What should Australia Do?

Building and strengthening safe and responsible AI in Australia

Table of Contents

General comments

The Government has published *Safe and responsible AI in Australia: Discussion paper* (2023). Australia is already funding and driving many important initiatives towards safe and responsible AI with a clear capability to improve over time.

While the paper covers a great deal of material relevant to safe and responsible AI, it will benefit from incorporating findings from government funded research efforts such as Trusted Autonomous Systems (TAS)[1]. In the report, 'Figure 2: Domestic and international governance responses to support the safe and responsible deployment of AI' does not mention the world-leading tools, resources and frameworks developed for both civilian and military robotics, autonomous systems and artificial intelligence funded by the Australian government through the NGTF; managed by DISR and the Queensland Government via TAS.

[1] https://tasdcrc.com.au/ and specifically https://tasdcrc.com.au/resources/

TAS offers a wide range of publicly available resources to help create responsible AI via the *Robotics and Autonomous Systems Gateway (RAS-Gateway)*[2]. In the immediate term, the government should recommend the resources of RAS-Gateway for The Australian public and consider the transition RAS-Gateway out of a temporary CRC into a new AI governance organisation[3] as explored below.

The government should reread the TAS submission to the AI Action Plan—see Box 1.[4] as its recommendations still apply. The current government should revisit all submissions to the AI Action plan 2020, as many of the recommendations are relevant and actionable in 2023.

Box 1. Key recommendations TAS AI Action Plan (2020) submission:

The government should ensure the development of Australian sovereign capability regarding AI expertise, data sets, algorithms and systems by:

- Identifying the data sets required by developers of AI-enabled systems or algorithms and then scoping how best to source and facilitate access to Australian organisations through a digital AI marketplace;
- Reducing barriers to entry for small and medium sized AI-centric businesses and prioritising implementation measures;
- Supporting development of guidelines, regulations and regulatory structures, education, with 'good data' to create a world leading Australian AI industry; and
- Implementing incentives, grants and funding to support industry-wide adoption of technical infrastructure for AI development, test, evaluation and delivery through, for example, funding low-risk digital and physical test environments in Australia.

The success of RAS-Gateway suggests that the Australian government ought to build a digital suite of information and tools based on TAS RAS-Gateway and scaled to incorporate guidance appropriate across the domains of Australian society. This portal ought to be built and managed by new AI governance organisation/s.

Additionally, the Australian Government should spearhead Australian-specific AI, such as creating an Australian LLM from all materials in Trove, Australia's national archive as

[2] https://rasgateway.com.au/ and ethics and law resources for Defence applications https://rasgateway.com.au/resources/ethics-and-law

[3] TAS also submitted a report to the Queensland Government regarding the establishment of a National Accreditation Support Facility (2022) funded through the Assurance of Autonomy Activity that DISR should request and review via info@tasdcrc.com.au

[4] https://tasdcrc.com.au/the-trusted-autonomous-systems-submission-to-the-australian-ai-action-plan/

an educational resource for all Australians and to allow knowledge of Australian history and culture to be accessed globally. Australia needs to build large scale AI in order to ensure governance requirements are fit-for-purpose and adapting to the real-world challenges of building actual AI systems at scale.

More importantly, Australian government departments need to mandate adoption of The Australian AI Ethics Framework and explain to the Australian public how they are implementing them within their own procurement processes. This is vital to build trust with The Australian public given Robodebt, CensusFail and other examples of elected government officials and public servants being unable to act responsibility in their roles to protect The Australian public.

Dual Use Technology

The Australian government should ensure that responsible AI efforts are coordinated across civilian, military and security branches of the government because AI is a true dual-use technology. The existing approach of separating military and civilian uses of AI misses the use of civilian technologies to disrupt global peace and security.

For example, Alex Karp (Palantir Technologies) in The New York Times, (25 Jul 2023)[5] compares the strategic urgency to develop AI weapons with the supposed importance of building the atom bomb (he draws on Einstein to build this case). Yet, AI is so broad in its applications, that to some degree, it's better to compare its enabling aspects to the rise of the internal combustion engine. AI is going to affect so many military capabilities connected to use of weapons.

So, AI is not like a nuclear bomb, and more like an enabler of weapons and thousands of other capabilities that can be clipped together. The manifestation of AI is ubiquitous, and that ubiquity creates its own cascades of risk. So, unlike the nuclear bomb—one of the most protected, if not the most protected and controlled of all weapons—AI weapons are some of the least protected weapons.

AI is in commercial-off-the-shelf (COTS) products stitched together—that is the threat the world faces. AI in weapons is completely unlike the Oppenheimer situation. We don't have a handful of physicists and only wealthy, coordinating governments in control anymore. Instead, we have small state and non-state actors being able to create new asymmetric AI enhanced capabilities as diverse as can be imagined by humans or created by AIs themselves. AIs are likely already designing and optimising weapons and weapons effects in modelling. simulations and live trials—just as AI is ubiquitous in almost every domain of innovation.

So, the government needs a process to regulate, influence and (in some cases) control the research, production, and use of AI across all domains. Because an AI weapon may

[5] https://www.nytimes.com/2023/07/25/opinion/karp-palantir-artificial-intelligence.html

draw on a myriad of technologies including public, multi-function artificial general intelligences easily accessible in the civilian domain.

Organisations

The National AI Centre is funded as well as the Responsible AI Network. These are both excellent initiatives and CSIRO/Data61 should be commended for their roles in establishing these entities. However, in the long term, it may be more appropriate for the National AI Centre and Responsible AI Network to be outside the governance of CSIRO/Data61 because its functions ought not simply be scientific ones, but regulatory, governance, advisory, service, coordinating and oversight functions. Or, potentially Australia needs a multi-faceted approach to AI governance and not rely on a singular authority either scientific or regulatory? In this way, the current Data61/CSIRO research function could remain similar to the way that the Australian Road Research Board (AARB) supports "Councils and transport agencies – consultants and contractors, and private sector organisations on any aspect of transport and mobility"[6] . And, just as the AARB also requires Austroads, AI governance may need multiple kinds of organisations dedicated to responsible AI in Australia.

The need for diverse expertise

Regardless, Australia's AI research body (whether Data61/CSIRO or not) needs to employ more human factors researchers, more scholars from the humanities including linguists, anthropologists, philosophers, psychologists, sociologists, social work, law, justice, communication, human-centred computing, and neuroscience. It is not human-centred to run AI for Australia through the lens of science-only, focussing on technical solutions via computer science and data. AI is a sociotechnical threat. The greatest harms to Australians from an automated decision technology to date, Robodebt, came from a lack of human-centricity and responsibility rather than technical incompetence.

A multi-disciplinary and transdisciplinary approach produces practical, evidence-based and respectful ways of using data[7] in Australian society including the use of that data for AI capabilities.

Should Australia invest in a new AI regulator or build AI regulatory capacity within existing regulators?

One option for the government is to create a new AI regulator within the Department of Infrastructure (following the EU approach for generalised regulations governing AI), based on the model of AMSA, CASA and the NTC. This regulator could provide generalised guidance and requirements on AI systems irrespective of the intended

[6] https://www.arrb.com.au/about-arrb

[7] See 'Good Data' https://eprints.qut.edu.au/125605/

context of use. The advantages of a centralising authority would be to enable coordination and shared processes across Australian society. The disadvantages are that AI is going to be used in such a broad range of contexts, that sector specific governance will need to be layered for any specific use case.

The new Automated driving system entities regulatory framework by the NTC is a great example of regulatory reform in action within a domain and the Government should accelerate adoption of the framework and immediately begin funding similar reforms in the air and maritime domains. See NTC Policy Paper, The regulatory framework for automated vehicles in Australia (28 Feb 2022)[8]

Other regulators also need to be brought into the AI regulatory conversation, e.g. the ACMA

Case Study: Artificial Intelligence in the EMS

Sensitive to the increasing role of AI in spectrum management, improving spectrum allocations to assist autonomous technologies to function as well as the role of spectrum awareness and prediction in counter-autonomy capabilities, Trusted Autonomous Systems created deliverables on an examination of the regulation of spectrum in Australia with regards to artificial intelligence.

The products are intended to be a resource to accelerate conversations between the regulator and the regulated developing new capabilities with novel uses of spectrum use, management, prediction and sharing—in particular multi-agent systems including autonomous swarming systems[9].

Regulatory reform for use of AI in the EMS is vital given the rapid rise of AI technologies that utilise Spectrum, e.g. see recent announcement[10]

> *South Australian research and development company, Consunet Pty Ltd, has successfully completed its Next Generation Technologies Fund (NGTF) project – Distributed aUtonomous Spectrum Management (DUST) – commenced through Trusted Autonomous Systems in 2019. The project utilises machine learning and artificial intelligence to plan and allocate radio spectrum usage to achieve optimised spectrum utilisation in congested and contested environments.*

[8] https://www.ntc.gov.au/sites/default/files/assets/files/NTC%20Policy%20Paper%20-%20regulatory%20framework%20for%20automated%20vehicles%20in%20Australia.pdf

[9] Part One: Regulation of EMS in Australia: The Legal and Ethical Framework and issues relating to the development and operation of artificial intelligence to manage the electromagnetic spectrum – The Australian context, Part Two Artificial intelligence in the EMS. Contact info@tasdcrc.com.au

[10] https://defencesa.com/news-events-and-media/news/consunet-delivers-ground-breaking-spectrum-capability/

The risk of depending on existing domain-specific regulators to solve the AI regulation problem is a lack of internal expertise to draw from to determine safety requirements for dynamically evolving AI technologies. The future likely needs to combine existing expertise with funded new positions to tackle specific safety risks from integration with AI.

A new AI Safety Commissioner

The government should establish a new AI Safety Commissioner as outlined in recommendations 22 & 23 *Human Rights and Technology Final Report* (2021).[11] It is essential that regular Australians are protected and that the most vulnerable and marginalised have recourse to challenge circumstances where they are victim of algorithmic decision making.

Comparison with other organisations

The Government should consider setting up a federal government-funded non-profit entity outside of government (examples of similar organisations include: Standards Australia, Austroads, National Association of Testing Authories (NATA)).

Standards Australia is trusted to support whole of society support for emerging technologies. But, applying standards can be expensive and out of reach for most Australians who are using AI daily and need guidance for how to manage their lives. Standards Australia should continue to offer best in global standards on AI.

Austroads is a non-profit, non-partisan organisation that provides high-quality, practical and impartial advice, information, tools and services for transport agencies to deliver efficient, reliable and safe mobility to their customers. Austroads employs about 70 people to solve problems for transport agencies in Australia and New Zealand. They focus on making mobility safer and more reliable for all users and transport infrastructure sustainable and future-proof. They also provide national services that help transport agencies to operate seamlessly across state borders and bring national efficiencies to their operations. They are a not-for-profit, nonpartisan organisation funded by Australian and New Zealand government transport agencies. Their work impacts a wide range of agencies including planning, service, infrastructure, health and safety, public health and policing. A similar body for AI could help governments at multiple levels across Australia. An AI organisation (AustAI?) built along similar lines to Austroads would be invaluable to the Australian ecosystem.

NATA is a trusted organisation with regards to measurement. NATA could be expanded to assist with AI test and evaluation, verification, and validation (TEVV). Autonomous

[11] https://tech.humanrights.gov.au/artificial-intelligence/ai-safety-commissioner

systems testing facilities and simulation environments could comply with requirements set by NATA.

Summary

Apart from sector specific AI governance, Australia may need a new AI Safety Authority (with an AI Safety Commissioner) and AustAI, a new Australian organisation capable of providing guidance and services to members in a manner similar to RAS-Gateway. NATA ought to be expanded to provide accreditation to AI and autonomous systems test and evaluation facilities.

AI Risk Considerations

RAID Toolkit

Trusted Autonomous Systems developed a Responsible AI in Defence (RAID) Toolkit[12] that provides an Australian model for an AI Toolkit worth consideration for government. Unlike the more technical approaches of Data61/CSIRO (that are needed and vital), the RAID Toolkit provides an entry point for a conversation between industry and government regarding a proposed AI capability.

Box 2. Motivating the Responsible AI in Defence (RAID) Toolkit

On the 16 Feb 2023 at the Responsible AI in the Military Domain (REAIM) Summit, The Hague, the Australian Minister for Foreign Affairs, HON Penny Wong, stated that Australia recognises the vital importance of ensuring AI technologies are developed and used responsibly in a civilian and military context. What does it mean for AI to be developed responsibly? The RAID Toolkit provides practical best-practise steps to design, build, test and deploy AI responsibly. Particularly to help industry facilitate faster communication with Defence about the ethical and legal risks for their AI technologies. The RAID Toolkit offers a uniquely Australian focus while aligning with international best practice, including two recent releases from the United States in the form of the United States Department of Defense Directive 3000.09 Autonomy in Weapons Systems 10 Year Update and National Institute of Standards and Technology (NIST) AI Risk Management Framework. The Toolkit adapts the OECD Classification of AI Systems Framework, making it suitable in consideration of ethical and legal risk in the military domain and incorporates Australia's legal obligations under international humanitarian law including Australia's commitment to Article 36 reviews of all new weapons, means and methods of warfare. Responsible AI conversations can be made suitable to different stages of the Defence acquisition process including how to model

[12] https://tasdcrc.com.au/responsible-ai-for-defence-consultation/

CONOPS with human-AI teams to demonstrate how identified risks can be managed and mitigated in simulated test environments. One of the key features of the Toolkit is that there are measurable elements, so that a test protocol can be developed and deployed within system testing.

While the TAS RAID Toolkit is specifically designed with military applications in mind, the structure of the Toolkit could be utilised across many sectors, with sector-specific versions created within the relevant bodies, regulators and government departments. Key features of the Toolkit include:

- an AI Checklist (modified from the OECD Classification of AI Systems),
- an AI Risk Register and
- a Legal and Ethical Assurance Program Plan (LEAPP).

The Toolkit successfully brings together both ethical and legal considerations into a format useful to diverse stakeholders of AI systems.

Managing Complex AI Risk for General Purpose AI

AI risk frameworks have a lot of attention (e.g. EU AI Act and NIST AI Risk Management Framework) Novelli et.al (2023) note the variable risk generated by many kinds of AI[13]. They suggest dealing with complex AI risks (particularly from general purpose AI (GPAI) like ChatGPT) using an adapted IPCC climate change risk framework. A risk framework suitable to complex risks will help the government manage the cascading risks associated with dual use AI technologies.

The Edinburgh Declaration on Responsibility for Responsible AI

From the meteoric trend of 'Responsible AI' as the overarching normative moniker (rather than the slightly musty and out-of-fashion 'trustworthy AI' or 'ethical AI') comes the recent *Edinburgh Declaration on Responsibility for Responsible AI* (July 2023)[14]. The government needs to know that there are at least five sorts of responsibility:

1. causal responsibility,
2. moral responsibility,
3. legal responsibility,
4. role responsibility and
5. virtue responsibility

[13] Novelli, C., Casolari, F., Rotolo, A. et. al. 'Taking AI risks seriously: a new assessment model for the AI Act', AI & Soc (2023). https://link.springer.com/article/10.1007/s00146-023-01723-z

[14] https://medium.com/@svallor_10030/edinburgh-declaration-on-responsibility-for-responsible-ai-1a98ed2e328b

The declaration announces Four Key Shifts on responsibility needed to achieve Responsible AI:

- Accepting responsibility over describing and ascribing responsibility
- Seeing responsibility as relational over seeing responsibility as an agent property
- Prioritising responsibility as attending to vulnerability over responsibility as blame
- Focusing on sustainability of responsible AI/AS innovation over pace of innovation

While all four are important the most profound are items 2 & 3 which point to significant human culture shift over regulation, new rules and requirements. That is, we are responsible when we care about how we relate to others including when our relationship is mediated by technology. Diligence, respect, mindfulness, care, concern for loss of relationship, desire to improve our relationships and reputation... all character traits to be fostered and rewarded in the absence of hard and fast rules and regulations. Prioritising concern over those affected by technologies rather than blameworthiness shifts the focus to what we do and how we do it, as we're doing it, as opposed to focussing on when things go wrong.

The Australian government needs to consider how to build a culture of responsibility beginning within the public service and leading by example when it comes to the acquisition and use of AI. Good case studies of responsible use can then be disseminated through the education system and influence best practice across Australia. Culturally responsible AI becomes a matter of etiquette, just something that is 'done' by Australians, intuitively and instinctively.

Risk and Value-Sensitive Design

The IEEE7000-2021 standard[15] moves between value-sensitive design and risk management. So DISR, in consultation with Standards Australia should consider whether IEEE7000-2021 is fit for purpose and/or whether adapted or modified versions might be developed to meet different stakeholder requirements.

Expanding Australia's AI Ethics Framework

Australia's AI Ethics Framework is a great start for DISR and more effort should be invested in the implementation of this framework within Government funded AI initiatives. For example, the Government could adopt the recommendations of IEEE7001

[15] IEEE. (2021). IEEE 7000™-2021 - IEEE Standard Model Process for Addressing Ethical Concerns During System Design. https://engagestandards.ieee.org/ieee-7000-2021-for-systems-design-ethical-concerns.html

Transparency Standard[16] to require AI explanations for five expert and non-expert AI stakeholders (end users, the wider public and bystanders, safety certifiers, incident/accident investigators, lawyers and expert witnesses).

One of the tools created by TAS to help abide by the transparency and explainability ethics principle is the *Autonomous Systems Demonstration Canvas*[17] that helps developers show how to best show cognitive and social decision making in complex autonomy demonstrations that makes both intelligent performance and errors understandable and how to design a demonstration to show the capability of a system to abide by decision making norms such as commander's intent, military objectives; plus ethical, legal and safety frameworks.

Conclusion

DISR has been given a very big task to consider with regards to safe and responsible AI. An interdepartmental AI Task Force might be the first best step to begin the parallel lines of effort required to develop efficient and responsible organisations who can ensure Australia has safe and responsible AI. Australia has a huge range of world-leading approaches to technology regulation and can be a global leader at implementing AI governance to help manage building sovereign AI capabilities and using international AI products in ways aligned to its values.

[16] Winfield, A. F. T., Booth, S., Dennis, L. A., Egawa, T., Hastie, H., Jacobs, N., Muttram, R. I., Olszewska, J. I., Rajabiyazdi, F., Theodorou, A., Underwood, M. A., Wortham, R. H., & Watson, E. (2021). IEEE P7001: A Proposed Standard on Transparency. Frontiers in Robotics and AI, 8(225). https://doi.org/10.3389/frobt.2021.665729

[17] https://rasgateway.com.au/resource-hub/autonomous-systems-demonstration-canvas