Submission in response to *Safe and responsible AI in Australia* Discussion paper

Dr. Theresa KD Anderson, Connecting Stones Consulting

Ruth Marshall, Hocone Pty Limited

4 August 2023

In this submission we wish to flag five key areas we believe need greater emphasis as part of the government's consideration of regulatory and governance frameworks for responsible AI:

- Misuse
- Data quality
- Inappropriate use
- Legitimacy
- Trust-building activities

The dynamics of AI innovations must be managed with full appreciation of the particulars of Australia's socio-political context. In our view this means putting the principles of safe and responsible AI into practice to mitigate the potential risks from AI and ADM in the form of evolving documents and processes. AI developments to date have shown that many dynamics and (often unintended) impacts will only come to light through use. We will therefore need robust review procedures in place that can dynamically inform our regulatory and governance frameworks.

While global focus on the potential harms of AI is driving an interest and enthusiasm of benevolent actors to ensure they develop AI responsibly, it must be acknowledged that, like splitting the atom, AI can be used to harm as well as help. An important aspect of safe and responsible AI should include taking measures to prevent powerful AI tools from getting into the wrong hands or being adapted to unethical purposes. There is an argument that the horse has already bolted on this front, but we disagree: though we can't undo the past, AI research will continue and new innovations in and applications of AI will continue to be deployed, so new ways of preventing it's use by bad actors, or ameliorating its harm must be developed alongside.

We also wish to emphasise that building trust takes more than technology, systems or regulatory frameworks. While guidance in these areas is important for demonstrating trustworthiness, building public trust in relation to the use of emerging AI technologies at its heart is about human connection. Trust is a local and contextual judgement that is shaped by personal experiences (both positive and destructive), and yet, **trust-building** requires a collective effort as no one sector or advocate can be expected to carry responsibility. Equally important is creating an AI ecosystem building on Australia's recognized strengths in governance and civil society advocacy for responsible AI. Doing so would contribute to the ongoing and genuine engagement with the community necessary to increase public trust and confidence in the development and use of such technologies.

The rest of this submission will respond to specific questions provided in the consultation document.

Potential gaps in approaches (Q 2-4)

Q2. 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?

a) The erosion of citizen's rights

With the predicted explosion in the availability of AI-generated models and analytics will come a growth of the information imbalance between technology-rich organisations and the public, and an erosion of citizens right to privacy. Predictive AI models can potentially now be used to generate insights about personal health, wealth and behaviour using relatively small amounts of information, and the more information the better the insights. Checks and balances are required to ensure that AI-generated personal insights do not further erode the information imbalance between population and government and private business, and, with it, public trust.

For example, an insurance company can use information available about its customers to learn via AI of an as-yet undiagnosed pre-existing condition. It might be a condition that would not otherwise have been detected for years to come. A number of questions arise from this scenario:

1. Who is the owner of this new information, the company or the individual to whom it pertains?
2. Is there an obligation to inform the individual of the insight?
3. Does the individual have any rights over how the information is used or preserved?
4. Should the individual have been informed that their Personal Information would be used to create this new insight and what rights should they have over the use of their Personal Information in this way?

In short, what are the rights of citizens and organisations in relation to the generation and use of personal information from an AI-based system and where will they be enumerated? The Privacy Act 1988 does answer some of these questions, though we suggest making the link to AI-generated personal information more explicit (below). The reason this is important is because of the risk to privacy, autonomy and self-determination.

There are two broader questions embodied in the above:

1. Whether we have a right as citizens **not** to have others learn private information about us (that we may not even know ourselves), and what, if any, knowledge and control we can and should have over these.
2. Whether anyone with the means and inclination can use AI in this way, or should they be required to meet a threshold of legitimacy?

There remains a question about how these rights will be enumerated, and how citizen rights will be protected in the effort to balance citizen rights against the need to mitigate risks and harms. There needs to be sustained attention devoted and plans for addressing concerns raised lest the information imbalance become too high. Although there are already some thresholds for the legitimate use of personal data in the Privacy Act and the concept of beneficial use in Australia's AI Ethics Principles, a wider discussion of legitimate use and the establishment of boundaries is required, particularly, for

example, as pre-built models for mass surveillance become prevalent, often originating from overseas. The wider community should be part of the conversation defining the parameters for legitimate use.

Beyond the AI system, the question of legitimacy also applies to the entities running the AI program. On what basis must the community accept the legitimacy of a particular entity to be in charge of such a project? A good analogy from the data world is the Indigenous Data Sovereignty initiative which seeks to put first-nation's data and its uses under their control, as the legitimate owners and controllers of their data. Any public-facing AI-based solutions should at a minimum have to establish the legitimacy and competency of the practitioners to do the work. This needs to be done to ensure that community trust and the empowerment that citizens in an open and democratic society should expect is not eroded.

In summary, we will need more thinking about

- what concerns these new contexts of data use raise and for whom;
- how citizens can be included in the discussion of these concerns and resulting decision making; and
- how this situation changes -- and will continue to change in the dynamic conditions within which these emerging technologies are deployed.

b) Change the wording of Australian Privacy law APPs and Guidance to explicitly address AI-generated Personal Information

Business and government have used their business knowledge to develop insights and opinions about their individuals from time immemorial. It is widely understood by the public that a certain amount of personal information will be kept by the organisations they do business with. AI will make it possible for increasing amounts of information to be created by an increasingly wide range of entities with increasingly loose and informal arrangements with the individual. This could be video based and facial recognition including analyses taken in shops, malls and other public areas, sentiment and other personal categorisations from written job applications or AI-bot interviews, or an infinite number of other scenarios where data is collected. There will often be no expectation by the individual that the information collected will be used by an AI to generate new knowledge or opinions about them. Although the threats of mass surveillance and AI classification reach well beyond privacy consideration, we believe that this is a worthwhile place to start. We believe existing privacy legislation does cover the products of AI from a privacy perspective, however this should be made more explicit in the wording.

According to our reading of the Privacy Act 1988, new predictions or observations about an identifiable individual derived from an AI model or Analytics process qualifies as Personal Information:

> *"Information or an opinion about an identified individual, or an individual who is reasonably identifiable: whether the information or opinion is true or not; and whether the information or opinion is recorded in a material form or not." (Privacy Act 1988).*

It is therefore believed that the Privacy Act and the Australian Privacy Principles (APP) governing the standards and obligations around the use of Personal Information are operative. This should be made

explicit for the avoidance of doubt, as AI-derived information will not generally enter the data lifecycle in the early "collection" stage of the project, but in a later "analyse/use" stage, and is considered the product of the AI project rather than a data input. The APP guidance focuses heavily on the terms under which data is collected and then used. It appears to assume a process of handling Personal Information which starts at the beginning of the data lifecycle.

Figure 1 – simple data lifecycle

For the avoidance of doubt it would be preferable to state unequivocally that new Personal Information may be created by an AI project and this should be treated as a form of "Collection". Furthermore, it is of course possible to generate Personal Information using models which were built from de-identified information collected from completely different data subjects (this is probably the more common scenario). Privacy protection extends beyond the individuals represented in the initial training (or model development) data set, but also to individuals who might be labelled or categorized using this model. This is not clearly reflected in the wording of the Privacy Act and APP's today.

APP 1 guidance talks about openness and transparency. The guidance on developing an APP Privacy Policy, consistent with the APP's overall, hinges on data collection. The guidance should make clear that insights generated from AI models and Analytics which may be linked to an individual should be listed in the Privacy Policy when describing the kinds of information collected/generated and held, how it is collected/generated and held, purpose of collection/generation, etc. Without providing the user with this information they will not be able to make a fully informed consent where relevant or ask to see this information and have it corrected or removed (rights under the Privacy Act), resulting in a material reduction in transparency. By making explicit reference to AI-generated Personal Information in the Privacy Act and APPs, important protections around the quality and use of the Personal Information generated by the AI/Analytics are also made explicit.

Q3. 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.

Regulation fits within a wider ecosystem that addresses both social and technical concerns. Building trust in any context is personal and local. So while being transparent about the way data sources are handled and AI/ADM systems are deployed can help demonstrate the trustworthiness of our AI practices, trust cannot be built without actively connecting with the communities who are represented in the data sets in use and potentially impacted by the AI technologies in use. If we want communities to feel secure in the knowledge that their government is regulating AI technologies in their best interests, the deployment of such technology and assurance frameworks needs to also include a strategy for communicating and connecting with community.

A growing public consciousness about the vulnerability of data to misinterpretation, misuse and misappropriation in data-driven decisioning prompts active risk mitigation **and** evidencing of good practice to the citizen population. Making visible the frameworks used for data protection and data governance is a critical step in building and maintaining trust. It encompasses appropriate data stewardship and "good practice" in terms of data management/governance located in the context of an organisation's decision making and policy processes. However, it also needs to move toward ensuring citizens can have a voice in the way that data and AI practices shape government activities, especially decisions that impact on the everyday life of its citizens. This line of thinking for gaining public trust in the data practices centres around four key themes: procedural fairness; distributional fairness; good governance; and demonstrating trustworthiness.

Participatory mechanisms to partner with the community can help ensure ongoing response to community needs/concerns. However, there remains a tension between passive vs active engagement of public opinions, obtaining "consents" or endorsements for activity involving the use of citizen-related/impactful data often including personally identifying information. Also, a focus on transparency about process and challenges can, as Ananny & Crawford caution, "... privilege seeing over understanding".[1] Consequently, making visible is not enough. Instead, co-design frameworks are needed, especially mechanisms for ongoing feedback with sufficiently diverse populations, especially including vulnerable groups.

Moving forward also calls for process-based rather than rule-based frameworks and a mindset of inclusive and constant consultation and "evolving design" as are common in deliberative democracy initiatives. To support these activities will mean creating conditions that give people time to think, experiment and safely share concerns and developments. These critical activities for promoting civil discourse and productive sharing of ideas and uncertainties allow trust to flourish within a community. However, they cannot be assumed to naturally unfold. These activities therefore need ongoing attention alongside the technical training the workforce will need to work with these new AI tools responsibly.

Q4. 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.

As mentioned in our introduction, creating an AI ecosystem building on Australia's recognized strengths in governance and civil society advocacy as well as the insights from indigenous voices[2] would contribute to the ongoing and genuine engagement with the community necessary to increase public trust and confidence in the development and use of such technologies.

---

[1] Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989.

[2] An excellent example is the 2021 paper Out of the Black Box: Indigenous Protocols for AI by Angie Abdilla, Megan Kelleher, Rick Shaw, Tyson Yunkaporta.

The NSW AI Assurance Framework provides a model that could be applied to a national context. Speaking as a member of the review committee responsible for creating the framework and seeing it applied to NSWGovernment projects (Anderson), we flag two key elements of the framework that could influence positive and responsible uptake of AI in Australia. First, the inclusive consultative process by which the assurance framework was devised – bringing together contributions and insights from key government departments with expertise drawn from industry, academia and civil society and encouraging full and frank discussions about priorities and pathways. Second, encouraging government departments to bring projects to the review committee as early as possible to talk through their interpretation of the framework with respect to the particulars of their project. Presentations to the review committee explored desired outcomes of each project, potential risks and ways forward. A key lesson from this NSW experience is the importance of creating an environment conducive to genuine engagement about risks as well as opportunities both within the committee itself and with project proposers. Another key lesson to apply to the Australian context is the appreciation that the framework is a dynamic, living document that needs socialising and review.

Target Areas (Q 9 & 11)

Q9. Given the importance of transparency across the AI lifecycle, please share your thoughts on:

a. where and when transparency will be most critical and valuable to mitigate potential AI risks and to improve public trust and confidence in AI?

An AI is dependent on human action at many points (to get it the data it needs, to define how that data is input, define how decisions are classified). However, in focusing energy on protecting the data that does exist (and is generated), there is a risk of losing sight of missing data (and population cohorts). The design of AI-technologies needs to give particular consideration to inequities that may/do get “baked in” to the data that fuels these technologies.

Transparency is very important, but collaboration and inclusion are better because people have a genuine experience of what is going on rather than simply being told what is going on. Making visible the value-laden chain of practices and policies shaping the insights gleaned from data -- human, social, organisational and technical - contributes to this reframing process. Our computational and decision support tools are made through practice. The decisions we make, as individuals and as a society, are based on our value judgments – judgments that are subjective and emotional rather than rational. This need not be problematic. Ethical data practice may mean learning to *make the invisible visible* by remaining alert to who (and what) is missing, under-represented or mis-represented in the data. Genuine public engagement and trust-building are essential for handling unintended consequences and impacts. Complaints and feedback mechanisms must be actively and meaningfully used – visible accountability, action and improvement offer opportunities for genuine ownership of the process. Doing so contributes not just to transparency but legitimacy.

We need to be more deliberate in our efforts to ensure that people and communities are appropriately represented in the data used within these AI systems. Distribution of benefits will not likely be even – particularly as unintended impacts reveal themselves. Experience has shown that favourable perceptions of a new technology are impacted by personal experience. And positive outcomes can be very quickly

diminished by detrimental/negative impacts. However, if we can build new data systems that start as two-way streets, we bring the individuals from whom the data comes more into the centre of the design process. Designing and genuinely (and consistently) making use of two-way street empowers the "subjects" [from whom data is so often taken but infrequently returned] to have a voice on the data's legitimate and beneficial use.

To elicit honest and open feedback about concerns and potential risks will require attention to be given to the "keystone practices"[3] of building trust.

Mapping the pathway from person to proxy as a two-way street is essential to building more representative - and thereby better quality - data. Returning to the community earlier in the process would elicit richer understandings about any data collected. Involving them in the analysis of the data in some way also offers an opportunity to enhance their data literacies, empowering them to speak about data that represents them. At the very least, better documentation of context (for instance in the descriptive metadata associated with a data set that will feed an AI system) could disambiguate the data and support community feedback.

Working to create such two-way streets in our data work would make sure that the community from whom data was taken also benefit from the process. It is important we develop ways to design **with** communities and not simply **for** them. The importance of a two-way street is all the more acute when dealing with vulnerable populations, who often have precious little power and influence over policy decisions that affect them. Creating two-way streets can help "catch" concerns and problem early. You cannot plan for specific unintended consequences, but you **can** build a diverse, informed and inclusive community to participate in planning and ongoing review.

b. mandating transparency requirements across the private and public sectors, including how these requirements could be implemented.

"At the core of any ethical approach is sound decision making, good housekeeping and open books (showing your work)."[4]

Transparency is not just for the AI products and output themselves, but also for the decisions that are made leading up to and during the AI project that will help to identify and correct errors at a later stage should an unintended consequence arise or the system not work as intended.

There are a number of important areas where pivotal decisions are made in the development or use of an AI system. They are:

- Project Justification – explaining the purpose and legitimacy of the project, and the skills, motivations and legitimacy of the team (including conflicts of interest). A record of all groups consulted in this process and included in the ongoing project should be included.

[3] for further discussions see page 60 in Trust building for data sharing – Understanding trust as a social relationship, Anderson, 2023.
[4] Implementing next generation privacy and ethics research in education technology R. Marshall et al 2021.

- Project method – explaining the detailed objective, hypotheses and success metrics and justification of the methodology chosen. If a third party AI product is procured, the selection process and vetting details should be included.
- Key data sources and selection decisions – including analysis of data quality and treatments which will be applied.
- Testing – strategy for testing the AI product and who will be included in the process of testing and final approval.
- Continuous monitoring protocol – strategy for collecting information about the workings of the system in production, including complaints and unanticipated behaviour or outcomes. Should include a mechanism for reporting adverse events and complaints, and mechanism for reporting AI product behaviour, complaints, adverse events and actions taken.
- Changes – changes to any of the above should also be recorded with explanations, keeping the original information available for the record.

Together, these will provide clarity to relevant parties on what behaviour can be expected from the system and the intended purpose as well as who participated in key decisions. For public service organisations, these should be readily available to the public during development and after the project has gone live, until it is retired. There should be mandatory public reporting of complaints and actions taken, similar to adverse event reporting in healthcare.

For private organisations they could be stored privately and made available only to project collaborators if commercially confidential. Other documentation such as Privacy Impact Assessments and AI Impact statements should be available in the same way.

To be worthy of trust takes more than authority – actions taken need to be explained along with evidence justifying those actions. The public needs to SEE HOW government is guiding safe and responsible use, not just be told that safeguards are in place. Again, questions of beneficence and improvement need to be asked in this context. Complaints and communications from the public and actions taken by the department to remedy the matter should be publicly available for scrutiny. Statistical reports on the number and types of communications and actions taken should be published. Publicly available information and metrics on such complaints mechanisms, for instance, makes visible how unintended and detrimental impacts discovered after deployment of a system are remedied.

Returning to our example of the NSW AI Assurance Framework: by documenting ways that a project relates to the various components of the framework, project owners are showing their hand. Documenting work at ethical choice points in a project not only helps developers to remain minded about the implications of their actions – it also serves as a useful guide for tracing back through decisions if and when any unintended consequences should occur. Not only can such documentation help provide assurance to the public about the care being taken in the deployment of new technologies and processes; it can also inform the review needed to put improvements and corrections in place.

Making that learning visible to the community is a powerful tool for demonstrating trustworthiness.[5]

[5] see discussion of the value of making time to think and test on page 58 in Trust building for data sharing – Understanding trust as a social relationship, Anderson, 2023.

Q11 What initiatives or government action can increase public trust in AI deployment to encourage more people to use AI?

Building public trust and encouraging positive/productive use of AI technologies requires active engagement with the wider social context shaping AI use. However well-intentioned and informed our frameworks to optimise principles of human value may be, we cannot avoid the impact of the wider socio-cultural context on the shaping and execution of those values. Principles of fairness and goodness (however defined) still rely on the social context within the system for their meaning. As we have mentioned in our responses to earlier questions, deliberate and deep engagement with all sectors of society is needed to mitigate the risks of AI systems contributing to injustices, misrepresentations or under-representations.

It is especially important to appreciate that the current global climate of distrust is fertile ground for abuse of AI in regards to disinformation. Regulation and assurance frameworks are not enough on their own when people feel disempowered, disillusioned and disengaged. There is a lesson to be learned from the transition of Cold War security politics in the 1980s about the value of deliberately public-facing engagements. Just as there was rising public sentiment within Europe calling out for greater voice in the decision making about nuclear weapons deployments, so too are data publics starting to demand more engagement, more socializing of their security and equity concerns. Policies have to be explained. Evidence has to be shared. Communities have to be involved. Actions and protections have to be socialised by foregrounding community benefit.

In Australia, all employees who serve alcohol must have a Responsible Serving of Alcohol(RSA) certificate. What would an RSA for responsible AI look like? Data has a social life that influences the way it is made and used. How it is shaped by researchers and analysts will reveal specific aspects of the critical social networks the data inhabits. In the words of Brown and Duguid:

> "…to participate in that shaping and not merely to be shaped requires understanding such social organization, not just counting on (or counting up) information."[6]

The responsible use of AI is tethered to responsible use of data. There are many dimensions to the ethical use of data. More often than not, discussions of data ethics pivot around data privacy and security, and specific applications such as Artificial Intelligence (AI) and automation. Yes, there are important ethical considerations in each of these areas, but, to be a responsible server of data, we should be thinking of a multitude of ethical considerations from the moment the data comes into our orbit. Doing so will make us all better data practitioners, generating more useful and reliable data products, ethically.

Data's value—and its power – is shaped by its social life. Responsible use of it, especially in connection to any AI system, also involves thinking beyond what you can do, and instead focusing on what you OUGHT to do. We need to find ways to engage productively, both individually and collectively, with risk and adversity. We need to ensure oversight so outcomes can be assured of serving all sectors of our communities -- particularly the most vulnerable. We need a culture of care. Trustworthiness would thus be demonstrated by focusing on communal well-being of people and their environment and taking responsibility for stewarding on behalf of future generations.

[6] Duguid, P., Brown, J. S. (2000). The social life of information. Boston: Harvard Business School Press.

References

[issues raised in this submission are evidenced and further expanded upon in the following papers]

Marshall & Anderson, 2023 The Responsible Service of Data | hocone

Marshall et al, 2021 Implementing next generation privacy and ethics research in education technology - British Journal of Educational Technology - Wiley Online Library

Anderson, 2023 "Trust building for data sharing – Understanding trust as a social relationship," in *Data and the Digital Self: What the 21st Century Needs*, Australian Computer Society, ch 4

Anderson, 2023, *Looking at Securitization as a Sociotechnical Activity: Lessons From a Cold War Past for AI Futures* IEEE Technology and Society Magazine