Submission by Professor Dan Jerker B. Svantesson to the Department of Industry, Science and Resources regarding:

The public consultation on Safe and responsible AI in Australia

July 2023

Professor Dan Jerker B. Svantesson

Faculty of Law, Bond University
Gold Coast, Queensland, 4229
Australia
dasvante@bond.edu.au

Summary of major points

- These submissions outline a framework consisting of 13 principles that ought to guide Australia's approach to mitigating the potential risks of AI.
- Australia should also consider adopting certain structures including an 'AI Safety Commissioner'.
- Where other States also adopt this approach, Australia may also wish to consider working towards establishing a structure for the active collaboration and cooperation between each State's respective AI Safety Commissioner; perhaps in the form of a 'Council of AI Safety Commissioners'.
- The impact of AI clearly depends on the context of its use. Thus, what may be a suitable approach in one context, may be highly unsuitable in another.
- AI brings with it risks of a scale that justify a move from the tradition of largely reactive law-making to the use of pre-emptive 'red lines'.
- AI-related work must proceed with realistic expectations and an acute awareness of the difference between, on the one hand, marketing samples and, on the other hand, products ready for safe and compliant, rights-respecting, transparent implementation.

1. General remarks

1. I welcome the initiative to seek input on the safe and responsible AI in Australia.

2. These submissions are intended to be made public.

3. These submissions deal only with a selection of the questions. No views are expressed on any other matters.

4. These submissions draw upon: Dan Svantesson, Cybercrime and the Adoption of Artificial Intelligence Systems for Judicial Decision-Making in Criminal Justice Systems, *Commonwealth Cybercrime Journal* (2023) Volume 1, issue 1, 152-179 https://thecommonwealth.org/publications/cybercrime-journal-1-1#svantesson.

2. Introductory observations

5. The overriding aim of the consultation is to seek advice on steps Australia can take to mitigate the potential risks of AI. Responding to that, I outline a framework consisting of 13 principles that ought to guide Australia's approach to mitigating the potential risks of AI. Some observations are also made about structures that may usefully be adopted.

6. However, before presenting those principles, and discussing the structural issues, it may be noted that, given that AI doubtlessly will continue to evolve, the task before us is not merely to ensure that regulation catches up with AI technology. Rather, it also involves setting rules and standards now that can guide and regulate the adoption of AI systems on an ongoing basis. This is not a task that can be addressed to completion at this stage; it will require an ongoing commitment, monitoring, and review.

7. Furthermore, the impact of AI clearly depends on the context of its use. Thus, what may be a suitable approach in one context, may be highly unsuitable in another.

8. Indeed, AI brings with it risks of a scale that justify a move from the tradition of largely reactive law-making to the use of pre-emptive 'red lines'. For example, Australia may already now wish to articulate a ban on AI as final arbiter in the context of judicial decision-making. Similarly, AI systems use for general biometric surveillance, and such systems used for 'social scoring' could, for example, be specifically banned.

9. Finally, by way of introduction, AI-related work must proceed with realistic expectations and an acute awareness of the difference between, on the one hand, marketing samples and, on the other hand, products ready for safe and compliant, rights-respecting, transparent implementation. AI hype is frequently created around 'almost' perfect products/solutions with the promise of perfection being around the proverbial corner. However, we must remain alert to the difference between near perfection and actual

perfection; progress to the point of near perfection is no guarantee for ever reaching the stage of actual perfection.

3. The 13 principles to guide Australia's approach to mitigating the potential risks of AI

10. A study of relevant policy documents, academic literature, ethical frameworks, and proposed legal instruments, shows a considerable degree of consistency in what is held to be required in the adoption of AI systems. Many of the same principles consistently appear across the various publications. I seek to summarise those principles here and advocate their use to guide Australia's approach to mitigating the potential risks of AI.

11. Most of the principles outlined below are relevant for both the public and the private sector adoption of AI.

12. Some of the principles outlined below are predominantly of relevance in the context of public sector adoption of AI. However, also for this latter type of principles, Australia may wish to explore ways to integrate aspects of them into the private sector adoption of AI.

a. The fundamental rights principle

A framework for the safe, accountable, and rights-respecting adoption of AI systems must be anchored in, and take care to integrate, international and domestic human rights law, and all other fundamental rights and values of free and democratic societies (the 'fundamental rights principle').

b. The rule of law principle

In addition, such a framework must be supportive of the multifaceted concept of the rule of law; both in the sense of directly supporting the rule of law and in the sense of working to enhance trust in the rule of law (the 'rule of law principle'). In fact, a rule of law focus may usefully be applied as a filter in the sense that any adoption of AI systems that supports the rule of law ought to be explored, and any adoption of AI systems that undermines the rule of law must be rejected even where they may otherwise prove effective.

c. The lifecycle principle

Steps to guarantee the adherence to, and support for, fundamental rights and the rule of law must be taken throughout the AI systems' entire 'lifecycle' (the 'lifecycle principle'), and

that lifecycle may be split into a number of stages.[1] Thus, the regulation of AI must be subject to ongoing monitoring, review and evaluation. This ongoing work should involve both public and private actors and benefits from extensive consultation, audits, democratic scrutiny,[2] and multistakeholder input.

d. The justification principle and the precautionary principle

Various tools may be pursued during the different stages of the AI lifecycle. For example, an important aspect in the 'design stage' is to promote responsible innovation through tools such as e.g., 'human-rights-by-design', 'privacy-by-design', and 'ethical-by-design'. In addition, in relation to the 'deployment stage', Australia may wish to consider and adopt the 'justification principle' and the 'precautionary principle'.

The 'justification principle' means that, for any proposed adoption of AI systems that proposal must be justified by reference to specific benefits and the achievability of the postulated benefits must be demonstrated. The justification principle will encourage a purposeful, rather than hype-driven adoption of AI.

The 'precautionary principle' signifies that, in any situation where it may reasonably be suspected that an AI system may cause harm, those proposing the adoption of the AI systems bears the burden to show that the system may be safely adopted. As noted by one leading commentator: "Ill-advised uses of AI need to be identified in advance and nipped in the bud, to avoid harm to important values".[3]

e. The appealability principle

Australia ought to consider and adopt the 'appealability principle'; that is, for any decision made by AI, it must always be possible to appeal the decision to a human. In fact, Australia

[1] For example, a document published by The Alan Turing Institute speaks of the 'design stage', the 'development stage', and the 'deployment stage' (Leslie, D., Burr, C., Aitken, M., Cowls, J., Katell, M., and Briggs, M. (2021). Artificial intelligence, human rights, democracy, and the rule of law: a primer. The Council of Europe., at 10-12). Similarly, the OECD speaks of four different phases: "AI system lifecycle phases involve: *i)* 'design, data and models'; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; *ii)* 'verification and validation'; *iii)* 'deployment'; and *iv)* 'operation and monitoring'." (OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449).

[2] See e.g.: the Montreal Declaration for a responsible development of artificial intelligence, https://www.montrealdeclaration-responsibleai.com/the-declaration, Principle 5.

[3] Roger Clarke, Guidelines for the Responsible Business Use of AI Foundational Working Paper Revised Version of 20 February 2019, Xamax Consulting Pty Ltd, http://www.rogerclarke.com/EC/GAIF.html.

may wish to consider embracing a ban on AI as final arbiter at least in certain specific contexts.

f. The explainability principle

Australia may wish to consider and adopt the 'explainability principle' essential for the above-mentioned appealability principle, for upholding justice and dignity for those affected by a decision, and for facilitating society's monitoring of justice and equality. Under this principle, any decision made by, or supported by, an AI system must be explainable to be valid. The principle covers both 'ex ante explainability' (i.e., the decision-making process being explainable prior to its use) and 'ex post explainability' (i.e., the decision-making process being explainable after its use).[4] Adherence to the 'explainability principle' may usefully incentivise further work on what has been termed 'explainable AI'.[5]

g. The transparency principle

Australia may wish to consider and adopt the 'transparency principle'. This principle is related to, and partly overlaps with, the explainability principle. However, it does not only relate to the need for transparency in the sense of explainability. It also calls for transparency in the sense of persons being made aware of the fact that AI has played a role in a decision made about them, what methods were used and, at least, what parameters were considered by the AI system.

Further, the 'transparency principle' emphasises the need for law to clearly identify what decisions may be made by, or partially made by, AI systems. The transparency principle may be in tension with intellectual property and trade secret protections afforded to the developers of AI systems. In such situations, the rule of law demands that only systems that fulfil the transparency principle are adopted. This may be considered in the rules governing the procurement process of AI systems and may even point to a need to explore the government playing a role in the design and creation of AI systems for sensitive matters such as judicial decision-making.

h. The non-discrimination principle

[4] Black and Murray notes that: "only some algorithmic methods lend themselves to ex ante transparency, notably those relying on decision trees. […] in the case of other algorithmic technologies, such as neural networks, the machine is learning as it processes the data and it is not possible to set out the reasoning in advance." (Julia Black & Andrew Murray, Regulating AI and Machine Learning: Setting the Regulatory Agenda, European Journal of Law and Technology, Vol 10 Issue 3 (2019)).

[5] See further: Ashley Deeks, "The Judicial Demand for Explainable Artificial Intelligence," Columbia Law Review 119, no. 7 (November 2019): 1829-1850.

Given the prominence of the risk of AI systems introducing, augmenting, or re-introducing, discrimination between individuals or groups of individuals, Australia may wish to specifically consider and adopt the 'non-discrimination principle' requiring an ongoing commitment to eliminate discrimination, and risks for discrimination, in the adoption of AI systems.

There is two dimensions to this principle. It aims to utilise AI systems to eliminate existing discrimination and it aims to prevent AI systems, one way or another, introducing discrimination. On a practical level, this may take several forms. For example, attention can be directed at what variables AI systems use as the basis for their decisions. Where the variables include examples of sensitive data such as gender, political opinions, or ethnicity, that may be used in a discriminatory manner, special steps may be required to ensure that the system does not unfairly discriminate between individuals or groups of individuals. Furthermore, as noted by European Commission For The Efficiency of Justice, "the use of machine learning and multidisciplinary scientific analysis to combat such discrimination should be encouraged."[6]

i. The quality assurance principle

Australia may wish to consider and adopt the 'quality assurance principle'. A reliable application of AI systems must be able to maintain quality assurance and should reliably operate in accordance with its intended purpose, over its lifecycle – while close enough may be good enough in some settings, that is not the case where AI is applied in a sensitive setting. This places quality and robustness requirements on both the AI system as such, and the data it uses. Further, it places quality requirements on the operation of the system by those using it. The use of certification schemes, external audits and the involvement of external, independent, expert assessment may be valuable tools in this context.

j. The resilience principle

While Australia has already commenced important work on cybersecurity, the world's cyber-dependence has by far outpaced efforts aimed at ensuring cyber-resilience. This has created serious societal vulnerabilities frequently exploited by criminals and hostile state actors. Thus, Australia may wish to consider and adopt the 'resilience principle'.

Under this principle, there should be no situation of full AI-dependence; at least not when it comes to sensitive systems. All systems must include back-up features ensuring continuous functionality of the judicial system even where a particular AI system is attacked or otherwise fails. The 'resilience principle' also imposes cybersecurity obligations on users of

[6] European Commission For The Efficiency of Justice, European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment (2018), https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c, at 9.

AI systems meaning that all reasonable steps must be taken to ensure system integrity, and to avoid manipulation, and unlawful access. In this context, users must be mindful that manipulation can take many forms. For example, also where the algorithms are operating properly, the data may have been manipulated so as to either cause undue outcomes in a specific instance or so as to impact the long-term operation of the system.

k. The human oversight principle

Many of the risks and challenges identified in relation to AI systems may be mitigated where the structures adopted include appropriate human oversight,[7] review, audits, and intervention. Thus, Australia may wish to consider and adopt a 'human oversight principle' mandating such oversight.

l. The accountability principle

Australia may wish to consider and adopt the 'accountability principle' that partly overlaps with some of the previously noted principles. The important role accountability can play in technology regulation is widely recognised,[8] and as outlined by the Australian Human Rights Commission:

> "Accountability involves ensuring that the law is followed in a decision-making process. It includes both a *corrective* function, facilitating a remedy for when someone has been wronged, as well as a *preventive* function, identifying which aspects of a policy or system are working and what needs adjustment."[9]

m. The human-centricity principle

Finally, Australia may wish to consider and adopt the 'human-centricity principle' often highlighted in works discussing the regulation and ethics of AI systems.[10] As noted by one such work: "Put simply, a human-centric approach to AI is placing humans and the human experience at the centre of design considerations and intended outcomes of AI

[7] See e.g.: New Zealand Government. (2020). Algorithm Charter for Aotearoa New Zealand. Statistics NZ. https://data.govt.nz/assets/data-ethics/algorithm/Algorithm-Charter-2020_Final-English-1.pdf.

[8] See e.g., the work of the Institute for Accountability in the Digital Age (I4ADA), https://i4ada.org/.

[9] Australian Human Rights Commission, Human Rights & Technology Final Report (March 2021), at 51. See also: UN Office of the High Commissioner for Human Rights, *Who Will be Accountable? Human Rights and the Post-2015 Development Agenda* (HR Pub/13/1, 2013) 10.

[10] See e.g.: Singapore's Model AI Governance Framework (Second Edition) (2020), https://www.pdpc.gov. sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf, at 15.

technologies."[11] Importantly, even where used to pursue legitimate goals, the adoption of AI systems must not undermine human dignity.

4. Structural considerations

13. In addition to the framework created by the principles canvassed above, Australia may wish to explore structural arrangements that support the safe adoption of AI systems. For example, the Australian Human Rights Commission has recommended that an 'AI Safety Commissioner' be established.[12] This is a sound idea and where other States also adopt this approach, Australia may also wish to consider working towards establishing a structure for the active collaboration and cooperation between each State's respective AI Safety Commissioner; perhaps in the form of a 'Council of AI Safety Commissioners'.

14. More broadly, there are clear benefits to be gained from collaboration, coordination, information sharing, sharing of best practices, and support, for example, amongst the Commonwealth member countries in the context of the adoption of AI systems. Such work can also help address inequality of resources and degrees of development amongst cooperating States. In this latter respect, Australia may wish to consider and adopt shared training and training resources, as well as enhanced digital literacy programmes e.g., for courts, but also for the legal communities more broadly.[13]

[11] AI Asia Pacific Institute, Trustworthy Artificial Intelligence in the Asia-Pacific Region (July 2021), https://aiasiapacific.org/wp-content/uploads/2021/07/2021-Trustworthy-Artificial-Intelligence-in-the-Asia-Pacific-Region.pdf, at 15.

[12] Australian Human Rights Commission, Human Rights & Technology Final Report (March 2021), at 127-135.

[13] For example, the European Commission For The Efficiency of Justice recommends that "Generally speaking, when any artificial intelligence-based information system is implemented there should be computer literacy programmes for users and debates involving professionals from the justice system." (European Commission For The Efficiency of Justice, European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment (2018), https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c, at 12). The emphasis on increasing digital literacy is also found in Commonwealth initiatives such as the *Commonwealth Cyber Declaration 2018* stating that "digital literacy can be a powerful catalyst for economic empowerment and inclusion, and commit to take steps towards expanding digital access and digital inclusion for all communities without discrimination and regardless of gender, race, ethnicity, age, geographic location or language." (*Commonwealth Cyber Declaration 2018* https://thecommonwealth.org/commonwealth-cyber-declaration).

Professor Dan Jerker B. Svantesson

Professor Svantesson is based at the Faculty of Law at Bond University. He is also a Researcher at the Swedish Law & Informatics Research Institute, Stockholm University (Sweden), a Visiting Professor, Faculty of Law, Masaryk University (Czech Republic) and serves on the editorial board on a range of journals relating to information technology law, cyber security, cybercrime, data privacy law and law generally.

He held an ARC Future Fellowship 2012-2016, has written extensively on Internet jurisdiction matters and has won several research prizes and awards including the 2016 Vice-Chancellor's Research Excellence Award. Professor Svantesson has been identified as the field leader in 'Technology Law' in The Australian RESEARCH magazine four years in a row (2021, 2020, 2019 and 2018).

The views expressed herein are those of the author and are not necessarily those of any organisation with which Professor Svantesson is associated.