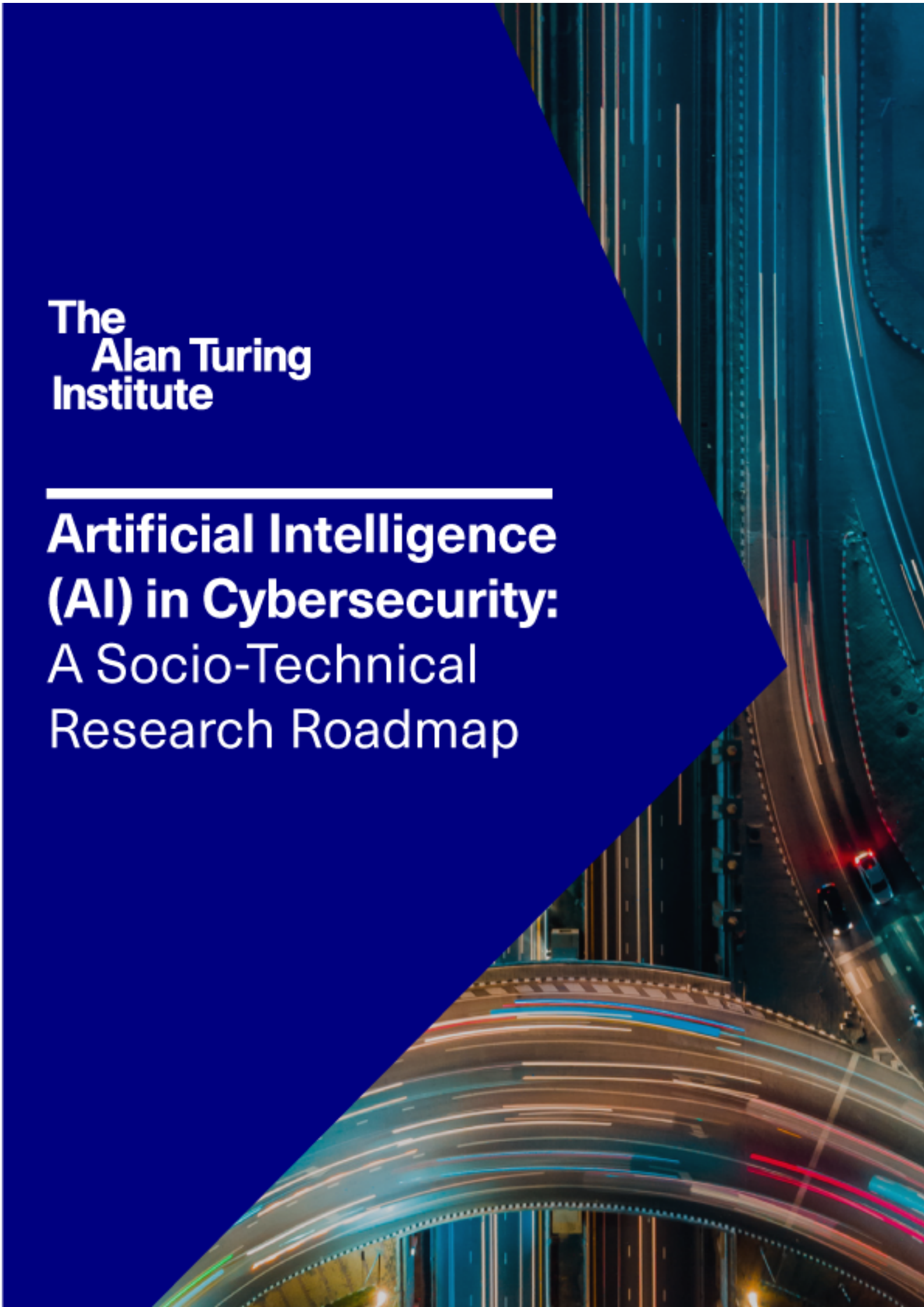
The Alan Turing Institute

Artificial Intelligence (AI) in Cybersecurity: A Socio-Technical Research Roadmap



The Alan Turing Institute

Artificial Intelligence (AI) in Cybersecurity: A Socio-Technical Research Roadmap

Authors (in alphabetical order):
Roba Abbas, Katina Michael, Jeremy Pitt, Kathleen M. Vogel, and Mariana Zafeirakopoulos

July 2023 (draft), October 2023 (final)

Perspective 1 contacts: Kathleen M. Vogel and Mariana Zafeirakopoulos
Perspective 2 contacts: Katina Michael, Roba Abbas and Jeremy Pitt

Contents

Executive Summary

Rapid progress in Artificial Intelligence (AI) is presenting both opportunities and threats that promise to be transformative and disruptive to the field of cybersecurity. The current approaches to providing security and safety to users are limited. Online attacks (e.g., identity theft) and data breaches are causing real-world harms to individuals and communities, resulting in financial instability, loss of healthcare benefits, or even access to housing, among other undesirable outcomes. The resulting challenges are expected to be amplified, given the increased capabilities of AI and its deployment in professional, public, and private spheres. As such, there is a need for a new formulation of these challenges that considers the complex social, technical, and environmental dimensions and factors that shape both the opportunities and threats for AI in cybersecurity. Through an exploration and application of the socio-technical approach, which highlights the significance and value of participatory practices, we can generate new ways of conceptualising the challenges of AI in cybersecurity contexts.

The purpose of this white paper is to explore the complex AI in cybersecurity landscape, employing a bottom-up and top-down approach that is focused on: (1) documenting and analysing the outcomes of six transdisciplinary workshop deliberations occurring between January and June 2021, with academia, industry, government, and NGO expert participation; and (2) a subsequent process of engagement with socio-technical literature to inform the reformulation and conceptualisation of the emerging AI in the cybersecurity landscape. The white paper will identify and elaborate on key issues, in the form of both gaps and opportunities, that need to be addressed by various stakeholders, while exploring substantive approaches to addressing the gaps and capitalising on the opportunities at the micro/meso/macro levels, which in turn will inform decision-making processes. The white paper offers approaches for responding to public interest security, safety, and privacy challenges arising from complex AI in cybersecurity issues in open socio-technical systems. The white paper begins with a qualitative thematic analysis of the six workshops, focussing on social, technical, organisational, environmental, and methodological issues. This is followed by an analysis of seminal socio-technical literature with a view to extract key themes that are then compared with the insights derived from the workshop analysis. The purpose of this exercise is to present a reformulation of AI in cybersecurity from a socio-technical perspective, while proposing a research roadmap that is expected to have material implications for stakeholders from a research, policy, and practice perspective.

Introduction

To introduce the context of Artificial Intelligence (AI) in cybersecurity, this section will firstly explore the cyberthreat environment as a means of providing the necessary backdrop, prior to addressing AI in cybersecurity in terms of what it is and relevant applications to date. The section concludes with emerging findings from the literature with respect to the socio-technical framing, identifying the audience and stakeholders that may benefit from this white paper.

The Cyberthreat Environment

The increasing prevalence of cybersecurity attacks on organisations focused on the provision of large-scale technical systems and the related critical infrastructure of nation states has been observed over time (Cyber Management Alliance, 2002). Traditionally, the impacts of cyber attacks have ranged from direct or indirect strikes on a) commercial organisations and their customers and their customer's customer, often having a financial impact in the form of ransom, brand damage, and organisational cybersecurity budgets; b) governments and their corresponding agencies whose citizen records have been compromised or whose website portals have been defaced; and even c) large-scale charities who maintain financial and verified address data on donors. Data breaches against some of the world's largest online platforms and service providers continue to increase in severity, scale, and frequency. We now have data breaches of customer records that are bigger than total populations of countries (Tyas Tunggal, 2022).

The stakes have continued to grow with more recent incidents demonstrating that sensitive personal identifiable information (PII), such as health records, have been stolen (Seh et al., 2020). Identity theft among other cybercrimes continues to proliferate as financial, insurance, and health institutions are targeted. In addition, there are other forms of cybercrime with wide-ranging motivations, from credential theft, to hacktivism, to insider threats, to industrial and political espionage, and even terrorism through breaches in cybersecurity defences. Moreover, cybersecurity issues can involve accidental publication of data to the web (i.e., through improper security settings); misconfiguration of security components or cloud computing infrastructure; zero-day vulnerabilities linked to service provider software (e.g., unprotected Application Programming Interface); disgruntled employees (e.g., insider attacks); social engineering (e.g., impersonation); lost data (e.g., on physical storage devices) that has not been disposed of properly, or has been misplaced in a public place; poor physical security perimeters (e.g., stolen computers); and more generally poor organisational security blueprints and employees who lack adequate cybersecurity training (Kolevski et al., 2021, pp. 3-4).

As almost all services have been digitally transformed, human dependence on these technological systems has grown (Bonaci et al., 2022) and continues to increase. This poses challenges whereby if given systems fail, a person or local community may not be granted access to a service and/or may be required to forgo a fundamental human right (e.g., access to drinking water) or to go without, albeit temporarily, a necessity to live and work in the modern world (e.g., access to money, to handheld devices for safety and other purposes, or even to the Internet). When there is a breach in any aspect of one's personal information, there are a range of cyber harms that may present due to the disruption: physical/digital, economic, psychological, reputational, and social/societal (Agrafiotis et al., 2018). A member of society can be vulnerable to attack in their workplace, on their own home network, on the personal devices that they carry on a day-to-day basis, and even the medical devices that they may bear. New targets include devices such as unsecured Internet of Things devices, wearable personal devices, and smartphones. Individuals may also fall victim to traditional social engineering attacks, phishing attacks, malware, ransomware, SMS scams, unencrypted email communications, and much more. Moreover, individuals may encounter challenges ascertaining what is disinformation versus fact, who they can

trust in specific contexts, in addition to indirect attacks that are increasingly automated like web scraping on social media platforms. Other challenges also exist, resulting in a complex cyberthreat environment that may be difficult to define and navigate by organisations, academic institutions, and other members of society alike (Smith, 2023).

Currently formulations of the cyberthreat environment, and of cybersecurity, tend to be technical in nature and accessible / comprehensible by certain stakeholders, resulting in ambiguous conceptions of security, including human security and cybersecurity more specifically. This is problematic, as security is a fundamental human requirement. Alkire (2003, p. 2) defined "human security" as: "[t]he objective … to safeguard the vital core of all human lives from critical pervasive threats, in a way that is consistent with long-term human fulfillment". Human security is subject to the reliability of technological systems, and those who govern the systems. The control capability can be highly asymmetric between those who can provision and those who can withdraw that provisioning through deliberate action, market forces, poor maintenance of physical systems, or sheer ignorance. Stolen digital assets can evoke feelings of distrust in once trusted service providers and systems of market exchange and interaction, as well as doubt in legal protections, and general disorientation about the government's ability to act on behalf of citizens. Cyber attacks, in whatever form, can also create feelings of anxiety, fear, helplessness, and anger across society (Bada and Nurse, 2020).

Responses to cyber harm must go deeper than simply band-aid solutions, such as asking an individual to change their password or order a new passport, health insurance, or credit card. Specific attention must be granted to the one who has been harmed. Equally, collective responses to cyber harms are critical for vulnerable communities, as these harms may further exacerbate existing sentiment toward power structures and amplify existing disparities. Select government agencies have focused their nation-wide educational campaigns on cybersafety (the protection of people) as opposed to cybersecurity (the protection of data). To date, government agencies have lacked mechanisms to engage and consult directly with the public about how to best combat the problem of individual attacks, attacks that destabilise organisations that citizens subscribe to as customers and disrupt flows of communication to critical government entities like Social Security that affect almost the entire population.

AI in Cybersecurity

The distinction between cybersafety and cybersecurity is critical, as is an assessment and exploration of cyberthreats and cybersecurity in the context of current technological developments that are no longer contained within the boundaries of government and industry and have entered the public sphere and public spaces. Amid these developments, there has recently been exponential growth of emerging technologies, such as AI. There is acceleration in the way AI interfaces with cybersecurity to pose new security challenges in addition to existing challenges. Simultaneously, AI provides opportunities for enhanced security and transformative potential across many application areas. While the emphasis on cybersecurity responses has traditionally been on the development of stronger technical capabilities through advanced encryption techniques and intrusion detection systems, current and emergent AI capabilities using machine learning approaches have begun to alter the

cybersecurity landscape. It is widely anticipated that AI-based hacking data breaches will have a significant impact on business operations and government agencies, as well as result in the exposure of the personal identifiable information (PII) of citizens, giving rise to even more complex negative externalities with a variety of social implications (Michael and Abbas, 2022). As noted by cybersecurity expert Ben Buchanan, “some of the most potent cyberattacks we have seen– including Stuxnet, the 2016 blackout in Ukraine, and the 2017 attack known as NotPetya that caused at least ten billion dollars in damage– feature some forms of automated propagation and attack capability” (Buchanan, 2019). This AI-enabled cyberthreat environment, however, is constantly changing. In a February 2023 survey of 1,500 IT and cybersecurity professionals conducted by BlackBerry, 51% of respondents believed that ChatGPT, the AI-enabled chatbot, will lead to a successful cyberattack in the next 12 months, 78% believe that such an attack will happen within two years, and 71% believe nation-states may already be leveraging ChatGPT for adverse reasons (Singh, 2023). Along with these concerns, it is also important to recognise that AI-enabled cybersecurity functions could also help with anticipating and responding to these varied attacks in ways that humans cannot. We should also remember that companies have used machine learning approaches for cyberdefence over the past couple of decades, particularly for detecting spam, malware, and intrusions (Musser and Garriott, 2021). Thus, there is an ongoing evolution that advances in AI technology bring to the cybersecurity landscape.

In recent years, the continued development of AI has resulted in new applications of machine learning and AI to the cybersecurity domain. AI is envisioned to enhance and automate the cybersecurity function in both offence and defence. For example, in offence, AI could enable adversarial reconnaissance, e.g., by speeding up the discovery of software and other computer system vulnerabilities. Then, AI could also be used to speed up the “kill chain”, i.e., the sequence of steps that a hacker must follow to conduct a cyber attack. AI could also be used to better tailor and scale spear-phishing attacks to increase their success rate on targets (Buchanan et al., 2020). AI might also be able to increase the mechanisms by which malicious code is spread in a system. Attackers may additionally use a variety of adversarial AI methods to target machine learning systems, e.g., “data poisoning” (Chen et al., 2017; National Institute of Standards and Technology, 2019). Data poisoning involves changing the data that is used to train the system to intentionally introduce errors into it that can then present new vulnerabilities to exploit (Hutson, 2018). This can make AI-systems susceptible to deception and manipulation through cyber attacks. For example, an iPhone’s “FaceID” function uses AI to recognize faces, making it susceptible to data poisoning by image altering that could bypass the FaceID security (Geng and Veerapaneni, 2018; Godage et al., 2023; Cukier, 2023). Also, an attacker might simply want to undermine confidence in a cybersecurity system.

In the area of defence, AI could help in speeding up the discovery (and repair) of software and system weaknesses, malicious code, intrusion detection, and other kinds of anomalous activities and insider threats, to better anticipate, find, and address threats (Michael et al., 2023b; Lohn et al., 2023). AI could also assist in selecting a quick and effective response strategy to mitigate or prevent cyber attacks (e.g., fix malicious code, isolate machines, reconfigure networks, impose user restrictions), and potentially aid with attribution of the attack by identifying, cataloguing, and analysing the cyber fingerprints or behaviours of intruders (National Science and Technology

Council, 2020). AI-enabled cyber defence systems could furthermore be used to deploy decoy systems called “honeypots” that lure attackers, allowing defenders to gather information about them, and deflect them from attacking their intended targets (AbuOdeh et al., 2021). Humans and existing software have often struggled to keep up with and anticipate cyber attacks and respond quickly. AI-enabled systems and sensors can provide support in this regard by more efficiently sorting through data from physical assets to predict areas of attack in a pre-emptive fashion. For example, there are cybersecurity companies that process trillions of data and then feed the data to machine learning models to predict new kinds of attacks (CrowdStrike, 2022). There are also cybersecurity companies that employ multiple machine learning methods to automatically mitigate attacks (Darktrace, 2022).

A survey of 850 executives on the role of AI in cybersecurity finds a strong business case for using AI in cybersecurity: “Three out of four executives say that using AI allows their organization to respond faster to breaches… Three in five firms say that using AI improves the accuracy and efficiency of cyber analysts… Most organisations say that AI lowers the cost of detecting and responding to breaches by 12%, on average” (Capgemini Research Institute, 2019). Experts have also noted that AI could change the strategy and speed of cyber operations among states, creating strategic stability (Reinhold et al., 2023). It is evident that countries are now investing in developing AI in cybersecurity capabilities (Hoffman, 2021). The UK has already signalled a prioritisation of AI in cybersecurity in its 2022 National Cyber Strategy (UK Government, 2022). Interestingly, some European countries have shown reticence in trusting AI in cybersecurity for cyber operations (Vercellone, 2020). The diverse geopolitical environment will shape how different actors address the growing role of AI in cybersecurity. All of this will demand public, government, academic and industry attention to mitigate harms, particularly with respect to the public interest, as well as create sustainable systems in the years to come. These examples illustrate how AI-enabled cybersecurity would create more obstacles and complexity for attackers and defenders alike, suggesting that AI in cybersecurity will remain a dual-use technology—creating both benefits and risks (Michael et al., 2023b). Thus, future AI-enabled systems (and their human operators) will have to learn and evolve to keep up with a constantly changing cyberthreat landscape.

A Socio-Technical Framing

A preliminary step in the process of learning and evolving in the context of the cyberthreat environment is to understand the system in question; the AI in cybersecurity system as an intricate and interconnected socio-technical ecosystem (IEEE TTS, 2023). This system is complex, is attempting to satisfy multiple criteria and objectives, and contains a multitude of components, subsystems and dimensions that interact and are linked together in many ways. Significantly, the nature of this socio-technical system is not entirely known. As such, this white paper seeks to reformulate AI in cybersecurity from a socio-technical perspective, with the intention of employing a socio-technical framing allowing for enhanced understanding of the opportunities and challenges of AI in cybersecurity. Broadly, in the context of security and cybersecurity, “socio-technical” refers to the interplay between users, technology, and processes (Stevens, 2020). This perspective draws largely from the foundational principles of socio-technical theory, a detailed review of which can be accessed in Abbas and Michael (2022). As has been noted by several authors, including

Appelbaum (1997) and Carayon et al. (2015), an organisation should be seen as a complex socio-technical system (Davis et al., 2014) that can be impacted by outside forces stemming from diverse players, and as such, it can be susceptible to the influence of the external environment within which these forces exist. Davis et al. (2014) point to more than just customers in that environment and give the example of a regulatory framework that has been enacted by the government and may affect how an organisation attains its goals. The socio-technical framing also applies beyond the organisational (meso) context, to the individual (micro) and societal (macro) levels and application. This multi-tiered, socio-technical perspective offers a rich understanding of socio-technical dimensions at the respective levels and allows for a more accurate depiction of the opportunities and challenges of AI in Cybersecurity.

The Way Forward

To date there have been a variety of technical papers (Sewak et al, 2022; Nguyen and Reddi, 2021; Kinyua and Awuah, 2021; Prebot et al., 2022; Myles et al., 2022; Piplai et al., 2022; Jin et al., 2022; Jiang et al., 2009; Silva et al., 2022; Applebaum et al., 2022; Wolk et al., 2022; Khan Adawadkar and Kulkarni, 2022; Veksler et al., 2020; Meier et al., 2021; Standen et al., 2021; Collyer et al., 2022) and policy-oriented papers (Buchanan, 2020; Hoffman, 2021; Ryan, 2020; Lohn, 2022; Congressional Research Service, 2020; Burke, 2020) that have looked at the growing developments and implications of AI in cybersecurity; however, there has been little focus on AI's socio-technical considerations from an integrated multi-layered and multi-stakeholder perspective. This has resulted in a lack of attention to human factors or larger socio-technical ecosystem concerns that shape whether or to what extent AI in cybersecurity yields risks or benefits in different contexts and for different members of society within communities of interest. It is critical to recognise that cyberharms persist at three levels: individual (micro), organisational (meso), and national / international / societal (macro) causing direct and immediate harms, indirect harms, and short- and long-term harms to people and property (Kowalski and Mwakalinga, 2011; Bauer and Dutton, 2016; Michael et al., 2023a). These have varying physical, economic, reputational, cultural, psychological, political, and other effects. Traditionally, the "human" has been identified as "the problem" and the "weakest link," but it is clear from data breaches over the past years that responsibilities for AI in cybersecurity will need to incorporate human-centred solutions within companies and government agencies (Schoenherr et al., 2023, pp. 12-13). Thus, applying a socio-technical lens may provide a better approach to both understanding and addressing AI in cybersecurity issues.

In response, this white paper employs a socio-technical approach to AI in cybersecurity, acknowledging that cybersecurity requires more than just a technical or policy dimension. It also requires incorporating with equal emphasis social and environmental considerations and their corresponding interrelationships, as well as patterns and trends in the interactions between micro, meso and macro level considerations. The white paper will provide an alternative multidisciplinary vision for understanding and anticipating the nexus of AI and cybersecurity and its effects on society at large and will provide guidance in the form of a research roadmap. This white paper emerged from a series of discussions and deliberative workshops in 2020-2021, organized by / contributed to by the authors, and staff from the Defence and Security Programme at The Alan Turing Institute and the UK's National Cyber Security Centre.

Who is the Primary Audience?

Stakeholders that belong to the cybersecurity socio-technical ecosystem can be considered at three levels: macro - society as it pertains to a local/ national/ international context inclusive of the governance structures and mechanisms; meso - organisations at any level involved in the provisioning of AI in cybersecurity infrastructure and services, or organisations seeking to update, introduce and or integrate cybersecurity infrastructure and services within their existing operations; and micro - individual human beings with commensurate rights who are personally identifiable in the context of AI in cybersecurity and may interact with/ be affected by AI-based socio-technical systems.

The primary audiences of this white paper are national and intergovernmental agencies and local and transnational business entities directly responsible for societal securitisation and fundamental human needs (e.g., safety), especially for the care of and provisioning for vulnerable members of the community. A secondary audience includes academic and research funders and councils, and other indirect stakeholders such as non-governmental organisations, who attempt to observe, study, warn, and respond to security-related incidents, share information at a national or international level, build technical standards and industry solutions, develop, and provide enforcement of laws and legislation, and protect consumers and their data. The tertiary audience of this white paper is individual members of society who have dual or triple roles within civil society; for instance, they may simultaneously assume multiple roles such as being members of the public, working with funding bodies and be employed in an organisation as a demonstrative example. These individuals are vital in the transmission and dissemination of information pertaining to cybersecurity both within their households and extended social (such as family and friendship) relations.

Thus, the research gaps are pertinent to academia, the roadmap is relevant to all stakeholders, and the recommendations section can be adopted and implemented, particularly by government and industry stakeholders. Additionally, the gaps, roadmap and recommendations are accessible to all stakeholders for information dissemination and other purposes.

Methodology

Perspective 1: Workshops (Bottom-Up Approach)

A socio-technical systems approach was used to explore the topic of AI in cybersecurity from a bottom-up perspective through a series of workshops (refer to Acknowledgements for further information regarding workshop presenters). The goals of this research included exploring systemic connections and relationships, the broader human implications of technical issues, and an openness to identifying unexpected or emergent themes from multidisciplinary dialogue. To achieve these goals, a Participatory Action Research (PAR) methodology was selected. The PAR methodology is known for its engagement of individuals in research that may either impact them or be of concern to them, within the context of a complex and emergent system, bringing together different disciplinary perspectives, practices and lived

experience through a process of iterative engagement, probing, reflection, and action, co-creating research questions and evolving the research through practice (Bergold and Thomas, 2012). Over six workshops from January to June 2021, participants were invited to explore the socio-technical-environmental dimensions of AI in cybersecurity settings. These participants were predominantly from the United Kingdom, United States of America, and Australia and worked in cybersecurity-related roles in government or academia. The six workshops followed a divergent-convergent design thinking process (Guilford, 1967). See Table 1 which presents each workshop and their phases, stages, and corresponding activities.

Table 1. Evolution of data collection and data analysis (Workshops 1-6)

Phase	**Diverge**				**Converge**	
Stage	***Intent***	***Discover***			***Define***	***Develop***
Workshop	1: Purpose: North Star	2: Socio: Stakeholder Analysis	3: Techno: Purposeful Technology that Meets Social Needs	4: Environmental: Legal and Policy Enablers	5: Reflexivity: Synthesis and Sensemaking	6: Next Steps: A Research Roadmap
Date	January 2021	February 2021	March 2021	April 2021	May 2021	June 2021
Description	Focus on 'why': Why is this work worth doing? What is the future worth wanting?	Focus on 'who': Who is involved/ impacted by/ influences this problem?	Focus on 'how': How do we produce purposeful technology that meets the needs of humans and does not do harm?	Focus on 'what': What legal, policy, regulatory or compliance mechanisms do we need to create robust protections in the face of new technology?	Focus on 'so what': What sense can we make from our collective explorations? What are our key questions and opportunities moving forward?	Focus on 'now what': What are our next steps? How might we move forward ethically and viably?
Presenters	Ian Levy (UK National Cyber Security Centre) Neil Zuring (National Security Agency, USA)	Roba Abbas (University of Wollongong) Genevieve Lively (University of Bristol)	Mariarosaria Taddeo (Oxford University) Jeremy Pitt (Imperial College London)	Gary Marchant (Arizona State University) Lyria Bennett Moses (University of New South Wales)	Ant Burke (The Alan Turing Institute)	Bruce Schneier (Harvard Kennedy School)
Activity	Bracketology Our North Star	Stakeholder Mapping Exploring Relationships, Power, Influence, and Impact	Empathy Mapping Life-life Persona "Grace" Life-like Scenarios/ Vignettes	Mapping Themes Contradictions Opportunities Magic Wand for Ideal Change	Evolving the Desired Future Thematic Areas of Change	Domains of Change Possible Actions to Implement Change Commitments

The systematic approach to data analysis involved two parallel activities: coding and thematic analysis of workshop data, which included the detailed notes from scribes, Zoom chat transcripts, and visual artefacts (post-it note contributions on Miro boards). Workshop data was coded based on patterns of agreement, disagreement, with repetition suggesting importance. These coded themes were mapped into a conceptual framework showing the relationships between these themes.

Perspective 2: Literature Review (Top-Down Approach)

A literature review was conducted to provide an additional perspective to supplement Perspective 1 from a top-down approach, with the purpose of determining the current landscape with respect to the study of AI in cybersecurity. The intention was to review existing literature to formulate a unique, socio-technical interpretation and framing of the prominent themes relevant to AI in cybersecurity, from a multi-stakeholder perspective. The literature review widened the scope of current understanding of AI in cybersecurity by moving beyond the organisational setting toward a societal and values-based view. This was achieved through the collection of select seminal peer reviewed literature across the organisational and information security corpuses inclusive of IEEEXplore, ACM, and ScienceDirect. In addition, recent studies were also gathered that focussed on the intersection of AI, cybersecurity and socio-technical scholarship more specifically, following a descriptive meta synthesis approach (Hoon, 2013) with a view to contribute to existing scholarship a socio-technical perspective or formulation of AI in cybersecurity.

The collected literature was thematically coded to uncover dominant topics that were subsequently distilled into five broad themes or areas of emphasis, covering both historical and contemporary accounts relevant to the study of AI in cybersecurity. Importantly, this reformulation of existing scholarship regarding cybersecurity, artificial intelligence, and societal implications is valuable in that it can be used to identify socio-technical considerations, implications, and gaps in existing scholarship, highlighting areas for future research. It can further serve to provide evidence of support or critique around the key insights derived from Perspective 1.

Comparative Analysis of Perspectives: Reformulation

The two perspectives brought distinct contributions. First the participation of representative stakeholders in dialogue regarding the emergent areas of AI in cybersecurity, and second the predominantly scholarly academic presentation of literature that has accumulated over time. The purpose of the comparative analysis was to consolidate the bottom-up (Perspective 1) and top-down (Perspective 2) approaches, after Islam et al. (2019), to enable a reformulation of AI in cybersecurity from a socio-technical perspective. The outcome is the identification of gaps, the delivery of policy and other recommendations, and a roadmap for future research. But first, we present the findings of Perspectives 1 and 2.

Findings

Perspective 1 Findings: Key Themes from Workshops

Analysis of the workshop data reveals that the social experience of AI in cybersecurity as a technical problem is under-explored. Perspective 1 uses tools to map the system

of stakeholders and explore the experiences of those stakeholders that are potentially most vulnerable to AI with respect to cybersecurity. Through future-oriented scenarios and a realistic persona named Grace; five major insights were drawn. A description of the realistic persona Grace is included in Inset 1. This persona could be a useful teaching aid for exploring the potential impact of emerging technologies on under-represented communities. First, trust is highlighted as an interconnected feature of socio-technical systems, and secondly, the need for diverse participation is noted as a necessary condition to forge this trust. Another insight highlighted the emotive and sensing nature of human experiences of concepts like privacy, safety, and security, emphasising the requirement to meet these needs in ways that differ from focusing on technical solutions alone. A third insight highlighted the gap in heuristics and language around the integration of socio-technical approaches. And finally, an insight related to the role of education as it pertains to human relationships and emerging technological challenges. In this context, education includes baseline AI literacy as well as reflective practices to better grapple with the unknown and emerging nature of AI in cybersecurity contexts.

Inset 1: Persona

Grace lives in London but migrated from Cote D'Ivoire in her youth. As a consequence of the COVID-19 pandemic she is out of work, and is finding it difficult to regain employment, partly due to her age - a 60-year-old - but also her complex health needs for which she requires ongoing care from her local GP. COVID-19 has made it challenging to maintain these in-person visits, but Grace also manages her health with natural and alternative remedies. Grace was quite social before the pandemic. She finds herself more reliant on online retail and government support services. Through experiences like Grace, we explored Cybersecurity scenarios around financial care and banking, health, and government support.

Insight 1: Trust is an interconnecting feature of socio-technical systems

The theme of trust was an enduring pattern identified throughout the six workshops, highlighting its role as an interconnecting feature of socio-technical systems. In this context, trust was explored in more than technical terms with consideration given also to the human user, peer-to-peer, and organisational conditions that are needed for trust to thrive.

Trust is a human experience and emotion that is experienced differently between people. This seemingly obvious finding contrasts with how the literature often discusses trust in technical terms such as "trustworthiness" or in relation to "formal methods." Trust in a software or policy context can be quantified or measured using formal methods, but these approaches do not necessarily align with the human experience of trust. Creating bonds of trust between people in communities, governments, or industries that create and use AI technology cannot be forced. Forging trust between humans, technologies, and organisations is especially challenging when there are emerging unknowns as in the case of AI in cybersecurity. One aspect of establishing trust involves how trust is made; it acknowledges that trust is felt and is created by consistent actions of trustworthy behaviour. If one way of interpreting the establishment of trust is that it is an act of making, another way is to think about trust as an act of faking, whereby users of technologies such as encryption are being exploited into feeling that these are authentically trustworthy systems. Socio-

technical systems are bound by the making of trust but are still vulnerable to the faking of this trust.

To engender a sense of trust with emerging technology, and the unknown, more diverse community participation is required. Practices of inclusive and participatory design may yield new opportunities for intervention.

Example quotes from the workshops:

"There are issues of transparency, trust. A lack of robustness requires governance as well. Tech solutions do not suffice."

"We need to develop appropriate trust. You need to understand when to not trust - when you need human intervention. We don't know how to rebuild trust b/w human and an AI system. AI doesn't have the same ability to rebuild relationships."

Insight 2: Forging trust requires diverse participation

In connection with insight 1, the data highlights the relational aspect of trust in technical contexts. Relationality can be thought of as a process that involves human participation, engagement, and collaboration with a diversity of perspectives. These human interactions give value, relevance, and meaning to all humans across society. The relationality of the system means that without diverse participation, biases can be introduced and reinforced in technologies. The result can affect broader sociopolitical factors whereby unchecked power tends to meet the needs of those who own the technology and can subsequently exploit or harm those who are using the technology. Although there was representation from government and academia during these workshops, the diversity of representation could be improved by including stakeholders from civil society and industry in dialogue about how different human needs and concerns may be met.

Diverse representation helps to redress power imbalances and mitigate the potential for bias in these technologies. This is especially the case for the industry stakeholder. As creators, visionaries, and implementers of technology, industry stakeholders hold power not only in regard to how their technology is created but for whom it is created. Industry can help influence and shape broader communities' thinking about how concepts like ethics and productivity or profitability can work without harming, or exacerbating existing harm, for people vulnerable to exploitation. Trust in the commercial supply chain of AI technologies and their users is likely to be quite low. Underpinning this perception is the belief that there is no profit motivator for industry to target technological developments beyond a market that represents mainstream needs. As such, inclusive design principles might not always be applied when developing technologies.

Recognising the varied users of this technology, diverse stakeholders should be given the opportunity to take part in discussions to help inform a future that ultimately affects them. This does not detract from the challenges associated with this type of civic participation. There are opportunities to learn more about how we might incentivise collaboration with industry and society, through governance and soft law.

Quote from one of the workshops:

"There are inequalities of knowledge, power, wealth that are relevant. People lack technical knowledge to understand AI in detail, while many technically knowledgeable people lack knowledge about human beings."

Insight 3: Lack of intellectual heuristics to integrate socio-technical ecosystems

Observations with respect to AI in cybersecurity by participants during the workshops showed a sharing of expert knowledge in the form of thematic discussion and agreement on issues deemed important. At times differences in professional opinions were offered about how action should be undertaken. However, there appeared to be limits to how expertise could be employed when exploring possible future scenarios, particularly in relation to potentially vulnerable stakeholders.

Participants' expert knowledge at times either seemed to reinforce existing problem/solution frames (i.e., drawing from good or best practice and using pre-existing past knowledge). However, that may not be applicable for framing future problems or possibilities. Alternatively, participants drew from their lived experience, exploring possibilities about AI in cybersecurity, asking questions about potential harms, risks, or consequences. The lived experience offers other ways to think about problems, raising interesting ideas and solutions by empathising with an individual's potential experiences and challenges arising from engaging with the technology.

There are opportunities to further research how to integrate the social lived experience knowledge (representing our ideals, hopes and desires for thriving in the world) with the expert-led knowledge (evidence-based and grounded in established knowledge of prior success). Although there are frameworks that can help develop these heuristics, they are not suited for the emerging AI in cybersecurity ecosystem context. There are opportunities for further research on how to operationalise socio-technical interventions to impact decision-making and for exploring new heuristics that bridge the gap between expertise and lived experience.

Example quotes from the workshops:

"We need to figure out how to train AI for the world we want, not the world we have."

"People failed to understand what was already science fact in AI while they debated sci fi scenarios. For example, we know persuasive systems have had a big effect on social and political discourse and we know now that having a completely unregulated information/disinformation infrastructure has profound social and political effects- some predicted this and were ignored while the tech industry insisted only they could innovate without constraints."

Insight 4: Diverse participation in the design of AI is necessary

Values like privacy, safety, and security are nuanced. These values are not static principles or features that emerging technologies need to consider as part of their design and engineering. These values are deeply personal and complex.

Quote from one of the workshops:

"Trust is a felt thing. Same with security. I think sometimes we use these terms without thinking about that. Sometimes people live in the most insecure places and yet they feel secure."

Participants reflected that these concepts are traditionally viewed in technical terms. The focus of design in the technical domain is often transactional, with requirements like privacy deemed to be non-functional, or a non-integral feature of the technology.

Diverse participation provides an opportunity to explore how human perceptions, experiences, and feelings toward concepts like safety, privacy, and security might be met. Meeting these needs is not just about the role of technology and its functionality, but also about how it sits within the broader ecosystem of organisational and political-environmental support. Whilst the technology itself may not be able to fulfil diverse user needs, exploring the various socio-environmental anchors to AI and cybersecurity might yield unexplored opportunities.

Meeting the broad range of human sentiments is complex. Exploring how we might rise to the challenge of integrating the diversity of human emotions, experiences, and perceptions into how we design technologies raises questions about how the complex socio-technical-environmental ecosystem might address privacy, safety, and security.

Insight 5: Education in the context of emerging technologies and social impact

Education as an overarching theme included the need to support different stakeholder groups with current information known about AI in cybersecurity, as well as to embrace education that encourages more exploration about what is not yet known about potential impacts of AI technology. The findings demonstrate that there are opportunities to further support broad-reaching and all-encompassing stakeholder groups to provide AI literacy in known disciplinary areas, enabled by improved literacy in data analysis, statistics, critical thinking, and futures thinking. This upskilling could begin in primary schools, and continue upwards throughout the formal education system, then outwards to professional circles, such as to policymakers, and in civic groups to users who are engaged with AI technology. Implementing a program for AI literacy also provides the opportunity to explore potential education strategies for raising awareness about cybersecurity, and for educating the public without exposing individuals to harm.

Quotes from the workshops:

"We need accessible language for explaining AI risks to policy makers."

"The human isn't the weakest link - they're the best defence we have. We just fail to give them actionable data. The NLP is dumb; people are uninformed."

The second aspect of education is how we approach what learning looks like in the context of AI in cybersecurity (e.g., the adoption of ChatGPT, Metaverse applications, Apple Vision etc). AI literacy is not just about the garnered expert knowledge about what we know about the risks and harms, i.e., the known and explicit aspects of AI. But AI literacy will also incorporate unexplored AI in cybersecurity contexts that are opaque, fuzzy, or less explicit. The concept of education as a future reflective practice may involve exploring how to consider one's own humanity when reflecting on intergenerational impacts, or about an individual's ability to opt-in or out of technological developments. It may involve awareness of potential bias and discrimination and how these factors may have repercussions on an individual's access to other systems, such as housing, health, and banking. There are also opportunities to consider policymaking or governance decision-making as a form of reflective practice and learning.

Quotes from the workshops:

> ***"If an AI already makes decisions about what to show people, and people then make choices, then who actually made the decision?"***

> ***"How we think about the future depends on interpretation of the present and the past, and these interpretations constantly change."***

Figure 1 presents an overarching conceptual map of the six workshops after data analysis was conducted in Miro.

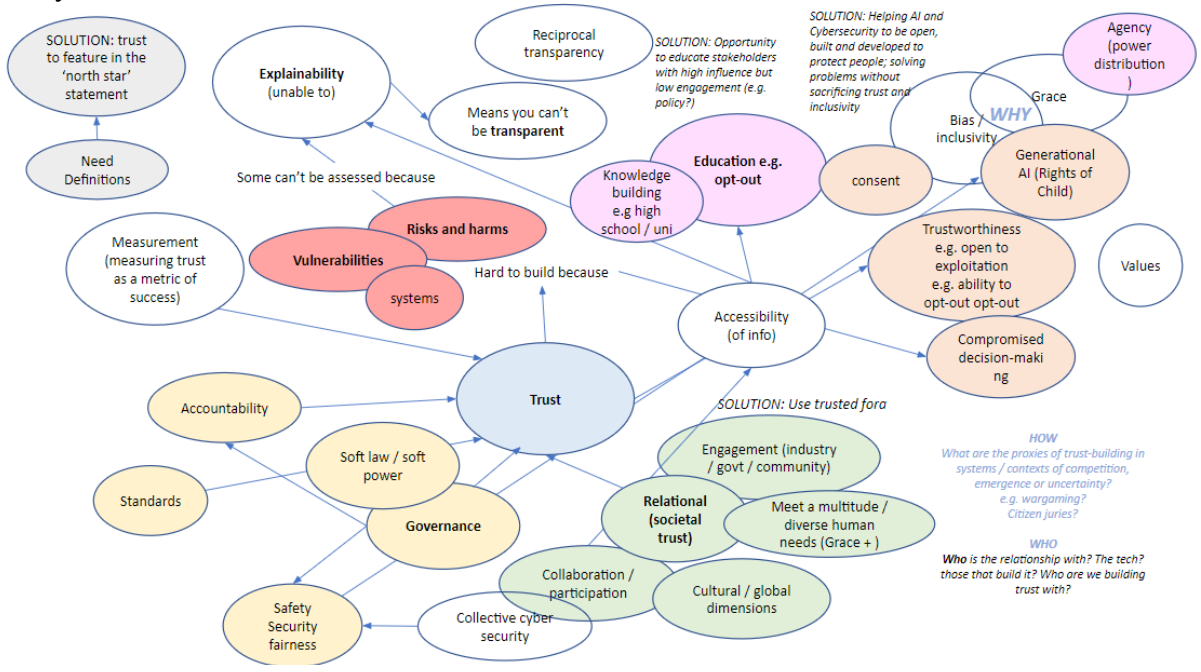


Figure 1. Key Findings of the Six Workshops- A Concept Map

Perspective 2 Findings: Key Themes from the Literature Review

The literature review takes a chronological view of cybersecurity, demonstrating how the field has changed over time, and will continue to change given the impact of artificial intelligence (AI), among other emerging technologies. Initially, the emphasis of the review is on the increasing dynamism and complexity, brought about by the Internet on organisations, supply chains, and operations. The latter part of the review is not merely how to understand the new threats that AI poses on cybersecurity, but how to problematise the intersection of the space using socio-technical theory, which seeks to jointly optimise issues pertaining to humans, technologies, and related processes. The review can thus be understood to go beyond AI, offering an underlying framing for how to address new threats that may be introduced by technologies and their emergent applications. Five themes are presented, culminating in the need for multidisciplinary and multi-paradigmatic responses to cybersecurity. These themes should be considered at the government/societal (macro), industry (meso), and individual (micro) levels.

Theme 1: Traditional cybersecurity scholarship

As we look at the changing landscape of cybersecurity, it is important to consider that before “cyber” as a concept became prevalent, “security” as a domain of study was firmly established, although the focus was typically on military tactics on the battlefield. According to Samtani et al. (2020), an organisation-centric perspective was largely adopted in computer security, and thus in relevant literature.

Organisations today, due to the growth of communications, may have a presence in more than one location, often in more than one country, requiring dedicated internal networks (intranets) to facilitate information access, exchange, and collaboration. This environment has supported the development of multinational and transnational entities that cross borders where different laws and regulations may apply. Increasingly individuals and companies are forming such “transnational networks that pay absolutely no heed to national boundaries and barriers” (Angell 1995, p. 10 quoted in Dhillon & Backhouse, 2000, p. 125).

With the rise of the public Internet, online services flourished, allowing for communications between users and providers that were distributed with a global reach. Security became focused on online applications, the storage of information in a digital format, and thus “cyber” security was born to respond to various forms of connectivity: intranets, extranets, and the Internet. Managed network services soon developed into Cloud solutions, and data demands grew exponentially through the increased use of personal devices, self-service business models, and government digital transformation initiatives. The idea of “vectors of attack” was born as the number of unsecured devices commensurately rose, as did the methods of attack with the introduction of wireless fidelity (wi-fi), smartphones, and the Internet of Things (IOT) (Dhanjani et al., 2012).

The CIA (confidentiality-integrity-availability) triad model, despite its many limitations, was used for decades to ensure organisational-centric security (Dhillon and

Backhouse, 2001). Confidentiality was required to ensure data remained private through the concealment of resources; integrity was required in order for data and software to only be changed in an authorised manner, ensuring trustworthiness; and availability was required for the proper functioning of a system by authorised users, free of attacks, ensuring reliability and robust systems design (Dhillon, 2007, p. 19; Bishop, 2005, pp. 2-4). But these "technical controls" were developed and intended for a very different setting when contrasted with the modern organisation (Samonas and Cross, 2014). Whereas once the emphasis was purely on the machine and the place in which the machine resided to guarantee security, there has been a departure from these lines of inquiry in scholarship "towards a wider socio-technical reconsideration of its core concepts" (Samonas and Cross, 2014, p. 23).

According to Samtani et al. (2020, p. 4) there are two types of cybersecurity data. These can be defined as internal cybersecurity data that pertain mainly to assets to the organisation (such as data storage, network-based fingerprint data, biometric data) and external sources of data that are available in the public domain (such as malware repositories, news media sources, carding shops). Knowing where data is stored or passed through is just as vital as knowing how to guard against data loss and data leakage. The first sign that an asset (a network, machine, device, or data), is under duress and may be compromised comes through the detection of anomalous traffic behaviour (e.g., too many login attempts, excessive upload and downloads based on historical patterns for benchmarking, and abnormal signal strength, among other signs). By bringing together internal and external data sources for the protection of an organisation, the organisation remains informed internally about the health of an asset, and externally about other examples that may forewarn about common attacks and changes to environmental settings.

The lessons are clear; we can no longer rely on just technical responses. The defences have proven too easy to overcome. According to Islam et al. (2019), it is evident that organisations still emphasise the "technical." Specifically, they emphasize the technological responses to cyber attacks and cybersecurity challenges at the expense of the social. Yet, the vectors of attack have grown so much that fool proof security blueprints with layers of security still suffer from what is known as "implementation gaps." Importantly, social engineering techniques can still play an important role in any hack.

Theme 2: Going beyond the organisation: supply chains and ecosystems

A supply chain is several organisations connected both logically and physically along the supply process, toward the production of goods and services for distribution to customers. Dhillon and Backhouse (2000, p. 125) emphasise that the structures of supply chains facilitate intense sharing of data and information and are characterised by "a high level of interpersonal and inter-organisational connectivity." This means a breach in defences in one organisation will be transferred across the supply chain and perpetuate the problem. A vulnerability in one layer of a single organisation is a vulnerability across the supply chain from producer to distributor to retailer and ultimately to the customer.

Cybersecurity has gone from an organisational-centric concept to a national and global affair and is increasingly about critical infrastructure that support individuals in society. The greater the number of cybersecurity attacks, the more the local and national contexts are undermined. Security is a communal good that requires participation from all sectors, systems, and structures that needs to be sustained. A socio-technical strategic focus to information security attempts to achieve effective security, holistically, "through the application of multiple organizational and social alignment mechanisms combined with competence in technology" (Kayworth and Whitten, 2010).

The need to respond to increasingly global security requirements that have local impacts, cuts across the three layers top-down or bottom-up. Upper-layer security analysis undoubtedly has an impact at the lower layers, and lower-layer security analysis undoubtedly has an impact on the upper layers (Li et al., 2018). Security vulnerabilities are not isolated incidents that can be "plugged"; the exposure in one layer carries across to other layers up and down the stack. The authors call this approach "multifaceted" where cybersecurity extends beyond being simply a "technical issue," towards being understood as a "business issue" that executives and senior management can no longer ignore because of fiscal and reputational brand repercussions of data breaches, not to mention the implications on people's privacy.

The ecosystems perspective relies on top-down and bottom-up approaches, which is generally referred to in the literature as a "hybrid" approach, that is in alignment with the methodology employed in this white paper (Islam et al. 2019, p. 6). Bauer and Dutton (2016) describe a range of actors in the "cybersecurity" ecosystem. Stevens (2020) describes these actor relationships as "complex assemblages" and names "players" such as military/ intelligence personnel, users/ citizens, hackers, organisations, and others.

Using an ecosystem view, stakeholders can come together to share their perspectives, and to voice the issues that are important to them and their constituents. Hodson and Marvin (2010) describe this very practically when they write: "'[e]ffective' responses to these pressures are thus predicated on multiple challenges, multiple actors and multiple levels that require effective coordination to inform control of infrastructure systems." Framing cybersecurity as a dynamic process within an ecosystems-based framework allows for those human-related risks to be better understood, exposing more complicated interactions at multiple views (cyber/physical/social) at the micro-meso-macro levels within an environmental context where events and actions have consequences. Understanding the complex processes taking place requires the adoption of theories from diverse fields including biological sciences, sociology, cultural studies, and computing/socio-technical systems (Islam et al. 2019, p. 5).

Theme 3: Humans, risk, uncertainty, complex and dynamic systems

As has been noted above, systems today may be described as dynamic. Farber and Pietrucha (2014) describe not just interconnectedness between organisations but interdependencies of large-scale, complex socio-technical infrastructural issues. To complete this picture, a security breach in a single socio-technical infrastructural system will have a ripple effect throughout the entire end-to-end system, albeit for a short time until the system returns to a steady state.

Wu et al. (2015) attempt to articulate where the complexity being experienced stems from and deduce that it comes about because of “interactions and interdependencies between a diverse range of social, technical and contextual elements in and around the system.” Modelling modern socio-technical systems in critical infrastructure and services such as transportation, organisational systems and energy infrastructure is a very challenging task. However, the ability to model is essential to the design, development, and delivery of modern systems, particularly in socio-technical systems engineering and decision support. Zimmermann and Renaud (2019), drill down further in describing the social, technical, and environmental elements in socio-technical systems. Pertaining to the issue of cybersecurity, they describe such elements as computers and networks (technical subsystem); human actors in different roles and with different levels of security expertise (social subsystem); and governance structures, operating systems, and the influences of the wider environment (environmental context).

Short of saying we cannot simply focus on the “technical” in cybersecurity, Zimmermann and Renaud (2019), pronounce that we have to do “Cybersecurity, Differently”. Zimmermann and Renaud (2019) emphasise that you cannot simply “home in” on a single component, hope to “fix it” and move on. That is not how socio-technical systems work given their complexity. Such an approach would be unrealistic because of “the emergent nature of the underlying system’s outcomes.” We would add that it would be unrealistic because the system works through interactions of components between subsystems and not on singular “anything.”

Samtani et al. (2020) describe the importance of cyber threat intelligence, inclusive of threat and actor identification in the interest of informed decision making. To become more resilient there must be a greater level of intelligence; this intelligence seeks to detect patterns and trends that might well serve to be effective in scenario planning. This is where AI can be incredibly useful in detecting anomalies in incoming and outgoing network traffic; patterns in Wi-Fi signalling; login attempts; pattern recognition (biometrics); etc. While these pattern detection techniques seek out exceptions, they are for the greater part emergent, but grant some mechanism with which to combat threats.

Stevens (2020) suggests that users can often be perceived as a threat vector. Insider attacks, referring here to members of an organisation, and users in general, have traditionally been called the “weakest link” in cybersecurity. Islam et al. (2019) concur that human behaviour and human error can be considered as threat vectors or sources. Yet, as noted by Zimmermann and Renaud (2019), merely “[l]abelling human actors as “the problem” does not acknowledge their ability to detect anomalies and halt attacks.” So, as much as humans are responsible for attacks on global networks, even through insider attacks (Stevens, 2020), humans are also responsible for devising responses to known attacks, or working in security teams to address unfolding attacks as they happen unexpectedly in an organisation.

Cybersecurity is not only a computer science or technical challenge, but increasingly (and in no small part driven by emerging AI technologies) it is a sociological, economic, and behavioural challenge. The act of securing our cyber existence is not yet a universal mindset. And the question is, how to make it so? How might we be able to utilise socio-technical theory to encourage the application of cybersecurity in every

facet of our digital and off-line realms? In effect, the hope is to change the mental models of users. It is proposed that one way to shift these mental models is through educational campaigns, although measuring what effective might mean is complex in its own right. Dupont (2013) grants a security mindset definition specifically for Internet users, defined as "a set of attitudes, beliefs and values that motivate individuals to continually act in ways to secure themselves and their network of users, such as by acquiring technical skills, new practices or changing their behaviour online." Yet, as Farber and Pietrucha (2014) point out, we must study closely why stakeholders may have different "mental models" of how infrastructural "sociotechnical systems function, even for supposedly the same systems, which is valuable knowledge for understanding "whole" systems of systems functioning." We need to develop information security capabilities at the management, operational and tactical levels as well as to continue to train competent security-centric personnel.

Theme 4: Socio-technical framing of the information security paradigm

Clearly there is a need to consider how we may be able to address the issues, concerns and dilemmas raised in the previous sections. One suggestion prevalent in the literature is to understand the information security paradigm through socio-technical framing. Paja et al., (2013) make the claim that "today's systems are Socio-Technical Systems (STSs)." The authors note that these STSs consist of participants—inclusive of humans, organisations, and software—that are autonomous and can interact with one another to achieve tasks. Security within socio-technical systems must not be seen merely as a technical challenge, but social components also need to be considered: "Today's systems are socio-technical, for they are an interplay of social actors (humans and organisations) and technical components (software and hardware) that interact with one another for reaching their objectives and requirements" (Paja et al., 2013). Following this research, Paja et al. (2015) and Mujinga et al. (2017) call for information systems design (ISD) strategies that can address both the social aspects and technical aspects, utilising the socio-technical systems (STS) approach.

Griffith and Dougherty (2001) further elaborate citing Rogers (1995) that the socio-technical perspective breaks down an organisation into a social system that is made up of people that utilise tools, techniques, and knowledge (technical system), to make something tangible or offer a service to a customer base. Customers/subscribers are defined as members of an organisation's external environment, as they sit outside the physical and logical boundary of an organisation. What is important is not that there are two individual systems, a social subsystem, and a technical subsystem, but how well these two systems are designed to interact with one another with respect to the demands of the external environment. The better the interaction between an organisation's products and services and the external environment (e.g., customers and other stakeholders), the more effective the organisation. However, turbulence in the external environment can impact an organisation as it keeps adding to the complexity already being experienced (Chen and Redar, 2014 cited in Malatji et al., 2019).

Without overemphasising the importance of the "social" over the "technical" or the "technical" over the "social," better understanding of human factors is vital for the

success of security management in the modern organisation. Worm et al. (2015) cite Johnsen and Veen (2013) who refer to the "human factor" as an interdependent network in "recognition of the importance of modelling the socio-technical system as a whole." Dupont (2013) similarly agreed that technologists and sociologists alike had to adopt a cybersecurity mindset. A mindset was not just about thinking and theorising but about actions. The social and cultural dimensions of cybersecurity were critical, Dupont (2013) argued, and needed to be "addressed alongside allied efforts to enhance educational, technical, organisational, business, policy, and regulatory approaches to cybersecurity".

Theme 5: Balanced multidisciplinary and multi-stakeholder approaches

We return to the fundamental premise that we need more than one discipline to respond to cybersecurity issues. As described by Beekun (1989) in Malatji et al. (2019), STS "seeks to optimise the alignment and correlation between the social and technical dimensions of a system, while considering the system's environment." We deduce a holistic approach is required. Cited in Samonas and Cross (2014), Dhillon and Backhouse (1996) draw on two empirical studies and warn that the result of an imbalance in the three subsystems of any socio-technical system will lead to uncertainty. This has the effect of creating complexity, which ultimately introduces inherent risk to an STS. Dhillon and Backhouse (1996) elaborate that this is "due to the continuous and out-of-control interactions of the technical, formal and informal sub-systems".

While traditionally the "human" was situated as the "problem" in security, Zimmermann and Renaud (2019) have highlighted in their seminal paper a movement toward viewing the "human" as the "solution." However, this perspective too can be seen as unbalanced as it pays more attention to the significance of the social subsystem rather than acknowledging that the social subsystem is just as important as the technical and environmental subsystems. The literature points to the short-sightedness of making a trade-off between social and technical issues. Enhanced psycho-social awareness of causes of cybersecurity breaches will not prevent an attack if the artefacts required to protect an organisation's data and network are so poor that they can be easily compromised. Here we return to the ideas already presented above of holism, balance, and interconnectivity, and stress the need for the incorporation of positivist, interpretivist, and critical methods to provide a clearer picture of how artificial intelligence may well impact the field of cybersecurity.

In Figure 1 depicted in Samtani et al. (2020, p. 9) a multidisciplinary perspective is presented, incorporating socio-technical, organisational, regulatory, cultural, cognitive, and psychological factors. Interrogated from a diverse array of perspectives it becomes possible to better understand how AI can be used to assist in decision-making of cybersecurity risks and responses that may need to be executed in near real-time. A multidisciplinary AI for cybersecurity roadmap includes a three-pronged approach incorporating (a) cybersecurity applications and data, (b) advanced AI methods, and (c) AI-enabled decision making. The process broadly considers (1) emerging application areas that have data source demands and whose data can be pre-processed for representation and analysis in a refined manner; (2) the gathered data then undergoes a multi-view and multi-modal analysis using explainable and

interpretable AI approaches and human-machine interfaces that are augmented to enable; and (3) AI-based cyber-defence and resilience toward automated cybersecurity predictions and dashboards that allow for the visualisation of events in real-time (Samtani et al. 2020, p. 9). In many ways this process as presented by Samtani et. al. (2020) is reminiscent of the Observe, Orient, Decide, Act (OODA) loop (Osinga, 2007), incorporating the power of artificial intelligence in defence. But the same principles may apply in offence.

Socio-technical systems by their very nature require multi-dimensionality because they are composed of multi-stakeholder relationships, and are based on knowledge stemming from multidisciplinarity, and are in fact a multi-paradigmatic approach. Kianpour et al. (2021) candidly consider and see the usefulness of adopting the multi-paradigmatic approach, which is to a degree pluralistic, in supporting boundaries and limits to analysing cybersecurity as a socio-technical phenomenon. Multi-paradigmatic approaches require the incorporation of multiple viewpoints from different disciplines inclusive of sociology, psychology, behavioural science, and social psychology. The real challenge may well not be the hackers, or the technical actors in the cybersecurity ecosystem, but the transforming of people's "consciousness to higher levels of awareness and understanding of oneself, others, and the complex interconnectedness of all things" (Kianpour et al., 2021).

Gaps and Opportunities for Future Research

The more complex our systems become, the greater the attack plane that can be targeted. A tit-for-tat, 'catch me if you can' attitude, will only lead to greater exposures, and misdirected cybersecurity attacks with mass-scale, even global implications. We need to discover and address the root causes of cybersecurity issues, which can only be achieved by exploring and analysing the complex socio-technical system within which AI in cybersecurity, and the related challenges, exist (Michael et al., 2023b). This does not require merely taking into consideration national and organisational-level risk assessment, but rather considering risk at the individual and or household level. Geopolitical pressures at the national level will have flow on effects, and governments must remain cognizant that interferences by state and non-state actors on critical infrastructure providers and major organisations, will have a direct impact on individual citizens and their respective households and communities at large (Michael et al., 2023b). This fragmentation will require new architectures for international AI governance (Cihon et al., 2020; Minkkinen and Mäntymäki, 2023).

Within an environment open to destabilisation, and factoring in the multiplicity of scenarios, it is easy to assume that the future of AI in cybersecurity is one void of human intervention: an entirely autonomous vision (Michael et al., 2023b). This is a consequential misconception when discussing the potential of AI in cybersecurity. That is, a human can, and in most instances should, be kept in the loop. Specialists must work with AI and keep striving for its appropriate and optimal use, and not become complacent or over-reliant on third party ML-based solutions (Michael et al., 2023b). External data sources can provide new sources of intelligence with respect to the latest cybersecurity attacks, the development of new information on the latest forms of

attack, and the construction of a customized cybersecurity knowledge repository that can act as an aid to decision-making for risk managers and security specialists (Zeadally et al., 2020).

Gap 1: Human factors are under-represented in cybersecurity research. We are advocating for the integration of human factors (i.e., affordances, cognition, visualisation, and perceptions) in socio-technical systems design, requiring a reframing from "humans as the problem" to "humans as the solution" and avoiding the scenario of the "human as exploitable." Importantly, human factors alone will not address cybersecurity concerns. Those concerns will be addressed by responding using human-machine teaming approaches, that is the human actor working alongside the technology.

Gap 2: Lack of emphasis on human values. These include things personal to us– trust/control, privacy/security, attention/safety, individual vs congruent shared values. Socio-technical systems design requires knowledge of the values of users of cyberspace to ensure cybersecurity and cybersafety are shared values. Trust is emphasised within an entity, between entities, and in the ecosystem at large. Trust in entities in the physical space cannot be auto-replicated or assumed in cyberspace, despite that trust acts as a binding agent in connectedness.

Gap 3: Single focus perspective of cybersecurity is limiting. The cybersecurity "problem" is seen through the eyes of a consumer, an organisation/ business, government agency, or national security entity. It may also be seen from the perspective of an individual member of a supply chain (end-user, retailer, wholesaler, etc.), value chain or care chain. We are advocating for an integrated view where everyone is responsible for cybersecurity. Responsibilisation does not mean that a consumer is used as a scapegoat, or an organisation is blamed for a major data breach. Accountability is paramount, especially in government.

Gap 4: Stakeholder mapping of the complex cybersecurity ecosystem is required. Stakeholders in the ecosystem are identified, as are the relationships and interdependencies between each entity. For each stakeholder, the key issues are articulated, as are the reasons for those issues, and how they might be overcome. An integrated view is needed with all stakeholders represented through not only engagement but consultation and participation. The complexity of the system map should show the external environment; the meshed physical and logical network, inclusive of the triple helix; the third sector and others.

Gap 5: Emphasis on educating members of society about the dynamic cybersecurity landscape. As threat vectors continue to increase, so does the nature of challenges pertaining to AI in cybersecurity as an emergent context. This gap has much to do with raising cybersecurity awareness among the populace, but also has to do with capacity building so that people instinctively know how to detect that an email or an SMS or an action request is suspicious. This gap extends to misinformation and disinformation online where members of society need to be able to conduct some basic assessments to determine validity of a piece of content.

Gap 6: Lack of attention to capabilities development and maturity models in organisations. This gap predominantly requires that businesses, governments, and

not-for-profits develop a capability maturity model that can be used as an investigative tool to support knowledge of a given process, and to support process improvement. The emphasis in this gap is in the configuration of capability maturity models in that they are made up of a designated set of elements that are structured and can be used to deliver on a security blueprint that directs organisations on how to improve their security capabilities.

Gap 7: Lack of emphasis on human-centricity, social securitisation, and security exposures. Securitisation of the person is fundamental at the macro, meso, and micro layers. For now, cybersecurity attacks aim to access personally identifiable information through unauthorised access. Attacks of the future will become increasingly sensitive (e.g., targeting implants), in addition to making use of behavioural analytics such as neurobiological processes through brain-to-computer interfaces (Tornas and Johnson, 2023). Responding to such security exposures is at the heart of social securitisation, human rights, dignity, and autonomy to counter human destabilisation.

Gap 8: Lack of regulatory and policy approaches and responses to cybersecurity issues. This gap focuses on the necessary support required for cybersecurity initiatives to govern emerging technologies such as artificial intelligence (e.g., illegal vs unlawful, legal vs unethical etc). Regulatory and policy sandboxes may be one approach to test solutions, enabling just-in-time responses to the pacing problem where advances in AI within the cybersecurity context outstrip the ability to defend against unknowns. This gap acknowledges that scenario planning can occur to consider ways forward, particularly in the context of autonomous cyber defence and AI security.

Gap 9: A process of socio-technical security design in conjunction with existing organisational cybersecurity practices. The gap promotes the need to go through a socio-technical security design process. Organisations should set security goals from the outset. After goals have been defined, an appropriate cybersecurity framework that aids in the continuous monitoring of mutual alignment between the social, technical, and environmental subsystems is required to maintain overall systems performance. The chosen framework is overlaid on top of existing cybersecurity practices in an organisation.

Gap 10: Development of cybersecurity models, simulations, and scenarios in the context of socio-technical systems from a micro, meso, and macro perspective relevant to the organisation/entity. The emphasis of this gap is on the need to conduct whole-of-systems modelling by better understanding the linkages between the micro, meso, and macro layers and the development of models that capture complexity through simulation. Approaches to the development of models must be multi-paradigmatic and multidisciplinary. This gap requires a diverse research community to work closely together to break down silos.

Table 2. Identification of Gaps in the Dynamic Cybersecurity Landscape

Gap	Description
1	Human factors are largely missing in cybersecurity research
2	Lack of emphasis on human values
3	Single focus perspective of cybersecurity is limiting
4	Stakeholder mapping of the complex cybersecurity ecosystem is required
5	Emphasis on educating members of society about the dynamic cybersecurity landscape
6	Lack of attention to capabilities development and maturity models in organisations
7	Lack of emphasis on human-centricity, social securitisation, and security exposures
8	Lack of regulatory and policy approaches and responses to cybersecurity issues
9	A process of socio-technical security design in conjunction with existing organisational cybersecurity practices
10	Development of cybersecurity models, simulations, and scenarios in the context of socio-technical systems from a micro, meso and macro perspective relevant to the organisation/entity

Reformulation: A Socio-Technical Approach to AI in Cybersecurity

To date, new ways with which to tackle the growing cybersecurity problem have been deliberated, and planned responses at the strategic level have been incorporated, e.g., at a variety of levels of government and education. The creation of the public interest technologist who is equipped with a multidisciplinary background to tackle emerging complex problems related to cybersecurity is beginning to gain some traction in the United States (Schneier, 2019). Beyond the emergence of a new transdisciplinary field of scholarship in public interest technology (PIT), and the embedding of “clinics” into core computer science university curricula, industry must create opportunities for workers to demonstrate the value of adopting diverse frameworks, approaches, techniques, and methods from a variety of disciplines. Demand must grow as should the respect for people who can assist in the fulfilment of socio-technical systems design toward better cybersecurity solutions. This requires opportunities for relevant exchange and the supply of information about critical intersecting spaces, on job boards, at conferences, meetups and more. We could say PIT has emerged because of the need to have balance within the social, technical, and environmental subsystems.

Using a multi-paradigmatic approach, there are ways to better design socio-technical security systems. We distinguish here between socio-technical systems that require cybersecurity to be embedded as a non-functional requirement, socio-technical systems built to fulfil a cybersecurity systems function, and industry-specific and

national cybersecurity strategies that securitize borders and citizens. While systems and entities should be considered within the contexts in which they have been conceived, the real scope for change in the field at large is in understanding the interrelationships and interdependencies not just horizontally across operations (e.g., supply chain, value chain, care chain), but vertically (i.e., macro, meso, micro). Malatji (2019: pp. 184-185) provides a long list of socio-technical systems (STS) security controls mapped against what they term "capability domains," that are defined as (1) organisational structure (functions); (2) actors; (3) technology (tools and resources); (4) and work activities (tasks) (reminiscent of Bostrom and Heinan's (1977) approach to analysing STS).

We have discussed throughout the white paper the importance of a balanced approach to social, technical, and environmental considerations in modern complex socio-technical systems where human actors, agents, and their corresponding relationships at the component level need to be mapped using a multi-stakeholder, multi-dimensional, multi-disciplinary, multi-paradigmatic approach toward interdisciplinarity and transdisciplinarity. No one paradigm has all the answers, but boundaries are still required as are stating underlying assumptions when aligning to a socio-technical systems cybersecurity framework. An overemphasis in one component of a singular subsystem will not result in better achievement of overall cybersecurity goals, but rather will come at the cost of another part of the sub-system that may well be prone to a socio-technical gap given the lack of attention. In fact, we make the claim it is "human and computer in the loop" that will best achieve an augmented capability (Clarke, 2023).

As Wall (2020, p.1) has stipulated, AI cannot oversee making the "hard decisions," but it can be there to aid practitioners, professionals, policy makers and politicians through informed analysis drawing out key concepts and directions and assisting in making sense of gathered intelligence. The responsibility of decision-making must always rest with the human on non-trivial matters. Stevens (2020, p. 164) elaborates that AI algorithms spur on knowledge production through new modes and locales of cybersecurity that, in turn, trigger the formation of new hybrid assemblages between humans (actors) and non-humans (artefacts). But this is not to say that the introduction of AI-driven "anything" (e.g., anomaly detection) is not without its own tensions and subjectivities. While AI can better detect network and data activity flows, it is not a substitute for human cognition and can create political problems in the workplace (Stevens 2020, p. 167). Furthermore, Stevens writes that the "core modality of offence–defence dynamics in the grey zone, remains open to contestation" (Stevens 2020, p. 168).

Awareness that the environmental subsystem cuts across the socio-technical subsystems is also important. External to the socio-technical system may well be pressures that impact the system as a whole, but many of these pressures are unpredictable. This unpredictability can be modelled using scenarios in multi-agent systems, if information can be gathered about the behaviours observed and fed back into the model (Worm et al., 2013). A cybersecurity framework must be agile enough to incorporate feedback, but also work in conjunction with existing technological processes. Though we have stated the importance of the sociological/ psychological/ cultural, we restate that this must not come at the expense of the techno-centric, nor

at the expense of the regulations, policies, rules, and guidelines that govern a socio-technical system.

Roadmap

In view of the anticipated impact that AI will have in the field of cybersecurity, and corresponding challenges understanding AI as more than just an "artefact" (but an interrelated system of artefacts in the form of hardware and software, and the human actors who are both responsible for and impacted by AI processes and outputs), a future roadmap based on a multidisciplinary perspective and relevant funding considerations is proposed (Samtani et al., 2020; Taddeo et al., 2023).

From the social dimension, it is important to restate the role of the human in decision making (Wall 2020) and ensure that the human is always over the loop in cybersecurity processes (Middleton et al., 2020). From the technical dimension, there is a growing need to remove siloes. It is also important to continually search for those internal and external sources of data stemming from machines and humans, on which to base decisions and develop proactive cybersecurity models for the prevention and detection of attacks. Additionally, there is a known lack of resources and support infrastructure with respect to cybersecurity. There is a lack of consideration for end-users, as new tools and techniques are introduced onto the market (Samtani et al., 2020). Finally, the environmental dimension cannot be ignored as it is a "grey zone" and represented by entangled assemblages (Stevens, 2020).

Bringing these dimensions together within an ecosystem, we can use the lens of the co-evolutionary perspective to identify the role of stakeholders, the nature of the risk in the ecosystem and possible ways to address this risk (Islam et al., 2019). Cybersecurity ecosystems are extended to incorporate "ubiquitous digital ecosystems" (Carillo et al. 2017) more broadly, which introduces yet another layer of complexity (National Cyber Security Centre, 2020). A way forward is to be hopeful in the benefits of AI in cybersecurity optimisation programs (after Malatji, 2019), but this in no way diminishes the responsibility of the human decision-maker. The path of "AI as part solution" together with "human at the helm" is also fraught with its own sets of risks, as either the human actor reverts to the AI to empower them, or the AI is riddled with a lack of data to power cybersecurity models and systems, or there is the phenomenon of internal bias with inconclusive results.

It is also important to attract a larger and more diverse pool of researchers into the cybersecurity field where philosophers, anthropologists, sociologists, and psychologists are engaged with how to better broach the bigger cybersecurity threats affecting our society at the macro, meso, and micro levels. Organisations must hire professionals who are able to approach the existing cybersecurity issues that have plagued us for the last decade in new ways, and who are able to deal with the emergent threats that are yet to be measured, as well as those that to an extent are still unknown, even to specialists in the field. This will not happen if we continue to engage the same scholars, with the same methodologies, and the same underlying motivations. How can we get more transdisciplinary teams working together where each member of the research team feels equally valued to contribute? The roadmap may, for example, encourage this transdisciplinarity by requiring certain types of

backgrounds to fill different parts of a research problem described in a grant application, or come together to devise an innovative whole national systems approach to cybersecurity. In addition, publishing outcomes in transdisciplinary journal outlets may be helpful, and engaging publics, government agencies, the third sector and small-to-medium industries in popular news, policy and trade publications, respectively, may be useful. Also ensuring that applicants are diverse in background and not just focusing on disciplinary types is a necessity.

Samtani et al. (2020: p. 13) provide a selection of National Science Foundation (NSF) funding opportunities to support AI for cybersecurity research and education programs. Each funding opportunity has been categorised into five funding types inclusive of: (1) early career status, (2) infrastructure-oriented, (3) core research, (4) transition to practice and (5) education-oriented. Over half the funding opportunities are listed as being “cross-cutting” with respect to the handling directorate and division, demonstrating that at least in the United States a great deal of emphasis is being placed on interdisciplinary and transdisciplinary research. To demonstrate the seriousness of the NSF funding opportunities in supporting AI for cybersecurity, we see that the funding ranges from $175K to up to $5M, with most of the funding opportunities being over $1M. Equally, governments must set aside support for industry creating innovative ways to combat cybersecurity threats through the application of AI, and support schools and colleges toward the development of economic information infrastructures surrounding cybersecurity. But most of all, citizen approaches are vital to shift the cybersecurity mindset and build capacity.

The initial phase of the roadmap in the first five years is to conduct public engagement with citizens around cybersecurity issues and the coming transformations from AI, inclusive of other sectors of society. The second phase of the roadmap is to target funders toward the generation of new multi/trans-disciplinary knowledge with respect to the changing cybersecurity paradigm, using a socio-technical framing. The third phase of the roadmap is to bring together members of the cybersecurity ecosystem and to define relationships and interdependencies with a long-term view of systems redesign and redevelopment, inclusive of the implementation of tools, techniques and methods, necessary standards and regulations, in addition to other resources. To that end, five recommendations are put forward to be satisfied over a ten-year horizon; these activities can be done in parallel approaches. These recommendations form the basis of a sociotechnical approach to AI-enabled cybersecurity.

Recommendations

1. Define and develop capacity building activities for citizens/ consumers / employees/ volunteers to institute a cybersecurity mindset that is empirically operationalised. (Dupont, 2013). Through active citizenship and guiding policy, create a set of concrete habits, values and attitudes that can be embraced by Internet users, and deal with the complexity of cyberspace.

2. Design and develop a socio-technical systems cybersecurity capability’s maturity model in the context of AI in cybersecurity that works in conjunction with existing cybersecurity frameworks (e.g., NIST 2017), and can be applied

to any workplace or context. Where there are known vulnerabilities and anticipated cybersecurity threats resulting from AI, the socio-technical gaps can be treated by using existing information and solutions (Malatji et al., 2019).

3. Identify and explain the subsystems and their relationships, at the macro-social level (national, intergovernmental, and societal dimensionality level boundaries), meso level (trans-organisational supply chain and corresponding linkages), and at the micro level (elements/components and their respective interconnected interfaces). As Griffith and Dougherty (2001) point out: "Can we build a broad socio-technical theory that explains the linkages at so many levels and/or for so many technology issues, or are there different kinds of socio-technical connections that require different theories?" The hope in this phase of the cycle is to move theory and research toward explication so that it may be more clearly relevant to practice.

4. Develop operationalisation methods to bridge the gap between theory and application; principles and action; security requirements and specifications within a given layer of inquiry (Abbas and Michael, 2022; Sanderson et al., 2023). In this recommendation we suggest security mechanisms (in whatever form suits an organisation or entity's existing practices) to satisfy security goals that are critical to one or more socio-technical system.

5. Conduct ongoing security analysis and design to gain more information about existing and future AI in cybersecurity threats which are rapidly evolving given the recent rise of Large Language Models (LLMs) and Generative AI. Security patterns should be identified and reused to address security problems, or socio-technical gaps and security models can be created and in turn we can embed these patterns using agents in anticipation of security breaches to understand plausible cyber, physical, and social responses and their measured effectiveness (Li et al., 2018).

In this way we may forge ahead by defining and developing capacity building activities and strategies for stakeholders in the AI-Cybersecurity Ecosystem. Then design and develop a socio-technical systems AI-cybersecurity capability's maturity model that will help us measure where various stakeholders are in terms of cybersecurity mindset engagement and more. The multilevel perspective here is paramount. By studying the subsystems at various levels, it becomes apparent that a whole-of-practice socio-technical approach is required. Importantly, we need to know where to begin to define these subsystems, the interconnections between systems, and then we need to map the component-level details at each level and how things will work toward the discovery of operationalisation methods. Finally, there is an ongoing requirement to scan the landscape for emergent threats, attempting to identify existing patterns that can be used in agent-based models to anticipate the types of security breaches that are possible, toward continual improvement of cybersecurity defences against AI or any other emergent social, technical, or environmental event or impact.

Acknowledgements

We acknowledge the time and effort of our contributors. In particular, we acknowledge Anthony Burke for his synthesis and analysis of the insights generated from this project, and George Balston from the Alan Turing Institute UK for nurturing this collaboration and helping to make this project a reality. We also thank Anna G and Paul J from the UK National Cyber Security Centre for their thoughtful support, expertise, and contributions throughout this initiative. Our gratitude also to Professor Genevieve Lively for reviewing the draft white paper and providing constructive feedback.

Thank you to our honorary presenters for their wisdom, experience which were key inputs to stimulating our thinking about socio-technical approaches to AI in Cybersecurity. These presenters were:
Workshop 1:
- Dr Ian Levy (National Cyber Security Centre UK)
- Mr Neil Zuring (National Security Agency USA)

Workshop 2:
- Dr Roba Abbas (University of Wollongong, Australia)
- Professor Genevieve Liveley (University of Bristol, UK)

Workshop 3:
- Professor Mariarosaria Taddeo (Oxford Internet Institute, UK)
- Professor Jeremy Pitt (Imperial College London, UK)

Workshop 4:
- Professor Gary Marchant (Arizona State University, USA)
- Professor Lyria Bennett Moses (University of New South Wales, Australia)

Workshop 5:
- Mr Anthony Burke (Camulos)

Workshop 6:
- Adjunct Lecturer Bruce Schneier (Harvard Kennedy School, USA)

For their steady hand in shepherding our explorations, we express our gratitude to our workshop facilitators Lucy Brownsdon, Dan Andrews and Liz Barlow from the Centre for Facilitation, UK.

Critically, we acknowledge and appreciate the diverse voices and contributions of the 40 workshop participants whose explorations, questions, and probing inspired and informed this White Paper. Without our presenters and participants, this call to action would not be possible.

References

Abbas, R. and Michael, K. (2022) Socio-Technical Theory: A review. In: S. Papagiannidis (Ed), TheoryHub Book. http://open.ncl.ac.uk. 978-1-7396044-0-0.

AbuOdeh, M., Adkins, C., Setayeshfar, O., Doshi, P., Kyu H. L. (2021). A Novel AI-Based Methodology for Identifying Cyber Attacks in Honey Pots. In: The Thirty-Fifth AAAI Conference on Artificial Intelligence, 35(17), pp. 15224-15231. https://doi.org/10.1609/aaai.v35i17.17786

Agrafiotis, I., Nurse, J. R. C., Goldsmith, M., Creese, S., Upton, D. (2018). A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate. Journal of Cybersecurity, 4(1), tyy006. https://doi.org/10.1093/cybsec/tyy006

Alkire, S. (2003). A conceptual framework for human security. Department of International Development, University of Oxford, Working Paper 2, 2. https://assets.publishing.service.gov.uk/media/57a08cf740f0b652dd001694/wp2.pdf

Applebaum, A., Dennler, C., Dwyer, P., Moskowitz, M., Nguyen, H., Nichols, N., Park, N., Rachwalski, P., Rau, F., Webster, A., Wolk, M. (2022). Bridging Automated to Autonomous Cyber Defense: Foundational Analysis of Tabular Q-Learning. In: AISec'22: Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security, November, pp. 149–159. https://doi.org/10.1145/3560830.3563732

Appelbaum, S.H. (1997). Socio-technical systems theory: an intervention strategy for organisational development. In: Management Decision, 35(6), pp. 452-463. https://doi.org/10.1108/00251749710173823

Bada, M. and Nurse, J.R.C. (2020). The social and psychological impact of cyberattacks. In: V. Benson and J. Mcalaney, Emerging Cyber Threats and Cognitive Vulnerabilities, pp. 73-92. https://doi.org/10.1016/B978-0-12-816203-3.00004-6

Bauer, J.M. and Dutton, W.H. (2016). The New Cybersecurity Agenda: Economic and Social Challenges to a Secure Internet. In: World Development Report 2016 Digital Dividends, pp. 1-35. https://openknowledge.worldbank.org/handle/10986/7735

Beekun, R.I. (1989). Assessing the effectiveness of sociotechnical interventions: Antidote or fad? In: Human Relations, 42(10), pp. 877-897. https://doi.org/10.1177/001872678904201002

Bella, G., Curzon, P. and Lenzini, G. (2015). Service security and privacy as a socio-technical problem, In: Journal of Computer Security, 23(5), pp. 563-585. https://doi.org/10.3233/JCS-150536

Bergold, J. and Thomas, S. (2012). Participatory Research Methods: A Methodological Approach in Motion. Historical Social Research / Historische Sozialforschung, 37(4), (142), pp. 191–222. http://www.jstor.org/stable/41756482.

Bishop, M. (2005). Introduction to Computer Security, Addison-Wesley, Boston. 978-0321247445

Bonaci, T., Michael, K., Rivas, P., Robertson, L.J. and Zimmer, M. (2022). Emerging Technologies, Evolving Threats: Next-Generation Security Challenges, In: IEEE Transactions on Technology and Society, 3(3), pp. 155-162. https://doi.org/10.1109/TTS.2022.3202323

Bostrom, R.P. and Heinen, J.S. (1977). MIS problems and failures: A socio-technical perspective. Part I: The causes. In: MIS Quarterly, pp. 17-32. https://doi.org/10.2307/248710

Buchanan, B. (2020). A National Security Research Agenda for Cybersecurity and Artificial Intelligence. CSET Issue Brief. May https://cset.georgetown.edu/wp-content/uploads/CSET-A-National-Security-Research-Agenda-for-Cybersecurity-and-Artificial-Intelligence.pdf

Buchannan, B. (2019). The Future of AI and Cybersecurity. Cypher Brief, 30 October https://cset.georgetown.edu/wp-content/uploads/The-Future-of-AI-and-Cybersecurity.pdf

Burgess, C. (2022). Congressional hearings focus on AI, machine learning challenges in cybersecurity. CSO Online. https://www.csoonline.com/article/3663688/congressional-hearings-focus-on-ai-machine-learning-challenges-in-cybersecurity.html

Burke, A. (2020). Robust artificial intelligence for active cyber defence. March. https://www.turing.ac.uk/sites/default/files/2020-08/public_ai_acd_techreport_final.pdf

Capgemini Research Institute (2019). Reinventing Cybersecurity with Artificial Intelligence: The new frontier in digital security. https://www.capgemini.com/wp-content/uploads/2019/07/AI-in-Cybersecurity_Report_20190711_V06.pdf

Carayon, P., Hancock, P., Leveson, N., Noy, I., Sznelwa, L. and Van Hootegem, G. (2015). Advancing a sociotechnical systems approach to workplace safety – developing the conceptual framework, In: Ergonomics, 58(4), pp. 548-564. https://doi.org/10.1080/00140139.2015.1015623

Carcary, M., Renaud, K., McLaughlin, S. and O'Brien, C. (2016). A framework for information security governance and management, In: IT Professional, 18(2), pp. 22-30. https://doi.org/10.1109/MITP.2016.27

Carillo, K., Scornavacca, E., & Za, S. (2017). The role of media dependency in predicting continuance intention to use ubiquitous media systems. In: Information & Management, 54(3), 317-335. https://doi.org/10.1016/j.im.2016.09.002

Carroll, N. and Helfert, M. (2015). Service capabilities within open innovation: revisiting the applicability, In: Journal of Enterprise Information Management, 28(2), pp. 275-303. https://doi.org/10.1108/JEIM-10-2013-0078

Chen, S.P. and Redar, J.M. (2014). Ageing workforce knowledge management and transactional and transformational leadership: a socio-technical systems framework and a Norwegian case study, In: International Journal of Business and Social Science, 5(5), pp. 11-21. https://doi.org/10.30845/ijbss

Chen, X., Liu, C., Li, B., Lu, K., & Song, D. (2017). Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:1712.05526.

Chung, L. (1993). Dealing with security requirements during the development of information systems. In: Rolland, C., Cauvet, C., Bodart, F. (eds.) CAiSE 1993. LNCS, 685, pp. 234–251. Springer, Heidelberg. https://doi.org/10.1007/3-540-56777-1_13

Cihon, P., Maas, M.M., and Kemp, L., (2020). Fragmentation and the future: Investigating architectures for international AI governance. Global Policy, 11(5), pp. 545–556.

Clarke, R. (2023). The Re-Conception of AI: Beyond Artificial, and Beyond Intelligence, In: IEEE Transactions on Technology and Society, 4(1), pp. 24-33. https://doi.org/10.1109/TTS.2023.3234051

Collyer, J., Andrew, A., Hodges, D. (2022). ACD-G: Enhancing Autonomous Cyber Defense Agent Generalization Through Graph Embedded Network Representation. Proceedings of the 39th International Conference on Machine Learning (ML4Cyber workshop). https://dspace.lib.cranfield.ac.uk/bitstream/handle/1826/18288/ACD-G-Enhancing_autonomous_cyber_defense-2022.pdf?sequence=1

Congressional Research Service. (2020). Artificial Intelligence and National Security (10 November) https://sgp.fas.org/crs/natsec/R45178.pdf

CrowdStrike. (2022). Machine Learning (ML) in Cybersecurity: How is ML used in Cybersecurity. 14 September, https://www.crowdstrike.com/cybersecurity-101/machine-learning-cybersecurity/

Cukier, K. (2023). Babbage: What if generative AI destroys biometric security? Babbage from The Economist [podcast]. May. [Online]. Available: https://shows.acast.com/theeconomistbabbage/episodes/babbage-what-if-generative-ai-destroys-biometric-security

Cyber Management Alliance. (2023). Recent Cyber Attacks, Data Breaches & Ransomware Attacks, 1 February. https://www.cm-alliance.com/cybersecurity-blog/recent-cyber-attacks-data-breaches-ransomware-attacks-november-2022.

Dakota C. (2022). Downrange: A Survey of China's Cyber Ranges/ Centre for Security and Emerging Technology, September. https://cset.georgetown.edu/publication/downrange-a-survey-of-chinas-cyber-ranges/

Darktrace. (2022). Darktrace AI: Combining Supervised and Unsupervised Machine Learning. https://darktrace.com/resources/darktrace-ai-combining-supervised-and-unsupervised-machine-learning

Davis, M.C., Challenger, R., Jayewardene, D.N.W., Clegg, C.W. (2014). Advancing socio-technical systems thinking: a call for bravery, In: Applied Ergonomics, 45(2), pp. 171-180. https://doi.org/10.1016/j.apergo.2013.02.009

van Deursen. N., Buchanan, W.J., Duff, A. (2013). Monitoring information security risks within health care. In: Computers & Security, 37, pp. 31-45. https://doi.org/10.1016/j.cose.2013.04.005

Dhanjani, N., Rios, B. and Hardin, B. (2009). Hacking: The Next Generation, O'Reilly Media, 978-1449379216.

Dhillon, G. (2007). Principles of Information Systems Security: Text and Cases, John Wiley and Sons. 978-1943153237.

Dhillon, G. and Backhouse, J. (1996). Risks in the use of information technology within organizations. In: International Journal of Information Management, 16(1), pp. 65-74. https://doi.org/10.1016/0268-4012(95)00062-3

Dhillon, G. and Backhouse, J. (2000). Technical opinion: Information system security management in the new millennium, In: Communications of the ACM, 43(7), pp. 125-128. https://doi.org/10.1145/341852.341877

Dhillon, G. and Backhouse, J. (2001). Current directions in IS security research: towards socio-organizational perspectives. In: Information Systems Journal, 11(2), pp. 127-153. https://doi.org/10.1046/j.1365-2575.2001.00099.x

Dupont, B. (2013). Cybersecurity futures: How can we regulate emergent risks? In: Technology Innovation Management Review, 3(7), pp. 6-11. https://doi.org/10.22215/timreview/700

Farber, D. and Pietrucha, M.T. (2014). A Socio-technical Analysis of Interdependent Infrastructures among the Built Environment, Energy, and Transportation Systems at the Navy Yard and the Philadelphia Metropolitan Region, In: Dolan, T. and Collins, B., (eds.) International Symposium for Next Generation Infrastructure Conference Proceedings: 30 September - 1 October 2014 International Institute of Applied Systems Analysis (IIASA), Schloss Laxenburg, Vienna, Austria. (pp. 151-156). UCL STEaPP: London, UK. http://www.ucl.ac.uk/steapp/isngi/proceedings

Foley, M. Hicks, C., Highnam, K., Mavroudis, V. (2002). Autonomous Network Defence using Reinforcement Learning. In: ASIA CCS '22: Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security, May, pp. 1252–1254. https://doi.org/10.1145/3488932.3527286

Geng, D. and Veerapaneni, R. (2018). Tricking Neural Networks: Create your own Adversarial Examples. Medium. 10 January. https://medium.com/@ml.at.berkeley/tricking-neural-networks-create-your-own-adversarial-examples-a61eb7620fd8

Godage, S.R., Løvåsdal, F., Venkatesh, S., Raja, K., Ramachandra, R., Busch, C. (2023). Analyzing Human Observer Ability in Morphing Attack Detection—Where Do We Stand? in IEEE Transactions on Technology and Society, June, 4(2), pp. 125-145. https://doi.org/10.1109/TTS.2022.3231450

Griffith, T. and Dougherty, D.J. (2001). Beyond socio-technical systems: introduction to the special issue. In: Journal of Engineering and Technology Management, 18(3-4), pp. 207-218. https://doi.org/10.1016/S0923-4748(01)00034-0

Guilford, J.P. (1967). The nature of human intelligence. McGraw-Hill.

Hadid, W., Mansouri, S.A. and Gallear, D. (2016). Is lean service promising? A socio-technical perspective, In: International Journal of Operations and Production Management, 36(6), pp. 618-642. https://doi.org/10.1108/IJOPM-01-2015-0008

Heckman, K.E., Stech, F.J., Thomas, R.K., Schmoker, B. and Tsow, A.W. (2015). Cyber Denial, Deception and Counter Deception: A Framework for Supporting Active Cyber Defense. Springer International, Cham. 978-3319251318.

Herrmann, P. and Herrmann, G. (2006). Security requirement analysis of business processes. In: Electronic Commerce Research, 6(3-4), pp. 305–335. https://doi.org/10.1007/s10660-006-8677-7

Hodson, M. and Marvin, S. (2010). Can cities shape socio-technical transitions and how would we know if they were? In: Research Policy, https://doi.org/10.1016/j.respol.2010.01.020

Hoffman, W. (2021). AI and the Future of Cyber Competition. CSET, January. https://cset.georgetown.edu/publication/ai-and-the-future-of-cyber-competition/

Hoffman, W. (2021). Making AI Work for Cyber Defense. CSET, December. https://cset.georgetown.edu/publication/making-ai-work-for-cyber-defense/

Holgate, J., Williams, S.P. and Hardy, C.A. (2012). Information Security Governance: Investigating Diversity in Critical Infrastructure Organizations. In: BLED Proceedings. 13, 17-20 June, Bled, Slovenia. https://aisel.aisnet.org/bled2012/13

Holton, R. and Boyd, R. (2021). 'Where are the people? What are they doing? Why are they doing it?'(Mindell) Situating artificial intelligence within a socio-technical framework. In: Journal of Sociology, 57(2), pp. 179-195. https://doi.org/10.1177/1440783319873046

Hoon, C. (2013). Meta-synthesis of qualitative case studies: An approach to theory building. Organizational Research Methods, April, 16(4), pp. 522-556. DOI: https://doi.org/10.1177/1094428113484969

Hutson, M. (2018). A turtle—or a rifle? Hackers easily fool AIs into seeing the wrong thing. Science. 19 July. https://doi: 10.1126/science.aau8383

IEEE TTS. (2023). Special Issue on Socio-Technical Ecosystem Considerations: An Emergent Research Agenda for AI in Cybersecurity. In: IEEE Transactions on Technology and Society, June, 4(2), pp. 1-194. https://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=10153436&punumber=8566059

Islam, T., Becker, I., Posner, R., Ekblom, P., McGuire, M., Borrion, H. and Li, S. (2019). A Socio-Technical and Co-evolutionary Framework for Reducing Human-Related Risks in Cyber Security and Cybercrime Ecosystems. In: Dependability in Sensor, Cloud, and Big Data Systems and Applications: 5th International Conference, DependSys, Guangzhou, China, November 12–15, Proceedings, 1123, 277-293. https://doi.org/10.1007/978-981-15-1304-6_22

Jiang, H., Choi, T., and Ko, R. K. L. K. Pandora: A Cyber Range Environment for the Safe Testing and Deployment of Autonomous Cyber Attack Tools. arXiv https://arxiv.org/ftp/arxiv/papers/2009/2009.11484.pdf

Jin, Z., Zhang, S., Hu, Y., Zhang, Y., and Sun, C. (2022). Security State Estimation for Cyber-Physical Systems against DoS Attacks via Reinforcement Learning and Game Theory. Actuators, 11(7), art. 192. https://doi.org/10.3390/act11070192

Johnsen, S. O., and Veen, M. (2013). Risk assessment and resilience of critical communication infrastructure in railways. In: Cognition, Technology & Work, 15, pp. 95-107. https://doi.org/10.1007/s10111-011-0187-2

Kayworth, T. and Whitten, D. (2010). Effective information security requires a balance of social and technology factors. In: MIS Quarterly Executive, 9(3), 5, pp. 2012-52. https://aisel.aisnet.org/misqe/vol9/iss3/5

Khan Adawadkar, A. M. and Kulkarni, N. (2022). Cyber-security and reinforcement learning — A brief survey. Engineering Applications of Artificial Intelligence, 114, September. https://doi.org/10.1016/j.engappai.2022.105116

Kianpour, M., Kowalski, S.J. and Øverby, H. (2021). Systematically Understanding Cybersecurity Economics: A Survey. In: Sustainability, 13, 13677. https://doi.org/10.3390/su132413677

Kinyua, J. and Awuah, L. (2021). AI/ML in Security Orchestration, Automation and Response: Future Research Directions. Intelligent Automation & Soft Computing, 28(2), pp. 527-545. https://doi.org/10.32604/iasc.2021.016240

Kline, R. (2015). The Cybernetics Moment, Johns Hopkins University Press. 978-1421424248.

Kolevski, D., Michael, K., Abbas, R. and Freeman, M. (2021). Cloud Data Breach Disclosures: the Consumer and their Personally Identifiable Information (PII). 2021 IEEE Conference on Norbert Wiener in the 21st Century (21CW), Chennai, India, pp. 1-9, doi: 10.1109/21CW48944.2021.9532579.

Kowalski, S. and Mwakalinga, J. (2011). Modelling the enemies of an IT security system - A socio-technical system security model, In: IMCIC 2011 - 2nd International Multi-Conference on Complexity, Informatics and Cybernetics, Proceedings, pp. 251–256.

van Lamsweerde, A. and Letier, E. (2000). Handling obstacles in goal-oriented requirements engineering. In: IEEE Transactions on Software Engineering, 26(10), pp. 978–1005. https://doi.org/10.1109/32.879820

Li, T., Horkoff, J. and Mylopoulos, J. (2018). Holistic security requirements analysis for socio-technical systems. In: Software & Systems Modeling, 17(4), pp. 1253-1285. https://doi.org/10.1007/s10270-016-0560-y

Lohn, A. (2022). Andrew Lohn's Testimony Before the House Homeland Security Subcommittee on Cybersecurity, Infrastructure Protection, and Innovation. 22 June. https://cset.georgetown.edu/publication/andrew-lohns-testimony-before-the-house-homeland-security-subcommittee-on-cybersecurity-infrastructure-protection-and-innovation/

Lohn, A., Knack, A., Burke, A., Jackson, K. (2023). Autonomous Cyber Defence: A roadmap from lab to ops, In: CETaS Research Reports, June, https://cetas.turing.ac.uk/publications/autonomous-cyber-defence

Malatji, M., Sune, V.S. and Marnewick, A. (2019). Socio-technical systems cybersecurity framework, In: Information and Computer Security, 27(2), pp. 233-272. https://doi.org/10.1108/ICS-03-2018-0031

Martinez Micah Musser, C. (2020). U.S. Demand for Talent at the Intersection of AI and Cybersecurity. CSET, November. https://cset.georgetown.edu/publication/u-s-demand-for-talent-at-the-intersection-of-ai-and-cybersecurity/

Martinez Micah Musser, C. and Garriott, A. (2021). Machine Learning and Cybersecurity: Hype and Reality. CSET, June. https://cset.georgetown.edu/publication/machine-learning-and-cybersecurity/

Meier, R., Heinäaro, K., Lenders, V., Lavrenovs, A. and Gambazzi, L. (2021). Towards an AI-powered Player in Cyber Defence Exercises. 13th International Conference on Cyber Conflict. https://nsg.ee.ethz.ch/fileadmin/user_upload/publications/roland-meier_ai-team_cycon21.pdf

Michael, K. and Abbas, R. (2022). Contribution 12 – technology, information systems and sustainability: a public interest research agenda. In Dwivedi, Y. et al. Climate change and COP26: Are digital technologies and information management part of the problem or the solution? An editorial reflection and call to action. International Journal of Information Management, 63, 102456, pp. 1-39. DOI: https://doi.org/10.1016/j.ijinfomgt.2021.102456

Michael, K., Abbas, R., Pitt, J., Vogel, K., Zafeirakopoulos, M. (2023a). Securitization for Sustainability of People and Place. In: IEEE Technology and Society Magazine, June, 42(2), pp. 22-28. DOI: 10.1109/MTS.2023.3283829

Michael, K., Abbas, R., Roussos, G. (2023b). AI in Cybersecurity: The Paradox. In: IEEE Transactions on Technology and Society, June, 4(2), pp. 104-109. DOI: 10.1109/TTS.2023.3280109.

Middleton, S.E., Lavorgna, A. and McAlister, R. (2020). STAIDCC20: 1st international workshop on socio-technical AI systems for defence, cybercrime and cybersecurity. In: 12th ACM Conference on Web Science Companion, July, (pp. 78-79). https://doi.org/10.1145/3394332.3402897

Minkkinen, M. and Mäntymäki, M. (2023). Discerning Between the “Easy” and “Hard” Problems of AI Governance. In: IEEE Transactions on Technology and Society, June, 4(2), pp. 188-194. doi: 10.1109/TTS.2023.3267382.

Mouratidis, H. and Giorgini, P. (2002). A natural extension of tropos methodology for modelling security. In: Proceedings Agent Oriented Methodologies Workshop, Annual ACM Conference on Object Oriented Programming, Systems, Languages (OOPSLA), Seattle - USA (OOPSLA2002). http://www.oopsla.org/2002/fp/index.html

Mujinga, M., Eloff, M.M. and Kroeze, J.H., (2017). A Socio-Technical Approach to Information Security. In: AMCIS2017: A Tradition of Innovation, Proceedings, 10. https://aisel.aisnet.org/amcis2017/SocialTechincal/Presentations/10

Musser, M. and Garriott, A. (2021). Machine Learning and Cybersecurity: Hype and Reality. CSET, June. https://cset.georgetown.edu/publication/machine-learning-and-cybersecurity/

National Academies of Sciences, Engineering, and Medicine. (2019). Implications of Artificial Intelligence for Cybersecurity: Proceedings of a Workshop. Washington, DC: The National Academies Press. https://doi.org/10.17226/25488

National Cyber Security Centre. (2020). Sociotechnical Security Group Problem Book: An outline of the StSG's future research in Cybersecurity. https://www.ncsc.gov.uk/blog-post/a-sociotechnical-approach-to-cyber-security#problem-book

National Institute of Standards and Technology. (2017). Framework for improving critical infrastructure cybersecurity, draft version 1.1, https://www.nist.gov/sites/default/files/documents////draft-cybersecurity-framework-v1.11.pdf

National Institute of Standards and Technology. (2019). A Taxonomy and Terminology of Adversarial Machine Learning. October. https://csrc.nist.gov/publications/detail/nistir/8269/archive/2019-10-30

National Science and Technology Council. (2020). Artificial Intelligence and Cybersecurity: Opportunities and Challenges: Technical Workshop Summary Report. Washington, DC: Executive Office of the President. March. https://www.nitrd.gov/pubs/AI-CS-Tech-Summary-2020.pdf

Nguyen, T.T. and Reddi, V. J. (2021). Deep Reinforcement Learning for Cyber Security. arXiv, 2 November https://doi.org/10.48550/arXiv.1906.05799

Orlikowski, W.J. and Barley, S.R. (2001). Technology and institutions: What can research on information technology and research on organizations learn from each other? In: MIS Quarterly, 25(2), pp. 145-165. https://doi.org/10.2307/3250927

Osinga, F.P. (2007). Science, strategy and war: The strategic theory of John Boyd. Routledge. 9780415459525.

Paja, E., Dalpiaz, F. and Giorgini, P. (2013). November. Managing security requirements conflicts in socio-technical systems. In: Ng, W., Storey, V.C. and Trujillo, J.C. (eds) Conceptual Modeling. ER 2013. Lecture Notes in Computer Science, 8217. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41924-9_23

Piplai, A., Anoruo, M., Fasaye, K., Joshi, A., Finin, T., Ridley, A. (2022). Knowledge Guided Two-player Reinforcement Learning for Cyber Attacks and Defenses. In International Conference on Machine Learning and Applications. https://ebiquity.umbc.edu/_file_directory_/papers/1173.pdf

Prebot, B., Du, Y., Xi, X. and Gonzalez, C. (2022). Cognitive Models of Dynamic Decisions in Autonomous Intelligent Cyber Defense. In: International Conference on Autonomous Intelligent Cyber-defense Agents, 2, October. https://www.researchgate.net/publication/364965185_Cognitive_Models_of_Dynamic_Decisions_in_Autonomous_Intelligent_Cyber_Defense

Reinhold, T., Kuehn, P., Günther, D., Schneider, T. and Reuter, C. (2023). ExTRUST: Reducing Exploit Stockpiles with a Privacy-Preserving Depletion System for Inter-State Relationships. In: IEEE Transactions on Technology and Society, June, 4(2), pp. 158-170. doi: 10.1109/TTS.2023.3280356.

Rogers, E.M. (1995). Diffusion of Innovations, 4th Edition. Free Press, New York. 978-0029266717

Ryan, K. L. K. (2020). Cyber Autonomy: Automating the Hacker–Self-healing, self-adaptive, automatic cyber defense systems and their impact to the industry, society and national security. arXiv, 8 December. https://doi.org/10.48550/arXiv.2012.04405

Samtani, S., Kantarcioglu, M. and Chen, H. (2020). Trailblazing the artificial intelligence for cybersecurity discipline: a multi-disciplinary research roadmap. In: ACM Transactions on Management Information Systems (TMIS), 11(4), pp. 1-19. https://doi.org/10.1145/3430360

Samonas, S. and Coss, D. (2014). The CIA strikes back: Redefining confidentiality, integrity and availability in security. In: Journal of Information System Security, 10(3), pp. 21-45.

Sanderson, C., Douglas, D., Lu, Q., Schleiger, E., Whittle, J., Lacey, J., Newnham, G., Hajkowicz, S., Robinson, C., Hansen, D. (2023). AI Ethics Principles in Practice: Perspectives of Designers and Developers. In: IEEE Transactions on Technology and Society, June, 4(2), pp. 171-187. doi: 10.1109/TTS.2023.3257303.

Schneier, B. (2019). Cybersecurity for the Public Interest. IEEE Security & Privacy, January/February, https://www.schneier.com/essays/archives/2019/02/public-interest_tech.html

Schoenherr, J.R., Abbas, R., Michael, K., Rivas, P. and Anderson, T.D. (2023). Designing AI Using a Human-Centered Approach: Explainability and Accuracy Toward Trustworthiness. In: IEEE Transactions on Technology and Society, 4(1), pp. 9-23, March. doi: 10.1109/TTS.2023.3257627.

She, A.H., Zarour, M., Alenezi, M., Sarkar, A.K., Agrawal, A., Kumar, R., Khan, R.A. (2020). Healthcare Data Breaches: Insights and Implications. Healthcare (Basel), May, 8(2), 133. doi: 10.3390/healthcare8020133.

Sewak, M., Sahay, S. K., and Rathore, H. (2022). Deep Reinforcement Learning for Cybersecurity Threat Detection and Protection: A Review. arXiv, 6 June. https://doi.org/10.48550/arXiv.2206.02733

Silva, R., Hickert, C., Sarfaraz, N., Brush, J., Silbermann, J., Sookoor, T. (2022). AlphaSOC: Reinforcement Learning-based Cybersecurity Automation for Cyber-Physical System., 2022

ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPS). https://conferences.computer.org/cpsiot/pdfs/ICCPS2022-ifhdJu28kaMK8qGYbf7d0/096700a290/096700a290.pdf

Singh, S. (2023). ChatGPT May Already Be Used in Nation State Cyberattacks, Say IT Decision Makers in BlackBerry Global Research, Blackberry, February. https://www.blackberry.com/us/en/company/newsroom/press-releases/2023/chatgpt-may-already-be-used-in-nation-state-cyberattacks-say-it-decision-makers-in-blackberry-global-research

Smith, C.B. (2023), The Semantic Attack Surface: A Systems-Dynamic Model of Narrative in Cyberspace. In: IEEE Transactions on Technology and Society, June, 4(2), pp. 146-157. doi: 10.1109/TTS.2022.3210782.

Standen, M., Lucas, M., Bowman, D., Richer, T. J., Kim, J., and Marriott, D. (2021). CybORG: A Gym for the Development of Autonomous Cyber Agents. arXiv, August. https://arxiv.org/pdf/2108.09118.pdf

Stevens, T. (2020). Knowledge in the grey zone: AI and cybersecurity. In: Digital War, 1, pp. 164-170. https://doi.org/10.1057/s42984-020-00007-w

Taddeo, M., Jones. P., Abbas, R., Vogel, K. and Michael, K. (2023). Socio-Technical Ecosystem Considerations: An Emergent Research Agenda for AI in Cybersecurity. In: IEEE Transactions on Technology and Society, June, 4(2), pp. 112-118. doi: 10.1109/TTS.2023.3278908.

Tournas, L.N. and Johnson, W.G. (2023). Regulating Brain–Computer Interfaces: Ensuring Soft Law Does Not Go Flat. In: IEEE Transactions on Technology and Society, June, 4(2), pp. 119-124. doi: 10.1109/TTS.2022.3208821.

Tyas Tunggal, A. (2022). The 68 Biggest Data Breaches. 12 December. https://www.upguard.com/blog/biggest-data-breaches

UK Government, National Cyber Strategy. (2022). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1053023/national-cyber-strategy-amend.pdf

Vercellone, C. (2020). European nations may be hesitant to trust AI for cybersecurity. C4ISRNet. 6 May. https://www.c4isrnet.com/cyber/2020/05/06/european-nations-may-be-hesitant-to-trust-ai-for-cybersecurity/

Veksler, V., Buchler, N., LaFleur, C. G., Yu, M. S., Lebiere, C., and Gonzalez, C. (2020). Cognitive Models in Cybersecurity: Learning from Expert Analysts and Predicting Attacker Behavior. Front Psychol. 11. doi: 10.3389/fpsyg.2020.01049

Vergun, D. (2019). Cyber Ops to Gain Speed, Accuracy from AI. USA Department of Defense. 5 September. https://www.defense.gov/News/News-Stories/Article/Article/1953183/cyber-ops-to-gain-speed-accuracy-from-ai/

Wall, D.S. (2020). The Challenges of Socio-Technical AI Systems: From a Criminological Perspective, In: WebSci'20 STAIDCC workshop, July, Southampton, UK, pp. 1-2. https://www.southampton.ac.uk/~sem03/STAIDCC20_wall_paper_07_07_2020.pdf

Walker, G.H., Stanton, N.A., Jenkins, D., Salmon, P., Young, M. and Aujla, A. (2007). Sociotechnical theory and NEC system design, In: Harris, D. (Ed.), Engineering Psychology and Cognitive Ergonomics, EPCE, 4562, Springer-Verlag, Berlin, pp. 619–628. https://doi.org/10.1007/978-3-540-73331-7_68

Wendler, R. (2012). The maturity of maturity model research: A systematic mapping study. In: Information and Software Technology, 54(12), pp. 1317-1339. https://doi.org/10.1016/j.infsof.2012.07.007

Whitworth, B. (2009). A brief introduction to sociotechnical systems, In: Khosrow-Pour, M. (Ed.), Encyclopedia of Information Science and Technology, 2nd edition, IGI Global, Hershey, pp. 394-400. https://doi.org/10.4018/978-1-60566-026-4.ch066

Winfield, A.F., Michael, K., Pitt, J. and Evers, V. (2019). Machine ethics: The design and governance of ethical AI and autonomous systems. In: Proceedings of the IEEE, 107(3), pp. 509-517. https://doi.org/10.1109/JPROC.2019.2900622

Wolk, M., Applebaum, A., Dennler, C., Dwyer, P., Moskowitz, M., Nguyen, H., Nichols, N., Park, N., Rachwalski, P., Rau, F., and Webster, A. (2022). Beyond CAGE: Investigating Generalization of Learned Autonomous Network Defense Policies. arXiv. 30 November. https://arxiv.org/pdf/2211.15557.pdf

Worm, D., Langley, D. and Becker, J. (2015). Modeling Interdependent Socio-technical Networks: The Smart Grid—An Agent-Based Modeling Approach. In: Obaidat, M., Koziel, S., Kacprzyk, J., Leifsson, L. and Ören, T. (eds.) Simulation and Modeling Methodologies, Technologies and Applications. Advances in Intelligent Systems and Computing, 319. Springer, Cham. https://doi.org/10.1007/978-3-319-11457-6_6

Wu, Paul P-Y., C. Fookes, J. Pitchforth and K. Mengersen. (2015). A framework for model integration and holistic modelling of socio-technical systems. In: Decision Support Systems, 71, C, March, pp. 14–27. https://doi.org/10.1016/j.dss.2015.01.006

Xinyun C., Liu, C., Li B., Lu, K., Song, D. (2017). Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. arXiv [cs.CR], 15 December. https://doi.org/10.48550/arXiv.1712.05526

Zeadally, S., Adi, E., Baig, Z., and Khan, I.A. (2020). Harnessing artificial intelligence capabilities to improve cybersecurity. IEEE Access, 8, pp. 23817–23837. DOI: 10.1109/ACCESS.2020.2968045

Zimmermann, V. and Renaud, K. (2019). Moving from a ‘human-as-problem” to a ‘human-as-solution” cybersecurity mindset. In: International Journal of Human-Computer Studies, 131, pp. 169-187. https://doi.org/10.1016/j.ijhcs.2019.05.005

About the Authors

Roba Abbas
Roba Abbas (Member IEEE, ACM) is a Senior Lecturer of Operations and Systems in the Faculty of Business and Law at the University of Wollongong (UOW), Australia. She was the Academic Program Director of the Bachelor of Business and Bachelor of Business Administration programs between 2021-22. She has served in research, teaching, and governance roles in multiple schools including Engineering and Informatics from 2007. Roba was also a Visiting Professor with the School for the Future of Innovation in Society in the College of Global Futures at Arizona State University in 2022. She is the Co-Editor-in-Chief of the *IEEE Transactions on Technology and Society*, and previously the Associate Editor and Administrator of the *IEEE Technology and Society Magazine*. Roba is also the Technical Committee Chair of the Socio-Technical Systems Committee of the IEEE. Prior to academia, Roba was employed by a web design and development company.

Katina Michael
Katina Michael BIT, MTransCrimPrev, PhD (Senior Member IEEE, ACM SIGCAS), is a Professor with Arizona State University and a Senior Global Futures Scientist with the Global Futures Laboratory. At ASU, she has a joint appointment with the School for the Future of Innovation in Society and School of Computing and Augmented Intelligence. Katina's research focuses on the social implications of emerging technologies. She was responsible for establishing the Human Factors Series in the Research Network for a Secure Australia (RNSA 2005-2009), was an external member of the Centre of Excellence in Policing and Security (CEPS 2009-2013), and ran the Social Implications of National Security (SINS) workshops from 2006 to 2022. Since 2021, Katina has advised DARPA on matters pertaining to ethics, law, and societal implications (ELSI) of complex socio-technical systems. She has been funded by the National Science Foundation, the Canadian Social Sciences and Humanities Research Council, and the Australian Research Council. She is the Director of the Society Policy Engineering Collective, the Founding Editor-in-Chief of the *IEEE Transactions on Technology and Society* and was formerly Editor-in-Chief of the *IEEE Technology and Society Magazine* and Editor at *Computers & Security*. She is the Founding Chair of the ASU Master of Science in Public Interest Technology, and Technical Committee Co-Chair of Socio-Technical Systems at IEEE. Prior to academia, Katina was employed by Nortel Networks, Anderson Consulting, and OTIS Elevator Company.

Jeremy Pitt
Jeremy Pitt is Professor of Intelligent and Self-Organising Systems in the Department of Electrical and Electronic Engineering at Imperial College London. He received a BSc in Computer Science from the University of Manchester and a PhD in Computing from Imperial College (University of London). He has been teaching and researching on Artificial Intelligence and Human-Computer Interaction for over thirty years, where his research programme has used computational logic to specify algorithmic models of social processes, with applications in cyber-physical and socio-technical systems, especially for sustainable, fair and legitimate self-governance. He has collaborated on research projects extensively in Europe as well as in India and New Zealand, and has held visiting professorial positions in Italy, Japan and Poland. He has published more

than 200 articles in journals, conferences and workshops, and this work has received several Best Paper awards. He is editor of *This Pervasive Day* (ICPress, 2012) and *The Computer After Me* (ICPress 2014), and author of *Self-Organising Multi-Agent Systems* (World Scientific, 2022). He is a trustee of AITT (the Association for Information Technology Trust), a Fellow of the BC5 (British Computer Society) and of the IET (Institution of Engineering and Technology), and from 2018-2023 was Editor-in-Chief of *IEEE Technology & Society Magazine*.

Kathleen M. Vogel
Kathleen M. Vogel is Professor in the School for the Future of Innovation and Senior Global Futures Scientist in the Global Futures Laboratory at Arizona State University. She is a 2023 Non-Resident Fellow with the Irregular Warfare Initiative, a joint production of Princeton's Empirical Studies of Conflict Project and the Modern War Institute at West Point. Previously Vogel was a Rutherford Fellow in the Defence and Security Programme at The Alan Turing Institute (2018-2019), a Jefferson Science Fellow in the U.S. Department of State (2016-2017), and a William C. Foster Fellow in the U.S. Department of State (2003). She has previously served on the faculty of the University of Maryland, College Park, North Carolina State University, and Cornell University. Vogel has also spent time as a visiting scholar at the Woodrow Wilson International Center for Scholars, Cooperative Monitoring Center, Sandia National Laboratories, and the Center for Nonproliferation Studies, Monterey Institute of International Studies. Vogel is author of *Phantom Menace or Looming Danger?: A New Framework for Assessing Bioweapons Threats* (Baltimore: The Johns Hopkins University Press, 2013). She is currently working on a book manuscript with Carl Ford on improving U.S. intelligence analysis.

Mariana Zafeirakopoulos
Mariana has 15 years of practice-based global experience working in strategic decision-making, insight and forecasting in national security and law enforcement contexts for government and private industry, particularly in areas of serious and organised crime and terrorism prevention. Over the last ten years, Mariana has focused her research and practice on innovation and design to progress national security efforts. Her transdisciplinarity approach unites practices from areas like futuring, social innovation, strategic design, systems thinking and strategic intelligence to emerging complex contexts. She is currently undertaking a PhD at the University of Technology, Sydney, exploring the role of design methods and processes for integrating knowledge in complex national security contexts. Mariana regularly teaches across a range of subjects related to futuring, strategic intelligence, and design innovation practices across several Australian universities. Previously, she worked as a lead Strategic Designer and Researcher at UTS's Design Innovation Research Centre, where she led projects on designing corruption-prevention for small-medium businesses, the prevention of sexual assault and harassment in educational settings and applying human-centred co-design to urban city and infrastructure projects.

The Alan Turing Institute

turing.ac.uk

@turinginst