. [Online]. Available: https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF. [Accessed 11 February 2020].

[44] Szpunar et al, "A taxonomy of prospection: Introducing an organizational framework for future-oriented cognition," Proceedings of the National Academy of Sciences, vol. 52, no. 111, p. 18414–18421, 2014.

[45] Amos-Binks et al, "Cognitive Systems for Anticipatory Thinking 2019," in Short Paper Proceedings of the Workshop on Cognitive Systems for Anticipatory Thinking (COGSAT 2019), Arlington, USA, 2019.

[46] Bresniker et al, "Grand Challenge: Applying Artificial Intelligence and Machine Learning to Cybersecurity," Transactions on Computers, vol. 52, no. 12, pp. 45-52, 2019.

[47] DARPA, "Cyber Grand Challenge," 2015. [Online]. Available: https://www.darpa.mil/program/cyber-grand-challenge. [Accessed 11 February 2020].

[48] NCSC, "Exercise in a Box," 16 April 2019. [Online]. Available: https://www.ncsc.gov.uk/information/exercise-in-a-box. [Accessed 15 March 2020].

[49] Director GCHQ, "Speech at CyberUK18," 18 April 2018. [Online]. Available: https://www.gchq.gov.uk/speech/director-cyber-uk-speech-2018. [Accessed 11 February 2020].

[50] Nguyen, "Multi-agent deep reinforcement learning with human strategies," in 2019 IEEE International Conference on Industrial Technology, Melbourne, 2019.

[51] Tong Chen, "Adversarial attack and defense in reinforcement learning-from AI security view," Cybersecurity, pp. 1-22, 2019.

[52] Elderman et al, "Adversarial Reinforcement Learning in a Cyber Security Simulation," ICAART, pp. 559-566, 2017.

[53] US DHS CISA, "WELCOME TO THE 2019 PRESIDENT'S CUP CYBERSECURITY COMPETITION!," 2019. [Online]. Available: https://www.cisa.gov/presidentscup. [Accessed 11 February 2020].

[54] Nielson, "Defence and Security Programme Impact Strategy," Alan Turing Institute, 2018.

[55] EC-Council, "Certified Network Defender Certification," International Council of E-Commerce Consultants, [Online]. Available: https://www.eccouncil.org/programs/certified-network-defender-cnd/. [Accessed 11 February 2020].

[56] Kick, "Cyber Exercise Playbook," The Mitre Corporation, January 2015. [Online]. Available: https://www.mitre.org/publications/technical-papers/cyber-exercise-playbook. [Accessed 11 February 2020].

[57] Exabeam, "The Modern Security Operations Center," Exabeam, [Online]. Available: https://www.exabeam.com/siem-guide/the-soc-secops-and-siem/. [Accessed 11 February 2020].

[58] NCSC, "Penetration Testing," NCSC, 8 August 2017. [Online]. Available: https://www.ncsc.gov.uk/guidance/penetration-testing. [Accessed 11 February 2020].

[59] Mitre Corporation, "Cybersecurity Situation Awareness," Mitre Corporation, [Online]. Available: https://www.mitre.org/capabilities/cybersecurity/situation-awareness. [Accessed 11 February 2020].

[60] Horneman, "Situational Awareness for Cybersecurity: An Introduction," Carnegie Mellon University Software Engineering Institute, 9 September 2019. [Online]. Available: https://insights.sei.cmu.edu/sei_blog/2019/09/situational-awareness-for-cybersecurity-an-introduction.html. [Accessed 11 Februarry 2020].

[61] Mitre Corporation, "Adversarial Tactics, Techniques, and Common Knowledge," Mitre, [Online]. Available: https://attack.mitre.org/. [Accessed 11 February 2020].

[62] NCSC, "How Cyber Attacks Work," NCSC, 15 October 2015. [Online]. Available: https://www.ncsc.gov.uk/information/how-cyber-attacks-work. [Accessed 11 February 2020].

[63] Husák et al, "Assessing Internet-wide Cyber Situational Awareness of Critical Sectors," in Proceedings of the 13th International Conference on Availability, Reliability and Security, 2018.

[64] OASIS OpenC2 Technical Committee, "OpenC2 - Open Command and Control," 01 July 2019. [Online]. Available: https://openc2.org/.

[65] IACD, "IACD Playbooks and Workflows," IACD, [Online]. Available: https://www.iacdautomate.org/intro-to-playbooks-and-workflows. [Accessed 11 February 2020].

[66] Dstl, "Prototype for Automated Network Defence Actions (PANDA)," Dstl, 9 October 2019. [Online]. Available: https://www.digitalmarketplace.service.gov.uk/digital-outcomes-and-specialists/opportunities/10809. [Accessed 11 February 2020].

[67] Veksler et al, "Simulations in Cyber-Security: A Review of Cognitive Modeling of Network Attackers, Defenders, and Users" Frontiers in Psychology , vol. 9, 2018.

[68] Wang et al, "A Bayesian network approach for cybersecurity risk assessment implementing and extending the FAIR model," Computers & Security, 2019.

[69] Cerotti et al, "A Bayesian Network Approach for the Interpretation of Cyber Attacks to Power Systems," in Proceedings of the Third Italian Conference on Cyber Security, Pisa, Italy, 2019.

[70] Ceren, "Optimal Decision-Making in Mixed-Agent Partially Observable Stochastic Environments via Reinforcement Learning," December 2018. [Online]. Available: https://arxiv.org/abs/1901.01325. [Accessed 11 February 2020].

[71] Källström et al, "Multi-Agent Multi-Objective Deep Reinforcement Learning for Efficient and Effective Pilot Training," in Proceedings of the 10th Aerospace Technology Congress, Stockholm, 2019.

[72] Hüttenrauch et al, "Deep Reinforcement Learning for Swarm Systems," Journal of Machine Learning Research, vol. 20, pp. 1-31, 2019.

[73] Baker et al, "Emergent Tool Use From Multi-Agent Autocurricula," OpenAI - Arxiv, 2019.

[74] Leibo et al, "Multi-agent Reinforcement Learning in Sequential Social Dilemmas," in Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems, Sao Paulo, 2017.

[75] Kuno, "Winners announced in multi-agent reinforcement learning challenge," Microsoft, 22 February 2019. [Online]. Available: https://www.microsoft.com/en-us/research/blog/winners-announced-in-multi-agent-reinforcement-learning-challenge/. [Accessed 11 February 2020].

[76] Gleave et al, "Adversarial Policies: Attacking Deep Reinforcement Learning," in International Conference on Learning Representations, 2020.

[77] Yen-Chen Lin et al, "Tactics of Adversarial Attack on Deep Reinforcement Learning Agents," in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), Melbourne, 2017.

[78] Zhao et al, "Blackbox Attacks on Reinforcement Learning Agents Using Approximated Temporal Information".

[79] Strom et al, "Finding Cyber Threats with ATT&CK™-Based Analytics," Mitre Corporation, 2017.

[80] Falco et al, "A Master Attack Methodology for an AI-Based Automated Attack Planner for Smart Cities," IEEE Access, vol. Volume 6, pp. 48360 - 48373, 2018.

[81] Schölkopf, "Causality For Machine Learning," 23 December 2019. [Online]. Available: https://arxiv.org/pdf/1911.10500.pdf. [Accessed 11 February 2020].

[82] Sprites, "Introduction to Causal Inference," Journal of Machine Learning Research, vol. 11, pp. 1643-1662, 2010.

[83] Royal Society, "Explainable AI: the basics," November 2019. [Online]. Available: https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf. [Accessed 11 February 2020].

[84] Silverman, "Explainable AI," Imperial College London, [Online]. Available: https://www.imperial.ac.uk/enterprise/long-reads/explainable-ai/. [Accessed 11 February 2020].

[85] University College London, University College London, 5 July 2019. [Online]. Available: https://blogs.ucl.ac.uk/hexai/. [Accessed 11 Febraury 2020].

[86] University of Bristol, "PhD studentship - Explainable AI for Interacting Autonomous Agents," [Online]. Available: http://www.bris.ac.uk/engineering/media/grad-school/scholarships/explainable_ai.pdf. [Accessed 11 February 2020].

[87] Kotheimer et al, "Using Honeynets and the Diamond Model for ICS Threat Analysis," May 2016. [Online]. Available: https://resources.sei.cmu.edu/asset_files/TechnicalReport/2016_005_001_454247.pdf. [Accessed 11 February 2020].

[88] Chismon, "Hunting With Honeypots," [Online]. Available: https://www.f-secure.com/en/consulting/our-thinking/hunting-with-honeypots. [Accessed 11 February 2020].

[89] Voris et al, "Fox in the Trap: Thwarting Masqueraders via Automated Decoy Document Deployment," in EuroSec '15: Proceedings of the Eighth European Workshop on System Security, 2015.

[90] Lyn Paul et al, "Opportunities and Challenges for Human-Machine Teaming in Cybersecurity Operations," in Human Factors and Ergonomics Society Annual Meeting, 2019.

[91] "Hypergame theory applied to cyber attack and defense," Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense IX, vol. 7666, 2010.

[92] Bakker et al, "Hypergames and cyber-physical security for control systems," arXiv preprint arXiv:1809.02240, 2019.

[93] Bakker et al, "Learning and Information Manipulation: Repeated Hypergames for Cyber-Physical Security," IEEE Control Systems Letters (Volume: 4 , Issue: 2 , April 2020), vol. 4, no. 2, pp. 295-300, 2019.

[94] Urias et al, "Cyber Range Infrastructure Limitations and Needs of Tomorrow: A Position Paper," 2018. [Online]. Available: https://www.osti.gov/servlets/purl/1594636. [Accessed 11 Februaary 2020].

[95] Sandia National Labs, "Emulytics," [Online]. Available: https://www.sandia.gov/emulytics/. [Accessed 15 March 2020].

[96] USCYBERCOM, "About DreamPort," [Online]. Available: https://dreamport.tech/about-us.php.

[97] NATO, "Locked Shields," 2019. [Online]. Available: https://ccdcoe.org/exercises/locked-shields/. [Accessed 11 February 2020].

[98] DARPA, "Cyber-Hunting at Scale (CHASE)," [Online]. Available: https://www.darpa.mil/program/cyber-hunting-at-scale. [Accessed 15 March 2020].

turing.ac.uk
@turinginst