Safe and Responsible AI in Australia: Response to the Discussion Paper

Mileva Security Labs

Firstly, we thank you for the opportunity to respond to this discussion paper. Mileva Security Labs provides advisory, training and research on AI Security and we would applaud the introduction of risk-based regulation of AI. We provide our responses here in addition to the online form.

1 Response to Questions

1.1 Do you agree with the definitions in this discussion paper? If not, what definitions do you prefer and why?

While there is no universally accepted definition of AI, we agree that aligning the definitions with ISO/IEC 22989:2022 is best.

1.2 What potential risks from AI are not covered by Australia's existing regulatory approaches? Do you have suggestions for possible regulatory action to mitigate these risks?

We agree with the summary of potential risks listed in the discussion paper but would add the risk of adversarial attacks on AI by nation-states and criminals. This mirrors closely the cyber security threat from nation states and proxies. We believe this highlights the pressing need for AI robustness and responsibility, to prevent AI systems being deployed that are vulnerable to attacks, and to ensure potential attacks can be identified and remediated early.

1.3 Are there any further non-regulatory initiatives the Australian Government could implement to support responsible AI practices in Australia? Please describe these and their benefits or impacts.

In our experience, most organisations want to ensure their AI systems align with all voluntary standards but they often lack confidence understanding or implementing high-level governance standards at the technical level. We would advocate for more training for both executives and practitioners to build confidence leading AI-driven businesses, and understand best practice for AI robustness.

1.4 Do you have suggestions on coordination of AI governance across government? Please outline the goals that any coordination mechanisms could achieve and how they could influence the development and uptake of AI in Australia.

In addition to the points already addressed, the Australian Cyber Security Centre (ACSC) would be an ideal organisation to take carriage of AI standards, and reporting risks and vulnerabilities, as an extension of their cyber security offering. We see the building of a community of practice across Government, academia and industry to be a vital component for increasing the awareness and adoption of standards by all organisations that might like to implement AI.

1.5 Are there any governance measures being taken or considered by other countries (including any not discussed in this paper) that are relevant, adaptable and desirable for Australia?

We recommend those systems that have opted for certification of higher-risk applications.

1.6 How can the Australian Government further support responsible AI practices in its own agencies?

We would again point to investment in a more mature sovereign AI community of interest. Increasing understand of responsible AI practices at the time of adoption (not as an afterthought) is an essential lesson from the rise of the cyber and information security threat.

1.7 Given the importance of transparency across the AI lifecycle, please share your thoughts on where and when transparency will be most critical and valuable to mitigate potential AI risks and to improve public trust and confidence in AI, and mandating transparency requirements across the private and public sectors, including how these requirements could be implemented.

We would not advocate for the total ban of any particular AI technology or use case, and rather point to stringent certifications or regulations on when and how technologies should be used. We recommend the increased up-skilling for all people when it comes to AI, but particularly in professional (even non-technical roles) and in educational institutions.

1.8 How would banning high-risk activities (like social scoring or facial recognition technology in certain circumstances) impact Australia's tech sector and our trade and exports with other countries?

At this stage we assess that banning of these technologies would most likely impact the academic and innovation communities. Social scoring and facial recognition are specific adaptions of more general capabilities and the investment in those general capabilities should not be stalled due to its potential uses.

1.9 Do you support a risk-based approach for addressing potential AI risks? If not, is there a better approach?

Yes we do, but suggest it could be more granular than a simple low/medium/high since similar applications of AI may rely on very different technical underpinnings, which could impact how vulnerable those technologies to adversarial attacks. We suggest an additional layer focusing not just on how the AI is used but also on how secure the actual AI systems are themselves (similar to cyber security).

1.10 What do you see as the main benefits or limitations of a risk-based approach? How can any limitations be overcome?

We advocate for the maturing of the AI Security to become more in line with cyber security, including the paradigm of risk-based management of mitigations and controls. However this must be supported by a technical and governance ecosystem, which is currently still emerging in the AI space and would benefit from increased investment.

1.11 Is a risk-based approach better suited to some sectors, AI applications or organisations than others based on organisation size, AI maturity and resources?

The risk-based approach is naturally more easily adopted by those organisations that already have mature cyber and information security risk management practices. Ideally such an approach should not put a great burden on innovation or small-medium enterprises, but up-skilling will likely be required by these groups.

1.12 What elements should be in a risk-based approach for addressing potential AI risks? Do you support the elements presented in Attachment C?

We support the elements included in Attachment C but the nature and scope of what would be required from an Impact Assessment should balance the need for transparency, while reducing unnecessary reporting burden.

1.13 How might a risk-based approach apply to general purpose AI systems, such as large language models (LLMs) or multimodal foundation models (MFMs)?

For these technologies in particular would arise questions around who is responsible for the assessment. For example, many organisations currently utilise open source technologies that might not make available all information to be able to fulfill the requirements of an assessment (ChatGPT for

example), unless the organisation that created technology provides this assessment. Further, this assessment may not be able to be verified by external sources since much of this information may be held as proprietary. This is especially the cause for high value LLMs and MFMs, and even then these technologies change rapidly. We would advocate for a flexible assessment approach that puts responsibility on both parties, but the balance of which should be guided by consultation.

1.14 Should a risk-based approach for responsible AI be a voluntary or self-regulation tool or be mandated through regulation?

We believe the risk-based approach should be mandatory for all organisations using AI, but the degree to how this is implemented should differ according to the use case, risk level, and type of organisation.

2 Discussion

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. 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 [1]. However, just a fraction of those businesses consider, or are aware of, AI Security [1]. 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.

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 [2].

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.

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. Funding for training related to AI risks at both the leadership and practitioner level;
2. Investment in the technical ecosystem to support AI assurance, mitigation, control and audit; and
3. Governance frameworks, policy and legislation that guides and mandates the security posture of AI systems.

Thank you again for the opportunity to respond to this paper.

References

[1] 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.

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