The Case for Including AI Security in the 2030 Cyber Security Strategy

Submission to the 2023-2030 Australian Cyber Security Strategy Discussion Paper

Harriet Farlow[1] and Dr. Julie Banfield[2]

[1]CEO of Mileva Security Labs, PhD Candidate at the University of New South Wales, Canberra
[2]CTO of Mileva Security Labs

Artificial Intelligence (AI) and Machine Learning (ML) have undergone rapid adoption across Government, Industry, and Academia around the world over the last few years. As intelligent systems are incorporated into cyber and information systems, their security should be considered an extension of cyber security. As the 2023 - 2030 Australian Cyber Security Strategy Discussion paper notes, trust in digital systems is essential in carrying out the goal of becoming the most cyber secure nation by 2030 [1]. Trust in those AI and ML components of the system are equally important. Currently, over 90% of business leaders surveyed report their business implements some kind of AI [2]. However, just a fraction of those businesses consider, or are aware of, AI Security [2]. AI Security refers to the technical and governance practices that aim to protect AI systems from deliberate subterfuge by an adversary. AI Security mirrors the field of cyber security in that it deals primarily with technical weaknesses in the design of an AI system that make it vulnerable to attack. Attacks on AI systems fall under the umbrella of Adversarial Machine Learning (AML). AML attacks could have catastrophic consequences on every productionised AI system, which currently spans all major industries including Health, Energy, and Defence. This informs two main recommendations for the inclusion of AI Security in the 2030 Cyber Security Strategy. First, we recommend investment in the technical ecosystem to support AI assurance, mitigation, control and audit. Secondly, we recommend governance frameworks, policy and legislation that guides and mandates the security posture of AI systems.

1 Introduction

Just as Artificial Intelligence (AI) does not have a universal definition, nor does AI Security. Those in the field generally refer to AI Security as the technical and governance considerations that pertain to hardening AI systems to technical exploits by an adversary. The offensive side of AI Security is Adversarial Machine Learning (AML), which represents the ability to hack Machine Learning (ML) algorithms through a range of methods that broadly exploit technical vulnerabilities inherent in the architecture of deep learning optimisation. These kinds of attacks are deliberately engineered by an adversary to compromise the AI system's ability to behave as intended, through leaking sensitive information about the training data, evading classification, or hijacking the model's functions [3].

AI Security is different to AI Safety, which is generally concerned with safety or ethical considerations borne out of biased data or poorly designed systems. AI Security, on the other hand, is an extension of the field of cyber security in that it deals primarily with technical weaknesses in the design of an AI system that make it vulnerable to attack. Therefore, governance considerations in AI Security have a different flavour to those that focus on AI Safety. They are inspired by the field of cyber security and recommend specific technical and governance controls to 'harden' AI systems, and incorporate these mitigations according to a risk-based approach.

2 AI Security Implications

As the 2023 - 2030 Australian Cyber Security Strategy Discussion paper notes, a safe and secure cyber and digital ecosystem is essential for enabling Australia's economic prosperity [1]. Trust in those cyber and information systems, of which AI is increasingly a key component, is extremely important. Much has changed since the last cyber security strategy, and one key transition is the rapidly increasingly capability and uptake of AI systems. Over 90% of business leaders surveyed report their business implements some kind of AI, and it has been adopted across a range of industries including Health, Defence, Finance and Manufacturing [2]. The potential for its application in more risky scenarios, such as facial recognition and lethal autonomous weapons systems (LAWS) is now very plausible. While regulations around how AI is used is increasingly discussed, a key challenge that is rarely addressed is how vulnerable AI systems are to adversarial attack [3, 4]. In 2014, researchers first demonstrated how specially crafted perturbations applied to an image can result in drastic changes to a model's ability to identify the subject of that image [5, 4]. What this means in practice is that a person could wear a pair of specially crafted adversarial glasses or jewellery and not just evade a facial recognition system but convince the system that they are someone else [6, 7]. This technology sounds like science fiction but is entirely possible, and can target any system that relies on computer vision, including LAWS, autonomous vehicles and image classification in healthcare and manufacturing. For example, systems that rely on object detection may be compromised through targeted perturbations camouflaging of the intended target, or causing some other object (or person) to be targeted instead. These methods can also apply cross-domain, with these techniques proven to work when transferred from computer vision to transaction-based fields like banking, medicine and communications, generating fake financial transactions that may crash the stock market, or communications signals to falsify messages [8, 9].

Existing AI Governance frameworks typically address AI Safety challenges like explainability, proportionality and fairness [10, 11, 12]. However, rarely do they address AI Security challenges. In cyber security there is a complex and interdependent ecosystem of organisations, regulations and tools that assure the security of cyber systems. There is no equivalent environment that exists for AI Security [13]. In fact, it is such an emergent discipline that it is rarely addressed in high-level

policy discussions [13, 14]. Therefore, we outline two recommendations for including AI Security in the 2030 Cyber Security Strategy.

2.1 Recommendation 1: Investment in the technical ecosystem to support AI assurance, mitigation, control and audit

We recommend more research at the intersection of cyber security and AI security. There are frameworks in cyber security that could be adapted to AI security to pivot this defensive mindset to one of assurance. MITRE's ATLAS framework is one that has been adapted from their CAPEC and ATT&CK frameworks for cyber security, and is the current industry leader [15]. However, cyber security has many other frameworks across a number of industries and authorities including those by standards organisations like ISO [16], governments [17, 18], and industry bodies [19]. There are also cyber security activities that occur in parallel across a community of interest that promote a security ecosystem. This might include ML pen-testing, red teaming and bug bounties. To complement this community of interest there could also be a mechanism to share discoveries between academia and those affected industry bodies.

The 2030 Cyber Security Strategy should play a role in stimulating this community of interest, much as it does for the cyber security ecosystem.

2.2 Recommendation 2: Governance frameworks, policy and legislation that guides and mandates the security posture of AI systems

While organisations like Microsoft led the charge for cyber security by design, many of these efforts would not have persisted if less resourced organisations were not incentivised to deploy secure code through regulation, governance, and more recently, legislation.

Many governments have AI Roadmaps [20], Initiatives [21, 10] or Action Plans [22]. However these focus on increasing AI use among the public and private sectors and on AI Safety, but not AI Security. Some companies, such as Microsoft, Facebook and Google, also have AI Red Teams, but their mandate or performance is not regulated, and therefore lies the risk that the main purpose of the Red Team is to ensure positive public sentiment.

There has recently been some progress with ISO standards for AI currently in the working group stage [23], and NIST noting "AI standards that articulate requirements, specifications, guidelines, or characteristics can help to ensure that AI technologies and systems meet critical objectives for functionality, interoperability, and trustworthiness—and that they perform accurately, reliably, and safely." The European Union's GDPR articulates requirements for data privacy and explainability in AI systems [24]. We would argue that the Australian Government needs to play a greater role in AI Security, and that aspects of its security should be guided by policy, and legislated. Given the cross-border nature of the Internet, tech companies, and ML models, international consistency would greatly benefit the ability of organisations to ensure compliance and for companies to comply with requirements.

To complement this, as there are authorities with the remit of cyber security, these organisations could similarly adopt AI Security. The Australian Cyber Security Centre (ACSC) could extend its role beyond cyber security to include AI Security, and over time mature its ability to manage AI Security standards, share information and insights with the community, and manage reporting obligations.

3 Conclusion

The current 'wild west' approach to AI mirrors a similar approach to computing and networking in the latter half of the twentieth century. The rapid adoption of AI/ML may represent a future security threat on par with the current cyber security threat. We posited AI Security as an extension of cyber security, making the following two recommendations:

1. Investment in the technical ecosystem to support AI assurance, mitigation, control and audit; and

2. Governance frameworks, policy and legislation that guides and mandates the security posture of AI systems.

Therefore, we recommend AI Security is included as a topic in the 2023-2030 Australian Cyber Security Strategy.

References

[1] 2023-2030 Australian Cyber Security Strategy.

[2] Ram Shankar Siva Kumar, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, and Sharon Xia. Adversarial Machine Learning – Industry Perspectives. *arXiv:2002.05646 [cs, stat]*, March 2021. arXiv: 2002.05646.

[3] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, December 2018.

[4] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*, 2015.

[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, February 2014. arXiv:1312.6199 [cs].

[6] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 1528–1540, New York, NY, USA, October 2016. Association for Computing Machinery.

[7] Tom Brown, Dandelion Mane, Aurko Roy, Martin Abadi, and Justin Gilmer. Adversarial Patch. 2017.

[8] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In *Proceedings of the 35th International Conference on Machine Learning*, pages 274–283. PMLR, July 2018. ISSN: 2640-3498.

[9] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust Physical-World Attacks on Deep Learning Visual Classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, Salt Lake City, UT, USA, June 2018. IEEE.

[10] National Artificial Intelligence Initiative - National and Federal Agency AI Strategy Documents.

[11] GCHQ | Pioneering a New National Security: The Ethics of Artificial Intelligence.

[12] Science Department of Industry. Australia's Artificial Intelligence Ethics Framework, June 2021. Publisher: Department of Industry, Science, Energy and Resources.

[13] Elisa Bertino, Murat Kantarcioglu, Cuneyt Gurcan Akcora, Sagar Samtani, Sudip Mittal, and Maanak Gupta. AI for Security and Security for AI. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, CODASPY '21, pages 333–334, New York, NY, USA, April 2021. Association for Computing Machinery.

[14] By Susan Miller and 2019 Jan 25. DARPA outlines adversarial AI defense - GARD (Guaranteeing AI Robustness against Deception) Program.

[15] The MITRE Corporation. MITRE | ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) Adversarial Attack Framework.

[16] ISO - ISO/IEC 27001 — Information security management.

[17] Homepage | CISA.

[18] Information Security Manual (ISM) | Cyber.gov.au.

[19] Cyber Kill Chain®, April 2022.

[20] AI Governance: Opportunity and Theory of Impact - EA Forum.

[21] NSCAI. The National Security Commission on Artificial Intelligence.

[22] Science Department of Industry. Australia's Artificial Intelligence Action Plan, June 2021. Publisher: Department of Industry, Science, Energy and Resources.

[23] ISO/IEC JTC 1/SC 42 - Artificial intelligence.

[24] Aloni Cohen and Kobbi Nissim. Towards Formalizing the GDPR's Notion of Singling Out. *Proceedings of the National Academy of Sciences*, 117(15):8344–8352, April 2020. arXiv: 1904.06009.